GCAT
TACG
GCAT

Grand Celebration: 10th Anniversary of the Human Genome Project

Volume 2

Edited by

John Burn, James R. Lupski,
Karen E. Nelson and Pabulo H. Rampelotto

Printed Edition of the Special Issue Published in *Genes*

MDPI

www.mdpi.com/journal/genes

John Burn, James R. Lupski, Karen E. Nelson and
Pabulo H. Rampelotto (Eds.)

Grand Celebration: 10th Anniversary of the Human Genome Project

Volume 2

MDPI

This book is a reprint of the special issue that appeared in the online open access journal *Genes* (ISSN 2073-4425) in 2014 (available at: http://www.mdpi.com/journal/genes/special_issues/Human_Genome).

Guest Editors
John Burn
University of Newcastle
UK

James R. Lupski
Baylor College of Medicine
USA

Karen E. Nelson
J. Craig Venter Institute (JCVI)
USA

Pabulo H. Rampelotto
Federal University of Rio Grande do Sul
Brazil

Editorial Office
MDPI AG
Klybeckstrasse 64
Basel, Switzerland

Publisher
Shu-Kun Lin

Assistant Editor
Rongrong Leng

1. Edition 2016

MDPI • Basel • Beijing • Wuhan

ISBN 978-3-03842-123-8 complete edition (Hbk)
ISBN 978-3-03842-169-6 complete edition (PDF)

ISBN 978-3-03842-124-5 Volume 1 (Hbk) ISBN 978-3-03842-170-2 Volume 1 (PDF)
ISBN 978-3-03842-125-2 Volume 2 (Hbk) ISBN 978-3-03842-171-9 Volume 2 (PDF)
ISBN 978-3-03842-126-9 Volume 3 (Hbk) ISBN 978-3-03842-172-6 Volume 3 (PDF)

Table of Contents

List of Contributors

Ana Alfirevic: Wolfson Centre for Personalised Medicine, Department of Molecular and Clinical Pharmacology, University of Liverpool, Block A Waterhouse Buildings, 1-5 Brownlow Street, Liverpool, L69 3GL, UK.

Fazil Aliev: Department of Psychiatry, Virginia Commonwealth University, P.O. Box 980126, Richmond, VA 23298, USA.

Jessica N. Cooke Bailey: Department of Epidemiology and Biostatistics, Case Western Reserve University Medical Center, Cleveland, OH 44106, USA.

Meike Bartels: Department of Biological Psychology, VU University Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands.

Oliver F. Bathe: Department of Surgery, Tom Baker Cancer Center, University of Calgary, 1331 29th St NW, Calgary, AB T2N 4N2, Canada; Department of Oncology, Tom Baker Cancer Center, University of Calgary, 1331 29th St NW, Calgary, AB T2N 4N2, Canada.

Dorret I. Boomsma: Department of Biological Psychology, VU University Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands.

Reinhard Buettner: Institute of Pathology, University Hospital Cologne, Kerpener Str. 62, Cologne D-50937, Germany; Center for Integrated Oncology Cologne-Bonn, Cologne D-50937, Germany.

Daniel F. Carr: Wolfson Centre for Personalised Medicine, Department of Molecular and Clinical Pharmacology, University of Liverpool, Block A Waterhouse Buildings, 1-5 Brownlow Street, Liverpool, L69 3GL, UK.

Gareth E. Davies: Avera Institute for Human Genetics, 3720 W. 69th Street, Sioux Falls, SD 57108, USA; Department of Psychiatry, University of South Dakota, 4400 W. 69th Street, Sioux Falls, SD 57108, USA.

Gianluca De Bellis: Institute for Biomedical Technologies, Consiglio Nazionale delle Ricerche, Milan 20132, Italy.

Danielle M. Dick: Department of Psychiatry, Virginia Commonwealth University, P.O. Box 980126, Richmond, VA 23298, USA.

Alexis C. Edwards: Department of Psychiatry, Virginia Commonwealth University, P.O. Box 980126, Richmond, VA 23298, USA.

Erik A. Ehli: Avera Institute for Human Genetics, 3720 W. 69th Street, Sioux Falls, SD 57108, USA; Department of Psychiatry, University of South Dakota, 4400 W. 69th Street, Sioux Falls, SD 57108, USA.

Radoslaw K. Ejsmont: VIB Center for the Biology of Disease, VIB, 3000 Leuven, Belgium; Center for Human Genetics, University of Leuven School of Medicine, 3000 Leuven, Belgium.

David M. Evans: School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, UK; Diamantina Institute/Translational Research Institute, University of Queensland, Level 7, 37 Kent Street, Woolloongabba, Brisbane QLD 4102, Queensland, Australia.

Farshad Farshidfar: Department of Surgery, Tom Baker Cancer Center, University of Calgary, 1331 29th St NW, Calgary, AB T2N 4N2, Canada.

Rebecca C. Fry: Curriculum in Toxicology/Lineberger Comprehensive Cancer Center/Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, The University of North Carolina, 135 Dauer Drive, CB 7431, UNC, Chapel Hill, NC 27599, USA.

Yunxin Fu: Division of Biostatistics and Human Genetics Center, The University of Texas Health Science Centerat Houston, 1200 Herman Pressler, Houston, TX 77025, USA; Laboratory for Conservation and Utilization of Bio-Resources, Yunnan University,Kunming 650091, China.

Francesco Gandolfi: Genethon, 1bis Rue de l'Internationale, 91020 Evry, France.

Michael Hadjidaniel: The Cyprus Institute of Neurology and Genetics, 6 International Airport Avenue, Ayios Dometios, Nicosia 2370, Cyprus.

Jonathan L. Haines: Department of Epidemiology and Biostatistics, Case Western Reserve University Medical Center, Cleveland, OH 44106, USA.

Bassem A. Hassan: VIB Center for the Biology of Disease, VIB, 3000 Leuven, Belgium; Center for Human Genetics, University of Leuven School of Medicine, 3000 Leuven, Belgium.

Bastiaan T. Heijmans: Department of Molecular Epidemiology, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, The Netherlands.

Matthew Hickman: School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, UK.

Richard S. Houlston: Division of Genetics and Epidemiology / Molecular and Population Genetics Team, Genetics and Epidemiology, The Institute of Cancer Research, Sutton, SM2 5NG, UK.

Cathrine Hoyo: Department of Biological Sciences, Center for Human Health and Environment, Campus Box 7633, NC State University, Raleigh, NC 27695, USA.

Jaakko Kaprio: Department of Public Health, Hjelt Institute, University of Helsinki, P.O. Box 41, Helsinki FI-00014, Finland; National Institute for Health and Welfare, Department of Mental Health and Substance Abuse Services, P.O. Box 30, Mannerheimintie 166, Helsinki FI-00300, Finland; University of Helsinki, Institute for Molecular Medicine (FIMM), P.O. Box 20, Tukholmankatu 8, Helsinki FI-00014, Finland.

Kenneth S. Kendler: Department of Psychiatry, Virginia Commonwealth University, P.O. Box 980126, Richmond, VA 23298, USA.

Christopher Ryan King: Department of Health Studies, University of Chicago, Chicago, IL 60637, USA.

Michael Kloth: Institute of Pathology, University Hospital Cologne, Kerpener Str. 62, Cologne D-50937, Germany; Center for Integrated Oncology Cologne-Bonn, Cologne D-50937, Germany.

Tellervo Korhonen: Department of Public Health, Hjelt Institute, University of Helsinki, P.O. Box 41, Helsinki FI-00014, Finland; National Institute for Health and Welfare, Department of Mental Health and Substance Abuse Services, P.O. Box 30, Mannerheimintie 166, Helsinki FI-00300, Finland.

George Koumbaris: NIPD Genetics Ltd., Neas Engomis 31, Engomi, Nicosia 2409, Cyprus; The Cyprus Institute of Neurology and Genetics, 6 International Airport Avenue, Ayios Dometios, Nicosia 2370, Cyprus.

Elena Kypri: NIPD Genetics Ltd., Neas Engomis 31, Engomi, Nicosia 2409, Cyprus.

Antti Latvala: Department of Public Health, Hjelt Institute, University of Helsinki, P.O. Box 41, Helsinki FI-00014, Finland.

Glyn Lewis: Division of Psychiatry, University College London, 67-73 Riding House St., London W1W 7EJ, UK.

Hua Ling: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

Gregory S. Liptak: Department of Pediatrics, SUNY Upstate Medical Center, Golisano Children's Hospital, Syracuse, NY 13210, USA.

Anu Loukola: Department of Public Health, Hjelt Institute, University of Helsinki, P.O. Box 41, Helsinki FI-00014, Finland.

Anneke Lucassen: Clinical Ethics and Law Unit, Wessex Clinical Genetics Service, The Princess Anne Hospital, Southampton, SO16 5YA, UK.

John Macleod: School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, UK.

Fulvio Mavilio: Genethon, 1bis Rue de l'Internationale, 91020 Evry, France.

David W. Mohr: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

Arianna Moiani: Genethon, 1bis Rue de l'Internationale, 91020 Evry, France.

Susan K. Murphy: Department of Obstetrics and Gynecology, Duke University Medical Center, B226 LSRC, Box 91012, Research Drive, Durham, NC 27708, USA.

Dan L. Nicolae: Departments of Medicine, Statistics, and Human Genetics, University of Chicago, Chicago,IL 60637, USA.

Monica D. Nye: Lineberger Comprehensive Cancer Center, The University of North Carolina, 450 West Street, CB 7295, UNC, Chapel Hill, NC 27599, USA; Department of Obstetrics and Gynecology, Duke University Medical Center, B226 LSRC, Box 91012, Research Drive, Durham, NC 27708, USA.

Elisavet A. Papageorgiou: NIPD Genetics Ltd., Neas Engomis 31, Engomi, Nicosia 2409, Cyprus; The Cyprus Institute of Neurology and Genetics, 6 International Airport Avenue, Ayios Dometios, Nicosia 2370, Cyprus.

Philippos C. Patsalis: The Cyprus Institute of Neurology and Genetics, 6 International Airport Avenue, Ayios Dometios, Nicosia 2370, Cyprus.

Margaret A. Pericak-Vance: Hussman Institute for Human Genomics, Miller School of Medicine, University of Miami, Miami, FL 33136, USA.
Munir Pirmohamed: Wolfson Centre for Personalised Medicine, Department of Molecular and Clinical Pharmacology, University of Liverpool, Block A Waterhouse Buildings, 1-5 Brownlow Street, Liverpool, L69 3GL, UK.
Ermanno Rizzi: Institute for Biomedical Technologies, Consiglio Nazionale delle Ricerche, Milan 20132, Italy.
Richard J. Rose: Department of Psychological and Brain Sciences, Indiana University, 1101 East 10th St., Bloomington, IN 47405, USA.
Jessica E. Salvatore: Department of Psychiatry, Virginia Commonwealth University, P.O. Box 980126, Richmond, VA 23298, USA.
Axel Schambach: Institute of Experimental Hematology, Hannover Medical School, Carl-Neuberg-Str.1, D-30625 Hannover, Germany.
Robert B. Scharpf: Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.
Alan F. Scott: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.
Marco Severgnini: Institute for Biomedical Technologies, Consiglio Nazionale delle Ricerche, Milan 20132, Italy.
P. Eline Slagboom: Department of Molecular Epidemiology, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, The Netherlands.
Roderick C. Slieker: Department of Molecular Epidemiology, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, The Netherlands.
Lisa Smeester: Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, The University of North Carolina, 135 Dauer Drive, CB 7431, UNC, Chapel Hill, NC 27599, USA.
Julia Debora Suerth: Institute of Experimental Hematology, Hannover Medical School, Carl-Neuberg-Str.1, D-30625 Hannover, Germany.
Jenny van Dongen: Department of Biological Psychology, VU University Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands.
Zachary M. Weber: Avera Institute for Human Genetics, 3720 W. 69th Street, Sioux Falls, SD 57108, USA.
Andrew E. Yosim: Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, The University of North Carolina, 135 Dauer Drive, CB 7431, UNC, Chapel Hill, NC 27599, USA.
Peng Zhang: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

Preface

In 1990, scientists began working together on one of the largest biological research projects ever proposed. The project proposed to sequence the three billion nucleotides in the human genome. The Human Genome Project took 13 years and was completed in April 2003, at a cost of approximately three billion dollars. It was a major scientific achievement that forever changed the understanding of our own nature. The sequencing of the human genome was in many ways a triumph for technology as much as it was for science. From the Human Genome Project, powerful technologies have been developed (e.g., microarrays and next generation sequencing) and new branches of science have emerged (e.g., functional genomics and pharmacogenomics), paving new ways for advancing genomic research and medical applications of genomics in the 21st century. The investigations have provided new tests and drug targets, as well as insights into the basis of human development and diagnosis/treatment of cancer and several mysterious humans diseases. This genomic revolution is prompting a new era in medicine, which brings both challenges and opportunities. Parallel to the promising advances over the last decade, the study of the human genome has also revealed how complicated human biology is, and how much remains to be understood. The legacy of the understanding of our genome has just begun. To celebrate the 10th anniversary of the essential completion of the Human Genome Project, in April 2013 *Genes* launched this Special Issue, which highlights the recent scientific breakthroughs in human genomics, with a collection of papers written by authors who are leading experts in the field.

John Burn, James R. Lupski,
Karen E. Nelson and Pabulo H. Rampelotto
Guest Editors

The Epigenome View: An Effort towards Non-Invasive Prenatal Diagnosis

Elisavet A. Papageorgiou, George Koumbaris, Elena Kypri, Michael Hadjidaniel and Philippos C. Patsalis

Abstract: Epigenetic modifications have proven to play a significant role in cancer development, as well as fetal development. Taking advantage of the knowledge acquired during the last decade, great interest has been shown worldwide in deciphering the fetal epigenome towards the development of methylation-based non-invasive prenatal tests (NIPT). In this review, we highlight the different approaches implemented, such as sodium bisulfite conversion, restriction enzyme digestion and methylated DNA immunoprecipitation, for the identification of differentially methylated regions (DMRs) between free fetal DNA found in maternal blood and DNA from maternal blood cells. Furthermore, we evaluate the use of selected DMRs identified towards the development of NIPT for fetal chromosomal aneuploidies. In addition, we perform a comparison analysis, evaluate the performance of each assay and provide a comprehensive discussion on the potential use of different methylation-based technologies in retrieving the fetal methylome, with the aim of further expanding the development of NIPT assays.

Reprinted from *Genes*. Cite as: Papageorgiou, E.A.; Koumbaris, G.; Kypri, E.; Hadjidaniel, M.; Patsalis, P.C. The Epigenome View: An Effort towards Non-Invasive Prenatal Diagnosis. *Genes* **2014**, *5*, 310-329.

1. Introduction

The discovery of free fetal DNA in maternal circulation [1] was a landmark towards the development of non-invasive prenatal diagnostic assays, and remarkable advances have taken place since then. The revolution was initiated in 1997 with the determination of the fetal fraction, which was estimated to be 3% during the early stages of the pregnancy [2]. In the following years, more advanced technologies were used (e.g., digital PCR) to re-evaluate the fetal DNA fraction, which is now estimated to be 10%–20% [3].

Deciphering the critical characteristics of the fetal genome has been the main goal for the development of non-invasive prenatal tests (NIPT). Studies have shown that the origin of maternal free DNA present in maternal peripheral blood is the hematopoietic system of the mother [4]. On the other hand, free fetal DNA (ffDNA) is derived from embryonic cell degradation in maternal peripheral blood [5,6] or from apoptotic placental cells [7–9]. More recent studies have confirmed the above, using bisulfite sequencing technologies and provided convincing evidence for the origin of both fetal and maternal free DNA in maternal plasma [10]. It has also been demonstrated that free fetal DNA from maternal plasma is cleared immediately (within a few hours) after pregnancy [11]. These findings were confirmed by more recent studies [12–15] and is a finding of great importance, since the presence of fetal DNA from previous pregnancies would interfere with the correct interpretation of subsequent pregnancies. A number of independent studies have also

demonstrated that the amount of fetal DNA released in maternal circulation increases with pregnancy progression [2,16].

Other studies characterizing ffDNA have found that the size of fetal DNA fragments were estimated to be <0.3 kb, whereas that of maternal DNA was >1 kb [17,18]. Follow-up studies have demonstrated that the release of fetal DNA is due to the apoptosis of no more than three nucleosomal complexes, and it has been shown that the average fetal fragment size is 286 ± 28 bp with a maximum ffDNA fragment size ranging from 219 to 313 bp [19]. However, better determination and characterization of free fetal DNA fragment sizes will allow further evaluation of the diagnostic limitations that are introduced because of fragment size.

The first attempts towards NIPT were based on the use of fetal-specific markers, which were easily distinguishable in maternal circulation, as they were fetal-specific. Such markers were Y-chromosome-specific loci for fetal sex determination, such as DYS14 [1,20], as well as fetal Rhesus D found in maternal circulation in pregnancies in which the mother was Rhesus D negative [21,22]. These methods were readily and rapidly introduced in the clinical setting of diagnostic laboratories worldwide [23], and within a few years, the field of NIPT evolved even further with the use of Y-chromosome-specific markers or paternally inherited polymorphic loci for the NIPT of X-linked inherited diseases, as well as through the identification of fetal-specific chromosomal translocations [24] and trinucleotide repeats in muscular dystrophy (DMPK) [25].

The above successful developments relied on the presence or absence of a fetal-specific marker. However, further developments and advances were needed for the identification of fetal specific-markers that are independent of gender and polymorphic sites and would allow direct discrimination of the free fetal DNA from the free maternal DNA [23,26]. The challenge of the field was the development of NIPT for the detection of chromosomal aneuploidies in the fetus. The need for the identification of fetal-specific markers that would enable the discrimination of a diploid pregnancy from an aneuploid pregnancy was urgent, because aneuploidies are among the most frequent fetal abnormalities, the most common of which are trisomy 21, trisomy 18, trisomy 13 and aneuploidies associated with chromosomes X and Y [23,27]. Major efforts took place from a number of independent research groups towards the NIPT of the most common chromosomal aneuploidies [23,26,28]. One such area that was extensively investigated was epigenetic modifications during development and how such changes could be taken into consideration for the identification of methylation fetal-specific markers that could potentially be used for the development of NIPT of fetal chromosomal abnormalities. In this review, we describe, compare and evaluate the different epigenetic-based approaches that have been implemented in the field of NIPT of fetal aneuploidies.

2. DNA Methylation in Fetal Development

DNA methylation is an enzymatic chemical modification of the genome, which includes the addition of a methyl group to the carbon-5 position of the cytokines of CpG dinucleotides [29].

The methylation pattern of the cell is reset during embryogenesis, and it is established early during development [30,31]. After its establishment, the methylation pattern is inherited from one cell generation to the next [29]. The methylation occurs in CpG dinucleotides non-uniformly

distributed in the genome. In contrast, areas rich in CpG dinucleotides (CpG Islands) are usually found in promoter regions of genes, and the majority of them are presented as non-methylated [29]. It is estimated that the human genome consists of approximately 30,000 CpG islands, of which, a proportion of 50%–60% lies within promoters [32]. Although the majority of these sequences are non-methylated, the CpG islands of imprinted genes and the inactive X chromosome are predominantly methylated [33].

DNA methylation is a dynamic process and may change during the post-developmental stage [34]. It is believed that 60% of tissue-specific differentially methylated regions (TDMRs) are methylated in embryonic cells, while during the differentiation of embryonic tissues to adult tissues, they undergo de-methylation [35–39]. More recent studies confirm the above, indicating that some of the methylated TDMRs undergo de-methylation in embryonic cells during the transformation into adult tissues, while a large proportion remains methylated in newborn tissues [40]. Therefore, the de-methylation of TDMRs occurs at a later developmental stage. In addition, the results indicated that specific regions of the genome show a different methylation pattern in different tissues and at different stages of development. The above findings provided convincing evidence that fetal DNA will present different methylation patterns from the methylation pattern of the maternal DNA.

Several independent research groups argued that methylation patterns are different between different tissues [41–44]. In 2008, a team of researchers led by Beck implemented a newly developed methodology known as MeDIP (methylated DNA immunoprecipitation), which was used in combination with whole genome microarray technologies to investigate the methylation status of all known promoter regions and CpG islands in different tissues [44]. Based on the above study, the phenomenon of CpG islands' methylation in normal cells and their contribution to normal cellular functions is more frequent than ever anticipated.

An epigenetic modification is a dynamic process and has been proven to play a very important role in the development of cancer cells [45,46]. More interestingly, the identification of tumor-specific DNA methylation patterns in the plasma of patients has led to great efforts towards the non-invasive diagnosis of cancer [47,48]. These developments in the field of cancer investigation have provided additional convincing support that epigenetic differences may be present between the fetal DNA and the maternal DNA in maternal circulation during pregnancy.

3. DNA Methylation Biomarkers Discovery

The aim of DNA methylation-based approaches was first to identify fetal-specific methylation markers that would allow the discrimination of fetal DNA from the maternal DNA in maternal circulation and that have the potential to be developed into non-invasive prenatal diagnostic markers. The approaches that have been used for investigating the DNA methylation patterns in fetal DNA and maternal DNA are of three main categories: sodium bisulfite-based approaches, restriction enzyme-based approaches and methylated DNA immunoprecipitation-based approaches.

3.1. Sodium Bisulfite-Based Approaches

Sodium bisulfite conversion leads to the transformation of an epigenetic modification into a genetic sequence change for further investigation. More specifically, the treatment of DNA with sodium bisulfite results in the conversion of unmethylated cytosines to uracils, leaving methylated cytosines unchanged [49]. The genetic composition of the converted sequences of interest could be investigated using methylation-specific PCR (MSP) in which the amplification process is separate for the methylated (non-converted) fragments and the non-methylated (converted) fragments [50]. Alternatively, the methylation status of bisulfite converted sequences could be assessed through the implementation of sequencing technologies [49,51]. In 2002, Poon and his colleagues demonstrated for the first time the potential for the presence of epigenetic differences between the fetus and the mother by performing sodium bisulfite conversion of placental DNA and female peripheral blood DNA followed by MSP [50,52]. The first differentially methylated region was identified in 2005 by the use of sodium bisulfite conversion in combination with MSP and sequencing. The differentially methylated gene, known as *SERPINB5*, was found to be hypomethylated in fetal DNA and hypermethylated in maternal DNA [12]. The identification of hypomethylated fetal-specific *SERPINB5* sequences was also achieved in maternal plasma during pregnancy. This genomic region was used to demonstrate that fetal DNA is not detectable in maternal plasma 24 h after delivery [28].

Since then, great efforts have taken place from independent groups towards the identification of fetal-specific methylation markers. The initial attempts were based on the investigation of promoter regions and CpG islands. In 2008, a bisulfite based systematic search for placental DNA methylation markers on chromosome 21 was described. In this study, the methylation-sensitive single nucleotide extension (Ms-SNuPE) method was used to assess the methylation differences of CpG sites [53,54]. The above study performed an evaluation of the methylation status of 114 CpG islands (based on bioinformatics criteria) in five first trimester placental tissues and two samples of non-pregnant female blood. Among them, 22 CpG islands were identified as having the potential to be developed into biomarkers for the NIPT of trisomy [54].

In 2010, a second study was performed with the aim of identifying a panel of fetal-specific hypermethylated markers on chromosome 21, and it used the methylation pattern of a previously characterized gene, *RASSF1A*. The *RASSF1A* gene is located on chromosome 3 and has been found to be completely methylated in fetal DNA and completely unmethylated in maternal DNA. This characteristic allowed the use of the *RASSF1A* gene as a fetal universal marker [28,55]. The study was performed using the combined bisulfite restriction analysis (COBRA) [56] to investigate 35 gene promoter regions on chromosome 21. The analysis demonstrated that the *HLCS* gene located on chromosome 21 is fully methylated in placenta and unmethylated in maternal blood cells [15].

A recent report published in 2013 illustrates the potential of retrieving the methylation profiles of placental tissues and maternal blood cells using sodium bisulfite in combination with next generation sequencing technologies [10]. The investigators were able to retrieve the fetal methylome through the identification of single nucleotide polymorphism (SNP) genotype differences between the mother and the fetus in maternal plasma and to identify differentially

methylated regions (DMRs). They identified 44,455 loci as being fetal-specific hypomethylated and 3081 regions as being fetal-specific hypermethylated. The above findings are in agreement with previous studies in which it was clearly evident that the fetal genome is mostly hypomethylated in contrast to the adult peripheral blood, which is greatly hypermethylated, indicating a regulatory role of the methylation patterns and gene expression profiles [44,57,58]. Interestingly, it has also been reported that hypomethylated sequences tend to be of a smaller fragment size. These findings could indicate a contribution of the fetal methylation status to the small fetal DNA fragments size in maternal plasma [10].

3.2. Restriction Enzyme-Based Approaches

Methylation patterns of CG dinucleotides can also be assessed using restriction enzymes, which have recognition sites containing CG sequences. Methylation-sensitive restriction enzymes can digest their recognition site only when unmethylated, whereas methylation insensitive restriction enzymes digest their recognition sites only when the cytokines of the CGs within their recognition site are methylated. In 2007, the team headed by Old reported for the first time the investigation and identification of a panel of differentially methylated regions on chromosome 21 using methylation-sensitive enzymes [59]. More specifically, the team used the HpaII enzyme, and the underlying idea was based on the fact that the enzyme would digest only the unmethylated type of its recognition site (CCGG). Therefore, this would allow them to identify regions containing the above recognition sites, which are differentially methylated between placenta and maternal blood cells. The study was focused on the investigation of promoters from highly expressed genes, randomly selected promoters, as well as randomly selected non-promoter regions. Among the 200 pre-selected regions, three promoter regions of the *AIRE*, *SIM2* and *ERG* genes were found to be methylated in the placenta and unmethylated in the maternal blood cells. The methylation status of those regions was confirmed by sodium bisulfite followed by MSP [59].

In 2011 a study performed by Peters and his team demonstrated that the use of methylation-based restriction enzymes, such as HpaII and MSpI, in combination with high-resolution arrays can distinguish differentially methylated regions between the placenta and maternal blood cells [58]. They presented a large panel of DMRs consisting of 6311 DMRs across chromosomes 13, 18 and 21 [58,60] and demonstrated that the fetal DNA is mostly hypomethylated, whereas the maternal blood cells are mostly hypermethylated, findings which are in agreement with previous reports [44,57]. Moreover, they illustrated that the majority of the hypomethylated regions of both fetal and maternal origin are located within CpG islands, promoters and exons, indicating a potential correlation with expression profiles [58].

3.3. Methylated DNA Immunoprecipitation-Based Approaches

One of the most modern methods of studying the levels of DNA methylation is the MeDIP (methylated DNA immunoprecipitation) approach. The method was first described in 2005 by Weber *et al*. with the aim of investigating the methylation pattern of cancer cells in a genome-wide fashion using microarray platforms [45]. In 2007, Beck and his team introduced linker-mediated

PCR amplification (LM-PCR) in combination with the MeDIP methodology. They obtained large amounts of immunoprecipitated DNA and generated the first whole genome mammalian methylome using a large panel of different tissues [44,61]. The principles of the MeDIP methodology includes fragmentation of the DNA (through sonication or enzymatic digestion) into short DNA fragments of 300–1000 bp. The sample is denatured and incubated with a monoclonal antibody, which recognizes and attaches to the 5-methylcytosines of CpG dinucleotides. Immunoprecipitation of methylated sequences is accomplished with the addition of magnetic beads. Through the implementation of the MeDIP methodology, you can achieve direct enrichment of methylated fragments. Enrichment of methylated target sequences is easily retrieved through the use of a large number of different technologies, such as PCR, qPCR (quantitative Polymerase Chain Reaction), microarray and sequencing. Since its development, MeDIP has been extensively used for the investigation of the methylation status/patterns of cancer tissues with great success either in combination with microarray technologies (MeDIP-chip) [42,44,45] or, more recently, in conjunction with next generation sequencing (MeDIP-seq) [62–65].

The MeDIP methodology was first introduced to the field of NIPT by our team in 2009 with the aim of investigating and identifying DMRs between placenta and female peripheral blood towards the development of NIPT for the identification of common aneuploidies [57]. Our team used MeDIP in combination with chromosome-specific high-resolution oligo arrays for the investigation of the methylation pattern of chromosomes 13, 18, 21, X and Y. Although previous studies solely investigated promoter regions and CpG islands for DMR identification, we were the first to screen entire chromosomes of interest irrespective of the genomic position or CG content. At the time, we reported the largest panel of DMRs with the potential to be developed into NIPT biomarkers for the most common fetal aneuploidies. More specifically, we identified around 2000 DMRs on each of the chromosomes investigated, and interestingly, we noticed that the vast majority of the DMRs were located within non-genic regions and in relatively poor CG regions. More specifically, we illustrated that 56%–83% of the DMRs were located within non-genic regions, whereas only 1%–11% were located within CpG islands. Our findings were concordant with previous studies performed by other groups investigating a panel of different tissues [44] and were also in agreement with more recent reports using bisulfite sequencing technologies [3,58]. We were also able to report the presence of inter-individual variability and the changes in the methylation patterns during the progression of the pregnancy, findings which have recently been confirmed by independent groups [10].

Following our study, the group headed by Chim used MeDIP in combination with a microarray platform targeting promoter regions and CpG islands. The group identified a panel of eight DMRs with the potential of being developed into biomarkers for diagnostic purposes [66], most of which are among the DMRs identified previously by our group [57]. Any discrepancies reported regarding the identification of DMRs, such as the failure to have the exact same methylation status of all DMRs reported by independent studies, are not uncommon, since different platforms and different methylation-based technologies were used.

4. Implementation of Methyl-Biomarkers in NIPT

The discovery of DMRs has mainly been focused on chromosomes 13, 18, 21, X and Y with the aim of identifying as a priority methylation-based biomarkers (methyl-biomarkers) suitable for the development of NIPT for the most common chromosomal fetal aneuploidies. The first attempt was reported back in 2006 for the NIPT of trisomy 18 (Edward's syndrome) [67]. In this study, the authors implemented a combination of sodium bisulfite conversion with MSP using maternal plasma samples from normal and trisomy 18 pregnancies. To achieve discrimination, they used the information of an SNP located within the *SERPINB5* gene. The cases were considered informative if the SNP was homozygous in the mother and heterozygous in the fetus, and only those cases could be used for NIPT of trisomy 18 (T18). To achieve this, the team introduced the so-called epigenetic allelic ratio (EAR) in which the chromosome 18 copy number was assessed based on the allele ratio calculation of an informative SNP. The challenge in this study was to have informative SNPs, and because there was only a single SNP in the target sequence, it was extremely difficult to be informative in all cases tested (Table 1). The results showed that among the 173 euploid placentas and 14 trisomy18 placentas genotyped for the polymorphism, only 31 and seven placentas, respectively, were informative. The rarity of having an informative SNP in this study does not allow this approach to be implemented population-wide [23,26].

To overcome the above limitations, in 2010, the same group developed an SNP-free methylation-based assay for NIPT of trisomy 21 (Down syndrome). Methylation-sensitive restriction digestion was used followed by digital PCR to investigate DMRs identified on chromosome 21 [15]. The copy number of chromosome 21 was determined through the epigenetic-genetic (EGG) chromosome dosage approach using the fetal-specific hypermethylated promoter region of the *HLCS* gene located on chromosome 21 and the *ZFY* locus on chromosome Y. The assay tested 24 maternal plasma samples from euploid pregnancies and five maternal plasma samples from trisomy 21 pregnancies. All but one euploid pregnancy were correctly classified (Table 1) [15].

The EGG chromosome dosage approach was also implemented for the NIPT of trisomy 18 in which the fetal-specific methylated *VAPA-APCDD1* loci on chromosome 18 and the *ZFY* on chromosome Y were quantified with digital PCR after HinP1I- and HpaII sample digestion [66]. The study was performed on nine maternal plasma samples from male trisomy 18 pregnancies and 27 maternal plasma samples from male euploid pregnancies. Among them, eight out of nine and one out of 27 trisomy 18 and euploid pregnancies, respectively, were correctly identified, which corresponds to 88.9% sensitivity and 96.3% specificity (Table 1) [66].

Table 1. Comparison of different methylation-based approaches towards the non-invasive prenatal tests (NIPT) of aneuploidies. EAR, epigenetic allelic ratio (EAR); EGG, epigenetic-genetic; SNP, single nucleotide polymorphism.

Assay	Technology	Sample size	Sensitivity/Specificity (%)	Advantages	Disadvantages	Reproduced by others
EAR on chromosome 18 [67]	Sodium bisulfite, digital PCR	2 normal 2 trisomy 18	Not defined/not applicable population-wide	Applicable irrespective of gender	Requires informative SNP, depends on the bisulfite conversion performance	No
EGG on chromosome 21 using *ZFY* [15]	* COBRA, digital PCR	24 normal 5 trisomy 21	95.8% specificity 100% sensitivity	SNP-free assay	Applicable only to male pregnancies, depends on the digestion and bisulfite conversion efficiency	No
EGG on chromosome 18 using *ZFY* [66]	* COBRA, digital PCR	27 normal 9 Trisomy 18	96.3% specificity 88.9% sensitivity	SNP-free assay	Applicable only to male pregnancies, depends on the digestion and bisulfite conversion efficiency	No
EGG on chromosome 21 using *TMED8* [68]	** MRED digestion, digital PCR	33 normal 14 trisomy 21	Variable depending on the fetal allele	Applicable irrespective of gender	Requires informative SNP, applicable only to male pregnancies, depends on the digestion efficiency	No
Fetal-specific DNA methylation ratio on chromosome 21 (1st study) [69]	*** MeDIP, real-time qPCR	40 normal 40 trisomy 21	100% specificity 100% sensitivity	Applicable irrespective of gender and SNPs	Depends on MeDIP performance	Yes [70,71]
Fetal-specific DNA methylation ratio on chromosome 21 (2nd study) [72]	*** MeDIP, real-time qPCR	125 normal 50 trisomy 21	99.2% specificity 100% sensitivity	Applicable irrespective of gender and SNPs	Depends on MeDIP performance	No
Bisulfite sequencing [10]	Sodium bisulfite, next generation sequencing	7 normal 5 trisomy 21	100% specificity 100% sensitivity	Applicable irrespective of gender and SNPs	Depends on bisulfite conversion efficiency	No

* Combined bisulfite restriction analysis; ** methylation restriction enzymatic digestion; *** methylated DNA immunoprecipitation.

Although the results from the studies using the EGG chromosome dosage approach were promising, the technology was restricted to male pregnancies, because the EGG calculation involved the use of the *ZFY* gene (Table 1). To overcome the above difficulties, a modification was introduced in the EGG calculation to be able to include the testing of female pregnancies, as well. The study was performed using 14 maternal plasma from trisomy 21 pregnancies and were compared to 33 cases with a euploid fetus [68]. For calculation purposes, the *ZFY* gene was replaced with an autosomal genetic reference marker. Interpretation of the results was achieved using a paternally-inherited SNP allele on the *TMED8* gene located on chromosome 14, which served as a baseline for the EGG chromosome dosage calculation. The sensitivity of the assay varied depending on which of the two alleles of an SNP was fetal-specific, making the evaluation of the assay performance even more challenging. Overall, although the limitation of testing only male pregnancies was overcome, the assessment of the copy number of chromosome 21 remained a challenge, as the presence of at least one informative SNP was necessary (Table 1).

A different approach was proposed by our group in 2011 and was based on using the MeDIP methodology in combination with real-time quantitative PCR (real time-qPCR) for the quantification of selected DMRs located on chromosome 21 [69]. We selected 12 previously identified DMRs located on chromosome 21 [57], which were hypermethylated in fetal DNA and hypomethylated in female peripheral blood cells. We used in our study a total of 40 maternal blood samples from euploid pregnancies and 40 maternal blood samples from trisomy 21 cases. We developed a diagnostic formula by calculating the DNA methylation ratio of the selected DMRs using 20 normal pregnancies and 20 trisomy 21 pregnancies. Eight specific DMRs were the most statistically significant markers in discriminating normal from trisomy 21 pregnancies. The MeDIP-qPCR methodology was used to then test 40 additional pregnancies, of which 20 were obtained from trisomy 21 pregnancies and showed 100% specificity and 100% sensitivity [69]. We also demonstrated that diagnostic accuracy can only be achieved through the combination of multiple DMRs from chromosome 21, which was an important finding for further NIPT developments [23].

Our team continued to improve the MeDIP-qPCR assay with a larger validation study of 175 pregnancies that included 50 trisomy 21 pregnancies [72]. In this larger-scale validation, we re-evaluated our diagnostic assay, taking into consideration the genomic composition of our DMRs and by selectively excluding those DMRs located in copy number variable (CNV) regions. Based on the above, we re-designed our diagnostic formula and then evaluated its performance using 100 new cases, which included 25 trisomy 21 pregnancies. The results demonstrated 100% sensitivity and 99.2% specificity (Table 1) [72]. Our group also investigated whether the variability of the fetal fraction present in maternal plasma has a negative effect in our assay's diagnostic efficiency. Although previous reports demonstrated an effect of different fetal amounts present in maternal plasma [73–75], our study has shown no significant association between cffDNA fraction, absolute fetal amount or the concentration present in maternal plasma with the test result classification using our diagnostic formula [20,72]. We speculate that this is due to the fact that maternal blood contains <1% of fetal DNA [20,72] in contrast to maternal plasma, which contains ~10%–15% fetal DNA [10,76].

More importantly, the results of our studies have been reproduced by two independent groups, which have reported their results using the MeDIP-qPCR methodology and the published diagnostic formula [70,71]. In addition, independent groups have also commented positively on the potential prospects or application of the MeDIP-qPCR assay towards the NIPT of chromosomal aneuploidies. The low cost of the technology and the ease of implementing it, in combination with the use of equipment common to every laboratory, allows its implementation in any diagnostic laboratory setting [77]. A major strength of the MeDIP-qPCR assay is that it is a gender- and polymorphism-independent assay that could be implemented population-wide. Nevertheless, a different independent group has failed to reproduce the MeDIP-qPCR results by performing a small scale validation study [78]. Lack of reproducibility of the results would not be a surprise to our team, since, as stated in our reply to the above manuscript, very stringent quality control criteria must be applied to critical reagents and conditions throughout the method [79].

A very interesting recent development of investigating DNA methylation for use in NIPT has been the implementation of sodium bisulfite DNA treatment in combination with next generation sequencing technologies (NGS) [10]. The study is presented as a proof of principle and demonstrates one use of the assay with the detection of trisomy 21. NGS technologies have already been introduced in the field of NIPT by different independent groups with the primary aim of detecting the most common chromosomal aneuploidies [73–76,80–82]. Biotechnology companies have already introduced in the market their NGS-based NIPT of the most common chromosomal fetal aneuploidies [83–85]. However, sequencing of maternal plasma can turn out to be very challenging, due to the restrictions of the very low amount of fetal DNA available. Furthermore, such technology is not yet available in all clinical laboratories. Sequencing technologies are still considered to be of a high cost, requiring significant infrastructure, are labor intensive and require highly trained personnel, and the bioinformatics analysis can be very challenging, especially when the target sequence is of a very low amount, such as fetal DNA present in maternal plasma.

5. Evaluating the Efficiency of Methylation Assays

Developments towards methyl-biomarker discovery and their applications in the NIPT of fetal chromosomal abnormalities were achieved through a number of independent groups, as described above, using different methylation-based approaches. Different analytical tools and a variety of quantitative approaches (e.g., MSP, digital PCR, real-time qPCR, microarray platforms and next generation sequencing) were used, of which the statistical power in discriminating normal from abnormal pregnancies has been extensively assessed [23,26,86]. Nevertheless, the statistical discriminating power of each of the end point analytical tools relies on the efficiency of the methylation-based technology used to enrich the fetal DNA in maternal circulation (Table 1). Therefore, the evaluation and assessment of the efficiency of the methylation-based enrichment technology used is of significant importance.

One of the most commonly used approaches is the treatment of DNA with sodium bisulfite. Sodium bisulfite conversion is considered the gold standard in the evaluation of the methylation status of different tissues and has been extensively used, especially in the field of cancer [87,88]. However, it is well known that this chemical treatment of the DNA is associated with a high degree

of DNA degradation, reaching >90% of the template DNA [89]. This major drawback of the technology is undesirable for its implementation in plasma samples of pregnant women. During pregnancy, the amount of fetal DNA in maternal plasma is very low [10,76], and further degradation will result in even fewer fetal DNA molecules available for quantification; therefore, the accuracy and sensitivity of the test will be reduced. To compensate for the degradation effect, much larger amounts of maternal plasma are required, which makes the testing of maternal plasma even more complicated. Furthermore, bisulfite conversion can be challenging, since 100% conversion of the unmethylated cytosines to uracils is rarely achieved, and purification is required to remove the sodium bisulfite [90]. Such an effect will bias the correct interpretation of the results [23]. On the other hand, bisulfite conversion strategies are not sensitive to low purity and low integrity samples, an advantage especially for samples with very low starting DNA amounts. Nevertheless, bisulfite conversion in combination with sequencing technologies can provide a comprehensive analysis of the methylation status at the base pair composition, which can make it a very powerful tool (Table 2).

Table 2. Comparison of different methylation assays.

Methylation assay	Advantages	Disadvantages	Analytical tool used for NIPT
Sodium bisulfite	Not sensitive to sample impurities, methylation analysis at the base pair level	DNA degradation (>90%), 100% conversion is rarely achieved	* MSP, microarrays, Digital PCR, ** COBRA, *** NGS
Restriction enzyme digestion	Easy to perform and low cost	Sensitive to sample impurities, requires high amount of starting DNA, applicable to a limited number of DNA sequences	** COBRA, digital PCR
**** MeDIP	Ideal for investigating low CG content regions, low cost assay, not sensitive to sample impurities, can be applied with low starting DNA amounts	Depends on antibody efficiency and ideal combination of affinity reagents	Real time-qPCR, microarrays

* Methylation-specific PCR; ** combined bisulfite restriction analysis; *** next generation sequencing; **** methylated DNA immunoprecipitation.

A different approach implemented by a number of independent groups towards methyl-biomarker discovery and methylation-based NIPT developments has been the use of methylation restriction enzymes, as described above. Through methylation restriction enzymatic digestions (MRED), the unmethylated maternal origin sequences, present in maternal plasma, are digested to achieve indirect enrichment for the corresponding sequences of fetal origin, which are methylated. The efficiency of the MRED assays depends on the purity of the sample, and for this reason, they require high purity and high integrity samples [90]. Additionally, MRED assays require fairly high quantities of starting material, which is a restriction to its implementation in plasma samples, because not only the target fetal DNA sequences are of a low amount, but also the total plasma DNA is very low (around 10 ng/4 mL plasma) [20]. An additional drawback of the assay is that it can only evaluate the methylation status of a specific and very limited number of genomic sequences. Only those sequences that include a recognition site of a methylation-dependent

restriction enzyme could be evaluated. Such inherent restrictions do not allow efficient and detailed genome-wide methylation assessment [23,26]. An example is the recognition sites of the HpaII restriction enzyme, which are presented in only 3.9% of CGs across non-repetitive sequences of the human genome [91]. Moreover, the efficiency of digestion should always be carefully evaluated for an unbiased interpretation of the results. Nevertheless, it is a very easy to perform assay and low cost.

The MeDIP assay, an affinity-enrichment method, was also utilized towards DMR identification and characterization to discriminate fetal DNA from maternal DNA in maternal circulation during pregnancy. Based on studies performed by several independent groups, it is clearly evident that the vast majority of DMRs identified between different tissues are located within non-genic and CG poor regions [44,58]. Based on recent reports, the MeDIP methodology is ideal for the investigation of low CpG density regions [92]. Indeed, the DMRs identified and selected for NIPT of trisomy 21 using MeDIP-qPCR are located in low CpG sites and are mostly found within non-genic regions [57,69,72]. Therefore, we strongly feel that MeDIP is the choice of selection for the investigation of DMRs towards NIPT. MeDIP is an efficient method for genome-wide methylation assessment [42,44,45], as it can evaluate the methylation levels irrespective of genomic composition and overcomes limitations of the previously described methodologies. The MeDIP assay can tolerate sample impurities, and thus, no prior sample purification is required. Furthermore, it has recently been proven to be applicable for low starting DNA templates, generating sufficiently enriched outputs [64,65], a development that simplifies and makes possible its implementation with plasma samples. Moreover, it is a technically robust methodology, easy to use and affordable. Nevertheless, the efficiency and performance of MeDIP greatly depends on determining the ideal combination of affinity reagents. This is very important, especially in regions with varying methylcytosine density, such as the DMRs identified for the NIPT of common chromosomal aneuploidies [57,69,72]. The advantages and disadvantages of all the different methylation-based assays implemented towards the NIPT of fetal chromosomal abnormalities are summarized in Table 2.

6. Conclusions and Future Directions

Deciphering the epigenome and understanding the underlying mechanisms that lead to epigenetic modifications has been one of the most interesting fields under investigation for the last decade. Since 2002, a large panel of DMRs has been identified by independent groups, with the potential of being developed into diagnostic markers having as a primary goal the development of NIPT for common fetal chromosomal abnormalities.

We speculate that epigenetic approaches towards NIPT will soon dominate the field of NIPT, because they are easy to perform, are fast and inexpensive compared to existing NIPT approaches, which are based on next generation sequencing technologies [73–75,81,82]. We speculate that one of the first epigenetic-based approaches that will be launched for the NIPT of common chromosomal fetal abnormalities will be a MeDIP-based approach. NIPD Genetics Ltd., a company in which three of the authors are employed, is dedicated to developing a MeDIP-qPCR-based

diagnostic assay. The company will be soon ready to launch the first epigenetic-based NIPT for trisomy 21 following completion of a large-scale validation study [23,72,93].

Methylation-based approaches could also be used for retrieving the methylation status of abnormal tissues, such as placental tissues from aneuploid pregnancies. A very recent study has shown that trisomy 21 placentas are characterized by a global hypermethylation in contrast to normal placentas, which are mainly hypomethylated [94]. Identifying such disease-associated characteristics can benefit and contribute to more robust and sensitive NIPT. Furthermore, methylation differences during fetal development have also been shown to be associated with transcription. It has been demonstrated that the early gestational placental methylome is significantly associated with gene expression [58]. Such structural and regulatory characteristics of the placental epigenome are of great importance and could be used to determine the role of aberrant or altered methylation in placental dysfunction.

In addition to the methods described in this review, the implementation of alternative methylation-based approaches, such as MBD (methylated binding domain) [92] and McrBC (a GTP-requiring, modification-dependent endonuclease of *Escherichia coli* K-12) fragmentation, as well as HELP (HpaII tiny fragment enrichment by ligation-mediated PCR) [95,96], in combination with the development of bioinformatics-based algorithms, will contribute to a better understanding of the fetal methylome. We envision that epigenetic-based enrichment methods will have a major contribution to fetal methylome analysis through direct testing of maternal plasma. Looking ahead, we predict that epigenetic-based approaches in combination with genetic-based approaches and advanced technological approaches, such as digital PCR and next generation sequencing, will contribute to the development of NIPT of more subtle fetal abnormalities, such as point mutations, microdeletion/microduplication syndromes, *etc*.

Acknowledgments

The work performed by the author's laboratories is supported by the Cyprus Institute of Neurology and Genetics, NIPD Genetics Ltd., EU 7th Framework Programme, as part of the ANGELAB (A New GEnetic LABoratory) project (#317635), and the European Research Council (ERC), as part of the European Research Council program, ERC-2012-AdG 322953-NIPD.

Author Contributions

Designed the structure and content of the manuscript: EAP, PCP. Contributed materials for writing the manuscript: GK, EK, MH. Wrote the manuscript: EAP, PCP.

Conflicts of Interest

The authors have filed patent applications on aspects on the use of free-fetal DNA in maternal circulation for non-invasive prenatal diagnosis.

References

1. Lo, Y.M.; Corbetta, N.; Chamberlain, P.F.; Rai, V.; Sargent, I.L.; Redman, C.W.; Wainscoat, J.S. Presence of fetal DNA in maternal plasma and serum. *Lancet* **1997**, *350*, 485–487.
2. Lo, Y.M.; Tein, M.S.; Lau, T.K.; Haines, C.J.; Leung, T.N.; Poon, P.M.; Wainscoat, J.S.; Johnson, P.J.; Chang, A.M.; Hjelm, N.M. Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *Am. J. Hum. Genet.* **1998**, *62*, 768–775.
3. Lun, F.M.; Chiu, R.W.; Chan, A.K.C.; Yeung Leung, T.; Kin Lau, T.; Lo, D.Y.M. Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma. *Clin. Chem.* **2008**, *54*, 1664–1672.
4. Partsalis, T.; Chan, L.Y.; Hurworth, M.; Willers, C.; Pavlos, N.; Kumta, N.; Wood, D.; Xu, J.; Kumta, S.; Lo, Y.M.; *et al*. Evidence of circulating donor genetic material in bone allotransplantation. *Int. J. Mol. Med.* **2006**, *17*, 1151–1155.
5. Bianchi, D.W.; Shuber, A.P.; DeMaria, M.A.; Fougner, A.C.; Klinger, K.W. Fetal cells in maternal blood: determination of purity and yield by quantitative polymerase chain reaction. *Am. J. Obstet. Gynecol.* **1994**, *171*, 922–926.
6. Lo, Y.M.; Lau, T.K.; Chan, L.Y.; Leung, T.N.; Chang, A.M. Quantitative analysis of the bidirectional fetomaternal transfer of nucleated cells and plasma DNA. *Clin. Chem.* **2000**, *46*, 1301–1309.
7. Alberry, M.; Maddocks, D.; Jones, M.; Abdel Hadi, M.; Abdel-Fattah, S.; Avent, N.; Soothill, P.W. Free fetal DNA in maternal plasma in anembryonic pregnancies: Confirmation that the origin is the trophoblast. *Prenat. Diagn.* **2007**, *27*, 415–418.
8. Tjoa, M.L.; Cindrova-Davies, T.; Spasic-Boskovic, O.; Bianchi, D.W.; Burton, G.J. Trophoblastic oxidative stress and the release of cell-free feto-placental DNA. *Am. J. Pathol.* **2006**, *169*, 400–404.
9. Smid, M.; Galbiati, S.; Lojacono, A.; Valsecchi, L.; Platto, C.; Cavoretto, P.; Calza, S.; Ferrari, A.; Ferrari, M.; Cremonesi, L. Correlation of fetal DNA levels in maternal plasma with Doppler status in pathological pregnancies. *Prenat. Diagn.* **2006**, *26*, 785–790.
10. Lun, F.M.; Chiu, R.W.; Sun, K.; Leung, T.Y.; Jiang, P.; Chan, K.C.; Sun, H.; Lo, Y.M. Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. *Clin. Chem.* **2013**, *59*, 1583–1594.
11. Lo, Y.M.; Zhang, J.; Leung, T.N.; Lau, T.K.; Chang, A.M.; Hjelm, N.M. Rapid clearance of fetal DNA from maternal plasma. *Am. J. Hum. Genet.* **1999**, *64*, 218–224.
12. Chim, S.S.; Tong, Y.K.; Chiu, R.W.; Lau, T.K.; Leung, T.N.; Chan, L.Y.; Oudejans, C.B.; Ding, C.; Lo, Y.M. Detection of the placental epigenetic signature of the maspin gene in maternal plasma. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 14753–14758.
13. Tsumita, T.; Iwanaga, M. Fate of injected deoxyribonucleic acid in mice. *Nature* **1963**, *198*, 1088–1089.
14. Emlen, W.; Mannik, M. Kinetics and mechanisms for removal of circulating single-stranded DNA in mice. *J. Exp. Med.* **1978**, *147*, 684–699.

15. Tong, Y.K.; Jin, S.; Chiu, R.W.; Ding, C.; Chan, K.C.; Leung, T.Y.; Yu, L.; Lau, T.K.; Lo, Y.M. Noninvasive prenatal detection of trisomy 21 by an epigenetic-genetic chromosome-dosage approach. *Clin. Chem.* **2010**, *56*, 90–98.
16. Smith, S.C.; Baker, P.N.; Symonds, E.M. Placental apoptosis in normal human pregnancy. *Am. J. Obstet. Gynecol.* **1997**, *177*, 57–65.
17. Chan, K.C.; Zhang, J.; Hui, A.B.; Wong, N.; Lau, T.K.; Leung, T.N.; Lo, K.W.; Huang, D.W.; Lo, Y.M. Size distributions of maternal and fetal DNA in maternal plasma. *Clin. Chem.* **2004**, *50*, 88–92.
18. Li, Y.; Zimmermann, B.; Rusterholz, C.; Kang, A.; Holzgreve, W.; Hahn, S. Size separation of circulatory DNA in maternal plasma permits ready detection of fetal DNA polymorphisms. *Clin. Chem.* **2004**, *50*, 1002–1011.
19. Kimura, M.; Hara, M.; Itakura, A.; Sato, C.; Ikebuchi, K.; Ishihara, O. Fragment size analysis of free fetal DNA in maternal plasma using Y-STR loci and SRY gene amplification. *Nagoya J. Med. Sci.* **2011**, *73*, 129–135.
20. Kyriakou, S.; Kypri, E.; Spyrou, C.; Tsaliki, E.; Velissariou, V.; Papageorgiou, E.A.; Patsalis, P.C. Variability of ffDNA in maternal plasma does not prevent correct classification of trisomy 21 using MeDIP-qPCR methodology. *Prenat. Diagn.* **2013**, *33*, 650–655.
21. Lo, Y.M.; Hjelm, N.M.; Fidler, C.; Sargent, I.L.; Murphy, M.F.; Chamberlain, P.F.; Poon, P.M.; Redman, C.W.; Wainscoat, J.S. Prenatal diagnosis of fetal RhD status by molecular analysis of maternal plasma. *N. Engl. J. Med.* **1998**, *339*, 1734–1738.
22. Daniels, G.; Finning, K.; Martin, P.; Summers, J. Fetal blood group genotyping: Present and future. *Ann. N. Y. Acad. Sci.* **2006**, *1075*, 88–95.
23. Papageorgiou, E.A.; Patsalis, P.C. Non-invasive prenatal diagnosis of aneuploidies: New technologies and clinical applications. *Genome Med.* **2012**, *4*, 46.
24. Chen, C.P.; Chern, S.R.; Wang, W. Fetal DNA analyzed in plasma from a mother's three consecutive pregnancies to detect paternally inherited aneuploidy. *Clin. Chem.* **2001**, *47*, 937–939.
25. Amicucci, P.; Gennarelli, M.; Novelli, G.; Dallapiccola, B. Prenatal diagnosis of myotonic dystrophy using fetal DNA obtained from maternal plasma. *Clin. Chem.* **2000**, *46*, 301–302.
26. Patsalis, P.C.; Tsaliki, E.; Koumbaris, G.; Karagrigoriou, A.; Velissariou, V.; Papageorgiou, E.A. A new non-invasive prenatal diagnosis of Down syndrome through epigenetic markers and real-time qPCR. *Exp. Opin. Biol. Ther.* **2012**, *12*, S155– S 161.
27. Driscoll, D.A.; Gross, S. Clinical practice. Prenatal screening for aneuploidy. *N. Engl. J. Med.* **2009**, *360*, 2556–2562.
28. Chiu, R.W.; Lo, Y.M. Non-invasive prenatal diagnosis by fetal nucleic acid analysis in maternal plasma: the coming of age. *Semin. Fetal Neonatal Med.* **2011**, *16*, 88–93.
29. Raedle, J.; Trojan, J.; Brieger, A.; Weber, N.; Schafer, D.; Plotz, G.; Staib-Sebler, E.;
30. Kriener, S.; Lorenz, M.; Zeuzem, S. Bethesda guidelines: Relation to microsatellite instability and MLH1 promoter methylation in patients with colorectal cancer. *Ann. Intern. Med.* **2001**, *135*, 566–576.
31. Bird, A.P. The relationship of DNA methylation to cancer. *Cancer Surv.* **1996**, *28*, 87–101.

32. Monk, M. Changes in DNA methylation during mouse embryonic development in relation to X-chromosome activity and imprinting. *Philos. Trans. R. Soc. Lond.* **1990**, *326*, 299–312.
33. Szyf, M. DNA methylation and demethylation as targets for anticancer therapy. *Biochemistry* **2005**, *70*, 533–549.
34. Costello, J.F.; Plass, C. Methylation matters. *J. Med. Genet.* **2001**, *38*, 285–303.
35. Reik, W.; Dean, W.; Walter, J. Epigenetic reprogramming in mammalian development. *Science* **2001**, *293*, 1089–1093.
36. Kawai, J.; Hirotsune, S.; Hirose, K.; Fushiki, S.; Watanabe, S.; Hayashizaki, Y. Methylation profiles of genomic DNA of mouse developmental brain detected by restriction landmark genomic scanning (RLGS) method. *Nucleic Acids Res.* **1993**, *21*, 5604–5608.
37. Watanabe, S.; Kawai, J.; Hirotsune, S.; Suzuki, H.; Hirose, K.; Taga, C.; Ozawa, N.; Fushiki, S.; Hayashizaki, Y. Accessibility to tissue-specific genes from methylation profiles of mouse brain genomic DNA. *Electrophoresis* **1995**, *16*, 218–226.
38. Shiota, K. DNA methylation profiles of CpG islands for cellular differentiation and development in mammals. *Cytogenet. Genome Res.* **2004**, *105*, 325–334.
39. Song, F.; Smith, J.F.; Kimura, M.T.; Morrow, A.D.; Matsuyama, T.; Nagase, H.; Held, W.A. Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 3336–3341.
40. Ching, T.T.; Maunakea, A.K.; Jun, P.; Hong, C.; Zardo, G.; Pinkel, D.; Albertson, D.G.; Fridlyand, J.; Mao, J.H.; Shchors, K.; *et al*. Epigenome analyses using BAC microarrays identify evolutionary conservation of tissue-specific methylation of SHANK3. *Nat. Genet.* **2005**, *37*, 645–651.
41. Song, F.; Mahmood, S.; Ghosh, S.; Liang, P.; Smiraglia, D.J.; Nagase, H.; Held, W.A. Tissue specific differentially methylated regions (TDMR): Changes in DNA methylation during development. *Genomics* **2009**, *93*, 130–139.
42. Eckhardt, F.; Lewin, J.; Cortese, R.; Rakyan, V.K.; Attwood, J.; Burger, M.; Burton, J.; Cox, T.V.; Davies, R.; Down, T.A.; *et al*. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* **2006**, *38*, 1378–1385.
43. Weber, M.; Hellmann, I.; Stadler, M.B.; Ramos, L.; Paabo, S.; Rebhan, M.; Schubeler, D. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **2007**, *39*, 457–466.
44. Illingworth, R.; Kerr, A.; Desousa, D.; Jorgensen, H.; Ellis, P.; Stalker, J.; Jackson, D.; Clee, C.; Plumb, R.; Rogers, J.; *et al*. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.* **2008**, *6*, e22.
45. Rakyan, V.K.; Down, T.A.; Thorne, N.P.; Flicek, P.; Kulesha, E.; Graf, S.; Tomazou, E.M.; Backdahl, L.; Johnson, N.; Herberth, M.; *et al*. An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res.* **2008**, *18*, 1518–1529.
46. Weber, M.; Davies, J.J.; Wittig, D.; Oakeley, E.J.; Haase, M.; Lam, W.L.; Schubeler, D. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **2005**, *37*, 853–862.

47. Jones, P.A.; Baylin, S.B. The epigenomics of cancer. *Cell* **2007**, *128*, 683–692.
48. Esteller, M.; Sanchez-Cespedes, M.; Rosell, R.; Sidransky, D.; Baylin, S.B.; Herman, J.G. Detection of aberrant promoter hypermethylation of tumor suppressor genes in serum DNA from non-small cell lung cancer patients. *Cancer Res.* **1999**, *59*, 67–70.
49. Lo, Y.M.; Wong, I.H.; Zhang, J.; Tein, M.S.; Ng, M.H.; Hjelm, N.M. Quantitative analysis of aberrant p16 methylation using real-time quantitative methylation-specific polymerase chain reaction. *Cancer Res.* **1999**, *59*, 3899–3903.
50. Frommer, M.; McDonald, L.E.; Millar, D.S.; Collis, C.M.; Watt, F.; Grigg, G.W.; Molloy, P.L.; Paul, C.L. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 1827–1831.
51. Herman, J.G.; Graff, J.R.; Myohanen, S.; Nelkin, B.D.; Baylin, S.B. Methylation-specific PCR: A novel PCR assay for methylation status of CpG islands. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 9821–9826.
52. Clark, S.J.; Harrison, J.; Paul, C.L.; Frommer, M. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.* **1994**, *22*, 2990–2997.
53. Poon, L.L.; Leung, T.N.; Lau, T.K.; Chow, K.C.; Lo, Y.M. Differential DNA methylation between fetus and mother as a strategy for detecting fetal DNA in maternal plasma. *Clin. Chem.* **2002**, *48*, 35–41.
54. Gonzalgo, M.L.; Jones, P.A. Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE). *Nucleic Acids Res.* **1997**, *25*, 2529–2531.
55. Chim, S.S.; Jin, S.; Lee, T.Y.; Lun, F.M.; Lee, W.S.; Chan, L.Y.; Jin, Y.; Yang, N.; Tong, Y.K.; Leung, T.Y.; *et al*. Systematic search for placental DNA-methylation markers on chromosome 21: Toward a maternal plasma-based epigenetic test for fetal trisomy 21. *Clin. Chem.* **2008**, *54*, 500–511.
56. Chan, K.C.; Ding, C.; Gerovassili, A.; Yeung, S.W.; Chiu, R.W.; Leung, T.N.; Lau, T.K.; Chim, S.S.; Chung, G.T.; Nicolaides, K.H.; *et al*. Hypermethylated RASSF1A in maternal plasma: A universal fetal DNA marker that improves the reliability of noninvasive prenatal diagnosis. *Clin. Chem.* **2006**, *52*, 2211–2218.
57. Xiong, Z.; Laird, P.W. COBRA: A sensitive and quantitative DNA methylation assay. *Nucleic Acids Res.* **1997**, *25*, 2532–2534.
58. Papageorgiou, E.A.; Fiegler, H.; Rakyan, V.; Beck, S.; Hulten, M.; Lamnissou, K.; Carter, N.P.; Patsalis, P.C. Sites of differential DNA methylation between placenta and peripheral blood: Molecular markers for noninvasive prenatal diagnosis of aneuploidies. *Am. J. Pathol.* **2009**, *174*, 1609–1618.
59. Chu, T.; Handley, D.; Bunce, K.; Surti, U.; Hogge, W.A.; Peters, D.G. Structural and regulatory characterization of the placental epigenome at its maternal interface. *PLoS One* **2011**, *6*, e14723.
60. Old, R.W.; Crea, F.; Puszyk, W.; Hulten, M.A. Candidate epigenetic biomarkers for non-invasive prenatal diagnosis of Down syndrome. *Reprod. Biomed. Online* **2007**, *15*, 227–235.

61. Chu, T.; Burke, B.; Bunce, K.; Surti, U.; Allen Hogge, W.; Peters, D.G. A microarray-based approach for the identification of epigenetic biomarkers for the noninvasive diagnosis of fetal disease. *Prenat. Diagn.* **2009**, *29*, 1020–1030.
62. Down, T.A.; Rakyan, V.K.; Turner, D.J.; Flicek, P.; Li, H.; Kulesha, E.; Graf, S.; Johnson, N.; Herrero, J.; Tomazou, E.M.; *et al*. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.* **2008**, *26*, 779–785.
63. Ruike, Y.; Imanaka, Y.; Sato, F.; Shimizu, K.; Tsujimoto, G. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC Genomics* **2010**, *11*, 137.
64. Feber, A.; Wilson, G.A.; Zhang, L.; Presneau, N.; Idowu, B.; Down, T.A.; Rakyan, V.K.; Noon, L.A.; Lloyd, A.C.; Stupka, E.; *et al*. Comparative methylome analysis of benign and malignant peripheral nerve sheath tumors. *Genome Res.* **2011**, *21*, 515–524.
65. Taiwo, O.; Wilson, G.A.; Morris, T.; Seisenberger, S.; Reik, W.; Pearce, D.; Beck, S.; Butcher, L.M. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat. Protoc.* **2012**, *7*, 617–636.
66. Borgel, J.; Guibert, S.; Weber, M. Methylated DNA immunoprecipitation (MeDIP) from low amounts of cells. *Methods Mol. Biol.* **2012**, *925*, 149–158.
67. Tsui, D.W.Y.; Lam, Y.M.D.; Lee, W.S.; Leung, T.Y.; Lau, T.K.; Lau, E.T.; Tang, M.H.Y.; Akolekar, R.; Nicolaides, K.H.; Chiu, R.W.K.; *et al*. Systematic Identification of Placental Epigenetic Signatures for the Noninvasive Prenatal Detection of Edwards Syndrome. *PLoS One* **2010**, *5*, e15069.
68. Tong, Y.K.; Ding, C.; Chiu, R.W.; Gerovassili, A.; Chim, S.S.; Leung, T.Y.; Leung, T.N.; Lau, T.K.; Nicolaides, K.H.; Lo, Y.M. Noninvasive prenatal detection of fetal trisomy 18 by epigenetic allelic ratio analysis in maternal plasma: Theoretical and empirical considerations. *Clin. Chem.* **2006**, *52*, 2194–2202.
69. Tong, Y.K.; Chiu, R.W.; Akolekar, R.; Leung, T.Y.; Lau, T.K.; Nicolaides, K.H.; Lo, Y.M. Epigenetic-genetic chromosome dosage approach for fetal trisomy 21 detection using an autosomal genetic reference marker. *PLoS One* **2010**, *5*, e15244.
70. Papageorgiou, E.A.; Karagrigoriou, A.; Tsaliki, E.; Velissariou, V.; Carter, N.P.; Patsalis, P.C. Fetal specific DNA methylation ratio permits non-invasive prenatal diagnosis of trisomy 21. *Nat. Med.* **2011**, *17*, 510–513.
71. Gorduza, E.V.; Popescu, R.; Caba, L.; Ivanov, I.; Martiniuc, V.; Nedelea, F.; Militaru, M.; Socolov, D.G. Prenatal diagnosis of 21 trisomy by quantification of methylated fetal DNA in maternal blood: Study on 10 pregnancies. *Rev. Rom. Med. Lab.* **2013**, *21*, 275–284.
72. Qin, H.; Bonifacio, M.; McArthur, S.; McLennan, A.; Boogert, T.; Bowman, M. Comment on "MeDIP real-time qPCR of maternal peripheral blood reliably identifies trisomy 21". *Prenat. Diagn.* **2013**, *33*, 403.
73. Tsaliki, E.; Papageorgiou, E.A.; Spyrou, C.; Koumbaris, G.; Kypri, E.; Kyriakou, S.; Sotiriou, C.; Touvana, E.; Keravnou, A.; Karagrigoriou, A.; *et al*. MeDIP real-time qPCR of maternal peripheral blood reliably identifies trisomy 21. *Prenat. Diagn.* **2012**, *32*, 996–1001.

74. Chiu, R.W.; Akolekar, R.; Zheng, Y.W.; Leung, T.Y.; Sun, H.; Chan, K.C.; Lun, F.M.; Go, A.T.; Lau, E.T.; To, W.W.; *et al*. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. *BMJ* **2011**, *342*, c7401.
75. Ehrich, M.; Deciu, C.; Zwiefelhofer, T.; Tynan, J.A.; Cagasan, L.; Tim, R.; Lu, V.; McCullough, R.; McCarthy, E.; Nygren, A.O.; *et al*. Noninvasive detection of fetal trisomy 21 by sequencing of DNA in maternal blood: a study in a clinical setting. *Am. J. Obstet. Gynecol.* **2011**, *204*, 205.e1–205.e11.
76. Palomaki, G.E.; Kloza, E.M.; Lambert-Messerlian, G.M.; Haddow, J.E.; Neveux, L.M.; Ehrich, M.; van den Boom, D.; Bombard, A.T.; Deciu, C.; Grody, W.W.; *et al*. DNA sequencing of maternal plasma to detect Down syndrome: An international clinical validation study. *Genet. Med.* **2011**, *13*, 913–920.
77. Chiu, R.W.; Chan, K.C.; Gao, Y.; Lau, V.Y.; Zheng, W.; Leung, T.Y.; Foo, C.H.; Xie, B.; Tsui, N.B.; Lun, F.M.; *et al*. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 20458–20463.
78. Ladha, S. A new era of non-invasive prenatal genetic diagnosis: Exploiting fetal epigenetic differences. *Clin. Genet.* **2012**, *81*, 362–363.
79. Tong, Y.K.; Chiu, R.W.; Chan, K.C.; Leung, T.Y.; Lo, Y.M. Technical concerns about immunoprecipitation of methylated fetal DNA for noninvasive trisomy 21 diagnosis. *Nat. Med.* **2012**, *18*, 1327–1328; author reply 1328–1329.
80. Patsalis, P.C. Reply to: Technical concerns about immunoprecipitation of methylated fetal DNA for noninvasive trisomy 21 diagnosis. *Nat. Med.* **2012**, *18*, 1328–1329.
81. Fan, H.C.; Quake, S.R. Sensitivity of noninvasive prenatal detection of fetal aneuploidy from maternal plasma using shotgun sequencing is limited only by counting statistics. *PLoS One* **2010**, *5*, e10439.
82. Palomaki, G.E.; Deciu, C.; Kloza, E.M.; Lambert-Messerlian, G.M.; Haddow, J.E.; Neveux, L.M.; Ehrich, M.; van den Boom, D.; Bombard, A.T.; Grody, W.W.; *et al*. DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. *Genet. Med.* **2012**, *14*, 296–305.
83. Chen, E.Z.; Chiu, R.W.K.; Sun, H.; Akolekar, R.; Chan, K.C.A.; Leung, T.Y.; Jiang, P.; Zheng, Y.W.L.; Lun, F.M.F.; Chan, L.Y.S.; *et al*. Noninvasive Prenatal Diagnosis of Fetal Trisomy 18 and Trisomy 13 by Maternal Plasma DNA Sequencing. *PLoS One* **2011**, *6*, e21791.
84. Aria Diagnostics, Inc. Available online: http://www.ariadx.com/ (accessed on 5 December 2013).
85. SEQUENOM, Inc. Available online: http://www.sequenom.com/ (accessed on 5 December 2013).
86. Verinata Health, Inc. Available online: http://www.verinata.com/ (accessed on 5 December 2013).
87. Tsui, D.W.; Chiu, R.W.; Lo, Y.D. Epigenetic approaches for the detection of fetal DNA in maternal plasma. *Chimerism* **2010**, *1*, 30–35.

88. Korshunova, Y.; Maloney, R.K.; Lakey, N.; Citek, R.W.; Bacher, B.; Budiman, A.; Ordway, J.M.; McCombie, W.R.; Leon, J.; Jeddeloh, J.A.; *et al*. Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. *Genome Res.* **2008**, *18*, 19–29.
89. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **2008**, *455*, 1061–1068.
90. Grunau, C.; Clark, S.J.; Rosenthal, A. Bisulfite genomic sequencing: Systematic investigation of critical experimental parameters. *Nucleic Acids Res.* **2001**, *29*, E65.
91. Laird, P.W. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* **2010**, *11*, 191–203.
92. Fazzari, M.J.; Greally, J.M. Epigenomics: beyond CpG islands. *Nat. Rev. Genet.* **2004**, *5*, 446–455.
93. Nair, S.S.; Coolen, M.W.; Stirzaker, C.; Song, J.Z.; Statham, A.L.; Strbenac, D.; Robinson, M.D.; Clark, S.J. Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics* **2011**, *6*, 34–44.
94. NIPD Genetics Ltd. Available online: http://www.nipd.com/ (accessed on 5 December 2013).
95. Jin, S.; Lee, Y.K.; Lim, Y.C.; Zheng, Z.; Lin, X.M.; Ng, D.P.; Holbrook, J.D.; Law, H.Y.; Kwek, K.Y.; Yeo, G.S.; *et al*. Global DNA hypermethylation in down syndrome placenta. *PLoS Genet.* **2013**, *9*, e1003515.
96. Stewart, F.J.; Panne, D.; Bickle, T.A.; Raleigh, E.A. Methyl-specific DNA binding by McrBC, a modification-dependent restriction enzyme. *J. Mol. Biol.* **2000**, *298*, 611–622.
97. Khulan, B.; Thompson, R.F.; Ye, K.; Fazzari, M.J.; Suzuki, M.; Stasiek, E.; Figueroa, M.E.; Glass, J.L.; Chen, Q.; Montagna, C.; *et al*. Comparative isoschizomer profiling of cytosine methylation: The HELP assay. *Genome Res.* **2006**, *16*, 1046–1055.

Polygenic Scores Predict Alcohol Problems in an Independent Sample and Show Moderation by the Environment

Jessica E. Salvatore, Fazil Aliev, Alexis C. Edwards, David M. Evans, John Macleod, Matthew Hickman, Glyn Lewis, Kenneth S. Kendler, Anu Loukola, Tellervo Korhonen, Antti Latvala, Richard J. Rose, Jaakko Kaprio and Danielle M. Dick

Abstract: Alcohol problems represent a classic example of a complex behavioral outcome that is likely influenced by many genes of small effect. A polygenic approach, which examines aggregate measured genetic effects, can have predictive power in cases where individual genes or genetic variants do not. In the current study, we first tested whether polygenic risk for alcohol problems—derived from genome-wide association estimates of an alcohol problems factor score from the age 18 assessment of the Avon Longitudinal Study of Parents and Children (ALSPAC; n = 4304 individuals of European descent; 57% female)—predicted alcohol problems earlier in development (age 14) in an independent sample (FinnTwin12; n = 1162; 53% female). We then tested whether environmental factors (parental knowledge and peer deviance) moderated polygenic risk to predict alcohol problems in the FinnTwin12 sample. We found evidence for both polygenic association and for additive polygene-environment interaction. Higher polygenic scores predicted a greater number of alcohol problems (range of Pearson partial correlations 0.07–0.08, all p-values $\leq$ 0.01). Moreover, genetic influences were significantly more pronounced under conditions of low parental knowledge or high peer deviance (unstandardized regression coefficients (b), p-values (p), and percent of variance (R^2) accounted for by interaction terms: b = 1.54, p = 0.02, R^2 = 0.33%; b = 0.94, p = 0.04, R^2 = 0.30%, respectively). Supplementary set-based analyses indicated that the individual top single nucleotide polymorphisms (SNPs) contributing to the polygenic scores were not individually enriched for gene-environment interaction. Although the magnitude of the observed effects are small, this study illustrates the usefulness of polygenic approaches for understanding the pathways by which measured genetic predispositions come together with environmental factors to predict complex behavioral outcomes.

Reprinted from *Genes*. Cite as: Salvatore, J.E.; Aliev, F.; Edwards, A.C.; Evans, D.M.; Macleod, J.; Hickman, M.; Lewis, G.; Kendler, K.S.; Loukola, A.; Korhonen, T.; Latvala, A.; Rose, R.J.; Kaprio, J.; Dick, D.M. Polygenic Scores Predict Alcohol Problems in an Independent Sample and Show Moderation by the Environment. *Genes* **2014**, *5*, 330-346.

1. Introduction

Alcohol consumption and related problems are classic examples of complex behavioral outcomes that likely involve many genes of small effect [1]. Twin studies, which infer genetic influences by comparing the phenotypic similarity between monozygotic (MZ) twins (who share all of their genetic variation) and dizygotic (DZ) twins (who share half of their genetic variation, on average), have been crucial for demonstrating that latent genetic influences account for a considerable amount of the variation in measures of alcohol consumption and problems, with

heritability estimates in the range of 50%–60% [2–5]. Twin studies have also been critical for demonstrating that environmental factors moderate the importance of genetic influences. In adolescents, for example, genetic influences on alcohol use and other closely related externalizing problems (e.g., conduct problems) increase under conditions of low parental knowledge (*i.e.*, the degree to which parents know about one's daily activities and associates) or high peer deviance (*i.e.*, the degree to which one's peer group engages in substance use and antisocial behavior) [6–9]. Thus, genetic influences appear to become more important under environmental conditions characterized by more social opportunity and less social control [10].

In contrast to the consistent evidence for the heritability of alcohol use and problems, no robust associations have been detected in genome-wide association studies (GWAS) to date. This is the case, in part, because the small samples typically used in alcohol research are underpowered to detect the very modest individual effect sizes that are generally observed in GWAS of complex behavioral outcomes. Large meta- and mega-analyses pooling across many studies are needed to obtain robust results in the substance use area [11]; only now are these studies underway for alcohol use and alcohol problems. In candidate gene studies, a few compelling associations have emerged within biologically plausible pathways. For example, polymorphisms in *ADH1B* and *ALDH2* genes, which code for alcohol-metabolizing enzymes, have well-replicated associations with alcohol dependence [12–15]. In another example, independent groups have found evidence that the α2 encoding subunit of the GABA-A receptor (GABRA2) is associated with alcohol dependence [16,17]. Likewise, despite consistent evidence from twin samples that environmental factors moderate latent genetic influences, measured gene-by-environment moderation effects for behavioral outcomes have been widely criticized on the grounds that they are underpowered and likely reflect Type I statistical error [18].

In the absence of success in identifying individual genes that account for a substantial proportion of the variance in alcohol outcomes, and lack of expectation that such genes will be found in the near future, polygenic approaches have emerged as one paradigm for examining aggregate measured genetic effects that can have predictive power when individual genes cannot [19]. This approach typically uses results from a genome-wide association study in a discovery sample. Using a *p*-value threshold much more liberal than what would be required for genome-wide significance, a polygenic risk score for each individual in an independent target sample is calculated by summing up the number of alleles for each single nucleotide polymorphism (SNP) weighted by the effect size drawn from a GWAS. The score then represents the composite additive effect of these multiple variants, which likely includes a mixture of true genetic signals and noise.

In the current study, we adopted a polygenic approach to examine alcohol problems in adolescence. Adolescence represents an important developmental period for the initiation of alcohol use [20], and, for some, the development of alcohol problems [21]. Longitudinal developmental studies indicate that the heritability of alcohol use increases across adolescence [4,22], making this an important period of the lifespan for beginning to identify the genetic predispositions toward alcohol problems, and how these predispositions interface with key environmental factors (e.g., low parental knowledge and affiliations with deviant peers) known to be associated with higher levels of alcohol problems. We tested the hypotheses that: (1) polygenic risk for alcohol

problems—derived from GWAS estimates in one population-based sample—would predict alcohol problems in adolescence in a second, independent, population-based sample; and (2) parenting and peer factors in adolescence would moderate polygenic risk to predict alcohol problems in the independent sample.

2. Experimental Section

We drew upon two population-based samples in the present study. GWAS results from the Avon Longitudinal Study of Parents and Children (ALSPAC) [23] were used to create polygenic risk scores in the independent FinnTwin12 sample [24]. The samples and measures are described in greater detail below.

2.1. Avon Longitudinal Study of Parents and Children

The ALSPAC sample included 15,247 pregnancies from women residing in Avon, UK with expected dates of delivery between April 1991 and December 1992, resulting in 15,458 fetuses. Of this total sample of 15,458 fetuses, 14,775 were live births and 14,701 were alive at 1 year of age. Additional details regarding the sample can be found in Boyd *et al.* [25]. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. In the present study, we used data from unrelated participants who completed an alcohol assessment at 16 and/or 18 years of age (5952 participants) for whom there were also genotypic data (n = 4304). Please note that the study website contains details of all the data that is available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary).

2.1.1. Alcohol Problems Factor Score

We measured alcohol problems using a factor score that included ten items from the Alcohol Use Disorders Identification Test (AUDIT) [26], seven DSM-IV Alcohol Dependence criteria [27], and three additional measures related to alcohol problems (getting into fights, police involvement, and drinking to alleviate withdrawal symptoms) that were collected as part of the age 18 assessment. To increase our sample size, we also imputed age 18 alcohol problems data for the participants who completed the age 16 alcohol assessment, but not the age 18 assessment (n = 1993) using imputation software IVEware [28]. Frequency and correlation checks after imputation showed that all imputations kept similar frequency distributions and that imputed and original variables were closely correlated. The results of an exploratory factor analyses indicated one main factor (eigenvalue = 6.78) that broadly measured heavy alcohol use and problems. We then ran a confirmatory factor analysis to calculate factor scores using Mplus 6.11 [29]. All items' factor loadings were >0.30, and the items with the greatest loadings were: frequency of heavy drinking (6 or more drinks on one occasion); drinks per day on drinking days; injuries as a result of drinking; and tolerance. In total, alcohol problems factor scores were calculated for 5952 participants.

2.1.2. Genotyping

ALSPAC participants were genotyped from blood samples using the Illumina 550K custom chip (San Diego, CA, USA). Multi-dimensional scaling modeling seeded with HapMap Phase II release 22 reference populations was used to identify individuals of non-European descent. To reduce bias introduced by population stratification, individuals of non-European descent were removed from subsequent analyses. Those of European descent were imputed to HapMap Phase II (release 22, NCBI build 36, hg18) using the Markov Chain Haplotyping software (MACH v.1.0.16) [30]. SNPs that were in Hardy-Weinberg equilibrium ($p > 5 \times 10^{-7}$) with a final call rate of >95%, and minor allele frequency >1% were used in the imputation procedure. The 2,450,300 autosomal SNPs that exceeded an Rsq metric of 0.3 and had a minor allele frequency >1% following imputation were used in the GWAS. Additional, detailed GWAS data cleaning information for this sample are available in Fatemifar *et al.* [31].

2.2. FinnTwin12

Our second, independent sample was FinnTwin12 [24]—a population-based twin sample identified through Finland's Population Register Center. Approximately 2700 pairs of twins were initially enrolled between ages 11–12 and have been contacted for multiple follow-up assessments of behavioral, emotional, and physical health. In the present study we used data from 1162 participants (467 MZ individuals, 684 DZ individuals, and 11 individuals of unknown zygosity; 53% female, 47% male) for whom there were genome-wide association (GWA) data. Relevant phenotypic data from a psychiatric interview and self-report measures of parental knowledge (n = 1115) and peer deviance (n = 1116) at age 14 were available for a subset of the GWA sample.

2.2.1. Alcohol Problems, Parental Knowledge, and Peer Deviance

Alcohol problems, parental knowledge, and peer deviance were assessed at age 14. The alcohol measure was a sum score of alcohol problems (range 0–30) from the Child version of the Semi-Structured Assessment for the Genetics of Alcoholism [32]. Sample items included needing 50% more alcohol to get an effect, being unable to cut down, reducing important activities to drink, and experiencing withdrawal symptoms.

The parental knowledge measure was the sum score of four adolescent self-report items adapted from Chassin and colleagues [33] about the degree to which their parents know about their daily plans, activities and whereabouts, how they spend their money, and where/who they are with when not at home. Responses were made on a 4-point scale ranging from *almost always* to *rarely or never*, and were summed such that high scores indicate low parental knowledge (more risk; range 4–16).

The peer deviance measure was the sum score of four adolescent self-report items regarding the number of friends/acquaintances who drink, smoke, use drugs, and get into trouble at school. Responses were made on a 4-point scale ranging from *none* to *more than five*, and were summed such that high scores indicate high peer deviance (more risk; range 4–16).

2.2.2. Genotyping

Genome-wide data were collected using blood samples obtained at the age 22 assessment. Genotyping was performed at the Wellcome Trust Sanger Institute (Hinxton, UK) on the Human670-QuadCustom Illumina BeadChip (Illumina, Inc., San Diego, CA, USA), as previously described in Broms *et al.* [34]. The data were checked for minor allele frequency (MAF > 1%), genotyping success rate per SNP and per individual (>95%; >99% for SNPs with MAF < 5%), Hardy-Weinberg Equilibrium (HWE $p > 1 \times 10^{-6}$), sex, and heterozygosity. In addition, to check whether any individuals were unexpectedly related to each other, a multidimensional scaling plot (using a pairwise-IBS matrix) with only one member of each known family was created. After the pedigree was checked for accuracy, the basic filters (MAF, genotyping success, HWE) were reapplied to the data.

Imputation was performed by using ShapeIT [35] in pre-phasing and IMPUTE2 [36] for genotype imputation, with the 1000 Genomes Phase I integrated variant set release (v3) reference panel. The posterior probability threshold for "best-guess" imputed genotypes was 0.9. Genotypes below the threshold were set to missing. Genotypes for altogether 6,729,635 SNPs were available for analysis.

2.3. Analytic Plan

2.3.1. Genome-Wide Association Analysis in the ALSPAC Sample

The GWAS was conducted using MACH2QTL [37] and was limited to individuals of European descent. Sex was included as a covariate.

2.3.2. Calculation of Polygenic Scores in FinnTwin12

We used ALSPAC GWAS estimates from the alcohol problems factor score to calculate polygenic scores for FinnTwin12 using the --score procedure in PLINK [38]. We computed a linear function of the number of score alleles an individual possessed weighted by the product of the sign of the SNP effect and the negative logarithm (base 10) of the associated GWAS *p*-value. This retains the same direction between calculated and original output values. Of the 2,450,300 autosomal SNPs that passed quality control in the ALSPAC sample, 2,221,783 (91%) were available in the FinnTwin12 sample.

There are no set criteria for creating maximally informative polygenic scores [39], and so we created a series of scores using *p*-value thresholds ranging from 0.05 to 0.50. Table 1 summarizes the number of SNPs meeting each threshold in the ALSPAC sample, as well as the number and percent of those SNPs that were available in the FinnTwin12 sample. Previous work using polygenic approaches indicates that pruning for linkage disequilibrium (LD) does not substantially change the results [19,40]. In view of this, we chose to incorporate all SNPs meeting each polygenic threshold into our scores.

Table 1. Autosomal single nucleotide polymorphisms (SNPs) contributing to each polygenic threshold in Avon Longitudinal Study of Parents and Children (ALSPAC) sample, and availability in FinnTwin12.

Polygenic threshold	Number of autosomal SNPs meeting threshold in ALSPAC	Number (percent) of SNPs available in FinnTwin12
$p \leq 0.05$	125,969	113,992 (90.5%)
$p \leq 0.10$	250, 244	226,789 (90.6%)
$p \leq 0.20$	495,760	449,273 (90.6%)
$p \leq 0.30$	739,758	670,293 (90.6%)
$p \leq 0.40$	984,167	891,782 (90.6%)
$p \leq 0.50$	1,231,165	1,115,557 (90.6%)

2.3.3. Polygenic Association and Moderation Analyses in FinnTwin12

We used partial Pearson correlations, controlling for sex, to test associations between the FinnTwin12 polygenic scores and alcohol problems. We used moderated multiple regression to test our gene-by-environment interaction hypotheses that parental knowledge and peer deviance would moderate the predictive association of polygenic scores with the age 14 alcohol problems measure. For these analyses, the parameters of interest were the statistical interactions between the environmental factors (parental knowledge and peer deviance) and the polygenic scores. The main effects of sex and the environmental factors were used as covariates in the relevant models. Parental knowledge, peer deviance, and polygenic scores were centered on their means prior to running moderation analyses to reduce co-linearity among predictor variables.

3. Results and Discussion

3.1. Descriptive Statistics and Zero-Order Correlations

Descriptive statistics for the focal variables and for an illustrative polygenic score (using the $p \leq 0.05$ threshold) are presented in Table 2. MZ twins' alcohol problems were correlated at $r = 0.53$ (232 pairs; $p < 0.01$), and DZ twins were correlated at $r = 0.36$ (277 pairs; $p < 0.01$). This pattern of twin correlations suggests that additive genetic effects accounted for approximately 34% of the variance in alcohol problems. Lower parental knowledge (indexed by higher scores on the parental knowledge scale used here) and higher peer deviance were associated with higher levels of alcohol problems [$r(1113) = 0.29$ and $r(1114) = 0.35$, both p-values < 0.01, respectively], which is consistent with previous work indicating that more permissive and deviant environments are associated with a greater amount of adolescent substance use [33,41,42].

Table 2. FinnTwin12 descriptive statistics for focal study variables.

Variable	M	SD	Min	Max
Alcohol problems (age 14), range 0–30	0.29	0.96	0	8
Parental knowledge (age 14), range 4–16	6.62	2.08	4	15
Peer deviance (age 14), range 4–16	7.91	3.14	4	16
Polygenic score ($p \leq 0.05$ threshold)	−0.07	0.02	−0.13	0.00

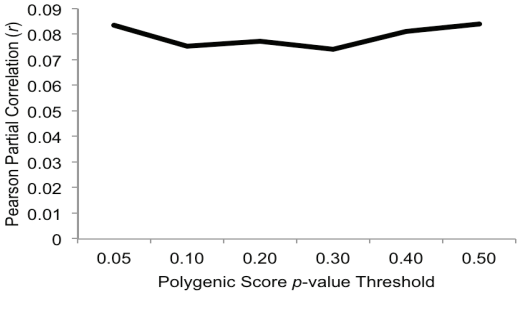
Abbreviations: M, mean; SD, standard deviation; Min, minimum observed value; Max, maximum observed value.

3.2. Polygenic Associations with Alcohol Problems

Partial correlations (controlling for sex) between the polygenic scores and alcohol problems are presented in Figure 1. As expected, higher polygenic scores predicted higher alcohol problems at age 14 (range of Pearson partial correlations 0.07–0.08, all *p*-values < 0.01). This is consistent with previous studies of other psychiatric conditions (such as bipolar disorder [19], schizophrenia [43] and externalizing disorders [40]) in showing that polygenic scores derived from GWAS weights from one sample can have predictive validity in an independent sample. Furthermore, our effect sizes were similar in magnitude to those observed in a polygenic analysis of a behavioral disinhibition measure (which included antisocial behavior, nicotine use/dependence, alcohol consumption and dependence, and drug use) [40].

The magnitude of the associations between polygenic scores and alcohol problems was fairly consistent across the range of selected *p*-value thresholds, and accounted for, on average, 0.63% of the variance in alcohol problems (range 0.55%–0.70%). To be sure that our effects were not driven by non-independence within the sample, we re-ran the association analyses after randomly dropping one member from each twin pair (n = 634) and found the same pattern of results. This is substantially lower than the estimate (derived from the pattern of MZ and DZ twin correlations in the same sample) that additive genetics effects account for 34% of the variance in alcohol problems. We note, however, that heritability estimates derived from twin models and the variance accounted for by a polygenic score are not directly comparable. Polygenic scores are composed of SNPs across a range of *p*-value thresholds, and thus their genetic informativeness is likely to be somewhere between a polygenic risk score based on genome-wide significant SNPs and SNP heritability as derived through methods that estimate the variance explained by genome-wide markers (e.g., GCTA; [44]). The limited amount of variance accounted for in our analyses may be attributable to the fact that GWAS-derived polygenic scores only account for common (*versus* rare; [45]) genetic variation; accordingly, incorporating rare genetic variation in polygenic scores may be an important direction for future research. In addition, the limited variance accounted for may also be attributable to the relatively small sample from which we derived our GWAS weights owing to the fact that smaller samples are likely to have a higher signal-to-noise ratio compared to larger samples.

Figure 1. Pearson partial correlations (controlling for sex) between polygenic scores and age 14 alcohol problems (all *p*-values ≤ 0.01) in FinnTwin12 (n = 1161).



We also tested whether there was evidence for gene-environment correlation by using Pearson correlations to examine the associations between polygenic scores and the parental knowledge and peer deviance environmental measures. As expected, higher polygenic scores were modestly associated with lower parental knowledge, although the effect was of a small magnitude and not significant [r (1113) = 0.05, p = 0.09]. Higher polygenic scores were also modestly associated with higher peer deviance [r (1114) = 0.08, p < 0.01]. This is consistent with previous evidence from twin studies showing that externalizing-spectrum behaviors such as alcohol use, tobacco use, and conduct problems are genetically correlated with environmental factors [7,8]. These findings highlight the complex interplay between genetic and "environmental" influences on behavioral outcomes such as alcohol problems [46].

3.3. Gene-by-Environment Interactions

The polygenic score using the $p \leq 0.05$ threshold accounted for the greatest proportion of variance (0.70%) in age 14 alcohol problems, and we carried this score forward for the gene-by-environment analyses in view of earlier suggestions that SNPs having a nominal association with a phenotype are likely to be enriched for gene-by-environment interaction [47].

Table 3. FinnTwin12 sample. Moderated multiple regression of age 14 alcohol problems on sex, polygenic score, parental knowledge, and the interaction of polygenic score and parental knowledge (top; n = 1115). Moderated multiple regression of age 14 alcohol problems on sex, polygenic score, peer deviance, and the interaction of polygenic score and peer deviance (bottom; n = 1116).

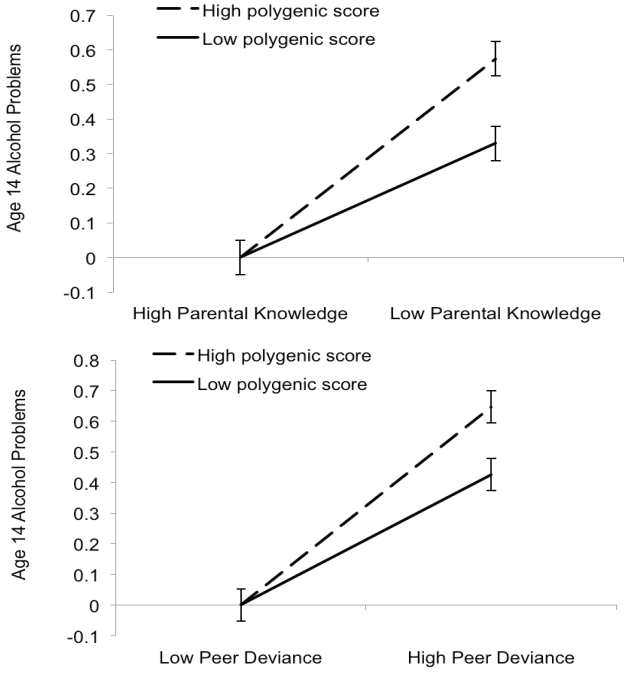
Parental Knowledge					
	b	*SE*	*t*	*P*	ΔR^2
Intercept	***0.16***	***0.04***	***3.97***	***<0.01***	**--**
Sex	***0.23***	***0.06***	***4.17***	***<0.01***	***0.006***
Polygenic score	**3.10**	**1.40**	**2.21**	**0.03**	***0.006***
Parental knowledge	***0.14***	***0.01***	***10.31***	***<0.01***	***0.088***
Polygenic score × Parental knowledge	**1.54**	**0.68**	**2.27**	**0.02**	**0.003**
Peer Deviance					
	b	*SE*	*t*	*P*	ΔR^2
Intercept	***0.19***	***0.04***	***4.88***	***<0.01***	**--**
Sex	***0.17***	***0.05***	***3.07***	***<0.01***	***0.006***
Polygenic score	2.75	1.38	1.99	0.05	0.006
Peer deviance	***0.11***	***0.01***	***12.43***	***<0.01***	***0.120***
Polygenic score × Peer deviance	**0.94**	**0.44**	**2.11**	**0.04**	**0.003**

Boldfaced statistics indicate $p < 0.05$. Boldfaced and italicized statistics indicate $p < 0.01$. Abbreviations: n = sample size, *b*, unstandardized regression estimates; *SE*, standard error for *b*; *t*, *t*-statistic; *P*, *p*-value; ΔR^2, step-wise change in variance accounted for by each parameter in model.

Moderated multiple regression analyses indicated that parental knowledge and peer deviance moderated the associations of polygenic scores with age 14 alcohol problems (Table 3). Genetic influences were more pronounced under conditions of low parental knowledge or high peer deviance compared to conditions of high parental knowledge or high peer deviance (Figure 2). The

interactions with parental knowledge and peer deviance accounted for 0.33% and 0.30% of the variance in alcohol problems, respectively. To verify that our effects were not driven by non-independence within the sample, we note that the same pattern of effects was found when we re-ran the moderation analyses after randomly dropping one member from each twin pair (n = 634).

Figure 2. Parental knowledge (top) and peer deviance (bottom) moderate polygenic risk to predict age 14 alcohol problems in FinnTwin12. Interactions are plotted as predicted values based on the moderated multiple regression equation for age 14 alcohol problems. Illustrative low and high values (±1 SD of mean) for the polygenic scores, parental knowledge, and peer deviance are shown. The predicted values for high parental knowledge and low peer deviance were out of bounds (negative values) and were set to zero—the lowest possible value for the alcohol problems measure. Error bars are equal to the standard deviation of the model residuals divided by the square root of the sample size. We note that high scores on the parental knowledge scale indicate low parental knowledge (*i.e.*, more risk). For ease of interpretation, we have formatted the axis for each figure so that the riskier environment appears on the right.



Although the effect sizes for the polygenic score X environment interactions were small, the pattern of effects is consistent with previous findings from the twin literature. Multiple independent twin studies find that parenting and peer environmental factors moderate latent genetic influences for alcohol use and related outcomes such that genetic influences increase under conditions of low parental knowledge and high peer deviance [6–9,48]. The convergence between the pattern of gene-environment interactions from twin studies and measured polygenic effects is encouraging,

and suggests that polygenic approaches may be a useful way to characterize gene-environment interplay for aggregate genetic risk using measured genotypic data.

In addition to these core analyses, we ran a series of supplementary analyses to examine the robustness of our effects after controlling for gene-environment correlation and after transforming our alcohol problems dependent variable to a logarithmic scale. Gene-environment correlation can produce spurious gene-environment interaction effects; likewise, interaction effects are known to be sensitive to scale. Accordingly, our supplementary analyses were intended to address concerns that our observed gene-environment interaction effects could be statistical artifacts.

To control for gene-environment correlation in our parental knowledge analyses, we used residualized polygenic score and parental knowledge variables in our model. To calculate residualized variables, we regressed polygenic scores onto parental knowledge (and *vice versa*) and saved the residuals for use in the moderation models. Using residualized variables in this way statistically eliminates gene-environment correlation from the model because the genetic and environmental effects have been partialled from one another. We used the same method to calculate residualized polygenic score and peer deviance variables for our peer deviance analyses. The moderation effect for parental knowledge continued to be statistically significant; however, the moderation effect for peer deviance trended in the same direction but was not statistically significant (unstandardized regression coefficients (*b*) and *p*-values (*p*) for interaction terms: $b = 1.33$, $p = 0.05$ and $b = 0.66$, $p = 0.14$, respectively) when we used residualized values in our analyses.

To test whether our interaction effects could be attributed to the scale of the alcohol problems measure, we used a log-transformed version of the measure (*i.e.*, $\log_{10}$ (alcohol problems + 1)) in our analyses. The interaction effects trended in the expected direction for parental knowledge and peer deviance, albeit failing to reach significance (unstandardized regression coefficients (*b*) and *p*-values (*p*) for interaction terms: $b = 0.51$, $p = 0.07$ and $b = 0.30$, $p = 0.10$, respectively). As a set, these supplementary analyses demonstrate that the moderation effects were modestly attenuated after controlling for gene-environment correlation and changing the scale of the alcohol problems outcome variable, but they continued to trend in the same direction and did not entirely go away.

Although the causal relationships among the genetic and environmental variables examined here are unknown, we note that early findings from genetically-informed randomized prevention studies suggest that efforts aimed at reducing environmental risk factors for adolescent alcohol use and related behavior problems may be particularly effective for those who are genetically predisposed toward developing such problems. For example, adolescents with either the short/short or short/long genotype of *SCL6A4(5-HTT)* who took part in a family-based prevention-intervention program aimed at increasing family cohesion were less likely to initiate risk behaviors (alcohol use, marijuana use, and sex) across a 29-month period compared to their counterparts in the control condition [49]. Examining whether efforts to bolster parental knowledge or reduce peer deviance attenuate polygenic risk for alcohol problems is an important direction for future research.

3.4. Set-Based Analyses Examining Enrichment for Gene-Environment Interaction among Top SNPs

The polygenic analyses indicated significant gene-environment interaction effects with parental knowledge and peer deviance, and we used set-based analyses to probe whether the individual top SNPs contributing to our polygenic scores were themselves enriched for gene-environment interaction. We examined this question using the set of top SNPs ($p \leq 0.0001$) from the ALSPAC GWAS. We selected this relatively stringent *p*-value threshold in view of the computing resources required to perform the permutation analyses described below. Of the 311 SNPs meeting this threshold in ALSPAC, 279 (90%) were available in FinnTwin12. We pruned by LD in the FinnTwin12 sample in order to reduce the set to include only independent ($r^2 < 0.50$) SNPs, which resulted in 76 SNPs. Because LD calculations should be made on independent individuals, we used a randomly-selected sample of independent individuals in the FinnTwin12 sample ($n = 634$) for this purpose. We then permuted the phenotypic and covariate information for these individuals 100,000 times while keeping the genotypic information (LD) unchanged. For each of these permuted datasets, we examined gene-environment interaction effects for parental knowledge and peer deviance. To calculate empirical *p*-values, we used the equation (R + 1)/(N + 1). R is the number of permutations where the sum of the absolute value of the *t*-scores for significant SNP interaction effects ($p < 0.05$) exceeded the sum of the absolute value of the *t*-scores for significant SNP interaction effects in the observed data. N is the number of permutations (100,000).

Our empirical *p*-values were 0.32 and 0.71 for parental knowledge and peer deviance, respectively. This indicates that the SNPs contributing to our polygenic scores were not individually enriched for gene-environment interaction, and further suggests that the polygenic moderation effects that we observed occur at the aggregate genetic level rather than at the level of individual SNPs. Attempts to replicate these effects in other independent datasets are critical for better understanding the contributions of individual SNPs to the aggregate effects observed for polygenic scores.

3.5. Limitations

Our results should be interpreted in the context of their limitations. First, the participants in our two samples were of European descent, the latter exclusively of Finnish descent, which may limit the generalizability of the present findings to samples from the same ancestral background. Second, although our association and moderation findings were in the expected direction, the effect sizes were quite small—often accounting for less than 1% of the variance. Although there is much enthusiasm for personalized medicine approaches [50] that use genome-wide information to identify for whom and under what conditions prevention and intervention efforts are likely to be effective, our results caution against using empirically-derived GWAS scores in a clinical setting for complex behavioral outcomes such as alcohol problems due to the fact that they account for a limited proportion of the variance [51]. Third, alcohol problems in FinnTwin12 were assessed at age 14. Accordingly, endorsements of alcohol problems at this age may represent a more severe

phenotype than those at age 18 in ALSPAC. The age and measurement differences across the ALSPAC and FinnTwin12 samples may explain, in part, the low percentage of variance accounted for by the polygenic score. Finally, the polygenic approach adopted here is limited in that it does not attempt to implicate the specific genes involved in alcohol problems. Additional methods, such as gene set approaches that examine whether SNPs included in a polygenic score are located in functionally related genes [52], are well suited to identify the potential biological mechanisms underlying polygenic effects.

4. Conclusions

Higher polygenic predispositions for alcohol problems (based on GWAS estimates from a population-based sample of young adults) predicted a higher number of adolescent alcohol problems in an independent, population-based sample. In addition, environmental factors in adolescence moderated these polygenic predispositions. Genetic predispositions were more important under conditions of low parental knowledge and high peer deviance. These gene-by-environment interactions, although small in magnitude, are consistent with previous findings from studies that show that environments low in social control or high in social opportunity permit the expression of genetic predispositions [10]. In contrast, environments high in social control or low in social opportunity may inhibit the expression of that same predisposition. Accordingly, prevention and intervention efforts that increase parental knowledge and decrease affiliations with deviant peers may be one strategy for reducing risk for adolescents with genetic predispositions toward alcohol problems; however additional study is needed before making strong claims about the potential effectiveness of such interventions.

Acknowledgments

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and the Wellcome Trust (Grant 092731) and the University of Bristol provide core support for ALSPAC. This research was also funded by the National Institute On Alcohol Abuse And Alcoholism of the National Institutes of Health under award R01AA018333 (to K.S.K. and D.M.D.). The Finnish Twin studies have been supported by the National Institute of Alcohol Abuse and Alcoholism (awards AA12502, AA00145, and AA09203 to R.J.R.; and AA15416 and K02AA018755 to D.M.D.), the Sigrid Juselius Foundation (to J.K.), the Academy of Finland (grants 100499, 205585, 141054, 264146 and 118555 to J.K.); and the Academy of Finland Centre of Excellence Programme (grants 213506, 129680 to J.K. and Lea Pulkkinen). We thank the Wellcome Trust Sanger Institute for genotyping support of the FinnTwin12 samples, and Kaisu Keskitalo-Vuokko, who participated in the genotyping of the FinnTwin12 sample; and Antti-Pekka Sarin, who performed pre-imputation quality controls and imputation in the FinnTwin12 sample. J.E.S. was supported by T32MH20030-14 and F32AA022269; A.C.E. was supported by K01AA021399; and K.S.K. was supported by

P20AA107828. This publication is the work of the authors, and J.E.S., F.A., and D.M.E. will serve as guarantors for the contents of this paper. The contents of the paper are solely the responsibility of the authors and do not necessarily represent the official views of the funders.

Author Contributions

Conceived and designed the study: JES, DMD, JK. Supervised collection and/or preparation of phenotypic data: JK, TK, ALa. Supervised genotypic data collection: JK, ALo. Analyzed the data: JES, FA, DME. Drafted and/or provided critical intellectual feedback on the paper: JES, FA, ACE, DME, JM, MH, GL, KSK, ALo, TK, ALa, RJR, JK, DMD.

Conflicts of Interest

Tellervo Korhonen and Jaakko Kaprio have acted as consultants of tobacco dependence for Pfizer in 2011–2013.

References

1. Manolio, T.A.; Collins, F.S.; Cox, N.J.; Goldstein, D.B.; Hindorff, L.A.; Hunter, D.J.; McCarthy, M.I.; Ramos, E.M.; Cardon, L.R.; Chakravarti, A.; *et al.* Finding the missing heritability of complex diseases. *Nature* **2009**, *461*, 747–753.
2. Heath, A.C.; Bucholz, K.K.; Madden, P.A.F.; Dinwiddie, S.H.; Slutske, W.S.; Bierut, L.J.; Statham, D.J.; Dunne, M.P.; Whitfield, J.B.; Martin, N.G. Genetic and environmental contributions to alcohol dependence risk in a national twin sample: Consistency of findings in women and men. *Psychol. Med.* **1997**, *27*, 1381–1396.
3. Kendler, K.S.; Heath, A.C.; Neale, M.C.; Kessler, R.C.; Eaves, L.J. A population-based twin study of alcoholism in women. *JAMA* **1992**, *268*, 1877–1882.
4. Rose, R.J.; Dick, D.M.; Viken, R.J.; Kaprio, J. Gene-environment interaction in patterns of adolescent drinking: Regional residency moderates longitudinal influences in alcohol use. *Alcohol. Clin. Exp. Res.* **2001**, *25*, 637–643.
5. Prescott, C.A.; Kendler, K.S. Genetic and environmental contributions to alcohol abuse and dependence in a population-based sample of male twins. *Am. J. Psychiatry* **1999**, *156*, 34–40.
6. Dick, D.M.; Viken, R.; Purcell, S.; Kaprio, J.; Pulkkinen, L.; Rose, R.J. Parental monitoring moderates the importance of genetic and environmental influences on adolescent smoking. *J. Abnorm. Psychol.* **2007**, *116*, 213–218.
7. Harden, K.P.; Hill, J.E.; Turkheimer, E.; Emery, R.E. Gene-environment correlation and interaction in peer effects on adolescent alcohol and tobacco use. *Behav. Genet.* **2008**, *38*, 339–347.
8. Button, T.M.M.; Corley, R.P.; Rhee, S.H.; Hewitt, J.K.; Young, S.E.; Stallings, M.C. Delinquent peer affiliation and conduct problems: A twin study. *J. Abnorm. Psychol.* **2007**, *116*, 554–564.

9. Hicks, B.M.; South, S.C.; DiRago, A.C.; Iacono, W.G.; McGue, M. Environmental adversity and increasing genetic risk for externalizing disorders. *Arch. Gen. Psychiatry* **2009**, *66*, 640–648.
10. Shanahan, M.J.; Hofer, S.M. Social context in gene-environment interactions: Retrospect and prospect. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* **2005**, *60B*, 65–76.
11. Hartz, S.M.; Short, S.E.; Saccone, N.L.; Culverhouse, R.; Chen, L.; Schwantes-An, T.H.; Coon, H.; Han, Y.; Stephens, S.H.; Sun, J.; *et al*. Increased genetic vulnerability to smoking at CHRNA5 in early-onset smokers. *Arch. Gen. Psychiatry* **2012**, *69*, 854–860.
12. Thomasson, H.R.; Edenberg, H.J.; Crabb, D.W.; Mai, X.L.; Jerome, R.E.; Li, T.K.; Wang, S.P.; Lin, Y.T.; Lu, R.B.; Yin, S.J. Alcohol and aldehyde dehydrogenase genotypes and alcoholism in Chinese men. *Am. J. Hum. Genet.* **1991**, *48*, 667–681.
13. Luczak, S.E.; Glatt, S.J.; Wall, T.L. Meta-analyses of ALDH2 and ADH1B with alcohol dependence in Asians. *Psychol. Bull.* **2006**, *132*, 607–621.
14. Whitfield, J.B. Alcohol dehydrogenase and alcohol dependence: Variation in genotype-associated risk between populations. *Am. J. Hum. Genet.* **2002**, *71*, 1247–1250.
15. Gelernter, J.; Kranzler, H.R.; Sherva, R.; Almasy, L.; Koesterer, R.; Smith, A.H.; Anton, R.; Preuss, U.W.; Ridinger, M.; Rujescu, D.; *et al*. Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. *Mol. Psychiatry* **2014**, *19*, 41–49.
16. Edenberg, H.J.; Dick, D.M.; Xuei, X.; Tian, H.; Almasy, L.; Bauer, L.O.; Crowe, R.; Goate, A.; Hesselbrock, V.; Jones, K.A.; *et al*. Variations in GABRA2, encoding the α2 subunit of the GABA-A receptor are associated with alcohol dependence and with brain oscillations. *Am. J. Hum. Genet.* **2004**, *74*, 705–714.
17. Covault, J.; Gelernter, J.; Hesselbrock, V.; Nellissery, M.; Kranzler, H.R. Allelic and haplotypic association of GABRA2 with alcohol dependence. *Am. J. Med. Genet. Part. B Neuropsychiatr. Genet.* **2004**, *129B*, 104–109.
18. Duncan, L.; Keller, M.C. A critical review of the first ten years of measured gene-by-environment interaction research in psychiatry. *Am. J. Psychiatry* **2011**, *168*, 1041–1049.
19. The International Schizophrenia Consortium Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **2009**, *460*, 748–752.
20. National Institute on Drug Abuse Monitoring the Future Study: Trends in Prevalence of Various Drugs (2009–2012). Available online: http://www.drugabuse.gov/related-topics/trends-statistics/monitoring-future/trends-in-prevalence-various-drugs/ (accessed on 10 December 2013).
21. Substance Abuse and Mental Health Services Administration. *Results from the 2012 National Survey on Drug Use and Health: Summary of National Findings*; SAMHSA: Rockville, MD, USA, 2013.
22. Kendler, K.S.; Schmitt, J.E.; Aggen, S.H.; Prescott, C.A. Genetic and environmental influences on alcohol, caffeine, cannabis, and nicotine use from adolescence to middle adulthood. *Arch. Gen. Psychiatry* **2008**, *65*, 674–682.

23. Golding, J.; Pembrey, M.; Jones, R.; ALSPAC Study Team. ALSPAC-The Avon Longitudinal Study of Parents and Children—I. Study methodology. *Paediatr. Perinat. Epidemiol.* **2001**, *15*, 74–87.
24. Kaprio, J.; Pulkkinen, L.; Rose, R.J. Genetic and environmental factors in health-related behaviors: Studies on Finnish twins and twin families. *Twin Res.* **2002**, *5*, 366–371.
25. Boyd, A.; Golding, J.; Macleod, J.; Lawlor, D.A.; Fraser, A.; Henderson, J.; Molloy, L.; Ness, A.; Ring, S.; Smith, G.D. Cohort profile: The Children of the 90s'—The index offspring of the Avon Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* **2012**, *42*, 1–17.
26. Babor, T.F.; Higgins-Biddle, J.C.; Saunders, J.B.; Monteiro, M.G. *AUDIT: The Alcohol Use Disorders Identification Test*, 2nd ed.; World Health Organization: Geneva, Switzerland, 2001.
27. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*; American Psychiatric Association: Washington, DC, USA, 1994.
28. Raghunathan, T.E.; Solenberger, P.W.; van Hoewyk, J. *IVEware: Imputation and Variance Estimation Software*; Institute for Social Research: Ann Arbor, MI, USA, 2002.
29. Muthén, L.K.; Muthén, B.O. *Mplus User's Guide*, 6th ed.; Muthén & Muthén: Los Angeles, CA, USA, 1998–2011.
30. Li, Y.; Willer, C.J.; Ding, J.; Scheet, P.; Abecasis, G.R. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **2010**, *34*, 816–834.
31. Fatemifar, G.; Hoggart, C.J.; Paternoster, L.; Kemp, J.P.; Prokopenko, I.; Horikoshi, M.; Wright, V.J.; Tobias, J.H.; Richmond, S.; Zhurov, A.I.; *et al*. Genome-wide association study of primary tooth eruption identifies pleiotropic loci associated with height and craniofacial distances. *Hum. Mol. Genet.* **2013**, *22*, 3807–3817.
32. Bucholz, K.K.; Cadoret, R.; Cloninger, C.R.; Dinwiddie, S.H.; Hesselbrock, V.M.; Nurnberger, J.I., Jr.; Reich, T.; Schmidt, I.; Schuckit, M.A. A new, semi-structured psychiatric interview for use in genetic linkage studies: A report on the reliability of the SSAGA. *J. Stud. Alcohol* **1994**, *55*, 149–158.
33. Chassin, L.; Pillow, D.R.; Curran, P.J.; Molina, B.S.G.; Barrera, M. Relation of parental alcoholism to early adolescent substance use: A test of 3 mediating mechanisms. *J. Abnorm. Psychol.* **1993**, *102*, 3–19.
34. Broms, U.; Wedenoja, J.; Largeau, M.R.; Korhonen, T.; Pitkaniemi, J.; Keskitalo-Vuokko, K.; Happola, A.; Heikkila, K.H.; Heikkila, K.; Ripatti, S.; *et al*. Analysis of detailed phenotype profiles reveals CHRNA5-CHRNA3-CHRNB4 gene cluster association with several nicotine dependence traits. *Nicotine Tob. Res.* **2012**, *14*, 720–733.
35. Delaneau, O.; Marchini, J.; Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **2012**, *9*, 179–181.
36. Howie, B.N.; Donnelly, P.; Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **2009**, *5*, e1000529.
37. Chen, W.M.; Abecasis, G.R. Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.* **2007**, *81*, 913–926.

38. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.; Daly, M.; *et al*. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575.
39. Evans, D.; Visscher, P.M.; Wray, N. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet.* **2009**, *18*, 3525–3531.
40. Vrieze, S.I.; McGue, M.; Miller, M.B.; Hicks, B.M.; Iacono, W.G. Three mutually informative ways to understand the genetic relationships among behavioral disinhibition, alcohol use, drug use, nicotine use/dependence, and their co-occurrence: Twin biometry, GCTA, and genome-wide scoring. *Behav. Genet.* **2013**, *43*, 97–107.
41. Dishion, T.J.; Owen, L.D. A longitudinal analysis of friendships and substance use: Bidirectional influence from adolescence to adulthood. *Dev. Psychol.* **2002**, *38*, 480–491.
42. Duncan, S.C.; Duncan, T.E.; Biglan, A.; Ary, D. Contributions of the social context to the development of adolescent substance use: A multivariate latent growth modeling approach. *Drug Alcohol Depend.* **1998**, *50*, 57–71.
43. Fanous, A.H.; Zhou, B.Y.; Aggen, S.H.; Bergen, S.E.; Amdur, R.L.; Duan, J.B.; Sanders, A.R.; Shi, J.X.; Mowry, B.J.; Olincy, A.; *et al*. Genome-wide association study of clinical dimensions of schizophrenia: Polygenic effect on disorganized symptoms. *Am. J. Psychiatry* **2012**, *169*, 1309–1317.
44. Yang, J.; Lee, S.H.; Goddard, M.E.; Visscher, P.M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **2011**, *88*, 76–82.
45. Gibson, G. Rare and common variants: Twenty arguments. *Nat. Rev. Genet.* **2012**, *13*, 135–145.
46. Kendler, K.S.; Baker, J.H. Genetic influences on measures of the environment: A systematic review. *Psychol. Med.* **2007**, *37*, 615–626.
47. Thomas, D. Gene-environment-wide association studies: Emerging approaches. *Nat. Rev. Genet.* **2010**, *11*, 259–272.
48. Chen, L.S.; Johnson, E.O.; Breslau, N.; Hatsukami, D.; Saccone, N.L.; Grucza, R.A.; Wang, J.C.; Hinrichs, A.L.; Fox, L.; Goate, A.M.; *et al*. Interplay of genetic risk factors and parent monitoring in risk for nicotine dependence. *Addiction* **2009**, *104*, 1731–1740.
49. Brody, G.H.; Beach, S.R.H.; Philibert, R.A.; Chen, Y.F.; Murry, V.M. Prevention effects moderate the association of 5-HTTLPR and youth risk behavior initiation: Gene x environment hypotheses tested via a randomized prevention design. *Child. Dev.* **2009**, *80*, 645–661.
50. Hamburg, M.A.; Collins, F.S. The path to personalized medicine. *N. Engl. J. Med.* **2010**, *363*, 301–304.
51. Yan, J.; Aliev, F.; Webb, B.T.; Kendler, K.S.; Edenberg, H.J.; Agrawal, A.; Kos, M.Z.; Almasy, L.; Nurnberger, J.I., Jr.; Schuckit, M.A.; *et al*. Using genetic information from candidate gene and genome wide association studies in risk prediction for alcohol dependence. *Addict. Biol.* **2013**, doi:10.1111/adb.12035.
52. Wang, L.; Jia, P.; Wolfinger, R.D.; Chen, X.; Zhao, Z. Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics* **2011**, *98*, 1–8.

Epigenetic Variation in Monozygotic Twins: A Genome-Wide Analysis of DNA Methylation in Buccal Cells

Jenny van Dongen, Erik A. Ehli, Roderick C. Slieker, Meike Bartels, Zachary M. Weber, Gareth E. Davies, P. Eline Slagboom, Bastiaan T. Heijmans and Dorret I. Boomsma

Abstract: DNA methylation is one of the most extensively studied epigenetic marks in humans. Yet, it is largely unknown what causes variation in DNA methylation between individuals. The comparison of DNA methylation profiles of monozygotic (MZ) twins offers a unique experimental design to examine the extent to which such variation is related to individual-specific environmental influences and stochastic events or to familial factors (DNA sequence and shared environment). We measured genome-wide DNA methylation in buccal samples from ten MZ pairs (age 8–19) using the Illumina 450k array and examined twin correlations for methylation level at 420,921 CpGs after QC. After selecting CpGs showing the most variation in the methylation level between subjects, the mean genome-wide correlation (rho) was 0.54. The correlation was higher, on average, for CpGs within CpG islands (CGIs), compared to CGI shores, shelves and non-CGI regions, particularly at hypomethylated CpGs. This finding suggests that individual-specific environmental and stochastic influences account for more variation in DNA methylation in CpG-poor regions. Our findings also indicate that it is worthwhile to examine heritable and shared environmental influences on buccal DNA methylation in larger studies that also include dizygotic twins.

Reprinted from *Genes*. Cite as: van Dongen, J.; Ehli, E.A.; Slieker, R.C.; Bartels, M.; Weber, Z.M.; Davies, G.E.; Slagboom, P.E.; Heijmans, B.T.; Boomsma, D.I. Epigenetic Variation in Monozygotic Twins: A Genome-Wide Analysis of DNA Methylation in Buccal Cells. *Genes* **2014**, *5*, 347-365.

1. Introduction

To date, hundreds of genetic risk variants for complex traits and diseases have been identified, although for most of these variants, the biological mechanisms remain to be elucidated [1]. Interestingly, the majority of disease-associated genetic variation is located in regulatory regions of the genome [2], including transcription-factor-occupied regions and DNase I hypersensitive sites (which correspond to open chromatin) [3]. This suggests that mechanisms that control the activity of genes, including epigenetic mechanisms, may represent an important link between DNA sequence variation and common disease susceptibility [4]. Trying to unravel the molecular biology underlying complex traits and disease, much attention has been drawn recently to these epigenetic mechanisms; non-DNA sequence-based regulation of gene expression by DNA methylation, histone modification, microRNAs, *etc*. [5]. DNA methylation is one of the most extensively studied epigenetic mechanisms in human populations and tissues and is the focus of this paper.

In humans, DNA methylation occurs almost exclusively at cytosines that are part of CpG dinucleotides. The relationship between DNA methylation and expression varies depending on the genomic context: CpG methylation at promoter regions is generally thought to repress gene expression, while gene body methylation is generally associated with active gene expression and has

been suggested to regulate splicing [6–8]. In most cell types, the majority of CpGs in the genome (on average, 70%–80%) is typically methylated [9]. Of the unmethylated CpG sites in the genome, most occur in areas of clustered CpGs, called CpG islands, which are often present in promoter regions. Yet, DNA methylation patterns may vary, and differential methylation has been demonstrated to occur across age [10], cell types, tissues [7,11] and disease states [12,13], and it has become clear that widespread variation in methylation patterns exist between individuals [14]. Accumulating evidence suggests that DNA methylation patterns can be affected by genetic variants (mQTLs) [15], environmental exposures [16] and stochastic factors [17,18], but it is largely unknown how much each of these factors account for overall variation between individuals in DNA methylation across the genome. Twin studies provide insight into the proportion of inter-individual variation in DNA methylation that is due to genetic variation, environmental effects and stochastic variation [19].

Because MZ twins derive from a single zygote and, therefore, have (nearly) identical DNA sequences (see, for example, Ye *et al.*, 2013 [20]), the comparison of DNA methylation patterns of MZ twins allows one to examine the extent to which differences in methylation between human individuals are related to environmental and stochastic events. Previous studies have highlighted that various tissues of MZ twins already show differences in DNA methylation at birth [21,22] and that differences between twins for average genome-wide DNA methylation, total histone acetylation levels and methylation at certain loci increase with age (referred to as "epigenetic drift") [23]. Although a cross-sectional study of DNA methylation discordance in saliva from 34 MZ pairs (age range: 21–55 years) found no evidence for larger differences in DNA methylation in older MZ pairs [24], results from a cross-sectional analysis based on 230 MZ pairs (age range: 18–89 years) suggested a gradual increase of DNA methylation discordance in MZ twins from early adulthood to advanced age at various candidate loci, which was supported by longitudinal data from 19 elderly MZ pairs [25].

In the past few years, various studies have examined DNA methylation at a set of candidate genes or particular genomic regions in MZ and dizygotic (DZ) twins [26–31], usually reporting greater similarity of MZ twins compared to DZ twins, suggesting that heritable influences contribute to DNA methylation variation at specific regions. While CpG sites at some imprinted loci showed evidence for moderate to high heritability in blood samples from adolescent and middle-aged twins [29], other genomic regions, including the major histocompatibility complex (MHC) region, showed little evidence for genetic influences on DNA methylation variation [28]. Twin studies also highlighted variation between tissues in the importance of genetic influences on methylation of candidate loci at birth [30]. A longitudinal classical twin study of three candidate genes (*DRD4*, *SLC6A4/SERT* and *MAOA*) based on buccal cells indicated that changes in the methylation of these genes within individuals between age five and 10 are mostly attributable to non-shared environmental influences and stochastic variation [31]. Clearly, twin studies of candidate regions suggest that there is broad variation in the importance of heritable influences and environmental or stochastic variation to DNA methylation at different regions.

To date, only a few genome-scale analyses of DNA methylation have been performed using the classical twin design, including a study of ~12,000 CpG sites within islands [32], two studies that

used a promoter-specific array targeting ~27,000 CpG sites (Illumina 27k) [21,33] and two studies that used the Infinium HumanMethylation450 array (Illumina 450k) [22,34], which assesses ~485,000 CpG sites across a variety of regions in the genome, including gene bodies and intergenic regions [35]. The studies that assessed heritability consistently reported that the average heritability of the methylation level at CpGs across the genome is low to moderate when all sites are considered, although the heritability of individual CpGs ranges between 0% and 100%. The following estimates of average heritability across genome-wide CpGs have been reported to date (based on all analyzed CpGs): 18% in blood from 32- to 80-year-old twins (21 MZ pairs and 31 DZ pairs) [33], 5% in placenta, 7% in human umbilical vascular endothelial cells (HUVEC) and 12% in cord-blood mononuclear cells (CBMC) from neonatal twins (22 MZ and 12 DZ pair [21]) and 19% in adipose tissue from adult female twins (97 MZ pairs and 162 DZ pairs) [34]. In two studies of neonatal twin tissues, methylation discordance in MZ and DZ twins increased with increasing distance from CpG islands (CGIs) for certain probes (Type I), *i.e.*, differences were larger in the shores and shelves that flank CGIs [21,22]. In the study of adipose tissue, it was noted that the average genome-wide heritability of DNA methylation was higher when restricting to the most variable CpG sites (for the top 10% CpGs of which the methylation level varied most between subjects, the average heritability was 37%) [34]. It was also found that gene body and intergenic regions showed higher average methylation levels, more variation between subjects and higher heritability compared to promoter regions in adipose tissue [34].

To summarize, there is great interest in unraveling the factors that contribute to variation in DNA methylation between persons, but most previous twin studies of DNA methylation have been limited to candidate genes or a subset of regulatory regions in the genome (mostly promoter regions and CGIs). Two earlier studies used the Illumina 450k to collect genome-wide data in MZ and DZ twins; one in adipose tissue in adults [34] and one in DNA isolated from buccal cells in infants (10 MZ pairs and five DZ pairs, longitudinal design) [22]. In line with earlier findings suggesting the divergence of DNA methylation profiles with age in MZ twins (mostly based on data from adult twins, cross-sectional comparisons and limited genomic coverage), Martino *et al.* [22] showed that widespread DNA methylation changes occur across the genome in buccal cells between birth and 18 months and that some MZ and DZ pairs already show divergence of DNA methylation profiles, whereas other pairs show stable difference levels or became more similar within the first 18 months after birth. In this paper, we analyzed genome-wide DNA methylation profiles (Illumina 450k) from buccal epithelium. We focused on 10 young and adolescent MZ twin pairs (age 8–19). The aim of our study was to examine how similar the DNA methylation profiles of buccal cells from genetically identical subjects are in childhood and adolescence and whether MZ twin similarity varies between different genomic regions.

Previous studies have highlighted differences in mean methylation level, differences in the effect of methylation level on gene expression and differences in the effect size and direction of the effect on methylation for disease associations across different regions in the genome [6]. These findings indicate that the establishment and maintenance of DNA methylation is differentially regulated in different regions and that a given change in methylation in different areas may have different downstream effects, suggesting that DNA methylation in some regions may be more

tightly controlled than in others. We questioned whether these regional differences are also accompanied by differences in the importance of environmental and stochastic influences *versus* familial factors (genetic variation and shared environment) to inter-individual variation in methylation levels. Therefore, we describe the MZ twin correlations of individual CpGs as a function of various genomic classifications, including the position relative to CGIs (CGI regions, shores, shelves and non-CGI regions), genes (distal to promoter, proximal to promoter, gene body and intergenic) and ENCODE regulatory regions (DNaseI hypersensitive sites (DHS) and transcription factor binding sites (TFBS)). Hereby, our study gives valuable insight into the factors influencing inter-individual genome-wide DNA methylation variation in buccal cells in childhood and adolescence and into the degree to which these influences vary across functional regions in the genome.

2. Experimental

2.1. Subjects

Ten monozygotic twin pairs, who take part in longitudinal studies of the Netherlands Twin Register (NTR), were selected for the current study. There were five young twin pairs [36] whose buccal samples were collected when the twins were between ages 8 and 10 years and five adolescent pairs [37] who were aged 18–19 years at the time of sample collection. In the young group, there were three male pairs and two female pairs, and in the adolescent group, there were two male pairs and three female pairs. The twins were unselected with respect to phenotypic characteristics. Informed consent was obtained from the parents (children) or from the twins themselves (adolescents). The study was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board certified by the U.S. Office of Human Research Protections (IRB number IRB-2991 under Federal-wide Assurance-3703; IRB/institute codes, NTR 03-180). Participants could indicate if they wished to be informed of the results of zygosity testing. Zygosity testing, based on a set of SNPs and VNTRs, as described in Van Beijsterveldt *et al.* 2013 [36], confirmed that all pairs were MZ. In addition to the twin samples, a single sample was used as a genomic DNA control. This DNA sample (CEPH) was derived from a stable cell line (female) from the HapMap project and was run in four replicates on the methylation BeadChip arrays.

2.2. Buccal DNA Collection

The procedures of buccal swab collection [38] and genomic DNA extraction [39] have been described previously. In short, 16 cotton mouth swabs were individually rubbed against the inside of the cheek by the participants and placed in four separate 15 mL conical tubes (four swabs in each tube) containing 0.5 mL STE buffer (100 mM sodium chloride, 10 mM Tris hydrochloride (pH 8.0) and 10 mM ethylenediaminetetraacetic acid) with proteinase K (0.1 mg/mL) and sodium dodecyl sulfate (SDS) (0.5%) per swab. Individuals were asked to refrain from eating or drinking 1 hour prior to sampling. High molecular weight genomic DNA was extracted from the swabs using a high salt (KAc) precipitation followed by a standard chloroform/isoamyl alcohol (24:1)

extraction. The DNA samples were quantified using absorbance at 260 nm with a Nanodrop ND-1000 (Nanodrop Technologies, Wilmington, DE, USA).

2.3. Infinium HumanMethylation450 BeadChip Data Generation

The epigenome-wide methylation data was generated using the Infinium HumanMethylation450 BeadChip Kit (Illumina Inc., San Diego, CA, USA). The Infinium HumanMethylation450 BeadChip is able to interrogate over 450,000 methylation sites across the entire genome, including 99% of RefSeq genes. Content was selected to include gene regulatory regions, such as the promoter, 5' UTR, first exon, gene body and the 3' UTR. Additionally, bead probes were also designed to cover regions adjacent to the CpG islands, such as the shores and shelves [35].

The Infinium DNA methylation assay was performed at the Avera Institute for Human Genetics. The assay was completed exactly as denoted in the manufacturer's protocol. The concentration of genomic DNA used in the Infinium DNA methylation assay was determined by comparing the binding of PicoGreen to known standards (λ DNA) and to the sample DNA. Briefly, 500 ng of genomic DNA was used for bisulfite conversion using the Zymo EZ DNA methylation kit (Zymo Research). Five microliters of bisulfite-converted DNA were whole genome amplified, which was followed by enzymatic end-point fragmentation. The resulting fragments were purified using an isopropanol precipitation, and the resuspended genomic DNA was denatured and hybridized to the BeadChip arrays for 18 hours. Extension, staining and washing were completed manually in flow cells followed by imaging using the iScan system (Illumina, Inc.). The raw data were extracted as *idat* files and were used in the downstream analysis.

2.4. Quality Control, Normalization and Data Processing

The raw intensity files (*idat*) were imported into the R environment [40], where further processing, quality control and normalization took place. The performance of bisulfite control probes confirmed successful bisulfite conversion for all samples. For each sample, we compared the overall (median) methylated signal intensity to the overall unmethylated signal intensity across all probes and compared the overall signal intensity from all CpG probes to the overall background signal ("noise"), as assessed using negative control probes. The overall signal from CpG probes was good and well-separated from the background signal for all samples. As a final quality check of the samples, cluster analysis was performed (cluster method = complete linkage) based on the Euclidean distance between samples, which was calculated from the pair-wise correlations between samples using the most variable probes (probes with an SD of the β-value across all 24 samples >0.10, with probes on the X and Y chromosomes and probes containing SNPs, as described in the next paragraph excluded; Nprobes = 38,359). The results of the cluster analysis were visualized in a dendrogram (see the "Results" section), which showed no outlier samples and illustrated tight clustering of the four replicate measures of control DNA.

Several probe-level QC steps were performed to filter out probes with low performance. For all samples, ambiguously-mapped probes were excluded, based on the definition of an overlap of at least 47 bases per probe from Chen *et al.* [41], and all probes containing an SNP, identified in the

Dutch population [42], within the CpG site (at the C or G position) were excluded, irrespective of minor allele frequency. For each sample individually, probes with an intensity value of zero (not present on the array of a particular sample), probes with a detection *p*-value > 0.01 (calculated using the function *detectionP* from the *minfi* package) and probes with a bead count <3 were excluded. After these steps, probes with a success rate <0.95 across samples were removed from all samples, and the success rate across probes for each sample was computed (range of per sample success rate: 0.9990–0.9998).

After QC, background and red/green color adjustment were applied to the raw probe intensity values using quantile normalization. Normalized intensity values were converted into beta-values (β). The β-value, which represents the methylation level at a CpG for an individual and ranges from 0 to 1, is calculated as:

$$\beta = \frac{M}{M + U + 100}$$

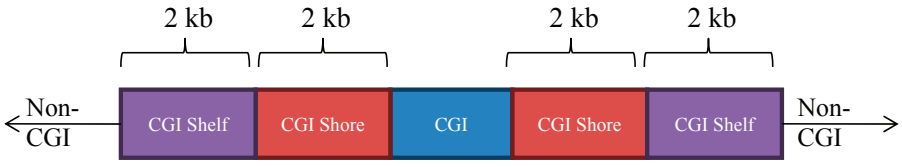
where M = methylated signal, U = unmethylated signal and 100 represents a correction term to control the β-value of probes with a very low overall signal intensity (*i.e.*, probes for which M + U ~ 0 after background subtraction).

Finally, in anticipation of our categorization of CpGs based on the mean β-value across samples, β-values were adjusted to account for (intra-sample) differences in the distributions of methylation values derived from Type I probes (two bead types per CpG site) *versus* Type II probes (one bead type per CpG site) using the beta-mixture quantile normalization method (BMIQ) [43].

2.5. Genomic Annotations

CpGs that passed QC criteria (N = 420,921) were mapped to genomic features, DNase I hypersensitive sites (DHS) and transcription factor binding sites (TFBS), as described by Slieker *et al.* [7]. The genomic feature annotation is based on first assigning CpGs to one of five gene-centric regions: intergenic region (>10 kb from the nearest transcription start site (TSS)), distal promoter (−10 kb to −1.5 kb from the nearest TSS), proximal promoter (−1.5 kb to +500 bp from the nearest TSS), gene body (+500 bp to 3' end of the gene) and downstream region (3' end to +5 kb from 3' end). Next, CpGs were mapped to CGIs (CG content >50%, length >200 bp and observed/expected ratio of CpGs >0.6; locations were obtained from the UCSC genome browser [44]), CGI shore (2-kb region flanking CGI), CGI shelf (2-kb region flanking CGI shore) or non-CGI regions (Figure 1). According to the gene-annotations, 14.4% of all CpGs were located in intergenic regions, 4.7% mapped to the distal promoter, 40.4% to the proximal promoter, 38.6% to the gene body and 1.9% to the downstream region. Thirty three percent of CpGs were located within CGIs, 23.8% in shores, 9.2% in shelves and 34.0% outside CGIs. The locations of DHS and TFBS, which were described by the ENCODE project [3], were downloaded from the UCSC genome browser. Finally, CpGs were mapped to imprinted genes that were described by Yuen *et al.* [45].

Figure 1. Illustration of a CpG island (CGI) with surrounding CGI shores, CGI-shelves and non-CGI regions.



2.6. Statistical Analysis of Twin Data

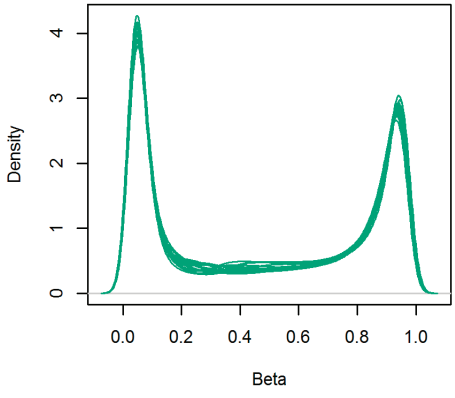
To examine the s imilarity of DNA methylation profiles of MZ twins, we computed correlations between the normalized β-values of MZ co-twins using the following two approaches: (1) for each MZ twin pair, the Spearman correlation (rho) was computed between the β-values of Twin 1 and the β-values of Twin 2 (across all CpGs, *i.e.*, CpGs are cases), as a measure of the overall similarity of the methylation profiles of each twin pair; (2) for each CpG, the Spearman correlation (rho) was computed between the β-value of Twin 1 and the β-value of Twin 2 (across all 10 MZ twin pairs, *i.e.*, MZ twin pairs are cases), as a measure of the similarity of the methylation level of a CpG in MZ twins. For Scenario 2, we describe the range of correlations for the most variable CpGs. The most variable CpGs were additionally grouped by genomic annotations and average methylation level. For each CpG, the average methylation level (β-value) and the standard deviation (SD) were computed across subjects (20 MZ twins). Based on the average β, CpGs were classified as hypomethylated (mean $\beta < 0.3$), intermediately methylated (mean $\beta \geq 0.3$–0.7), or hypermethylated (mean $\beta \geq 0.7$). Based on the SD, CpGs were classified as "most variable CpGs" if they had an SD ≥ 0.05.

3. Results and Discussion

3.1. DNA Methylation Level across the Genome

After QC of the methylation data, 420,921 CpGs from 10 monozygotic twin pairs were analyzed. The methylation level across genome-wide CpGs showed the typical bimodal distribution for each subject (Figure 2). Based on our β-value cut-offs (see the "Experimental" section); 184,765 CpGs (43.9%) were classified as hypomethylated, 64,829 CpGs (15.4%) were intermediately methylated and 171,327 CpGs (40.7%) were hypermethylated. CGIs were on average hypomethylated, with CGIs in proximal promoter regions showing a narrow range of average methylation levels across individual CpGs and CGIs in gene bodies, downstream regions and intergenic regions showing a broader range of methylation levels across individual CpGs (see Figure 3). Compared to CGIs, the shores, shelves and non-CGI regions on average had a higher methylation level, except for proximal promoter shores. Shores generally showed the widest range of average methylation levels across individual CpGs, when compared to CGIs, shelves and non-CGI regions (Figure 3).

Figure 2. Density of β-values after normalization for all twin samples.



3.2. Similarity of Genome-Wide Methylation Profiles of MZ Twins

A cluster analysis of the methylation data revealed that all but one MZ twin clustered closely together with their co-twin (Figure 4), which could be related to differences in the cellular composition of the samples of this twin pair. Buccal swab samples are mainly composed of buccal epithelial cells with a small proportion of leukocytes, but the exact proportions may vary between persons. Using information from a reference 450k methylation dataset [7], we examined potential variation between twin samples in the proportion of buccal *versus* blood cells, by clustering the twin data based on methylation values at CpGs that showed a large difference in methylation between blood and buccal samples in the reference dataset (see the Supplementary Methods). Although some variation was indicated by this approach, the exclusion of twin samples with putatively deviant cellular proportions yielded similar results for the correlation analyses (see Table S1 and Figure S1), and we therefore decided to keep all samples in the analyses reported in this paper.

Figure 5 shows a typical scatterplot of genome-wide CpG methylation levels in buccal cells from an MZ twin pair. It illustrates that overall, the buccal DNA methylation profiles of MZ pairs are highly concordant when all CpGs are considered (rho = 0.981–0.994 for different MZ pairs, mean rho = 0.991); however, these correlations are to a large extent driven by invariable CpGs that are hypomethylated or hypermethylated in both twins. For pairs of unrelated subjects, the mean correlation was 0.983 (range: 0.970–0.992). When comparing only the most variable CpGs (SD of $\beta \geq 0.05$), the correlations ranged from 0.869 to 0.989 (mean rho = 0.966) in MZ twins (and mean rho = 0.859, range: 0.608–0.963 for unrelated subjects). Thus, when looking only at CpGs that may vary between individuals, the overall pattern of methylation across CpGs is still highly similar within MZ pairs on average, but more variation between individual pairs becomes visible, as the methylation level at variable CpGs overall was more strongly correlated for some MZ pairs than for others. This finding is in line with the results from Martino *et al.* based on buccal cells from twins at birth and at the age of 18 months [22], which also indicated that some MZ pairs are more similar than other pairs with respect to their DNA methylation profiles.

Figure 3. Average methylation level of individual CpGs across gene regions (**a**), CpG islands (CGI) and non-CGI regions (**b**) and for each genomic feature separately (**c**).

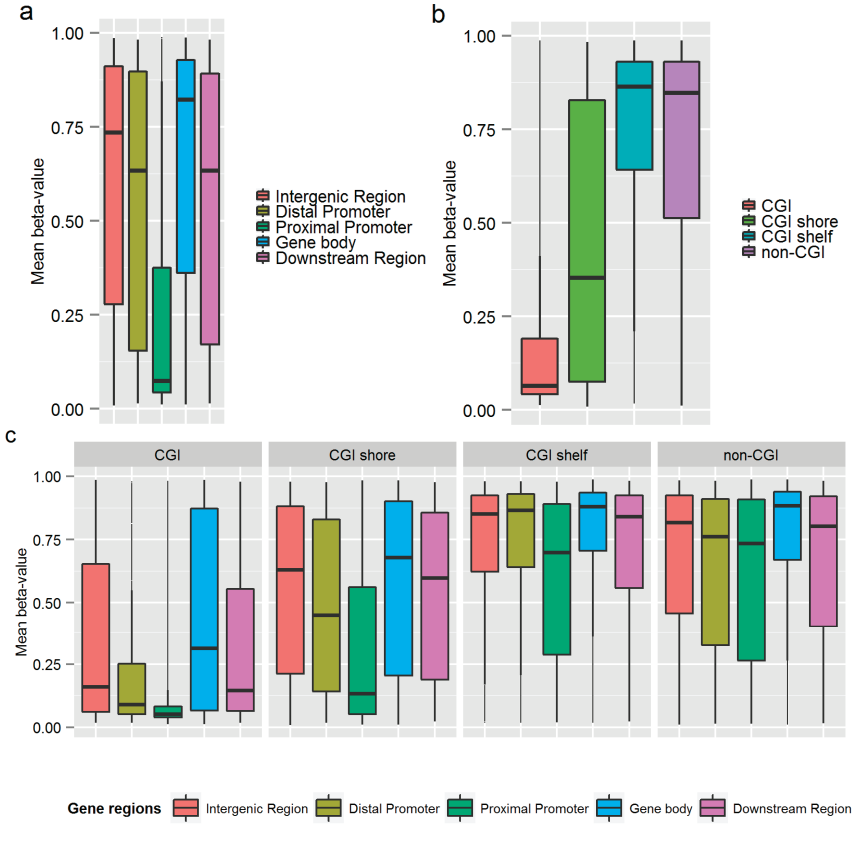


Figure 4. Cluster dendrogram of all twin and control samples. From left to right, the first two branches separate the control samples (HapMap cell line DNA) from the buccal samples from twins.

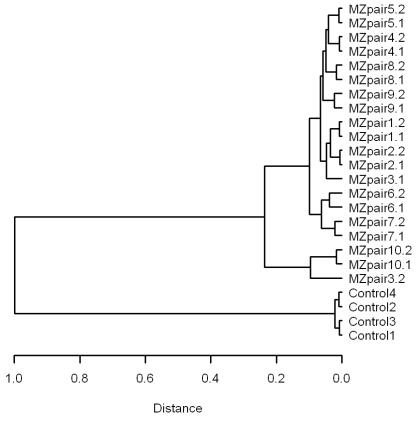
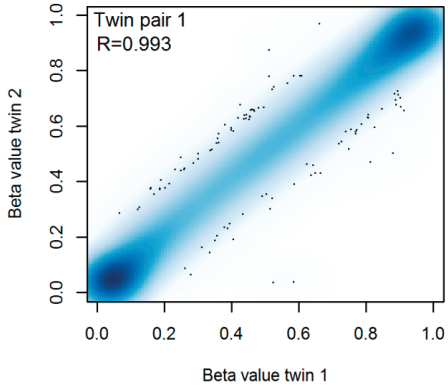


Figure 5. Smooth scatterplot of DNA methylation levels (β-values) at 420,921 CpGs in buccal cells from a monozygotic twin pair.



3.3. Similarity of the Methylation Level at Individual CpGs in MZ Twins

Although all ten MZ twin pairs showed high overall similarity of methylation across genome-wide CpGs, some CpGs differed within MZ twin pairs (Figure 5), and we questioned how similar the methylation level at individual CpGs is when summarized across all MZ pairs. To this end, we computed for each CpG the correlation between methylation values of MZ twins. A high MZ twin correlation for a CpG suggests that MZ co-twins consistently show similar methylation levels at this CpG, indicating little stochastic and environmental variation (including measurement error) at this site, whereas a low MZ twin correlation for a CpG suggests dissimilar methylation levels in co-twins, which is indicative of a large degree of stochastic and environmental influences.

Summarizing the individual CpG correlations over all 420,921 CpGs, the average MZ twin correlation was 0.31 (median = 0.35, range: −0.963–1), which is in line with the low heritability across genome-wide CpGs reported by previous studies [21,33,34]. However, as the majority of CpGs showed very little variation in the methylation level between subjects, all subsequent analyses were conducted using only the most variable sites (N = 59,041), which showed an average genome-wide correlation of 0.54 (median = 0.54, range: −0.661–1) in MZ twins. These findings suggest that while the large majority of CpGs are either hyper- or hypo-methylated and show little between-individual variation in DNA methylation in buccal samples, a small portion does vary markedly, and these CpGs are on average moderately to strongly correlated in MZ twins.

Table 1 describes the MZ twin correlations separately for various genomic regions and separately for hypomethylated, intermediately methylated and hypermethylated CpGs. Comparing the different gene-centric classifications, the average MZ twin correlation was highest for CpGs in proximal promoter areas (mean rho = 0.57) and lowest for gene body CpGs (mean rho = 0.51). The MZ twin correlation of methylation values was also lower on average in CGI shores (mean rho = 0.54), shelves (mean rho = 0.50) and non-CGI regions (mean rho = 0.49) compared to CGIs (mean rho = 0.66). Looking at the MZ twin correlations across genome annotations separately for hypomethylated (29.8% of variable CpGs), intermediately methylated (50.0% of variable CpGs) and hypermethylated CpGs (20.2% of variable CpGs), the median MZ twin

correlation was consistently lower in the shelves, shores and non-CGI regions compared to CGIs, for all genic and intergenic regions, and this difference was most pronounced for hypomethylated CpGs (Figure 6). This observation suggests that the relative influence of familial *versus* individual-specific influences differs between these regions, with regions of low CpG density showing more variation due to individual-specific environmental and stochastic factors compared to CpG dense regions. Larger methylation discordance of MZ twins in CGI shores and shelves was also previously indicated by studies of neonatal twins [21,22]. Our results thus replicate previous findings and add to these findings that the pattern previously observed in MZ twins at birth is also visible in childhood and adolescence.

Figure 6. MZ twin correlations for individual CpGs grouped by genomic region and average methylation level. Hypo = Hypomethylated. Inter = intermediate methylation. Hyper = Hypermethylated. Results are based on the most variable CpGs (N = 59,041).

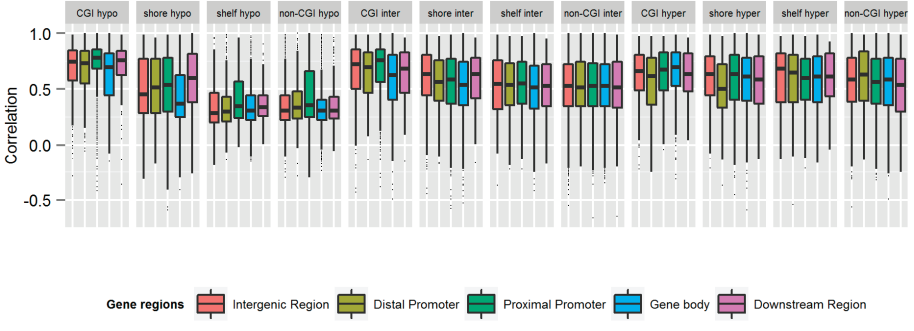


Table 1. Spearman correlation between the methylation values of monozygotic (MZ) twins for individual CpGs. Results are based on the most variable CpGs (N = 59,041).

Category	N CpGs	Mean rho	Median rho	Min rho	Max rho
All CpGs	59,041	0.54	0.54	−0.661	1
Gene-centric annotations	N CpGs (%)	Mean rho	Median rho	Min rho	Max rho
Intergenic (>10 kb from TSS)	11,430 (19.4%)	0.52	0.53	−0.56	1
Distal Promoter (−10 kb to −1.5 kb from TSS)	3193 (5.4%)	0.53	0.53	−0.54	1
Proximal Promoter (−1.5 kb to +500 bp from TSS)	17,880 (30.3%)	0.57	0.62	−0.66	1
Gene Body (+500 bp to 3' end)	25,163 (42.6%)	0.51	0.50	−0.59	1
Downstream region (3' end to +5 kb from 3' end)	1375 (2.3%)	0.55	0.55	−0.66	1
CGI annotations	N CpGs (%)	Mean rho	Median rho	Min rho	Max rho
CGI	10,576 (17.9%)	0.66	0.73	−0.49	1
CGI shore	14,803 (25.1%)	0.54	0.55	−0.59	1
CGI shelf	6001 (10.2%)	0.50	0.49	−0.54	1
Non-CGI	27,661 (46.9%)	0.49	0.47	−0.66	1
Methylation level	N CpGs (%)	Mean rho	Median rho	Min rho	Max rho
Hypomethylated (average beta <0.3)	17,581 (29.8)	0.48	0.42	−0.59	1
Intermediately methylated (average beta ≥0.3–0.7)	29,519 (50.0)	0.55	0.56	−0.66	1
Hypermethylated (average beta ≥0.7)	11,941 (20.2)	0.58	0.61	−0.59	1

The most strongly correlated CpGs in MZ twins (mean rho = 0.73) were hypomethylated CpGs located in proximal promoter CGIs (N = 2547 CpGs, constituting 4.3% of the most variable CpGs, and 2.9% of all CpGs located in proximal promoter CGIs), while MZ twin correlations on average were lowest in hypomethylated non-CGI gene body CpGs (N = 2972 CpGs constituting 5.0% of the most variable CpGs, Mean rho = 0.34). In combination with our observation that most proximal promoter CpGs are on average hypomethylated (Figure 3), these findings indicate that DNA methylation variation is generally depleted in proximal promoter CGIs. Yet, a small proportion of CpGs in proximal promoter CGIs does show marked variation in young and adolescent individuals, and the high average MZ twin correlations at these sites suggest that this variation may be to a large extent under genetic control.

3.4. MZ Twin Resemblance at CpGs in ENCODE Regulatory Regions

To further examine DNA methylation at regulatory regions in the genome, we focused specifically on CpGs located within DNase I hypersensitive sites (DHS) and CpGs within transcription factor binding sites (TFBS) identified by the ENCODE project. It has previously been described that these regions are enriched among disease-associated genetic variants [3], but it has not yet been studied to which extent heritable *versus* other sources of variation account for variation in DNA methylation in these regions. We found that both DHS and TFBS were on average hypomethylated, as expected for transcriptionally active DNA (DHS: mean $\beta = 0.27$, median = 0.09; TFBS: mean $\beta = 0.24$, median = 0.08). The most variable CpGs in these areas (representing 16.2% of all CpGs in DHS and 13.7% of CpGs in TFBS) showed a mean correlation of 0.52 (DHS) and 0.53 (TFBS), respectively, in MZ twins. These results suggest that buccal cells overall show little variation in the methylation level at the majority of CpGs within DHS and TFBS. A small proportion of CpGs in DHS and TFBS, however, does show variation between individuals, and these sites were moderately to strongly correlated in MZ twins, suggesting that these sites may be of particular interest for follow-up in future studies of heritability.

3.5. MZ Twin Resemblance at CpGs in Imprinted Genes

At imprinted genes, one of the alleles is typically methylated to repress expression, while the other allele is unmethylated, depending on the parent from whom the allele was inherited. This results in a methylation level of around 50% at imprinted CpGs when the two alleles are measured simultaneously. A previous twin study demonstrated moderate to high heritability at CpGs at two imprinted loci [29], suggesting that CpGs within imprinted genes may on average show more heritable variation compared to most other genome-wide CpGs. In our dataset, 346 CpGs were located in DMRs (differentially methylated regions) of 59 imprinted genes, described by Yuen *et al.* [45]. These genes were identified as imprinted in human placental tissue, and although some of these genes showed similar methylation patterns in one or multiple fetal tissues, including muscle, brain and kidney, it is unknown whether these genes are also imprinted in buccal cells. From the Yuen *et al.* set, 144 CpGs in 46 genes (see Table S2) showed a methylation level indicative of imprinting in our data (intermediate methylation; mean $\beta \geq 0.3$–0.7). The average MZ

twin correlation for this set of CpGs was 0.47 (median rho = 0.50), suggesting that MZ twin correlations at imprinted gene CpGs on average are comparable to the MZ twin correlation at intermediately methylated CpGs in general.

3.6. Interpretation and Future Directions

The average twin correlation of methylation values for MZ twins at individual CpGs was low across all measured genome-wide CpGs, but it was moderate to large on average when focusing only on variably methylated CpGs. This is in line with results from a heritability analysis of DNA methylation in adipose tissue, which showed that the average heritability across all CpGs was higher for the top 10% of CpGs with the largest standard deviation of methylation level across subjects [34].

Importantly, in addition to the effects of environmental and stochastic influences, differences in DNA methylation within MZ twin pairs may result from variation in the cellular composition of samples and from technical variation (including measurement error). Buccal swab samples are mainly composed of buccal epithelial cells with a small proportion of leukocytes, but the exact proportions may vary between persons, which could lead to methylation variation within MZ twin pairs that mainly tag differences in cell type composition. We examined the impact of variation in the proportion of buccal *versus* leukocytes on our data by studying the methylation patterns of all twin samples at CpGs with a large methylation level difference between buccal and blood samples (see Supplementary Methods). Exclusion of four twin pairs, for which this approach indicated a more deviant cellular composition in one or both twins (lower proportion of buccal epithelial cells; see Figure S2 and Table S3), however, had very little impact on the average MZ twin correlations reported in this paper and led to the same conclusions (See Table S1 and Figure S1 for the results based on the exclusion of the putatively more heterogeneous samples).

With respect to technical variation, it is important to note that if the actual methylation status at a particular site is either completely unmethylated (0%) or completely methylated (100%) without true biological variation between subjects, some variability between the measured values of individuals is expected due to technical variation [46]. It is therefore likely that at sites that were on average hypomethylated or hypermethylated in our data, technical variation may account for a large part of the observed variation (although true biological variation may of course also account for part of the variation at these sites). An interesting question that largely remains to be examined is what types of environmental influences can induce changes in DNA methylation and thereby possibly impact on gene expression. Although our study design does not provide insight with regard to which of the observed differences between twins are the result of different environmental exposures and which differences have arisen due to stochastic variation in molecular processes, future studies of MZ twins who are discordant for environmental exposures should allow one to examine the effects of such influences on DNA methylation. Our finding that many CpGs in the genome show dissimilar methylation levels in young and adolescent MZ twins indicates that it is of interest for further studies to specifically search for regions in the genome where differential methylation in MZ twin coincides with differential exposures. As we observed that DNA methylation in MZ twins is overall less similar at CpGs in non-CGI regions, CGI shores and

shelves, these regions are of particular interest to studies examining environmental exposures, as these regions may show the strongest effects of environmental influences.

To check whether the lower average MZ twin correlation at hypomethylated sites is not merely related to the distribution of β-values being truncated at zero (and one) by definition, we also ran the analyses on M-values ($M = \log_2\left(\frac{\beta}{1-\beta}\right)$), which have better statistical properties, but reduced biological interpretability, compared to β-values [47]. The MZ twin correlations based on M-values were highly similar to those based on β-values and showed a similar genome-wide average (Table S4) and a similar pattern across regions and mean methylation categories (Figure S3). Irrespective of whether the lower resemblance of MZ twins mainly reflects that these sites harbor more biological variation that is unique to MZ twins or reflects that more variation at these sites is related to measurement error, our findings provide useful information for future heritability and mQTL studies. CpGs that are very weakly correlated between MZ twins are not likely to show high heritability or strong effects of DNA variants on the methylation level.

A limitation of our study is the modest study size, which limited the scope of our analyses to the description of the major patterns (*i.e.*, averages) of twin correlations across the genome. A second limitation is that we did not include DZ twins. The correlation between the phenotypes of MZ twins summarizes the contribution of heritable influences and shared environmental factors to phenotypic variation. It thus remains to be established whether CpGs that were strongly correlated in MZ twin pairs are strongly affected by heritable influences or whether shared environmental influences are also important at these sites. Of interest, a previous twin study of DNA methylation in adipose tissue identified a number of CpGs with evidence for shared environmental effects on DNA methylation [34]. Future studies that include data from both MZ and DZ twin pairs are needed to separate the effects of heritable effects and shared environment on genome-wide DNA methylation profiles in buccal cells. Our results indicate that such studies are worthwhile, as we have shown that methylation at a number of CpGs is strongly correlated between MZ twins in buccal cells.

We studied DNA methylation extracted from buccal samples, which may be easier to collect than blood samples in, e.g., young children, and are therefore well-suited for large-scale studies in humans. A relevant question is how representative DNA methylation extracted from these samples is for DNA methylation variation in other tissues and whether methylation studies of buccal *vs.* blood-derived DNA would lead to similar insights. Although DNA methylation patterns are to a large extent tissue-specific [7] and epigenetic changes arising later in life in one tissue may not be detectable in others, epigenetic variation that is established early in development is more likely to be reflected in multiple tissues [4]. Yet, the methylation patterns of buccal cells are likely to be more informative to the methylation state of other ectoderm-derived tissues, whereas methylation patterns in blood may be more comparable to other mesoderm-derived tissues. Finally, it may be regarded as an advantage that compared to blood, which consists of many different cell types, buccal samples represent a relatively homogenous sample type [48], in the sense that it consists of only two major cell types, which potentially makes correction for cell types more straightforward. On the other hand, an advantage of blood samples is that they may provide more insight into DNA methylation variation related to immune system-mediated processes in the body, which are

important in many diseases. To conclude, blood and buccal samples are both valuable for gaining insight into the overall importance of heritable and environmental factors to DNA methylation variation in the genome, and our study showed that the average genome-wide MZ twin correlation for DNA methylation in buccal cells is similar to the average correlation previously reported for peripheral blood [33].

4. Conclusions

To summarize, we computed genome-wide MZ twin correlations for the buccal DNA methylation level at individual CpGs. Methylation levels in MZ twins were moderately to strongly correlated at CpGs with the largest inter-individual variation, which constituted a relatively small proportion of the CpGs that were measured. The average MZ twin correlation across all CpGs was relatively low (mean rho = 0.31), which is similar to findings from previous twin studies [21,33]. Although most CpGs within CGIs were on average hypomethylated, some of them showed large variation in methylation levels. We observed that CpGs with variable methylation levels were more strongly correlated in MZ twins when located in CGIs compared to CpGs in shores and shelves. CpGs in DHS and TFBS were generally hypomethylated, as expected for regulatory active DNA, but CpGs in these regions that were more variably methylated were moderately to strongly correlated in MZ twin pairs, in line with our findings for variably methylated CpGs in general. To conclude, we have shown that in buccal samples from young and adolescent MZ twins, most CpGs show an average methylation level close to zero or 100% and little inter-individual variation, and a subset of CpGs show larger variability with evidence for a familial component (DNA sequence variation or shared environment). These findings are relevant for future heritability studies of DNA methylation and for mQTL studies.

Acknowledgments

This study was funded by Genetics of Mental Illness: A lifespan approach to the genetics of childhood and adult neuropsychiatric disorders and comorbid conditions (European Research Council (ERC-230374)); and by the Avera McKennan Hospital and University Health Center. We would like to thank the twins who participated in this study.

Author Contributions

Gareth E. Davies, Meike Bartels, and Dorret I. Boomsma conceived of the study. Erik A. Ehli, Zachary M. Weber and Gareth E. Davies performed the laboratory work. Jenny van Dongen and Erik A. Ehli performed the data analysis. Roderick C. Slieker, Bastiaan T. Heijmans, and P. Eline Slagboom provided advice and gave assistance with the data processing and analysis. Jenny van Dongen drafted the manuscript with significant input from all the other authors.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Hindorff, L.A.; MacArthur, J.; Morales, J.; Junkins, H.A.; Hall, P.N.; Klemm, A.K.; Manolio, T.A. A Catalog of Published Genome-Wide Association Studies. Available online: http://www.genome.gov/gwastudies/ (accessed on 27 January 2014).
2. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J.P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367.
3. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74.
4. Mill, J.; Heijmans, B.T. From promises to practical strategies in epigenetic epidemiology. *Nat. Rev. Genet.* **2013**, *14*, 585–594.
5. Goldberg, A.D.; Allis, C.D.; Bernstein, E. Epigenetics: A landscape takes shape. *Cell* **2007**, *128*, 635–638.
6. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **2012**, *13*, 484–492.
7. Slieker, R.C.; Bos, S.D.; Goeman, J.J.; Bovee, J.V.; Talens, R.P.; van der Breggen, R.; Suchiman, H.E.; Lameijer, E.W.; Putter, H.; van den Akker, E.B.; *et al*. Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics Chromatin* **2013**, doi:10.1186/1756-8935-6-26.
8. Maunakea, A.K.; Nagarajan, R.P.; Bilenky, M.; Ballinger, T.J.; D'Souza, C.; Fouse, S.D.; Johnson, B.E.; Hong, C.; Nielsen, C.; Zhao, Y.; *et al*. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **2010**, *466*, 253–257.
9. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **2002**, *16*, 6–21.
10. Horvath, S.; Zhang, Y.; Langfelder, P.; Kahn, R.S.; Boks, M.P.; van Eijk, K.; van den Berg, L.H.; Ophoff, R.A. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* **2012**, doi:10.1186/gb-2012-13-10-r97.
11. Ziller, M.J.; Gu, H.; Muller, F.; Donaghey, J.; Tsai, L.T.; Kohlbacher, O.; de Jager, P.L.; Rosen, E.D.; Bennett, D.A.; Bernstein, B.E.; *et al*. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **2013**, *500*, 477–481.
12. Kuehnen, P.; Mischke, M.; Wiegand, S.; Sers, C.; Horsthemke, B.; Lau, S.; Keil, T.; Lee, Y.A.; Grueters, A.; Krude, H. An Alu element-associated hypermethylation variant of the POMC gene is associated with childhood obesity. *PLoS Genet.* **2012**, *8*, e1002543.
13. Dempster, E.L.; Pidsley, R.; Schalkwyk, L.C.; Owens, S.; Georgiades, A.; Kane, F.; Kalidindi, S.; Picchioni, M.; Kravariti, E.; Toulopoulou, T.; *et al*. Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. *Hum. Mol. Genet.* **2011**, *20*, 4786–4796.
14. Talens, R.P.; Boomsma, D.I.; Tobi, E.W.; Kremer, D.; Jukema, J.W.; Willemsen, G.; Putter, H.; Slagboom, P.E.; Heijmans, B.T. Variation, patterns, and temporal stability of DNA methylation: Considerations for epigenetic epidemiology. *FASEB J.* **2010**, *24*, 3135–3144.

15. Bell, J.T.; Pai, A.A.; Pickrell, J.K.; Gaffney, D.J.; Pique-Regi, R.; Degner, J.F.; Gilad, Y.; Pritchard, J.K. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **2011**, doi:10.1186/gb-2011-12-1-r10.
16. Heijmans, B.T.; Tobi, E.W.; Stein, A.D.; Putter, H.; Blauw, G.J.; Susser, E.S.; Slagboom, P.E.; Lumey, L.H. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 17046–17049.
17. Jeffries, A.R.; Perfect, L.W.; Ledderose, J.; Schalkwyk, L.C.; Bray, N.J.; Mill, J.; Price, J. Stochastic choice of allelic expression in human neural stem cells. *Stem Cells* **2012**, *30*, 938–1947.
18. Waterland, R.A.; Dolinoy, D.C.; Lin, J.R.; Smith, C.A.; Shi, X.; Tahiliani, K.G. Maternal methyl supplements increase offspring DNA methylation at Axin Fused. *Genesis* **2006**, *44*, 401–406.
19. Van Dongen, J.; Slagboom, P.E.; Draisma, H.H.; Martin, N.G.; Boomsma, D.I. The continuing value of twin studies in the omics era. *Nat. Rev. Genet.* **2012**, *13*, 640–653.
20. Ye, K.; Beekman, M.; Lameijer, E.W.; Zhang, Y.; Moed, M.H.; van den Akker, E.B.; Deelen, J.; Houwing-Duistermaat, J.J.; Kremer, D.; *et al*. Aging as accelerated accumulation of somatic variants: Whole-genome sequencing of centenarian and middle-aged monozygotic twin pairs. *Twin Res. Hum. Genet.* **2013**, *16*, 1026–1032.
21. Gordon, L.; Joo, J.E.; Powell, J.E.; Ollikainen, M.; Novakovic, B.; Li, X.; Andronikos, R.; Cruickshank, M.N.; Conneely, K.N.; Smith, A.K.; *et al*. Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. *Genome Res.* **2012**, *22*, 1395–1406.
22. Martino, D.; Loke, Y.J.; Gordon, L.; Ollikainen, M.; Cruickshank, M.N.; Saffery, R.; Craig, J.M. Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance. *Genome Biol.* **2013**, doi:10.1186/gb-2013-14-5-r42.
23. Fraga, M.F.; Ballestar, E.; Paz, M.F.; Ropero, S.; Setien, F.; Ballestar, M.L.; Heine-Suner, D.; Cigudosa, J.C.; Urioste, M.; Benitez, J.; *et al*. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 10604–10609.
24. Bocklandt, S.; Lin, W.; Sehl, M.E.; Sanchez, F.J.; Sinsheimer, J.S.; Horvath, S.; Vilain, E. Epigenetic predictor of age. *PLoS One* **2011**, *6*, e14821.
25. Talens, R.P.; Christensen, K.; Putter, H.; Willemsen, G.; Christiansen, L.; Kremer, D.; Suchiman, H.E.; Slagboom, P.E.; Boomsma, D.I.; Heijmans, B.T. Epigenetic variation during the adult lifespan: Cross-sectional and longitudinal data on monozygotic twin pairs. *Aging Cell* **2012**, *11*, 694–703.
26. Boks, M.P.; Derks, E.M.; Weisenberger, D.J.; Strengman, E.; Janson, E.; Sommer, I.E.; Kahn, R.S.; Ophoff, R.A. The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS One* **2009**, *4*, e6767.

27. Coolen, M.W.; Statham, A.L.; Qu, W.; Campbell, M.J.; Henders, A.K.; Montgomery, G.W.; Martin, N.G.; Clark, S.J. Impact of the genome on the epigenome is manifested in DNA methylation patterns of imprinted regions in monozygotic and dizygotic twins. *PLoS One* **2011**, *6*, e25590.
28. Gervin, K.; Hammero, M.; Akselsen, H.E.; Moe, R.; Nygard, H.; Brandt, I.; Gjessing, H.K.; Harris, J.R.; Undlien, D.E.; Lyle, R. Extensive variation and low heritability of DNA methylation identified in a twin study. *Genome Res.* **2011**, *21*, 1813–1821.
29. Heijmans, B.T.; Kremer, D.; Tobi, E.W.; Boomsma, D.I.; Slagboom, P.E. Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. *Hum. Mol. Genet.* **2007**, *16*, 547–554.
30. Ollikainen, M.; Smith, K.R.; Joo, E.J.; Ng, H.K.; Andronikos, R.; Novakovic, B.; Abdul Aziz, N.K.; Carlin, J.B.; Morley, R.; Saffery, R.; *et al*. DNA methylation analysis of multiple tissues from newborn twins reveals both genetic and intrauterine components to variation in the human neonatal epigenome. *Hum. Mol. Genet.* **2010**, *19*, 4176–4188.
31. Wong, C.C.; Caspi, A.; Williams, B.; Craig, I.W.; Houts, R.; Ambler, A.; Moffitt, T.E.; Mill, J. A longitudinal study of epigenetic variation in twins. *Epigenetics* **2010**, *5*, 516–526.
32. Kaminsky, Z.A.; Tang, T.; Wang, S.C.; Ptak, C.; Oh, G.H.; Wong, A.H.; Feldcamp, L.A.; Virtanen, C.; Halfvarson, J.; Tysk, C.; *et al*. DNA methylation profiles in monozygotic and dizygotic twins. *Nat. Genet.* **2009**, *41*, 240–245.
33. Bell, J.T.; Tsai, P.C.; Yang, T.P.; Pidsley, R.; Nisbet, J.; Glass, D.; Mangino, M.; Zhai, G.; Zhang, F.; Valdes, A.; *et al*. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet.* **2012**, *8*, e1002629.
34. Grundberg, E.; Meduri, E.; Sandling, J.K.; Hedman, A.K.; Keildson, S.; Buil, A.; Busche, S.; Yuan, W.; Nisbet, J.; Sekowska, M.; *et al*. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.* **2013**, *93*, 876–890.
35. Bibikova, M.; Barnes, B.; Tsan, C.; Ho, V.; Klotzle, B.; Le, J.M.; Delano, D.; Zhang, L.; Schroth, G.P.; Gunderson, K.L.; *et al*. High density DNA methylation array with single CpG site resolution. *Genomics* **2011**, *98*, 288–295.
36. Van Beijsterveldt, C.E.; Groen-Blokhuis, M.; Hottenga, J.J.; Franic, S.; Hudziak, J.J.; Lamb, D.; Huppertz, C.; de Zeeuw, E.; Nivard, M.; Schutte, N.; *et al*. The Young Netherlands Twin Register (YNTR): Longitudinal twin and family studies in over 70,000 children. *Twin Res. Hum. Genet.* **2013**, *16*, 252–267.
37. Estourgie-van Burk, G.F.; Bartels, M.; Hoekstra, R.A.; Polderman, T.J.; Delemarre-van de Waal, H.A.; Boomsma, D.I. A twin study of cognitive costs of low birth weight and catch-up growth. *J. Pediatr.* **2009**, *154*, 29–32.
38. Willemsen, G.; de Geus, E.J.; Bartels, M.; van Beijsterveldt, C.E.; Brooks, A.I.; Estourgie-van Burk, G.F.; Fugman, D.A.; Hoekstra, C.; Hottenga, J.J.; Kluft, K.; *et al*. The Netherlands Twin Register biobank: A resource for genetic epidemiological studies. *Twin Res. Hum. Genet.* **2010**, *13*, 231–245.

39. Meulenbelt, I.; Droog, S.; Trommelen, G.J.; Boomsma, D.I.; Slagboom, P.E. High-yield noninvasive human genomic DNA isolation method for genetic studies in geographically dispersed families and populations. *Am. J. Hum. Genet.* **1995**, *57*, 1252–1254.
40. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online: http://www.R-project.org/ (accessed on 27 January 2014).
41. Chen, Y.A.; Lemire, M.; Choufani, S.; Butcher, D.T.; Grafodatskaya, D.; Zanke, B.W.; Gallinger, S.; Hudson, T.J.; Weksberg, R. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **2013**, *8*, 203–209.
42. Boomsma, D.I.; Wijmenga, C.; Slagboom, E.P.; Swertz, M.A.; Karssen, L.C.; Abdellaoui, A.; Ye, K.; Guryev, V.; Vermaat, M.; van Dijk, F.; *et al*. The Genome of the Netherlands: Design, and project goals. *Eur. J. Hum. Genet.* **2014**, *22*, 221–227.
43. Teschendorff, A.E.; Marabita, F.; Lechner, M.; Bartlett, T.; Tegner, J.; Gomez-Cabrero, D.; Beck, S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. *Bioinformatics* **2013**, *29*, 189–196.
44. Kent, W.J.; Sugnet, C.W.; Furey, T.S.; Roskin, K.M.; Pringle, T.H.; Zahler, A.M.; Haussler, D. The human genome browser at UCSC. *Genome Res.* **2002**, *12*, 996–1006.
45. Yuen, R.K.; Jiang, R.; Penaherrera, M.S.; McFadden, D.E.; Robinson, W.P. Genome-wide mapping of imprinted differentially methylated regions by DNA methylation profiling of human placentas from triploidies. *Epigenetics Chromatin* **2011**, doi:10.1186/1756-8935-4-10.
46. Pan, H.; Chen, L.; Dogra, S.; Teh, A.L.; Tan, J.H.; Lim, Y.I.; Lim, Y.C.; Jin, S.; Lee, Y.K.; Ng, P.Y.; *et al*. Measuring the methylome in clinical samples: improved processing of the Infinium Human Methylation450 BeadChip Array. *Epigenetics* **2012**, *7*, 1173–1187.
47. Du, P.; Zhang, X.; Huang, C.C.; Jafari, N.; Kibbe, W.A.; Hou, L.; Lin, S.M. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinform.* **2010**, doi:10.1186/1471-2105-11-587.
48. Thiede, C.; Prange-Krex, G.; Freiberg-Richter, J.; Bornhauser, M.; Ehninger, G. Buccal swabs but not mouthwash samples can be used to obtain pretransplant DNA fingerprints from recipients of allogeneic bone marrow transplants. *Bone Marrow Transpl.* **2000**, *25*, 575–577.

Characterization of the Genomic Architecture and Mutational Spectrum of a Small Cell Prostate Carcinoma

Alan F. Scott, David W. Mohr, Hua Ling, Robert B. Scharpf, Peng Zhang and Gregory S. Liptak

Abstract: We present the use of a series of laboratory, analytical and interpretation methods to investigate personalized cancer care for a case of small cell prostate carcinoma (SCPC), a rare and aggressive tumor with poor prognosis, for which the underlying genomic architecture and mutational spectrum has not been well characterized. We performed both SNP genotyping and exome sequencing of a Virchow node metastasis from a patient with SCPC. A variety of methods were used to analyze and interpret the tumor genome for copy number variation, loss of heterozygosity (LOH), somatic mosaicism and mutations in genes from known cancer pathways. The combination of genotyping and exome sequencing approaches provided more information than either technique alone. The results showed widespread evidence of copy number changes involving most chromosomes including the possible loss of both alleles of CDKN1B (p27/Kip1). LOH was observed for the regions encompassing the tumor suppressors TP53, RB1, and CHD1. Predicted damaging somatic mutations were observed in the retained TP53 and RB1 alleles. Mutations in other genes that may be functionally relevant were noted, especially the recently reported high confidence cancer drivers FOXA1 and CCAR1. The disruption of multiple cancer drivers underscores why SCPC may be such a difficult cancer to manage.

Reprinted from *Genes*. Cite as: Scott, A.F.; Mohr, D.W.; Ling, H.; Scharpf, R.B.; Zhang, P.; Liptak, G.S. Characterization of the Genomic Architecture and Mutational Spectrum of a Small Cell Prostate Carcinoma. *Genes* **2014**, *5*, 366-384.

1. Introduction

The promise of the Human Genome Project (HGP), for which we mark the tenth anniversary, was that individualized genomics would become a reality for medical diagnosis and care. However, only recently have methods for sequencing, data analysis and the interpretation of variation with respect to medically relevant sequence information become sufficiently robust to make this approach useful. In this paper we compared different methods to investigate the genomic architecture and mutational spectrum of a rare tumor, small cell prostate cancer (SCPC). Our goal was to identify which methods were most informative and what information might provide the best guidance to the patient and his physician. Secondarily, we hoped to provide further characterization for this tumor type that may be of use to the community.

SCPC is a high-grade malignant tumor with neuroendrocrine differentiation sometimes referred to as neuroendocrine prostate cancer (NEPC) [1]. SCPC is often discovered after the occurrence of metastases, has been reported to account for 0.5%–2% of all prostate carcinomas and has a median survival from diagnosis of approximately 12.5 months [2]. The largest SCPC series was published by Wang and Epstein [3] who histologically examined 95 cases of which 92% showed expression

of the neuroendocrine marker CD56 (NCAM1) and of which approximately 80% failed to show elevated PSA levels. Aparicio *et al.* [1] noted that, although rare as a primary diagnosis, NEPC may be more common than appreciated and could account for as much as 25% of lethal prostate cancer.

A few studies have looked at the genomic events characterizing NEPC. Beltran *et al.* [4] measured gene expression using NGS RNA-sequencing and oligonucleotide arrays in NEPC tumors and observed a correlation between overexpression of MYCN and AURKA both of which were amplified at the gene level in 40% of NEPCs. The authors also noted evidence for a TMPRSS2-ERG gene fusion, a lack of the ERG protein marker, high expression of the neuroendocrine genes CGA and SYP, and low expression of the androgen-regulated genes KLK3 (PSA), TMPRSS2 and NXK3.1. Beltran *et al.* [4] further showed that NEPC cell lines were sensitive to the AURKA inhibitor danusertib which produced a suppression of neuroendocrine expression. However, phase II clinical trials of men with castration-resistant prostate cancer were disappointing [5]. Tzelepi *et al.* [6] produced SCPC xenografts and performed expression studies and genomic profiling using array-CGH (comparative genomic hybridization) which showed up-regulation of UBE2C and other mitotic genes along with the absence of expression of the androgen receptor (AR), RB1, and cyclin D1. A subset of tumors showed microdeletions of RB1. Grasso *et al.* [7] sequenced 50 lethal metastatic castration-resistant prostate cancers (CRPC) which include SCPC. The authors identified subsets of tumors with either disruptions in CHD1 (chromodomain helicase DNA-binding protein 1) or in ETS2 (usually from fusions of ETS2 with TMPRSS2). The authors also found mutations in multiple genes whose protein products physically interact with androgen receptors such as the ERG gene fusion, the chromatin modifying protein MLL2, and FOXA1 among others. Grasso *et al.* [7] further showed that mutated FOXA1 repressed androgen signaling and enhanced tumor growth. The importance of FOXA1 in tumor progression was also demonstrated by Imamura *et al.* [8] who were able to reduce proliferation in cell culture with an siRNA directed against FOXA1. Van Allen *et al.* [9] performed whole exome sequencing on a CRPC bone metastasis and identified a homozygous deletion in PTEN and a nonsense mutation in BRCA2, both of which suggested clinical treatment strategies.

The diversity of findings in the studies cited above likely results from both tumor heterogeneity and the different laboratory methods used. In this study we have used array based genotyping to examine the overall genomic architecture, exome sequencing to identify somatic mutations, various software tools to analyze the resulting data and both public and commercial interpretation tools to attempt to understand the findings.

2. Experimental

2.1. Sample

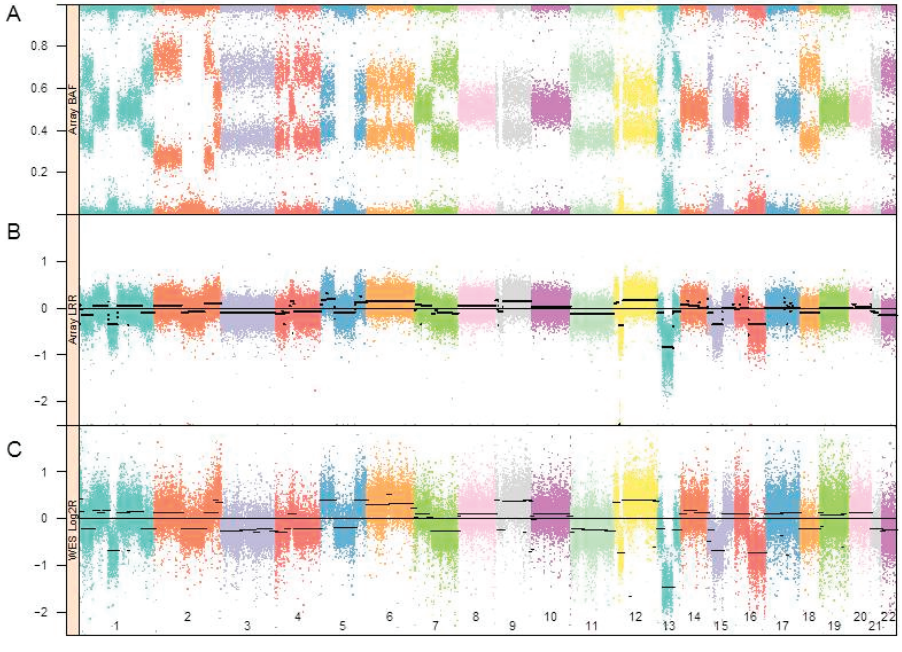
The patient was a consented 63-year old male of European ancestry who presented with hematuria and without elevated levels of prostate specific antigen (PSA). The cancer was detected after the development of metastases and the diagnosis of SCPC was made at the patient's primary care hospital and confirmed at the Johns Hopkins Hospital. High molecular weight DNA was isolated from two needle aspirates of a metastatic Virchow node and from saliva (Scope™

mouthwash, Procter & Gamble) using standard methods but with extended proteinase K digestion time for the biopsies. Cells from the aspirates were examined by a pathologist at the time of collection and the remaining tissue was transferred, unfixed, to the laboratory for DNA isolation. A total of 25.5 µg of DNA was obtained from the tumor and 60 µg from the mouthwash collection.

2.2. Genomic SNP Array and Analysis

The DNA from the metastasis was adjusted to 50 ng/µL and 5 µL (250 ng) were genotyped on an Illumina HumanOmni2.5S BeadChip™ array at the SNP Center of the Genetic Resources Core Facility [10] at the Johns Hopkins School of Medicine. Illumina GenomeStudio software (Illumina Inc., San Diego, CA, USA) was used to process the array data and calculate B-allele frequencies (BAF) and log R Ratios (LRR). The BAF and LRR values generated by GenomeStudio ver. 1.7.4 (Illumina, Inc.) are plotted in Figure 1 (panels A and B, respectively) for all the autosomes. The LRR values were segmented using the circular binary segmentation algorithm implemented in the R package DNAcopy (ver. 1.36.0) [11]. The black lines in the LRR plots are the average LRRs for those segments of the chromosome.

Figure 1. Allele frequencies (**A**) and log R ratios (**B**) estimated by GenomeStudio. Autosomal log R ratios were segmented by circular binary segmentation [11] as indicated in black; (**C**) Log R values from whole exome sequencing were obtained by the EXCAVATOR program [12] and aligned to panels A and B, providing a qualitatively similar profile of the copy number alterations. Black lines depict the segmentation of the log R values.



2.3. Exome Capture and Sequencing

Exome-sequencing was performed at the High-Throughput Sequencing facility of the GRCF. DNA (3 µg) from the tumor and saliva were sheared to a size of 150 to 200 bp using a Covaris E210 system (Covaris Inc., Woburn, MA, USA). End repair and addition of an overhanging "A" base was performed using a NEBNext™ reagent kit (New England Biolabs, Ipswich, MA, USA). DNA fragments were ligated to library adapters (Illumina). The ligated fragments were then size selected through purification using SPRI beads and PCR amplified to prepare the libraries. An Agilent Bioanalyzer DNA1000 assay was used for quality control of the libraries to ensure adequate concentration and appropriate fragment size. Sequencing was performed on an Illumina HiSeq™ 2000 following library capture with an Agilent SureSelect All Exon v3 kit. Sample indexing was applied to distinguish the source of the libraries. Sequence data was processed using CIDRSeqsuite v2.3.0 [13] as follows. Sequence reads were processed through Illumina software generating base calls and corresponding base-call quality scores. Resulting data was aligned to hg19 with the Burrows-Wheeler Alignment (BWA; [14]) tool resulting in a SAM/BAM file. Molecular and optical duplicate reads were flagged using software from the Picard program suite [15]. Post-processing of the aligned data included local realignment around SNPs and indels and base-call quality score recalibration using the Genome Analysis Tool Kit ver. 2 (GATK2; [16]). Single sample calling was done using GATK2 HaplotypeCaller with hard filtering and outputted in VCF 4.0 format. Analyses were performed in accordance with GATK Best Practices recommendations [17,18]. All positions reported are with respect to the hg19 reference sequence.

2.4. Sequence Interpretation

Differences of SNVs and indels between the tumor and normal exomes were computed using both open source and commercial software to identify somatic mutations. The open source programs included ANNOVAR [19], SG-adviser, a suite of web-based tool offered by The Scripps Translational Science Institute [20] Strelka [21] and Seurat [22] which were used with their default settings. In addition, we used the IntOGen-mutations platform [23,24] to identify genes mutated in the TCGA/ICCG cancer genome projects. The Condel online tool [25] was used to obtain scores for missense mutations [26] shown in the exome sequencing data tables. We also used the Ingenuity Variant Analysis web-based application to compare sequence between the tumor and matched normal exomes with differing filtering parameters (see Supplementary Methods: Analysis and Variant Filtering).

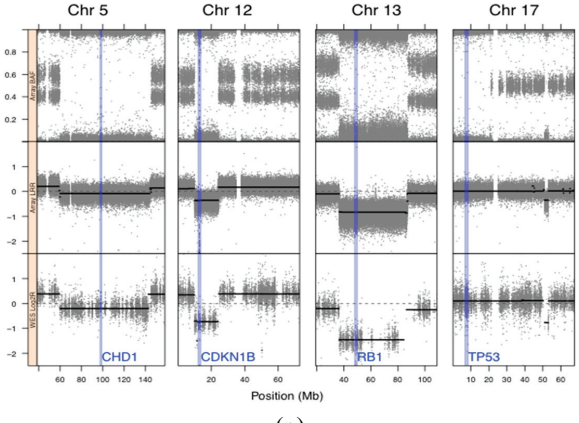
3. Results and Discussion

3.1. Genomic Landscape Detailed from Genotyping Data

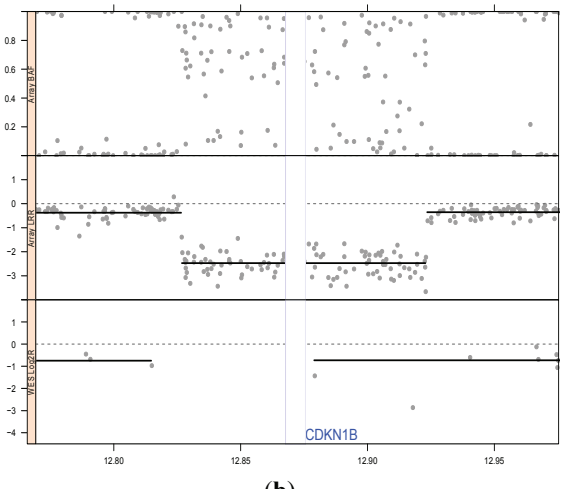
The genotyping array showed a large degree of copy number variation of chromosomal segments and loss of heterozygosity with 17 of the 22 autosomes grossly affected (Figure 1). The genotypes from the Illumina 2.5 M BeadChip processed by circular binary segmentation indicated a modest range in copy number. Visual inspection of the BAF plots clearly identifies blocks of

homozygosity on chromosomes 1, 2, 5, 9, 12, 13, 15, 16, and 17. At least four critical tumor suppressors are within these regions of LOH; TP53 on Chr 17, RB1 on Chr 13, CDKN1B on Chr 12 and CHD1 on chromosome 5 (Figure 2a). While the regions that include TP53 and CHD1 are essentially copy neutral, the LRR plot and segmentation showed a reduced copy number for the chromosome 13 block containing RB1 and a likely homozygous deletion of the chromosome 12 region containing CDKN1B (Figure 2b).

Figure 2. The location of the key genes described in the text. CHD1 (Chr 5: 98.190–98.265 Mb) was not mutated but occurs in a block of LOH; CDKN1B (Chr 12: 12.870–12.875 Mb) has low copy number, RB1 (Chr 13: 48.878 Mb–49.056 Mb) occurs in a block of LOH, has reduced copy number and the retained allele is predicted to be damaging; TP53 (Chr 17: 75.712 Mb–75.909 Mb) occurs in a copy neutral block of LOH and the retained allele is predicted to be damaged. Figure 2b, magnified view of the CDKN1B region showing a likely deletion of the gene supported by both the array (top and middle panels) and exome-sequencing platforms (bottom).



(**a**)



(**b**)

Grasso *et al.* [7] recently described the mutational landscape of castration-resistant prostate cancer (CRPC) based on exome sequencing of 50 lethal metastatic cases. An important finding of their study was that tumors involving CHD1 lacked ETS2 gene fusions and ETS2 mutations. CHD1 is an ATP-dependent chromatin-remodeling enzyme that recognizes histone H3 lysine 4 methylation and is associated with the promoters of active genes where it presumably acts in nucleosome disassembly [27]. In this tumor the regions around ETS2 and TMPRSS2 have normal copy number although SNP arrays are not capable of identifying contiguous chromosomal events and a translocation in non-coding DNA is certainly possible. In contrast, CHD1 clearly falls within a region of LOH and although the exon sequences for the gene are the same as reference we do not know if there might have been mutations in regulatory regions. Unfortunately, we were unable to obtain RNA from the limited biopsy specimen so we could not measure changes in CHD1 expression.

The genotyping data identified a region of LOH and markedly reduced copy number on chromosome 12p (~Chr 12: 10–24 Mb) that includes CDKN1B (Chr 12: 12.870–12.875 Mb) the gene which encodes the p27 cyclin-dependent kinase inhibitor also referred to as p27(KIP1). The exome reads were also significantly reduced and it is possible that the gene is completely absent in the tumor and that the observed reads represent those from contaminating or infiltrating normal cells. CDKN1B blocks cell division in G_0/G_1, regulates cell motility and apoptosis and is classified as a tumor suppressor [28]. Although CDKN1B was not mutated in this tumor, decreased copy number has been associated with tumor pathology in mice (e.g., [29]) and in lethal human epithelial cancers with a poor outcome [30].

Tumor Purity

Several regions of LOH identified by genotyping showed very few spurious mutant sequencing reads indicating that normal cells were not present in the tumor to any significant degree (e.g., TP53 Chr 17:5,578,394). Nevertheless, the mean read depth for the tumor library was 353X and 138X for the normal exome. Although the range in the frequency of mutant reads (Table 1) varies considerably, the fact that 97% of the reads for TP53 are mutant confirms that the tumor was unlikely to have been contaminated with a significant number of normal cells and that differences in allelic fraction at other positions most likely represent tumor heterogeneity. Because of this high level of tumor purity we felt that our exome sequencing provided a good representation of the genomic events without the need to sequence to extraordinary depth to distinguish tumor from infiltrating non-tumor cells.

3.2. Exome Variant Interpretation

We used open source tools to generate variant calls and a mixture of open-source and commercial programs to evaluate the significance of the somatic mutations. All variant filtering is a trade-off between sensitivity and specificity and the risk of missing variants of biological significance must be weighed against a larger number of false calls. The variant calling programs included Haplotype Caller [16,17], Strelka [21] and Seurat [22]. Haplotype Caller was run under GATK best practices with

hard filtering [18], producing single sample calls for both tumor and normal. Strelka and Seurat are somatic variant callers that identify SNVs and indels present in a tumor but not the matched normal sample. Both were run using default settings. The default filter for Seurat removes sequences with a mapping quality score less than 10 while Strelka removes all read pairs with a mapping quality below 40. Seurat identified 3577 somatic SNVs and 2290 indels. In comparison, Strelka found 535 SNVs and 11 indels. Lists from both somatic callers were submitted to the Integrative Onco Genomics single tumor analysis web tool [23] which searches somatic mutations, genes and pathways identified, at the time of the analysis, from 4623 tumor/normal exomes by the International Cancer Genome Consortium (ICCG; [31]) and The Cancer Genome Anatomy (TCGA; [32]) initiatives. Because the Seurat results appeared to have high sensitivity but low specificity we decided to focus on the Strelka list. We manually inspected the Strelka /IntOGen dataset by examining each of the positions using the Integrative Genomic Browser (IGV, ver. 2.3.23; [33]).

Ingenuity Variant Analysis (QIAGEN, Redwood City, CA, USA) was also used to filter and interpret somatic variants under different filtering criteria and single-sample variant call files were uploaded, parsed, and comparatively queried initially for rare (<3% allele frequency in public genome/exome datasets) missense, nonsense, coding indel, or clinically classified (pathogenic/likely pathogenic) variants, confidently called (PHRED-scaled variant call quality >20 in either sample) in genes directly or indirectly (within 2 upstream interaction hops) implicated in “prostate cancer” or “small cell adenocarcinoma” (interactive supplement at https://variants.ingenuity.com/Scott-etal-2014). Other filtering parameters (e.g., 1 upstream interaction hop, frequency in 1000 genomes or Complete Genomics data of less than 0.001%, and broader disease terms including “small cell adenocarcinoma”, “castration-refractory prostate cancer”, and “metastases”) were also evaluated (interactive supplement at https://variants.ingenuity.com/Scott2014ver2).

We grouped SNVs into four categories: (1) The principal findings (Table 1) that we speculate have a strong likelihood of causing or contributing to SCPC; (2) Potentially implicated genes (Table 2) for which there is some evidence of an involvement in cancer but are less certain; (3) Genes with probable passenger mutations (Table S1) whose involvement in cancer is less obvious or lacking and (4) Possibly inherited risk factors (Table S2) for cancer susceptibility. The distinction between each of the first three categories is somewhat arbitrary.

Table 1 shows the top six genes based on their designation as high-confidence or candidate drivers (HCD, CD) of cancer by Tamborero *et al.* [34], their consensus deleteriousness (Condel) scores [26] or their reduced copy number. Premature nonsense mutations were presumed to be deleterious. The fraction of mutant reads at each position was also calculated from the BWA alignment. As discussed above, the predicted copy numbers in Table 1 and whether the gene fell into a region of LOH was based on both the genotyping array data as well as normalized exome capture read depth.

Table 1. Principal findings: Protein coding genes with somatic mutations. The copy number and location within regions of LOH are noted. The SNV in the tumor *vs.* normal cells is indicated in the allele column and the percent of variant reads or mosaicism is given. Condel scores (D = deleterious, N = neutral) and protein changes are shown for each predicted isoform. Premature stop mutations are indicated and assumed to be damaging. Genes confirmed by the Ingenuity analysis are indicated along with the Ingenuity "assessment".

Chr	Position	Gene	Driver	CN	EXC	LOH	Ref	SNV	% Var	Condel Score	Protein change	Ingenuity Assesment	Gene Description	Comments
17	7,578,394	TP53	HCD	2.02	2.12	YES	C	T	0.97	D(0.97–1.0)	H179R, H86R, H47R	Pathogenic	Tumor protein p53	Damaging in all alternate translation products
13	48,941,657	RB1	HCD	1.12	0.72	YES	G	T	0.75	STOP GAIN	E323*	Pathogenic	Retinoblastoma 1	Premature termination codon is inferred damaging
14	38,061,334	FOXA1	HCD	2.11	2.25	NO	G	T	0.44	D(1.0)	R219S, R186S	Likely Pathogenic	Forkhead box A1	Damaging in two alternate translation products
10	70,508,917	CCAR1	HCD	2.05	2.13	NO	G	A	0.24	N(0.02), D(0.81), D(0.83), N(0.02)	R269H, R258H, R284H, R89H	No Assessment	Cell division cycle and apoptosis regulator 1	Probably damaging in two alternate translation products
12	12.870 Mb–12.875 Mb	CDKN1B	HCD	0.36	0.13	YES		None		Same as hg19 reference	DELETION	Not flagged	Cyclin-dependent kinase inhibitor 1B (p27/KIP1)	No mutations in coding regions

Table 2. Potentially implicated genes. Selection criteria included whether the gene function is cancer-related, a member of a gene family with established cancer drivers, or assessed as damaged by Condel or the Ingenuity Variant Caller.

Chr	Position	Gene	CN	EXC	LOH	Ref	SNV	% Var	Condel Score	Protein change	Ingenuity Assesment	Gene Description	Comments
2	106,498,240	NCK2	1.89	1.70	YES	C	G	1.00	D(0.91)	P228R	Uncertain	NCK adaptor protein 2	Promotes melanoma cell proliferation, migration and invasion
2	107,041,278	RGPD3	1.89	1.70	YES	C	A	0.90	SIFT = Damaging	E1049*	Likely Pathogenic	RANBP2-like and GRIP domain containing 3	Reported expression in testis and HeLa cells
1	109,742,795–109,742,798	KIAA1324, EIG21	1.57	1.24	YES	G	4 bp del	0.85	Frameshift	G829fs*10	Uncertain	KIAA1324; Estrogen induced gene 121	High expression is associated with shorter survival in ovarian cancer
2	102,407,183	MAP4K4	1.89	1.70	YES	G	T	0.49	D(0.88) or N(0.02)	G42V, G4V	Likely Pathogenic	Mitogen-activated protein kinase kinase kinase kinase 4	Often overexpressed in cancer and has roles in various cancer processes
19	50,247,621	TSKS	2.00	2.14	NO	C	T	0.45	N(0.05)	E410K	Likely Pathogenic	Testis-specific kinase substrate	Low expression in some embryonal carcinoma lines
15	88,678,358	NTRK3	2.02	2.14	NO	C	T	0.41	D(0.72–0.87)	G295D, G393D	Likely Pathogenic	Neurotrophic tyrosine kinase, receptor, type 3	Potential tumor suppressor, often fused with ETV6 in thyroid cancer
16	7,759,062	RBFOX1	2.10	2.14	NO	G	A	0.40	D(0.91–1.0)	G307R, G334R, G355R, G339R, G377R	Uncertain	RNA binding protein, fox-1 homolog (*C. elegans*) 1	Related gene RBFOX2 is a Candidate Driver (CD)
19	38,865,389	PSMD8	2.00	2.10	NO	C	T	0.38	N(0.00)	R50C	Uncertain	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 8	Upregulated in a choriocarcinoma cell line

Table 2. *Cont.*

Chr	Position	Gene	CN	EXC	LOH	Ref	SNV	% Var	Condel Score	Protein change	Ingenuity Assesment	Gene Description	Comments
12	31,254,871	DDX11	2.25	2.61	NO	C	G	0.34	N(0.37)	H693Q, H317Q	Likely Pathogenic	DEAD/H box helicase 11	Associated with small-cell carcinoma
1	36,290,920	AGO4	1.83	1.71	NO	G	A	0.24	N(0.00)	M105V	Uncertain	Argonaute RISC catalytic component 4	Down-regulated in hepatocellular cancer
12	31,250,875	DDX11	2.25	2.61	NO	G	C	0.14	D(0.49)	A607P	Pathogenic	DEAD/H box helicase 11	Expressed at high levels in melanoma
19	4,689,651	DPP9	2.00	2.09	NO	G	T	0.12	D(0.50)	S560R	Uncertain	Dipeptidyl-peptidase 9	Expressed in breast and ovarian cancers
10	81,921,760	ANXA11	2.05	2.13	NO	G	A	0.04	D(0.85–0.95)	R338C, R371C, R4C	Not flagged	Annexin A11	May enhance metastasis and invasion; related gene ANXA6 is a Candidate Driver (CD)
9	100,843,284	TRIM14	2.23	2.59	NO	C	T	0.03	D(0.82–0.94)	R264W	Not flagged	Tripartite motif containing 14	Related gene TRIM7 is a High Confidence Driver (HCD)

The most obvious findings from the genotyping and sequencing data are that the classic tumor suppressors TP53 and RB1 both occur in blocks of LOH and the retained alleles were mutated. In the case of TP53 the His to Arg missense substitution is damaging by both SIFT and Condel. The RB1 mutation was a premature stop codon at amino acid 323. As with other treatment resistant cancers, mutations in TP53 and RB1 have been reported in lethal prostate cancers [6,7]. Mutations in MLL2 have been reported in about 9% of CRPC while mutations in FOXA1 occurred in about 3.4% of tumors [7]. In this case of SCPC we did not find mutations in MLL2 but did detect a mutation in FOXA1. FOXA1 is a nuclear protein that promotes tumor progression through its interaction with the androgen receptor, which in turn, induces several prostate-specific genes. FOXA1 levels are positively correlated with PSA, Gleason scores and AR expression [8]. The damaging FOXA1 mutation reported here occurred in about half of the sequence reads and is presumed to result in lower activity which may, in part, explain the fact that the patient's PSA levels were not elevated. Grasso *et al.* [7] also showed that FOXA1 mutations repressed androgen signaling and increased tumor growth.

Barbieri *et al.* [35] examined 112 prostate cancer tumor-normal pairs by exome sequencing and found recurrent somatic mutations in the genes FOXA1 and MED12 (~5% of tumors each) and in SPOP (~13% of tumors) in individuals with metastatic disease. The authors observed three different FOXA1 missense mutations in the forkhead (FH) domain, the DNA-binding domain of the protein [35,36]. FOXA1 binds to the androgen receptor and regulates the transcription of prostatic genes and is required for development of the prostate. The damaging mutation reported here also occurred in the FH domain. We did not observe mutations in SPOP and it does not occur in the region of LOH we observed on chromosome 17. Likewise, MED12 on the X chromosome did not appear to have somatic mutations when compared to the normal DNA sample.

Recurrent deletions of 5q21 have been reported in prostate cancer [35,37], and correlated with loss of the tumor suppressor CHD1. Further, Burkhardt *et al.* [37] showed a strong correlation between the loss of CHD1 and the biochemical failure to detect prostate-specific antigen. Similarly, we observed a region of LOH on chromosome 5 (~60–145 Mb, 5q12.1–31.3) which includes the CHD1 gene (Chr 5: 98,188,908–98,264,238) and, as noted, PSA levels were also not elevated in this cancer. While the remaining allele of CHD1 appears to have a normal sequence the genotyping array shows a reduced LRR. We were unable to perform studies to determine if RNA or protein levels were concomitantly reduced. The LOH region in our patient also includes the PIK3R1 gene, mutations in which are associated with various tumors (e.g., [9,38]). We observed no somatic mutations in the retained PIK3R1 allele.

The list of potentially implicated genes with somatic mutations is shown in Table 2 and ordered by the percent of mosaicism of the variant. These were selected based on the Ingenuity assessment, being a member of a gene family in which a related gene is a known or candidate cancer driver or from published literature implicating them in some aspect of cancer. Among these are NCK2 whose potentially damaging mutation occurs in 100% of reads. NCK2 is reported to promote melanoma cell proliferation, migration and invasion [39]. RGPD3 is expressed in the testis and HeLa cells [40], KIAA1324 or estrogen-expressed gene 21 is associated with ovarian cancer survival [41] and overexpression of EIG121 was observed to cause "profound suppression" of cell

growth [42]. Presumably, reduced expression would have the opposite effect. MAP4K4 is a serine/threonine kinase that is overexpressed in many cancers where it is implicated in migration and invasion [43] and RBFOX1 is related to the candidate driver RBFOX2 [44]. Zhou *et al.* [45] reported a mutation in RBFOX1 in a colorectal adenoma. Decreased expression of testis-specific kinase substrate, TSKS has been observed in cancerous testicular tissue and in very low levels in various embryonal carcinomas [46]. NTRK3 is a potential tumor suppressor [47] often fused with ETV6 in thyroid cancer [48]. PSMD8 is up-regulated in a choriocarcinoma cell line [49]. AGO4 or EIF2C4 is down-regulated in hepatocellular cancer [50]. DDX11 is required for sister chromatid cohesion and is expressed at high levels in primary and metastatic melanomas [51] and DDP9 is expressed in breast and ovarian cancer [52]. ANXA11 plays an important role in cell division and disruption of the gene "may lead to or enhance the metastasis, invasion and drug resistance of cancers" [53]. A literature survey for TRIM14 did not identify a link to cancer but the protein shares homology to the reported high confidence driver TRIM7 [34].

We noted LOH and slightly reduced copy number for the region on chromosome 2 (181.5–181.8 Mb) containing the long non-coding RNAs SChLAP1 (LINC00913) [54] and for PCGEM1 (193.6 Mb) [55] both of which have been reported to be overexpressed in aggressive prostate cancer. We did not find evidence that CDKN2A was deleted or that CCNE1, E2F3, UBE2C, or MYCC were amplified as seen in other cancers (e.g., [56]). Our patient did not receive castration therapy and the androgen receptor gene (AR) was not deleted. We did not find evidence for a fusion between TRPSSC2 and ERG based on exon sequences although, as noted above, we cannot rule out a translocation outside of coding regions. Cyclin D1 (CCND1), a gene often altered in cancer and a modifier of androgen receptor function [57], may have reduced copy numbers based on the array data but had no obvious sequence differences from the normal sample. We also found no evidence for copy number or somatic mutations in AURKA, KLK3, CGA, SYP, NXK3.1, NCAM1, CD56, ETS or UBE2C. In fact, the total estimated mutational burden is low (<1/50,000 bp).

The patient's normal genome was also studied for inherited SNPs that might confer an increased risk for cancer (Table S2). A heterozygous SNP in FOXC1 that creates a P321Q variant that is predicted to be deleterious by SIFT and Condel (score = 0.92) was observed. Overexpression of FOXC1 has been correlated with poor outcome (e.g., [58,59]) and as a promoter of invasion in breast cancer [60]. The patient was also heterozygous for a known rare variant in DND1 (rs72800920) that is predicted to be damaging by both SIFT and CONDEL (score = 0.96). In mice, a premature stop mutation in Dnd1 has been shown to markedly increase the risk of testicular germ line tumors [61]. We do not know if rs72800920 is a cancer risk factor or simply a private rare variant. Other possible risk factors for which the patient was heterozygous were ALK, NCK2, DDX11 and CBWD3. Somatic mutations in ALK (anaplastic lymphoma receptor tyrosine kinase) have been seen in neuoblastomas [62]. NCK adaptor protein 2, NCK2, has been reported to promote melanoma cell proliferation [39]. DEAD/DEAH box helicase 11, DDX11, is required for sister chromatid cohesion and has been reported to be essential for the survival of advanced melanomas [51]. It is curious that the patient was a carrier for a likely pathogenic inherited variant and his tumor showed two somatic mutations in DDX11. Each of the germline alleles in these

genes remained heterozygous in the tumor (*i.e.*, did not show evidence for selection) and we have no formal evidence that they conferred risk for disease or its progression.

4. Conclusions

The main goal of this study was to assess how the new genomic technologies, analysis methods and interpretation tools might be used to provide clinical utility. Secondarily, we hoped to better characterize a SCPC metastasis in a single case using these approaches. The combination of genotyping arrays, to provide a broad overview of the genomic landscape, and exome sequencing, to identify specific mutations, was more useful than either method alone. The genotyping array highlighted key regions of the genome that showed abnormal copy number or loss of heterozygosity. In general, these changes in genomic architecture are clues to underlying genes that may be implicated in cancer. LOH is commonly associated with the loss of tumor suppressors and by identifying those regions first on an array we were able to focus attention on the somatic sequence variants found there.

Because of limited sample we were unable to perform karyotyping or expression studies (either arrays or RNAseq). However, such limitations are likely to be expected in routine clinical testing so maximizing information from samples is critical. Going forward it would be preferable to do dual RNA and DNA isolations from fresh needle biopsies and perform RNA sequencing to measure relative expression levels, identify the main splice variants and any fusion transcripts. Given the good correlation between exome read depth and copy number from the array shown here it may be unnecessary to perform high-density genotyping in the future. However, we would likely replace exome-capture with PCR-free whole genome sequencing in order to eliminate biases related to the capture reagents and be able to potentially identify chromosomal translocation events and other genomic rearrangements. Currently, we feel it is important to manually review the sequence data at key positions and perform Sanger sequencing as confirmation for actionable mutations. However, improvements in laboratory methods and analysis may soon make this unnecessary (e.g., [63]).

We found that both open source and commercial tools were invaluable for interpreting somatic variants although it is essential that the analysis pipeline that produces the variants for interpretation be as rigorous as possible. Interpretation software essentially performs two tasks: it matches lists of variants within a study to those reported in various databases or in the literature in a way that is meaningful for the disease or mode of inheritance and it uses one or more algorithms to predict the effects of mutations. The first function will only be as good as the databases referenced and for commercial databases the details are usually not available. In general, there was excellent concordance between the open source tools and the Ingenuity Variant Analysis although the latter identified several somatic mutations in genes that were not flagged by the IntOGen analysis. This is not surprising given that IntOGen is based on data from large cancer sequencing projects while Ingenuity also includes literature-based gene information and predictive algorithms that infer change in protein function.

A current limitation to variant interpretation is, as seen in Table 1, that many genes produce multiple alternative transcripts and may have deleterious mutations predicted in some isoforms but

not others. In the absence of RNASeq data we do not know which isoforms may predominate in a given cancer type. Information about the splice variants and fusion transcripts will have to be included in a comprehensive analysis. Further, because the interpretation tools used in this analysis were based on VCFs they did not take copy number or LOH into account. As shown from the Excavator analysis this is something that could certainly be added. A clear advantage of the Ingenuity Variant Analysis tool was the ability of a user to easily link to the biomedical literature and pathway information that included potential drugs for targets it identified. Providing such information will be a valuable adjunct to physicians acting on exome and genome test results.

As noted above, as with other types of genetic testing, NGS approaches need to be standardized, accurate and have practical utility. Perhaps more than other genetic tests, whole genome or whole exome sequencing blurs the borders between clinical testing and research. This was also true of other methods when they first appeared (e.g., FISH, comparative hybridization arrays, *etc.*) and only with the accumulation of large datasets and more standardized methods in the laboratory and during analysis will the utility of genome sequencing become routine. As more correlations are made between patterns of the genomic landscape and mutational profiles we should be better able to tailor treatments or predict the course of disease. Already, sequencing data are being used to design patient-specific tests to follow response to treatment [64] and gene or mutation-specific treatments have been and are being developed.

SCPC is a lethal cancer with a poor prognosis. Ciriello *et al.* [56], in summarizing the ICGC and TCGA oncogenic signatures from over 3000 tumors, concluded that cancers generally fall into one of two classifications; “M” class cancers with, often large numbers of somatic mutations, and the “C” class with chromosomal abnormalities and fewer variants but which often involve somatic mutations in TP53, the likely cause of the genomic instability [65]. This tumor clearly falls into the C class. It is remarkable that while the overall somatic mutation rate was relatively low given that so many cancer driver genes were mutated. Perhaps the rarity of SCPC reflects the need to accumulate many separate deleterious mutations. Unfortunately, nothing in our sequencing or interpretation analysis offered a useful treatment strategy but, hopefully, the approaches and results reported here will be of use to others studying this and other aggressive cancers. By redefining cancers based on their genomic, expression and mutational architectures we may be able to markedly improve cancer diagnosis and therapy.

Acknowledgments

The authors thank Rachel Pattiglio and Laura Kasch of the Johns Hopkins Genetic Resources Core Facility for isolation of DNA from the biopsy and mouthwash samples and Nathan Pearson and Shela Shah from Ingenuity for assistance with the Ingenuity Variant Caller and recommended filtering criteria.

This work is dedicated to the memory of Gregory S. Liptak, physician, academic and friend.

Author Contributions

Designed study: Alan F. Scott and Gregory S. Liptak; Provided clinical information: Gregory S. Liptak; Wrote manuscript: Alan F. Scott, Robert B. Scharpf, David W. Mohr, Hua Ling, and Peng Zhang; Performed laboratory work: David W. Mohr; Implemented and ran software: David W. Mohr, Robert B. Scharpf, Hua Ling, and Peng Zhang.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Aparicio, A.; Logothetis, C.J.; Maity, S.N. Understanding the lethal variant of prostate cancer: Power of examining extremes. *Cancer Discov.* **2011**, *1*, 466–468.
2. Spiess, P.E.; Pettaway, C.A.; Vakar-Lopez, F.; Kassouf, W.; Wang, X.; Busby, J.E.; Do, K.A.; Davuluri, R.; Tannir, N.M. Treatment outcomes of small cell carcinoma of the prostate: A single-center study. *Cancer* **2007**, *110*, 1729–1737.
3. Wang, W.; Epstein, J.I. Small cell carcinoma of the prostate. A morphologic and immunohistochemical study of 95 cases. *Am. J. Surg. Pathol.* **2008**, *32*, 65–71.
4. Beltran, H.; Rickman, D.S.; Park, K.; Chae, S.S.; Sboner, A.; MacDonald, T.Y.; Wang, Y.; Sheikh, K.L.; Terry, S.; Tagawa, S.T.; *et al.* Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets. *Cancer Discov.* **2011**, *1*, 487–495.
5. Meulenbeld, H.J.; Bleuse, J.P.; Vinci, E.M.; Raymond, E.; Vitali, G.; Santoro, A.; Dogliotti, L.; Berardi, R.; Cappuzzo, F.; Tagawa, S.T.; *et al.* Randomized phase II study of danusertib in patients with metastatic castration-resistant prostate cancer after docetaxel failure. *BJU Int.* **2013**, *111*, 44–52.
6. Tzelepi, V.; Zhang, J.; Lu, J.F.; Kleb, B.; Wu, G.; Wan, X.; Hoang, A.; Efstathiou, E.; Sircar, K.; Navone, N.M.; *et al.* Modeling a lethal prostate cancer variant with small-cell carcinoma features. *Clin. Cancer Res.* **2012**, *18*, 666–677.
7. Grasso, C.S.; Wu, Y.M.; Robinson, D.R.; Cao, X.; Dhanasekaran, S.M.; Khan, A.P.; Quist, M.J.; Jing, X.; Lonigro, R.J.; Brenner, J.C.; *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **2012**, *487*, 239–243.
8. Imamura, Y.; Sakamoto, S.; Endo, T.; Utsumi, T.; Fuse, M.; Suyama, T.; Kawamura, K.; Imamoto, T.; Yano, K.; Uzawa, K.; *et al.* FOXA1 promotes tumor progression in prostate cancer via the insulin-like growth factor binding protein 3 pathway. *PLoS One* **2012**, *7*, e42456.
9. Van Allen, E.M.; Foye, A.; Wagle, N.; Kim, W.; Carter, S.L.; McKenna, A.; Simko, J.P.; Garraway, L.A.; Febbo, P.G. Successful whole-exome sequencing from a prostate cancer bone metastasis biopsy. *Prostate Cancer Prostatic Dis.* **2014**, *17*, 23–27.
10. Genetic Resources Core Facility. Available online: http://grcf.jhmi.edu/ (accessed on 2 January 2014).
11. Olshen, A.B.; Venkatraman, E.S.; Lucito, R.; Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **2004**, *5*, 557–572.

12. Magi, A.; Tattini, L.; Cifola, I.; D'Aurizio, R.; Benelli, M.; Mangano, E.; Battaglia, C.; Bonora, E.; Kurg, A.; Seri, M.; *et al.* EXCAVATOR: Detecting copy number variants from whole-exome sequencing data. *Genome Biol.* **2013**, doi:10.1186/gb-2013-14-10-r120.
13. Barnhart, M.G.S.; Hetrick, K.; Goldstein, J.; Marosy, D.; Mohr, D.; Craig, B.; Watkins, L., Jr.; Doheny, K. CIDRSeqSuite 2.0: An automated analysis pipeline for next-generation sequencing. In Proceedings of the presented at the 61st annual meeting of the American society for human genetics, Montreal, QC, Canada, 12 October 2011.
14. Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **2010**, *26*, 589–595.
15. Picard. Available online: http://picard.sourceforge.net/ (accessed on 2 January 2014).
16. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303.
17. DePristo, M.A.; Banks, E.; Poplin, R.; Garimella, K.V.; Maguire, J.R.; Hartl, C.; Philippakis, A.A.; del Angel, G.; Rivas, M.A.; Hanna, M.; *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **2011**, *43*, 491–498.
18. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; *et al.* From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **2013**, *11*, 1–33.
19. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164.
20. Scripps Genome Adviser. Available online: http://genomics.scripps.edu/ADVISER/ (accessed on 2 January 2014).
21. Saunders, C.T.; Wong, W.S.; Swamy, S.; Becq, J.; Murray, L.J.; Cheetham, R.K. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **2012**, *28*, 1811–1817.
22. Christoforides, A.; Carpten, J.D.; Weiss, G.J.; Demeure, M.J.; von Hoff, D.D.; Craig, D.W. Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC Genomics* **2013**, doi:10.1186/1471-2164-14-302.
23. IntOgen: Interactive Onco Genomics Mutations Server. Available online: http://www.intogen.org/mutations/ (accessed on 4 January 2014).
24. Gonzalez-Perez, A.; Perez-Llamas, C.; Deu-Pons, J.; Tamborero, D.; Schroeder, M.P.; Jene-Sanz, A.; Santos, A.; Lopez-Bigas, N. IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **2013**, *10*, 1081–1082.
25. Condel: CONsensus DELeteriousness Score of Missense SNVs Server. Available online: http://bg.upf.edu/condel/analysis/ (accessed on 4 January 2014).
26. Gonzalez-Perez, A.; Lopez-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.* **2011**, *88*, 440–449.

27. Petty, E.; Pillus, L. Balancing chromatin remodeling and histone modifications in transcription. *Trends Genet.* **2013**, *29*, 621–629.
28. Cuesta, R.; Martinez-Sanchez, A.; Gebauer, F. miR-181a regulates cap-dependent translation of p27(kip1) mRNA in myeloid cells. *Mol. Cell. Biol.* **2009**, *29*, 2841–2851.
29. Fero, M.L.; Randel, E.; Gurley, K.E.; Roberts, J.M.; Kemp, C.J. The murine gene p27Kip1 is haplo-insufficient for tumor suppression. *Nature* **1998**, *396*, 177–180.
30. Chu, I.M.; Hengst, L.; Slingerland, J.M. The Cdk inhibitor p27 in human cancer: Prognostic potential and relevance to anticancer therapy. *Nat. Rev. Cancer* **2008**, *8*, 253–267.
31. Hudson, T.J.; Anderson, W.; Artez, A.; Barker, A.D.; Bell, C.; Bernabe, R.R.; Bhan, M.K.; Calvo, F.; Eerola, I.; Gerhard, D.S.; *et al.* International network of cancer genome projects. *Nature* **2010**, *464*, 993–998.
32. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120.
33. Robinson, J.T.; Thorvaldsdottir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative genomics viewer. *Nat. Biotechnol.* **2011**, *29*, 24–26.
34. Tamborero, D.; Gonzalez-Perez, A.; Perez-Llamas, C.; Deu-Pons, J.; Kandoth, C.; Reimand, J.; Lawrence, M.S.; Getz, G.; Bader, G.D.; Ding, L.; *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **2013**, doi:10.1038/srep02650.
35. Barbieri, C.E.; Baca, S.C.; Lawrence, M.S.; Demichelis, F.; Blattner, M.; Theurillat, J.P.; White, T.A.; Stojanov, P.; van Allen, E.; Stransky, N.; *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **2012**, *44*, 685–689.
36. Clark, K.L.; Halay, E.D.; Lai, E.; Burley, S.K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **1993**, *364*, 412–420.
37. Burkhardt, L.; Fuchs, S.; Krohn, A.; Masser, S.; Mader, M.; Kluth, M.; Bachmann, F.; Huland, H.; Steuber, T.; Graefen, M.; *et al.* CHD1 is a 5q21 tumor suppressor required for ERG rearrangement in prostate cancer. *Cancer Res.* **2013**, *73*, 2795–2805.
38. Quayle, S.N.; Lee, J.Y.; Cheung, L.W.; Ding, L.; Wiedemeyer, R.; Dewan, R.W.; Huang-Hobbs, E.; Zhuang, L.; Wilson, R.K.; Ligon, K.L.; *et al.* Somatic mutations of PIK3R1 promote gliomagenesis. *PLoS One* **2012**, *7*, e49466.
39. Labelle-Cote, M.; Dusseault, J.; Ismail, S.; Picard-Cloutier, A.; Siegel, P.M.; Larose, L. Nck2 promotes human melanoma cell proliferation, migration and invasion *in vitro* and primary melanoma-derived tumor growth *in vivo*. *BMC Cancer* **2011**, doi:10.1186/1471-2407-11-443.
40. Ciccarelli, F.D.; von Mering, C.; Suyama, M.; Harrington, E.D.; Izaurralde, E.; Bork, P. Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* **2005**, *15*, 343–351.
41. Schlumbrecht, M.P.; Xie, S.S.; Shipley, G.L.; Urbauer, D.L.; Broaddus, R.R. Molecular clustering based on ERalpha and EIG121 predicts survival in high-grade serous carcinoma of the ovary/peritoneum. *Mod. Pathol.* **2011**, *24*, 453–462.

42. Deng, L.; Feng, J.; Broaddus, R.R. The novel estrogen-induced gene EIG121 regulates autophagy and promotes cell survival under stress. *Cell Death Dis.* **2010**, doi:10.1038/cddis.2010.9.
43. Qiu, M.H.; Qian, Y.M.; Zhao, X.L.; Wang, S.M.; Feng, X.J.; Chen, X.F.; Zhang, S.H. Expression and prognostic significance of MAP4K4 in lung adenocarcinoma. *Pathol. Res. Pract.* **2012**, *208*, 541–548.
44. Venables, J.P.; Brosseau, J.P.; Gadea, G.; Klinck, R.; Prinos, P.; Beaulieu, J.F.; Lapointe, E.; Durand, M.; Thibault, P.; Tremblay, K.; *et al.* RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. *Mol. Cell. Biol.* **2013**, *33*, 396–405.
45. Zhou, D.; Yang, L.; Zheng, L.; Ge, W.; Li, D.; Zhang, Y.; Hu, X.; Gao, Z.; Xu, J.; Huang, Y.; *et al.* Exome capture sequencing of adenoma reveals genetic alterations in multiple cellular pathways at the early stage of colorectal tumorigenesis. *PLoS One* **2013**, *8*, e53310.
46. Scorilas, A.; Yousef, G.M.; Jung, K.; Rajpert-De Meyts, E.; Carsten, S.; Diamandis, E.P. Identification and characterization of a novel human testis-specific kinase substrate gene which is downregulated in testicular tumors. *Biochem. Biophys. Res. Commun.* **2001**, *285*, 400–408.
47. Luo, Y.; Kaz, A.M.; Kanngurn, S.; Welsch, P.; Morris, S.M.; Wang, J.; Lutterbaugh, J.D.; Markowitz, S.D.; Grady, W.M. NTRK3 is a potential tumor suppressor gene commonly inactivated by epigenetic mechanisms in colorectal cancer. *PLoS Genet.* **2013**, *9*, e1003552.
48. Kralik, J.M.; Kranewitter, W.; Boesmueller, H.; Marschon, R.; Tschurtschenthaler, G.; Rumpold, H.; Wiesinger, K.; Erdel, M.; Petzer, A.L.; Webersinke, G. Characterization of a newly identified ETV6-NTRK3 fusion transcript in acute myeloid leukemia. *Diagn. Pathol.* **2011**, doi:10.1186/1746-1596-6-19.
49. Kobayashi, Y.; Banno, K.; Shimizu, T.; Ueki, A.; Tsuji, K.; Masuda, K.; Kisu, I.; Nomura, H.; Tominaga, E.; Nagano, O.; *et al.* Gene expression profile of a newly established choriocarcinoma cell line, iC3-1, compared to existing choriocarcinoma cell lines and normal placenta. *Placenta* **2013**, *34*, 110–118.
50. Kitagawa, N.; Ojima, H.; Shirakihara, T.; Shimizu, H.; Kokubu, A.; Urushidate, T.; Totoki, Y.; Kosuge, T.; Miyagawa, S.; Shibata, T. Downregulation of the microRNA biogenesis components and its association with poor prognosis in hepatocellular carcinoma. *Cancer Sci.* **2013**, *104*, 543–551.
51. Bhattacharya, C.; Wang, X.; Becker, D. The DEAD/DEAH box helicase, DDX11, is essential for the survival of advanced melanomas. *Mol. Cancer* **2012**, doi:10.1186/1476-4598-11-82.
52. Wilson, C.H.; Abbott, C.A. Expression profiling of dipeptidyl peptidase 8 and 9 in breast and ovarian carcinoma cell lines. *Int. J. Oncol.* **2012**, *41*, 919–932.
53. Wang, J.; Guo, C.; Liu, S.; Qi, H.; Yin, Y.; Liang, R.; Sun, M.Z.; Greenaway, F.T. Annexin A11 in disease. *Clin. Chim. Acta* **2014**, *431*, 164–168.
54. Prensner, J.R.; Iyer, M.K.; Sahu, A.; Asangani, I.A.; Cao, Q.; Patel, L.; Vergara, I.A.; Davicioni, E.; Erho, N.; Ghadessi, M.; *et al.* The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat. Genet.* **2013**, *45*, 1392–1398.

55. Yang, L.; Lin, C.; Jin, C.; Yang, J.C.; Tanasa, B.; Li, W.; Merkurjev, D.; Ohgi, K.A.; Meng, D.; Zhang, J.; *et al.* lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. *Nature* **2013**, *500*, 598–602.
56. Ciriello, G.; Miller, M.L.; Aksoy, B.A.; Senbabaoglu, Y.; Schultz, N.; Sander, C. Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **2013**, *45*, 1127–1133.
57. Comstock, C.E.; Augello, M.A.; Schiewer, M.J.; Karch, J.; Burd, C.J.; Ertel, A.; Knudsen, E.S.; Jessen, W.J.; Aronow, B.J.; Knudsen, K.E. Cyclin D1 is a selective modifier of androgen-dependent signaling and androgen receptor function. *J. Biol. Chem.* **2011**, *286*, 8117–8127.
58. Xu, Y.; Shao, Q.S.; Yao, H.B.; Jin, Y.; Ma, Y.Y.; Jia, L.H. Up-expression of FOXC1 correlates with poor prognosis in gastric cancer patients. *Histopathology* **2013**, doi:10.1111/his.12347.
59. Wei, L.X.; Zhou, R.S.; Xu, H.F.; Wang, J.Y.; Yuan, M.H. High expression of FOXC1 is associated with poor clinical outcome in non-small cell lung cancer patients. *Tumour Biol.* **2013**, *34*, 941–946.
60. Sizemore, S.T.; Keri, R.A. The forkhead box transcription factor FOXC1 promotes breast cancer invasion by inducing matrix metalloprotease 7 (MMP7) expression. *J. Biol. Chem.* **2012**, *287*, 24631–24640.
61. Youngren, K.K.; Coveney, D.; Peng, X.; Bhattacharya, C.; Schmidt, L.S.; Nickerson, M.L.; Lamb, B.T.; Deng, J.M.; Behringer, R.R.; Capel, B.; *et al.* The Ter mutation in the dead end gene causes germ cell loss and testicular germ cell tumors. *Nature* **2005**, *435*, 360–364.
62. Chen, Y.; Takita, J.; Choi, Y.L.; Kato, M.; Ohira, M.; Sanada, M.; Wang, L.; Soda, M.; Kikuchi, A.; Igarashi, T.; *et al.* Oncogenic mutations of ALK kinase in neuroblastoma. *Nature* **2008**, *455*, 971–974.
63. Schmitt, M.W.; Kennedy, S.R.; Salk, J.J.; Fox, E.J.; Hiatt, J.B.; Loeb, L.A. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 14508–14513.
64. Leary, R.J.; Sausen, M.; Kinde, I.; Papadopoulos, N.; Carpten, J.D.; Craig, D.; O'Shaughnessy, J.; Kinzler, K.W.; Parmigiani, G.; Vogelstein, B.; *et al.* Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl. Med.* **2012**, doi:10.1126/scitranslmed.3004742.
65. Rausch, T.; Jones, D.T.; Zapatka, M.; Stutz, A.M.; Zichner, T.; Weischenfeldt, J.; Jager, N.; Remke, M.; Shih, D.; Northcott, P.A.; *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **2012**, *148*, 59–71.

The Little Fly that Could: Wizardry and Artistry of *Drosophila* Genomics

Radoslaw K. Ejsmont and Bassem A. Hassan

Abstract: For more than 100 years now, the fruit fly *Drosophila melanogaster* has been at the forefront of our endeavors to unlock the secrets of the genome. From the pioneering studies of chromosomes and heredity by Morgan and his colleagues, to the generation of fly models for human disease, *Drosophila* research has been at the forefront of genetics and genomics. We present a broad overview of some of the most powerful genomics tools that keep *Drosophila* research at the cutting edge of modern biomedical research.

Reprinted from *Genes*. Cite as: Ejsmont, R.K.; Hassan, B.A. The Little Fly that Could: Wizardry and Artistry of *Drosophila* Genomics. *Genes* **2014**, *5*, 385-414.

1. Introduction

The Human Genome Project, on its way to producing an assembled genome of *Homo sapiens*, has gone through several test runs yielding sequenced genomes of other organisms of high relevance for research into human development and disease. The first published genome of a free-living organism was that of the proteobaterium *Haemophilus influenzae* [1], followed by sequencing of the genome of *Saccharomyces cerevisiae* yeast, the first eukaryotic genome sequenced [2], and the genome of *Caenorhabditis elegans*, the first genome of a multicellular organism and the first animal genome [3]. The second animal genome sequenced was that of the fruit fly *Drosophila melanogaster* [4]. In this review, we discuss the significance of the sequencing of the *Drosophila* genome as well as the technical advances and new research avenues that have accompanied it.

2. Drosophila as a Model

2.1. In Development

The fruit fly has been studied for over a century and the lessons learned from fly research makes it almost impossible to enumerate but a few of the most notable cases. The pioneering studies that identified genes involved in *Drosophila* embryo segmentation [5,6] and establishment of segment polarity [6] were seminal for understanding conserved developmental strategies in the animal kingdom. The discovery of homeotic genes is one of the best-known examples of genes discovered in the fruit fly, and these were found to be conserved and play analogous roles in humans [7–9]. *Drosophila* has played a seminal role in sensory organ development research. The discovery of the *eyeless* gene [10], a fly homolog of human and mouse PAX6 [11,12], and determination of its targets [13] shed light on vertebrate eye development and led to discovery of novel disease related genes in humans [14]. The proneural gene atonal plays a crucial role in the development of *Drosophila* photoreceptor neurons [15] and chordotonal organs [16]. Its function is conserved in

mammals, where its homologs Math5 and Math1 were shown to be involved in regulating formation of retinal ganglion cells [17] and inner ear mechanosensory hair cells [18].

2.2. In Signaling

Drosophila has been extensively used for studies of signaling pathways. In Hedgehog signaling, both the Hedgehog ligand itself [6,19,20] and its receptor Patched [6,21,22] were first identified in the fly, though the link between the two was first established in mammals [23,24]. The ligand of the Wnt signaling pathway turned out to be a well-known *Drosophila* segment polarity protein, *i.e.*, Wingless. The Wnt receptor, Frizzled [25], and several other signal transduction cascade members were identified in the fly as members of the Wnt pathway [26–28]. The planar cell polarity (PCP) pathway is yet another example of a signaling cascade in which key players and mechanisms of action have been, to a large extend, identified in *Drosophila* [29,30]. The Notch signaling pathway, associated with cell fate control, lateral inhibition, and signal integration during development, has been discovered and extensively studied in fruit flies [31–33]. Finally, major components and mechanisms of action of the Hippo signaling pathway have been described in *Drosophila* [34–36]. All these pathways play major roles in human development and disease.

2.3. In Disease

Over the past two decades the fruit fly became an increasingly popular model organism for the study of human disease, with focus on neurodegenerative [37] and neuromuscular [38] diseases as well as cancer [39]. Neurological diseases that have been modeled in *Drosophila* include trinucleotide repeat disorders [40–42], Alzheimer's disease [43–46], Parkinson's disease [47,48], amyotrophic lateral sclerosis [49,50], and dystrophy [51]. Other examples that include use of the fruit fly model are studies of alcohol abuse [52,53], cocaine addiction [54], obesity [55] and diabetes [56], cardiac diseases [57], and asthma [58]. *Drosophila* has been demonstrated to be a great model to identify tumor suppressor genes [59] or genes involved in metastasis [60]. Thanks to the conservation of major signaling pathways, tumor suppressors and oncogenes, various fly cancer models have been established. Understanding how signal transduction pathways like Hippo, Notch, Dpp or JAK-STAT affect tumor formation was aided by research in fruit flies [61–63]. *Drosophila* has been used as a model for tumor invasion and metastasis [64], and as a platform to identify novel therapeutic targets [65].

3. Meet the Drosophila Genome

The *Drosophila* genome is estimated to be approximately 200 Mb, with one third of it forming pericentric heterochromatin [66]. It is organized on three autosomes (numbered 2, 3 and 4) and sex chromosomes, X (also referred to as the first chromosome) and Y. The initial assembly of the fruit fly genome was published in March 2000, after almost a year of whole genome shotgun sequencing. The first published assembly, referred to as Release 1 of the genome, included 13,991 genes encoding for 14,080 peptides. Over two thirds of annotated genes were assigned gene

ontology (GO) terms upon annotation. The initial assembly contained ~1300 gaps in mapped sequences [4] that were filled with subsequent releases.

The third release of the genome was the first that included pericentric heterochromatin sequences [67]. The mutations indicated in the sequenced strain's genotype, as well as several other identified mutations, have been corrected with wild-type sequence [68]. With that release, a comprehensive set of resources were published, including a library of full-length cDNAs for 40% of genes [69] and an atlas of gene expression patterns during embryogenesis [70]. Sequence analysis provided insights into transposable elements within the genome [71], core promoter structures [72], and largely improved annotation of gene models [68].

The current, fifth assembly of the genome has closed all but 9 gaps in the main assembly. The sequenced genome covers over 120 Mb of euchromatin, and over 9 Mb of mapped and over 10 Mb of unmapped heterochromatin. The current annotation revision contains 13,942 protein coding genes and over 2354 non-coding RNA genes, including ribosomal (rRNAs), transport (tRNAs), micro- (miRNAs), and small nuclear (snRNA) and small nucleolar (snoRNA) RNAs [73]. Through genome analysis, fruit flies have been found to contain complex gene structures. Approximately 7.5% of all genes, including non-coding RNAs, are located within the introns of other genes. Messenger RNAs for about 15% of genes overlap with mRNAs of genes on the opposite strands. Over 30 genes have been identified as dicistronic, *i.e.*, producing single mRNA encoding for two separate protein products through independent translation initiation events [68]. Over 30% of *Drosophila melanogaster* genes were found to be alternatively spliced [74], yielding a diverse set of almost 30,000 protein-coding transcripts [73]. The next release of the genome assembly (Release 6) is expected this year (2014).

Improved assembly and annotation of the fruit fly genome was possible not only due to new sequencing data, but also thanks to advances in bioinformatics tools. An integrated computational pipeline and a tailored database schema have been developed to facilitate genomic data storage and automated sequence annotation [75]. Computed annotations have been manually curated by experts and to aid in this task, a dedicated annotation editor was developed [76]. Finally, automated genome annotation in general requires the use of computational tools, some of which were first applied in the *Drosophila* genome project [72,77].

4. Genomes by the Dozen

Analysis of coding parts of the genome can be facilitated by comparison of genomic sequences with sequences of cDNAs originating from the same species. Most of the DNA in the majority of species, however, is non-coding. One approach to identify functional non-coding DNA segments, such as *cis*-regulatory elements, relies on finding conserved regions or motifs across related species. This naturally requires having more than one genome sequenced and was a driving force behind sequencing of the genomes of *Schizosaccharomyces pombe* [78] and *Caenorhabditis briggsae* [79] in the yeast and worm research communities, respectively. In the *Drosophila* genus, the comparative genomics era began with sequencing of the *Drosophila pseudoobscura* genome [80]. The two genomes were found to be very similar, despite 25–55 million years of evolutionary divergence. Synteny is preserved in blocks containing 10.7 genes on average, which

corresponds to ~83 kb. The vast majority of synteny breaks were caused by intrachromosomal rearrangements. On average, ~48% of the base pairs are conserved between these two species.

The next advance in *Drosophila* comparative genomics came with sequencing of ten further species, *Drosophila sechellia*, *simulans*, *yakuba*, *erecta*, *ananassae*, *persimilis*, *willistoni*, *mojavensis*, *virilis*, and *grimshawi*. These species span a broad spectrum of morphologies, ecologies, and behaviors, yet have identical body plans and very similar life cycles [81]. Furthermore, these species share approximately 70% of their genes. Genome sizes estimated by flow cytometry vary between 130 Mb in *D. mojavensis* to 364 Mb in *D. virilis* [66]. The synteny conservation between sequenced species varies with an average of 122 genes per block between *D. melanogaster* and *simulans* down to 8 genes per block between *D. melanogaster* and *grimshawi*. Overall genome size, number of genes, distribution of transposable element classes, and patterns of codon usage are all very similar across the 12 sequenced genomes. At a finer scale, however, the number of structural changes and rearrangements is larger, including rearrangements of genes within the Hox cluster or highly dynamic sizes and content of multigene families [81].

Together, the 12 *Drosophila* genomes provide a solid platform for annotation and analysis of both coding and non-coding DNA. This unprecedented dataset enabled the use of evolutionary signatures—specific patterns of change in DNA elements upon selection—for *de novo* prediction and correction of previously annotated protein-coding gene models [82], non-coding RNAs, and transcription factor (TF) binding sites [83]. Identification of TF binding motifs has traditionally been based on DNA alignments. Alignment-based methods can also be used for the identification of *cis*-regulatory modules (CRMs), which are comprised of a number of TF binding motifs [84]. In many cases, however, the number and order of individual motifs varies between species, especially when these are distant, while preserving regulatory outcome. To address such cases, alignment-free approaches have also been developed [85,86].

5. Genomes by Population

Drosophila provides an unmatched set of resources for studying quantitative traits [87]. In the post genomic era, genome-wide association studies (GWAS) have become a preferred method for analyzing complex traits. The GWAS methodology is now routinely and successfully applied in the identification of human disease-associated genes [88]. Two fruit fly resources, the *Drosophila* Genetic Reference Panel (DGRP) [89] and the *Drosophila* Synthetic Population Resource (DSPR) [90], offer large sets of sequenced and mapped fly lines tailored for GWAS and quantitative trait loci (QTL) mapping.

The DGRP is a collection of more than 200 fully sequenced recombinant inbred lines (RILs) that were established from mated females collected from a market in Raleigh, North Carolina, USA. The genomic sequences of these lines contain over 4.5 million single nucleotide polymorphisms (SNPs), over one hundred thousand polymorphic microsatellites, and over 36 thousand transposable elements [89]. The DGRP has been extensively characterized and in addition to detailed genomic sequence analysis includes microarray [91] and RNA-seq [92] datasets for selected lines. To date, numerous genome-wide association studies have been published on various

traits using DGRP, including oxidative stress [93], mitochondrial function [94], viral infection resistance [95], and sleep [96].

The *Drosophila* Synthetic Population Resource uses a different approach. Over 1,700 DSPR RILs were established from 15 isogenic founder lines created from geographically distinct *Drosophila* populations. The founder lines were split in two groups of eight (with one line in both groups) and mixed for 50 generations to create two synthetic populations, from which two sets of RILs were established. The founder lines were fully sequenced and each RIL was mapped using restriction-site associated DNA (RAD) markers onto the founders' sequence with 17 kb median resolution. The number of SNPs in the founder lines exceeds 1.6 million [90]. The DSPR is complementary to DGRP and both resources can be used together for cross-validation and to increase the mapping power [97].

6. Decoding the Genome's Secrets

6.1. The modENCODE Project

The information encoded by genomes goes far beyond a simple trinucleotide code used to translate nucleic acid sequence into protein. A plethora of information is hidden within introns, UTRs, non-coding RNAs, *cis*-regulatory elements, and chromatin marks. These elements are known to regulate where and when a gene product is expressed. The human ENCODE (ENCyclopedia Of DNA Elements) project [98] aims to identify and understand the information carried by the human genome. The modENCODE project is the model organism counterpart of ENCODE with focus on two species *C. elegans* [99] and *D. melanogaster* [100]. The fruit fly modENCODE data includes high-throughput transcriptome sequencing (RNA-seq), chromatin immunoprecipitation followed by sequencing (ChIP–seq) for transcription factor binding sites and histone modifications, DNA replication patterns, and nucleosome occupancy. The samples have been collected from 12–30 developmental time points of the sequenced *D. melanogaster* strain and from several cell lines [100].

6.2. The Transcriptome

Comprehensive transcriptomics data has redefined gene models for 75% of fly genes by adding new exons or splice variants. The majority of annotation changes were supported by direct cDNA evidence. Analysis of transcription start sites (TSSs) for over half of *Drosophila* genes resulted in identification of over 1500 novel promoters. The structural analysis of RNA-seq-identified transcripts that did not seem to encode proteins revealed that a majority of them has no thermodynamically stable secondary structure, suggesting structure-independent functions. Among structural non-coding RNAs, several hundred novel small regulatory RNAs (miRNAs, siRNAs, and piRNA) have been identified. Additionally, transcription start sites for both protein coding and non-coding RNAs have been derived from the presence of chromatin marks characteristic of transcriptionally active regions, such as H3K4me3 enrichment, H3K9ac, and presence of RNA polymerase II in TSS-proximal regions [100].

6.3. Chromatin Landscape

Eukaryotic genomes are organized into large domains that exhibit distinct chromatin properties [100]. Analysis of large-scale organization of the chromatin landscape has revealed unexpected complexity and plasticity among different cell types. Some regions in the usually silent pericentric heterochromatin exhibited surprisingly high gene expression activity. Conversely, large regions of normally transcriptionally active euchromatin harbored histone marks (H3K9me2) typical for heterochromatin [100,101]. Chromatin signatures characteristic of various functional elements have been identified by ChIP-chip for 18 histone modifications (both activating, such as H3K4me, H3K9/18/27ac, H2B ubiquitination and repressive, such as H3K9me2/3 or Polycomb associated H3K27me3) and variants (H1, H4) from several cell lines and developmental stages. Correlating chromatin signatures with transcriptome and protein binding data (replication factors, insulator-binding proteins, and transcription factors) helped identify marks specific for promoters, actively transcribed regions, introns, insulators, and origins of replication [100]. The presence of specific chromatin marks was found to correlate with the physical properties of chromatin, where transcriptionally active chromatin exhibited high solubility and high nucleosome-turnover rates [100]. Computational analysis of combinatorial patterns of histone modifications revealed distinct chromatin states associated with active TSSs, exons, introns, and other open chromatin as well as closed chromatin states [100,102,103].

6.4. Transcriptional Regulation

The modENCODE project has identified binding sites for almost 40 transcription factors through both ChIP-chip and ChIP-seq. The analysis has revealed that out of nearly 40,000 identified unique binding sites found, 5% are bound by 8 different transcription factors or more and are considered High Occupancy Target (HOT) regions. Furthermore, almost 40% of the sites can be bound by more than two factors [103]. The HOT regions exhibit decreased nucleosome density, increased nucleosome turnover and often colocalize with TSS and ORC (origin recognition complex) binding sites, suggesting interplay between chromatin regulation, TF binding, and DNA replication [100,103]. In total, modENCODE ChIP experiments revealed over 500 silencers, 2300 new promoters, over 14 candidate CBP-bound *cis*-regulatory elements, and over 7500 putative insulators [103]. Pairwise analysis of binding site co-occurrence has revealed over 800 known and putative transcription factor co-binding interactions. Binding sites for transcription factors regulating biologically opposing roles exhibited negative associations. The modENCODE TF binding data sets combined with external data were used to construct a network covering over 80 transcription factors and characterizing over 800, largely novel, regulatory interactions. Binding site co-occurrence among various analyzed promoters corresponded to temporal co-expression of the respective target genes, supporting the existence of combinatorial transcription factor codes [103].

7. The Fruit Fly Toolkit

7.1. Getting Constructs in

Drosophila is famous for its extensive range of forward and reverse genetics tools. The powerful toolkit was primed by the discovery of P transposable element-based germline transformation [104]. This revolutionary development allowed, for the first time, efficient delivery of foreign DNA into the genome. Development of the vast majority of *Drosophila* tools required at some stage use of P-element or other transposon systems. P-elements were used for gene cloning [105], genetic rescue [106], and as potent mutagens by their insertion [107] or excision [108]. P-element insertions have enabled the creation of enhancer traps, thus allowing visualization of gene expression patterns using genetically encoded reporters [109]. The *Drosophila* Gene Disruption Project used P-elements to create single transposon insertions in over 30% of fly genes [110]. The remaining genes, due to target sequence bias of P-elements, are currently targeted using other transposons [111]. The catalog of transposons that can be used for fly transformation has been expanded over the years and includes mariner [112,113], Minos [114,115], and piggyBac [116,117]. Each of these transposable elements, except for Minos that seems to insert randomly, has its hot and cold spots, but failure to target a certain region can often be addressed by using a different transposon [111].

While random, transposon-mediated transgenesis is desirable for gene disruption or genomic targeting, but integration of reporter or rescue constructs calls for more control over the locus where these integrate, thus reducing the chance of position effects that can strongly influence gene expression [118,119]. Early attempts to repeatedly target a specific locus in the fruit fly genome were based on transposon homing [120]. Short regulatory sequences from *Polycomb* target genes or from the *linotte* locus included in the transposon were shown to increase the likelihood of such transposon landing in the vicinity of genomic regions bearing these sequences [120,121]. The low resolution (30 bp) and efficiency (20% of insertions) of this homing technique prompted further developments in the field. The introduction of an irreversible, site-specific recombinase from the phiC31 phage ushered in a new era in fly transgenesis. The phiC31 integrase catalyzes unidirectional recombination between two attachment sites, attP and attB, leading to the formation of attL and attR sites [122]. A circular construct harboring an attB site can be efficiently and specifically integrated into an attP site located on the genome [123]. The phiC31 integrase system has, for the first time, enabled transformation of flies with BAC-sized constructs [124]. The integrase can be expressed from mRNA co-injected with the construct [123] or from the genome under a germ-line specific promoter, the latter method being more efficient [125]. Several dozens of attP landing lines have been created and tested [123–126], creating unlimited possibilities to combine transgenes.

7.2. Express What You Want, Where You Want, and When You Want

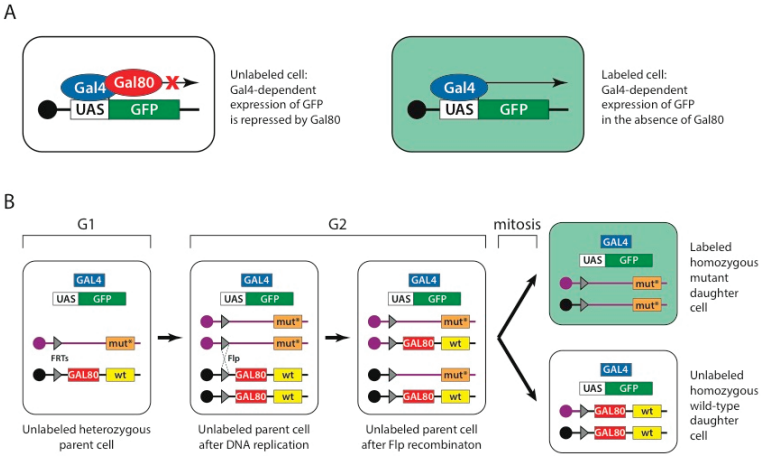
P-element transgenesis has enabled the creation of a plethora of other *Drosophila* tools, of which the Gal4/UAS system is the most notable example. The system is based on yeast

transcription factor gene *GAL4* fused to a minimal promoter. This construct is randomly inserted into the genome, for instance by means of P-element transgenesis, hitchhiking nearby enhancers and creating an enhancer trap. Alternatively, enhancer sequences can be cloned upstream of the minimal promoter and other arrangements, with enhancers cloned elsewhere, for example within introns, are also possible. The second component of the system is the Gal4 binding site, known as the upstream activation sequence (UAS), driving expression of the target gene. The combination of a *GAL4* enhancer trap with a UAS-driven target enables expression of the gene of interest in the desired tissues or cell types [127]. Several collections of *GAL4* enhancer traps have been created using P-element [128,129] and piggyBac [110,130] insertions. The enhancer trap resources have recently been supplemented by large collections of cloned enhancers driving expression of Gal4 [131,132]. The Gal4 expression pattern can be refined spatially [133] or temporarily [134] using the Gal80 repressor [135]. Further control over Gal4-driven expression can be obtained using variants requiring drugs for activation [136–139]. Today, the Gal4-UAS system is one of several binary expression systems available in *Drosophila*. Other examples include the LexA transactivator that binds LexOp sites [140] and the Q system with QF transactivator, QUAS binding sites and the QS repressor whose activity can additionally be drug controlled [141]. The existing binary systems can be combined to provide fine control over target expression pattern or for simultaneous targeting of different cellular populations [142].

7.3. Mutant Tissue on Demand

The Gal4-UAS system is an important component of yet another powerful fruit fly tool, the mosaic analysis with a repressible cell marker or MARCM. Induction of mosaicism in *Drosophila* is used either for studying an otherwise lethal phenotype within a tissue of interest [143] or for marking a clone of cells within a tissue of interest [144]. Mosaics can be created using flippase (Flp) mediated mitotic recombination between homologous chromosomes [145]. In this technique, homologous chromosomes carry an insertion (usually P-element mediated) of a flippase recognition target (FRT) site. One of the chromosomes carries a wild type and the other a mutant allele of the gene of interest. In the presence of flippase, recombination events between homologous chromosomes can occur during cell division, leading to the generation of homozygous mutant cells from heterozygous precursors. The MARCM technique (Figure 1) enhances Flp-mediated mitotic recombination by uniquely labeling mutant cells using a genetically encoded marker. The mutant clone is marked with a UAS-GFP (green fluorescent protein) construct, driven by ubiquitously expressed *GAL4*. These two transgenes are usually inserted together on any chromosome, except the wild type chromosome that carries the Gal4 repressor—*GAL80* under the control of a ubiquitous promoter. The presence of the Gal4 repressor on the wild-type chromosome prevents GFP expression in both heterozygous and homozygous wild type cells [135]. The MARCM technique was later extended to label wild type cells as well [146].

Figure 1. Mosaic analysis with a repressible cell marker (MARCM). (**A**) Gal4 transcription factor (blue oval) drives expression of green fluorescent protein (GFP) gene (green box) by binding the upstream activation sequence (UAS) (white box). This expression is repressed when Gal80 (red oval) is present. As a consequence cells that do not carry a gene encoding Gal80 but carry genes encoding Gal4 and UAS-GFP are marked green. (**B**) In MARCM, the *GAL80* repressor gene (red box) is carried on a chromosome that bears the wild-type allele of a gene (yellow box) of interest and a flippase recognition target (FRT) site (grey triangle) placed pericentricaly. The homologous chromosome carries a FRT site in exactly the same position and a mutant allele (orange box), but does not carry the *GAL80* gene. Cells also carry the *GAL4* gene and UAS-GFP on the other chromosomes. During G2 phase (after DNA replication), flippase mediates recombination between two FRT sites of homologous chromosomes, thus generating sister chromatids; one of which carries the wild-type allele and *GAL80* repressor and the other the mutant allele. During mitosis, sister chromatids are distributed to daughter cells, generating cells that are homozygous wild-type or homozygous mutant. Cells that are homozygous mutant are the only cells lacking the *GAL80* gene and thus are labeled with GFP. Reproduced with permission from MacMillan: Nature Protocols ©2007 [181].



MARCM, among other mitotic-recombination-based approaches, has enabled the creation of tissue specific mutant cells for genes where a mutant exists. This, however, is not yet [111,147] the case for all fruit fly genes. Post-transcriptional gene silencing by double-stranded RNA (dsRNA) [148], commonly known as RNA interference (RNAi), allows the silencing of virtually any transcript encoded by the genome [149]. *Drosophila* is not only one of the first organisms where RNAi has been used to silence genes [150,151], but it has also played an important role in studying the mechanism of dsRNA dependent gene silencing [152]. Injections of dsRNA into the *Drosophila* embryos were used to pioneer RNAi in the fruit fly. However, this mode of delivery has limited use for studying gene function in the late stages of development or in a tissue specific

manner. A combination of genetically encoded hairpin-loop RNAs with a Gal4/UAS system has been introduced to address these issues and place *Drosophila* RNAi under spatio-temporal control [153]. Efficient transformation techniques developed for *Drosophila* Schneider (S2) cells [154] combined with *Drosophila* dsRNA libraries allowed RNAi screens in cell culture on a genome-wide scale [155–158]. With genome-wide libraries of fly lines carrying UAS-driven hairpin RNAs, tissue-specific RNAi screens in the whole animals became possible [159,160]. While the first library used P-element for transgenesis, thus leading to variability of hairpin RNA expression levels in different lines, the next generation of libraries followed, using phiC31-mediated insertions into a defined locus [161,162]. RNAi in flies has proven very effective and allowed for a number of large scale screens to be performed, including ones targeting muscle development [162], heart function [163], obesity [164], pain [165], glial function [166], or piRNA pathways [167]. The off-target effect, a well-known pitfall of RNA interference, has been addressed in flies by specificity control using either cross-species rescue [168,169] or engineered RNAi-refractory transgenes [170].

7.4. Bright Rescue

Modern classical and reverse-genetic approaches often call for reliable sources of transgenes, both to induce new and rescue induced phenotypes. Classically, clones from cDNA libraries [69] combined with the Gal4/UAS system [127] have been used to specifically express a gene of interest in target tissue. These constructs could be used either to ectopically express a gene of interest [127], rescue a mutant phenotype [171], or by using a fusion of cDNA with a fluorescent protein coding sequence to visualize the localization of a protein of interest [172]. These approaches, however, do not allow simultaneous modification, such as introduction of point mutations, truncation, tagging, and expression of a protein of interest under native or nearly native control. This usually requires a larger genomic context.

Genome-wide libraries of fruit fly genomic DNA cloned in bacterial artificial chromosomes (BACs) or fosmids, spanning between 20 and over 100 kb, have been constructed for the purpose of genome sequencing [4]. The p[ACMAN] system (Figure 2B,C) has enabled turning them into reliable sources of modifiable genomic inserts, tailored for fly transgenesis. The centerpiece of the system is a single copy vector harboring a second, inducible medium copy origin of replication (oriV), a fly selectable marker (*white*), and attachment site (attB) for phiC31-mediated transgenesis [124]. Site-specific-recombinase-based transformation enables the insertion of constructs over 100 kb in size. Genomic inserts are subcloned into the backbone using Red/ET homologous recombination (Figure 3), also known as recombineering [173–176]. The ability to arbitrarily modify and transform large genomic constructs has fostered the development of transformation ready genomic libraries of *Drosophila melanogaster* and other fly species. Two such resources have been created so far, the p[ACMAN] [177] and FlyFos [178]. The p[ACMAN] features BAC libraries with average insert sizes of 21 and 83 kb. The vector used is similar to the one in the p[ACMAN] subcloning kit. The FlyFos system (Figure 2A,C) features 36 kb fosmid libraries for *Drosophila melanogaster* and *pseudoobscura* [169,178]. The library vector also includes an inducible oriV, attB site, and a dominant fluorescent marker, selectable in diverse

insect species [179]. The liquid culture recombineering pipeline [180] introduced in the system enables high-throughput gene tagging with a variety of tags in 96-well format [178].

Figure 2. FlyFos and p[ACMAN] genomic libraries. (**A**) FlyFos library is cloned in a fosmid vector, pFlyFos. Genomic inserts were cloned into the *Pml*I site. pFlyFos features an inducible origin of replication (oriS for single copy and oriV for arabinose-inducible moderate copy maintenance), attB site for fly transgenesis, and 3xP3-dsRed as a fly-selectable marker. (**B**) p[ACMAN] libraries are cloned into the *Bam*HI site of a p[ACMAN] bacterial artificial chromosome (BAC) vector. This vector also features inducible oriS/oriV and attB site, but uses white as fly selectable marker. In addition to phiC31-mediated transgenesis, p[ACMAN] vector carrying small inserts can theoretically be used for P-element transformation. (**C**) Size distribution of FlyFos and p[ACMAN] *D. melanogaster* libraries. There are two p[ACMAN] libraries: CHORI-321 with average clone size of 83.3 kb and CHORI-322 with average clone size of 21 kb. The FlyFos library has an average clone size of 36 kb.

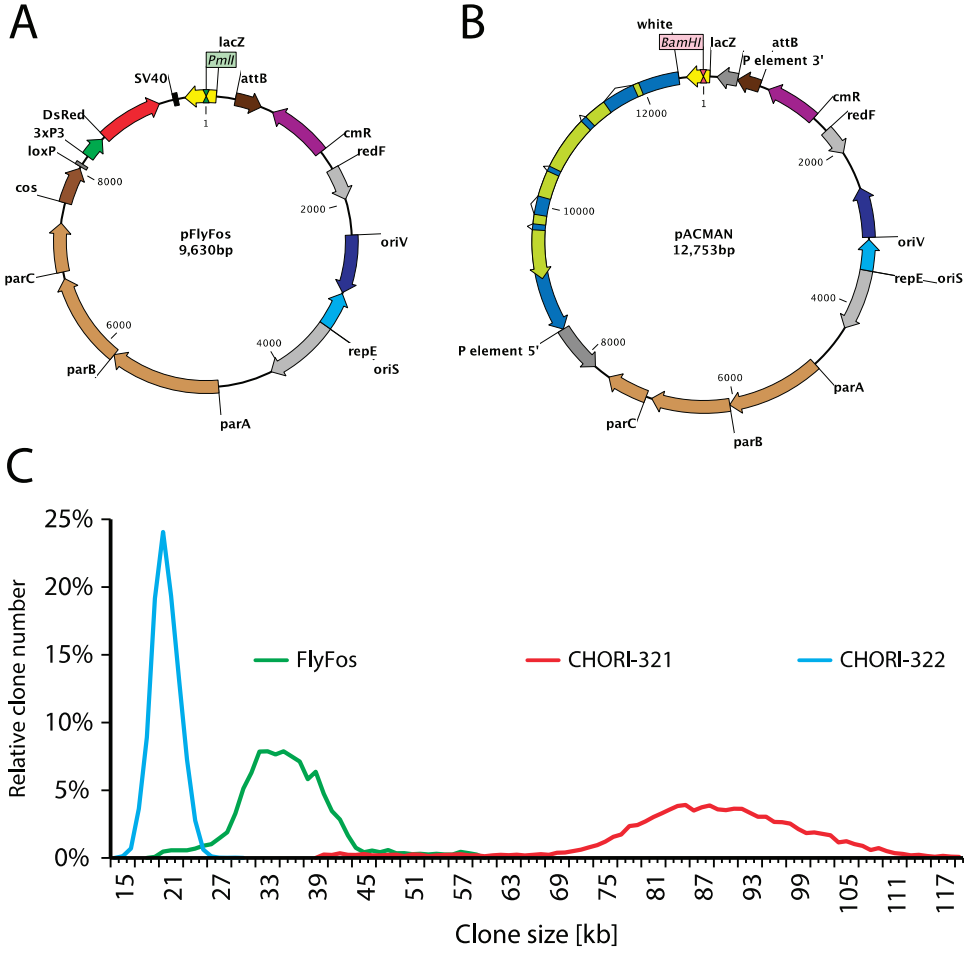
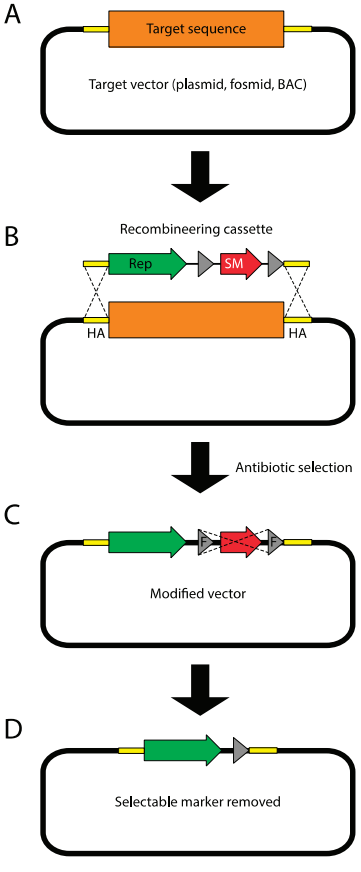


Figure 3. Recombineering principles. (**A**) The target sequence (orange box) is carried on a single copy fosmid or bacterial artificial chromosome (BAC) vector. The 50 bp fragments flanking the target sequence, called homology arms (HA), are depicted as yellow boxes. In this example, the target sequence will be replaced with the recombineering cassette. However, when homology arms are designed to directly follow each other, the cassette can be simply inserted into the target vector. (**B**) The PCR-amplified recombineering cassette harboring homology arms (introduced as primer overhangs) on its termini is electroporated into bacteria carrying the target vector. In the depicted example, the cassette contains a reporter (green arrow) and a flippase recognition target (FRT)-flanked (grey triangles) bacterial selectable marker (red arrow). Homologous recombinase, transiently expressed in bacteria mediates recombination between homology arms replacing the target sequence with the recombineering cassette. (**C**) Recombinant bacterial cells are selected using the selectable marker encoded in the recombineering cassette. If the selectable marker is flanked by FRT sites, it can now be removed (flipped-out) through transient expression of flippase. (**D**) The final recombineering product contains the desired sequence and a 34 bp FRT scar flanked by the homology arms.



BAC and fosmid-based recombineering has enabled the introduction of modified “third alleles” of genes of interest. The powerful fruit fly genetics toolkit also allows for modifications of genes *in situ*, in their native loci. The first *in situ* genomic targeting in *Drosophila* was performed using the ends-in technique [182]. Ends-in genomic targeting relies on double strand break (DSB) repair through homologous recombination. The targeting construct contains homology arms, one of which is antiparallel to the genomic sequence, and leads to duplication of the targeted locus upon recombination. Initial targeting attempts involving linear DNA injection into the germline were unsuccessful. Inserting the targeting construct into a random locus first, via P-element transgenesis, has solved the issue. FRT sites present on the flanks of the construct were used to mobilize the targeting construct from the genome before generating DSB using I-*Sce*I nuclease [182]. Ends-out targeting uses very similar basic logic, but relies on homology arms that are both parallel to the genomic locus, therefore leading to a clean insertion or replacement [183]. Both ends-in and ends-out have provided reliable means to target genomic loci; however, at a cost of relatively low efficiency. This has made targeting the same locus with different cassettes a labor-intensive task. The integrase-mediated approach for gene knock-out (IMAGO) technique (Figure 4) combines ends-out targeting with phiC31-mediated recombinase-mediated cassette exchange (RMCE) [184]. IMAGO uses ends-out to replace the targeted locus with an attP-flanked selectable marker, which can subsequently be replaced with any desirable construct, thus enabling *in situ* gene tagging, conditional knock-outs, or functional analysis of orthologs. An alternative strategy uses a single attP site and a loxP-flanked selectable marker as the knock-out cassette [185]. Rescue constructs can then be integrated into the target locus using phiC31-mediated transgenesis, just like into any other landing site.

Genomic targeting techniques using DSBs induced in the targeting construct have proven to be robust tools. However, these approaches have a quite high price tag, because of their low efficiency. Homologous recombination with genomic loci is known to be much more effective if DSBs are introduced in the chromosome [186]. Induction of chromosomal DSBs in specific genomic loci requires designer nucleases that can target a sequence of choice. Currently three custom nuclease systems are in broad use: zinc finger nucleases (ZFNs) [187], transcription activator-like effector nucleases (TALENs), and the bacterial clustered regularly interspaced short palindromic repeat (CRISPR) system and its RNA-driven Cas9 nuclease. Double strand breaks are repaired using one of two cellular mechanisms: non-homologous end joining (NHEJ) and homologous recombination (HR). NHEJ involves processing and ligation of broken strands and usually leads to insertions and deletions [188]. However; it has also been shown to mediate efficient knock-ins in zebrafish [189]. HR requires a sequence homologous to the locus in which DSB has occurred, either from a sister chromatid, paralogous locus; or provided linear or plasmid DNA, and can, therefore, be exploited to insert or replace a genomic sequence with custom constructs [190].

Figure 4. Integrase-mediated approach for gene knock-out (IMAGO). (**A**) A targeting construct harboring *white* gene flanked by attP sites, 1 kb–5 kb homology arms, I-*Sce*I meganuclease site, and flippase recognition target (FRT) sites is inserted into the fly genome using transposition or site-specific integration. (**B**) The targeting cassette is mobilized into the circular episome by flippase and subsequently linearized by meganuclease. This linear fragment induces cellular double strand break repair mechanisms and (with certain frequency) replaces the genomic locus flanked by homology arms. (**C**) Recombinant progeny are selected for *white* dominant marker. The attP sites flanking the *white* gene can be used for recombinase-mediated cassette exchange. (**D**, **F**) A plasmid containing the attB-flanked construct (cKO or a mutant allele) is injected into phiC31-expressing fly embryos and exchanges the attP-flanked *white* gene. (**E**, **G**) Recombinant progeny carrying modified alleles are selected for loss of the *white* dominant marker.

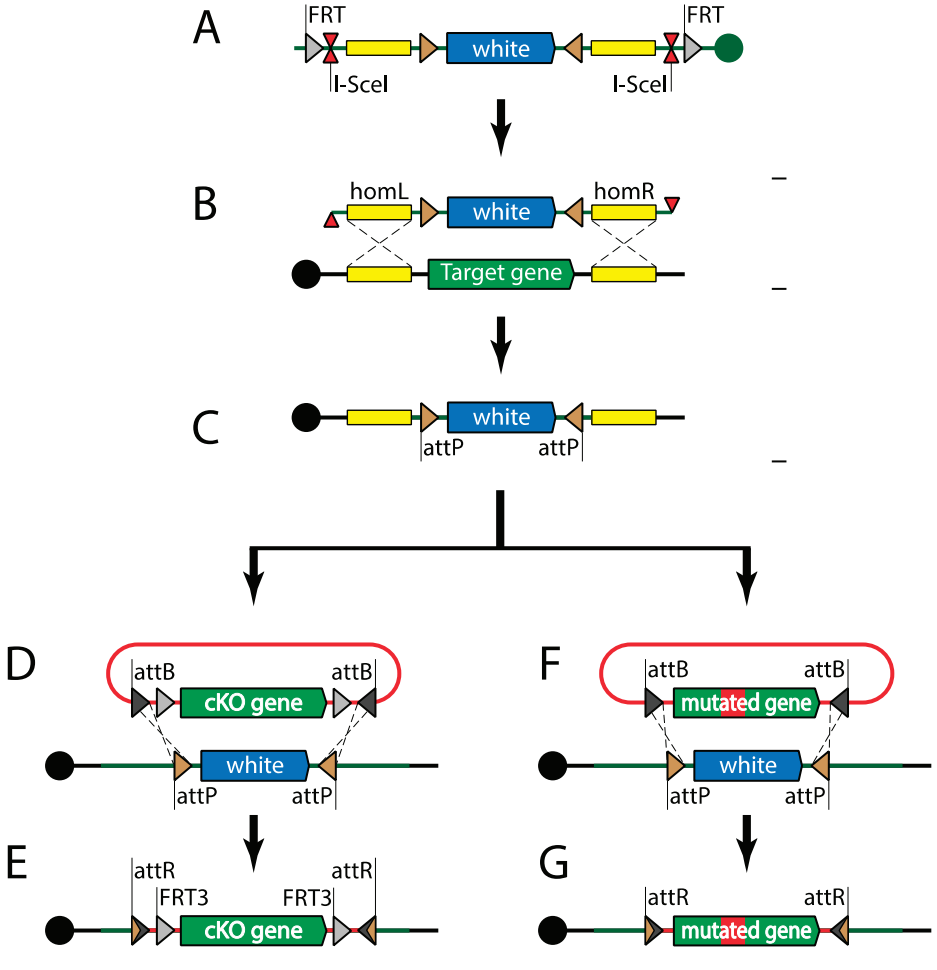
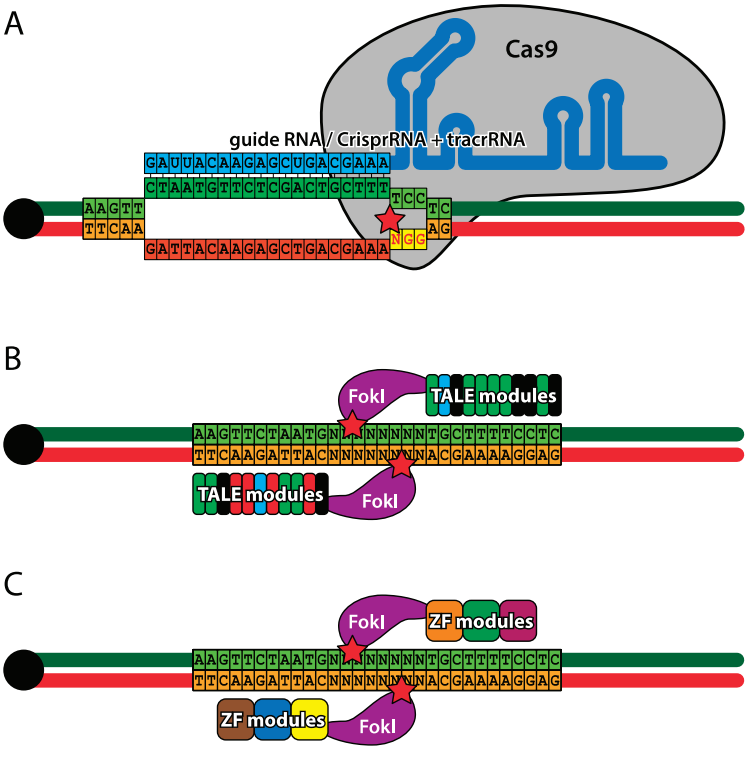


Figure 5. Designer nucleases. (**A**) Zink finger nucleases (ZFNs) combine a zinc finger DNA binding domain with a FokI nickase. Each zinc finger recognizes a triplet of bases and usually three to six zinc fingers are present in the targeting domain. Cleavage occurs outside the target sequence and requires a pair of ZFNs, each binding one DNA strand. (**B**) Transcription activator-like effector nucleases (TALEN), similarly to ZFNs have two domains: a DNA binding domain and a FokI nickase domain. The targeting domain is composed of 33–35 amino acid repeats, each binding a single nucleotide. The cleavage mechanism of TALENs is identical to ZFNs. (**C**) Clustered regularly interspaced short palindromic repeats (CRISPR) is a RNA driven double-stranded DNA endonuclease system. Cleavage specificity is provided by crRNA (cyan) that hybridizes with the target sequence (green). Cleavage is performed by the Cas9 protein that, in addition to crRNA, requires tracrRNA for activity. The cleavage site (star) is located between the target sequence and NGG protospacer adjacent motif, complimentary to the sequence immediately downstream of the target. crRNA and tracrRNA can be fused to form guide RNA of similar activity.



ZFNs were the first designer nuclease system (Figure 5A) to be introduced in flies [188]. They are comprised of three to four zinc-finger DNA binding modules, each recognizing three base pairs, and a FokI endonuclease. Since FokI needs to dimerize for activity, a pair of ZFNs is required for DNA cleavage [187]. The specificity and affinity of zinc-finger modules is context dependent,

therefore, several strategies have been developed to achieve assembly of optimal DNA binding domains [191–194]. TALENs (Figure 5B), similar to ZFNs, are hybrids of DNA binding domains derived from transcription factors and FokI endonuclease and, as a consequence, two TALENs are required to form a functional nuclease [195–197]. The TALE (transcription activator-like effector) domains contain a tandem array of 15.5–19.5 repeats, each made of 34 residues, two of which provide DNA-binding specificity against a single nucleotide [198]. Due to the highly repetitive coding sequence of the TALE domain, special approaches have been developed for its efficient assembly using type IIs endonucleases [197]. CRISPR (Figure 5C), a defensive nuclease system from *Streptococcus pyogenes*, takes a completely different approach to DNA cleavage. The specificity is provided by a crRNA pairing with a 20 nt complimentary sequence within the DNA target. The cleavage is performed by Cas9 nuclease that requires trans-activating CRISPR RNA (tracrRNA) in addition to crRNA for activity. The complimentary region of the DNA target must be followed by a 3 bp PAM (protospacer adjacent motif) [199,200]. A pair of crRNA and tracrRNA can be replaced by a single hybrid guide RNA (sgRNA), thus reducing the system to two components [201]. To date, several implementations of the CRISPR system have been created in *Drosophila* [202–204], including transgenic flies with genomically encoded sources of Cas9 [205–207] and tracrRNA/sgRNA [208]. The CRISPR system has been combined with classical ends-out targeting and site-specific integrase approaches, resulting in a versatile toolkit for genome engineering [209]. It should be stated that at this stage the efficiency and specificity of all designer-nuclease-based approaches *in vivo* remains to be fully established, although CRISPR is showing great promise.

8. Conclusions

Drosophila occupies a paramount position among model organisms, largely due to the variety of genetic tools unique to the fruit fly, its short generation time, and ease of transformation. The *Drosophila* classic Gal4/UAS two-component expression system and its counterparts, LexA and Q, can be combined with one another and with site-specific recombination systems like Flp/FRT, Cre/LoxP or phiC31, yielding novel combinatorial systems for even tighter spatio-temporal gene expression control, clonal analysis, and lineage tracing [210]. The fruit fly genome is easily accessible using a broad range of genome engineering tools, including those based on classic transposition, site-specific recombinases and fosmid/BAC recombineering [211], as well as the emerging field of genome editing using designer nucleases [212]. Availability of an almost complete genomic sequence for over 12 species from genus *Drosophila* and dozens of various *D. melanogaster* strain genomes make fruit flies an excellent model for comparative genomics and population genetics. A large number of human disease-related genes that have homologs in the fruit fly [213,214] connected with powerful resources for QTL mapping and GWAS [89,90] make *Drosophila* an attractive model for studying the genetic basis of human disease.

Rapid development of genome engineering techniques, especially those introducing synthetic approaches using designer DNA binding domains of TALEs [215,216] and the CRISPR system [217], will undeniably affect the *Drosophila* field in the next years. Completing the genome

and transcriptome sequencing effort for additional fly species [218] will aid in further functional annotation of the *Drosophila* genome and fuel the evolutionary developmental biology field.

Acknowledgments

Work in the Hassan lab is supported by VIB and grants form FWO, BELSPO, and KU Leuven. Radoslaw K. Ejsmont was supported by an EMBO ALTF 1056-2011 long-term fellowship and currently by a Marie Curie Actions co-funded Omics@VIB postdoctoral fellowship.

Author Contributions

Wrote the manuscript: Radoslaw K. Ejsmont and Bassem A. Hassan.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Fleischmann, R.D.; Adams, M.D.; White, O.; Clayton, R.A.; Kirkness, E.F.; Kerlavage, A.R.; Bult, C.J.; Tomb, J.F.; Dougherty, B.A.; Merrick, J.M.; *et al*. Whole-genome random sequencing and assembly of haemophilus influenzae Rd. *Science* **1995**, *269*, 496–512.
2. Clayton, R.A.; White, O.; Ketchum, K.A.; Venter, J.C. The first genome from the third domain of life. *Nature* **1997**, *387*, 459–462.
3. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **1998**, *282*, 2012–2018.
4. Adams, M.D.; Celniker, S.E.; Holt, R.A.; Evans, C.A.; Gocayne, J.D.; Amanatides, P.G.; Scherer, S.E.; Li, P.W.; Hoskins, R.A.; Galle, R.F.; *et al*. The genome sequence of drosophila melanogaster. *Science* **2000**, *287*, 2185–2195.
5. Lewis, E.B. A gene complex controlling segmentation in drosophila. *Nature* **1978**, *276*, 565–570.
6. Nusslein-Volhard, C.; Wieschaus, E. Mutations affecting segment number and polarity in drosophila. *Nature* **1980**, *287*, 795–801.
7. Gehring, W.J.; Hiromi, Y. Homeotic genes and the homeobox. *Annu. Rev. Genet.* **1986**, *20*, 147–173.
8. Lawrence, P.A.; Morata, G. Homeobox genes: Their function in drosophila segmentation and pattern formation. *Cell* **1994**, *78*, 181–189.
9. Krumlauf, R. Hox genes in vertebrate development. *Cell* **1994**, *78*, 191–201.
10. Quiring, R.; Walldorf, U.; Kloter, U.; Gehring, W.J. Homology of the eyeless gene of drosophila to the small eye gene in mice and aniridia in humans. *Science* **1994**, *265*, 785–789.

11. Ton, C.C.; Hirvonen, H.; Miwa, H.; Weil, M.M.; Monaghan, P.; Jordan, T.; van Heyningen, V.; Hastie, N.D.; Meijers-Heijboer, H.; Drechsler, M.; *et al*. Positional cloning and characterization of a paired box- and homeobox-containing gene from the aniridia region. *Cell* **1991**, *67*, 1059–1074.
12. Hill, R.E.; Favor, J.; Hogan, B.L.; Ton, C.C.; Saunders, G.F.; Hanson, I.M.; Prosser, J.; Jordan, T.; Hastie, N.D.; van Heyningen, V. Mouse small eye results from mutations in a paired-like homeobox-containing gene. *Nature* **1991**, *354*, 522–525.
13. Halder, G.; Callaerts, P.; Flister, S.; Walldorf, U.; Kloter, U.; Gehring, W.J. Eyeless initiates the expression of both sine oculis and eyes absent during drosophila compound eye development. *Development* **1998**, *125*, 2181–2191.
14. Kumar, J.P. The sine oculis homeobox (six) family of transcription factors as regulators of development and disease. *Cell. Mol. Life Sci.* **2009**, *66*, 565–583.
15. Jarman, A.P.; Grell, E.H.; Ackerman, L.; Jan, L.Y.; Jan, Y.N. Atonal is the proneural gene for drosophila photoreceptors. *Nature* **1994**, *369*, 398–400.
16. Jarman, A.P.; Grau, Y.; Jan, L.Y.; Jan, Y.N. Atonal is a proneural gene that directs chordotonal organ formation in the drosophila peripheral nervous system. *Cell* **1993**, *73*, 1307–1321.
17. Brown, N.L.; Patel, S.; Brzezinski, J.; Glaser, T. Math5 is required for retinal ganglion cell and optic nerve formation. *Development* **2001**, *128*, 2497–2508.
18. Bermingham, N.A.; Hassan, B.A.; Price, S.D.; Vollrath, M.A.; Ben-Arie, N.; Eatock, R.A.; Bellen, H.J.; Lysakowski, A.; Zoghbi, H.Y. Math1: An essential gene for the generation of inner ear hair cells. *Science* **1999**, *284*, 1837–1841.
19. Tabata, T.; Eaton, S.; Kornberg, T.B. The drosophila hedgehog gene is expressed specifically in posterior compartment cells and is a target of engrailed regulation. *Genes Dev.* **1992**, *6*, 2635–2645.
20. Lee, J.J.; von Kessler, D.P.; Parks, S.; Beachy, P.A. Secretion and localized transcription suggest a role in positional signaling for products of the segmentation gene hedgehog. *Cell* **1992**, *71*, 33–50.
21. Hooper, J.E.; Scott, M.P. The drosophila patched gene encodes a putative membrane protein required for segmental patterning. *Cell* **1989**, *59*, 751–765.
22. Nakano, Y.; Guerrero, I.; Hidalgo, A.; Taylor, A.; Whittle, J.R.; Ingham, P.W. A protein with several possible membrane-spanning domains encoded by the drosophila segment polarity gene patched. *Nature* **1989**, *341*, 508–513.
23. Stone, D.M.; Hynes, M.; Armanini, M.; Swanson, T.A.; Gu, Q.; Johnson, R.L.; Scott, M.P.; Pennica, D.; Goddard, A.; Phillips, H.; *et al*. The tumour-suppressor gene patched encodes a candidate receptor for sonic hedgehog. *Nature* **1996**, *384*, 129–134.
24. Marigo, V.; Davey, R.A.; Zuo, Y.; Cunningham, J.M.; Tabin, C.J. Biochemical evidence that patched is the hedgehog receptor. *Nature* **1996**, *384*, 176–179.
25. Bhanot, P.; Brink, M.; Samos, C.H.; Hsieh, J.C.; Wang, Y.; Macke, J.P.; Andrew, D.; Nathans, J.; Nusse, R. A new member of the frizzled family from drosophila functions as a wingless receptor. *Nature* **1996**, *382*, 225–230.

26. Siegfried, E.; Chou, T.B.; Perrimon, N. Wingless signaling acts through zeste-white 3, the drosophila homolog of glycogen synthase kinase-3, to regulate engrailed and establish cell fate. *Cell* **1992**, *71*, 1167–1179.
27. Wehrli, M.; Dougan, S.T.; Caldwell, K.; O'Keefe, L.; Schwartz, S.; Vaizel-Ohayon, D.; Schejter, E.; Tomlinson, A.; DiNardo, S. Arrow encodes an ldl-receptor-related protein essential for wingless signalling. *Nature* **2000**, *407*, 527–530.
28. Brunner, E.; Peter, O.; Schweizer, L.; Basler, K. Pangolin encodes a lef-1 homologue that acts downstream of armadillo to transduce the wingless signal in drosophila. *Nature* **1997**, *385*, 829–833.
29. Adler, P.N. The genetic control of tissue polarity in drosophila. *BioEssays* **1992**, *14*, 735–741.
30. Eaton, S.; Julicher, F. Cell flow and tissue polarity patterns. *Curr. Opin. Genet. Dev.* **2011**, *21*, 747–752.
31. Wharton, K.A.; Johansen, K.M.; Xu, T.; Artavanis-Tsakonas, S. Nucleotide sequence from the neurogenic locus notch implies a gene product that shares homology with proteins containing egf-like repeats. *Cell* **1985**, *43*, 567–581.
32. Artavanis-Tsakonas, S.; Matsuno, K.; Fortini, M.E. Notch signaling. *Science* **1995**, *268*, 225–232.
33. Artavanis-Tsakonas, S.; Rand, M.D.; Lake, R.J. Notch signaling: Cell fate control and signal integration in development. *Science* **1999**, *284*, 770–776.
34. Wu, S.; Huang, J.; Dong, J.; Pan, D. Hippo encodes a STE-20 family protein kinase that restricts cell proliferation and promotes apoptosis in conjunction with salvador and warts. *Cell* **2003**, *114*, 445–456.
35. Harvey, K.F.; Pfleger, C.M.; Hariharan, I.K. The drosophila mst ortholog, hippo, restricts growth and cell proliferation and promotes apoptosis. *Cell* **2003**, *114*, 457–467.
36. Udan, R.S.; Kango-Singh, M.; Nolo, R.; Tao, C.; Halder, G. Hippo promotes proliferation arrest and apoptosis in the salvador/warts pathway. *Nat. Cell Biol.* **2003**, *5*, 914–920.
37. Fortini, M.E.; Bonini, N.M. Modeling human neurodegenerative diseases in drosophila: On a wing and a prayer. *Trends Genet.* **2000**, *16*, 161–167.
38. Lloyd, T.E.; Taylor, J.P. Flightless flies: Drosophila models of neuromuscular disease. *Ann. N. Y. Acad. Sci.* **2010**, *1184*, e1–e20.
39. Potter, C.J.; Turenchalk, G.S.; Xu, T. Drosophila in cancer research. An expanding role. *Trends Genet.* **2000**, *16*, 33–39.
40. Jackson, G.R.; Salecker, I.; Dong, X.; Yao, X.; Arnheim, N.; Faber, P.W.; MacDonald, M.E.; Zipursky, S.L. Polyglutamine-expanded human huntingtin transgenes induce degeneration of drosophila photoreceptor neurons. *Neuron* **1998**, *21*, 633–642.
41. Warrick, J.M.; Paulson, H.L.; Gray-Board, G.L.; Bui, Q.T.; Fischbeck, K.H.; Pittman, R.N.; Bonini, N.M. Expanded polyglutamine protein forms nuclear inclusions and causes neural degeneration in drosophila. *Cell* **1998**, *93*, 939–949.
42. Morales, J.; Hiesinger, P.R.; Schroeder, A.J.; Kume, K.; Verstreken, P.; Jackson, F.R.; Nelson, D.L.; Hassan, B.A. Drosophila fragile x protein, dfxr, regulates neuronal morphology and function in the brain. *Neuron* **2002**, *34*, 961–972.

43. Luo, L.Q.; Martin-Morris, L.E.; White, K. Identification, secretion, and neural expression of appl, a drosophila protein similar to human amyloid protein precursor. *J. Neurosci.* **1990**, *10*, 3849–3861.
44. Luo, L.; Tully, T.; White, K. Human amyloid precursor protein ameliorates behavioral deficit of flies deleted for appl gene. *Neuron* **1992**, *9*, 595–605.
45. Ye, Y.; Fortini, M.E. Apoptotic activities of wild-type and alzheimer's disease-related mutant presenilins in drosophila melanogaster. *J. Cell Biol.* **1999**, *146*, 1351–1364.
46. Wittmann, C.W.; Wszolek, M.F.; Shulman, J.M.; Salvaterra, P.M.; Lewis, J.; Hutton, M.; Feany, M.B. Tauopathy in drosophila: Neurodegeneration without neurofibrillary tangles. *Science* **2001**, *293*, 711–714.
47. Feany, M.B.; Bender, W.W. A drosophila model of parkinson's disease. *Nature* **2000**, *404*, 394–398.
48. Whitworth, A.J.; Theodore, D.A.; Greene, J.C.; Benes, H.; Wes, P.D.; Pallanck, L.J. Increased glutathione s-transferase activity rescues dopaminergic neuron loss in a drosophila model of parkinson's disease. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 8024–8029.
49. Parkes, T.L.; Elia, A.J.; Dickinson, D.; Hilliker, A.J.; Phillips, J.P.; Boulianne, G.L. Extension of drosophila lifespan by overexpression of human sod1 in motorneurons. *Nat. Genet.* **1998**, *19*, 171–174.
50. Shahidullah, M.; Le Marchand, S.J.; Fei, H.; Zhang, J.; Pandey, U.B.; Dalva, M.B.; Pasinelli, P.; Levitan, I.B. Defects in synapse structure and function precede motor neuron degeneration in drosophila models of fus-related als. *J. Neurosci.* **2013**, *33*, 19590–19598.
51. Van der Plas, M.C.; Pilgram, G.S.; Plomp, J.J.; de Jong, A.; Fradkin, L.G.; Noordermeer, J.N. Dystrophin is required for appropriate retrograde control of neurotransmitter release at the drosophila neuromuscular junction. *J. Neurosci.* **2006**, *26*, 333–344.
52. Moore, M.S.; DeZazzo, J.; Luk, A.Y.; Tully, T.; Singh, C.M.; Heberlein, U. Ethanol intoxication in drosophila: Genetic and pharmacological evidence for regulation by the camp signaling pathway. *Cell* **1998**, *93*, 997–1007.
53. Kaun, K.R.; Azanchi, R.; Maung, Z.; Hirsh, J.; Heberlein, U. A drosophila model for alcohol reward. *Nat. Neurosci.* **2011**, *14*, 612–619.
54. McClung, C.; Hirsh, J. Stereotypic behavioral responses to free-base cocaine and the development of behavioral sensitization in drosophila. *Curr. Biol.* **1998**, *8*, 109–112.
55. Al-Anzi, B.; Sapin, V.; Waters, C.; Zinn, K.; Wyman, R.J.; Benzer, S. Obesity-blocking neurons in drosophila. *Neuron* **2009**, *63*, 329–341.
56. Baker, K.D.; Thummel, C.S. Diabetic larvae and obese flies-emerging studies of metabolism in drosophila. *Cell Metab.* **2007**, *6*, 257–266.
57. Wolf, M.J.; Amrein, H.; Izatt, J.A.; Choma, M.A.; Reedy, M.C.; Rockman, H.A. Drosophila as a model for the identification of genes causing adult human heart disease. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 1394–1399.
58. Roeder, T.; Isermann, K.; Kallsen, K.; Uliczka, K.; Wagner, C. A drosophila asthma model—What the fly tells us about inflammatory diseases of the lung. *Adv. Exp. Med. Biol.* **2012**, *710*, 37–47.

59. Xu, T.; Wang, W.; Zhang, S.; Stewart, R.A.; Yu, W. Identifying tumor suppressors in genetic mosaics: The drosophila lats gene encodes a putative protein kinase. *Development* **1995**, *121*, 1053–1063.
60. Pagliarini, R.A.; Xu, T. A genetic screen in drosophila for metastatic behavior. *Science* **2003**, *302*, 1227–1231.
61. Januschke, J.; Gonzalez, C. Drosophila asymmetric division, polarity and cancer. *Oncogene* **2008**, *27*, 6994–7002.
62. Brumby, A.M.; Richardson, H.E. Using drosophila melanogaster to map human cancer pathways. *Nat. Rev. Cancer* **2005**, *5*, 626–639.
63. Vidal, M.; Cagan, R.L. Drosophila models for cancer research. *Curr. Opin. Genet. Dev.* **2006**, *16*, 10–16.
64. Miles, W.O.; Dyson, N.J.; Walker, J.A. Modeling tumor invasion and metastasis in drosophila. *Dis. Models Mech.* **2011**, *4*, 753–761.
65. Gonzalez, C. Drosophila melanogaster: A model and a tool to investigate malignancy and identify new therapeutics. *Nat. Rev. Cancer* **2013**, *13*, 172–183.
66. Bosco, G.; Campbell, P.; Leiva-Neto, J.T.; Markow, T.A. Analysis of drosophila species genome size and satellite DNA content reveals significant differences among strains as well as between species. *Genetics* **2007**, *177*, 1277–1290.
67. Hoskins, R.A.; Smith, C.D.; Carlson, J.W.; Carvalho, A.B.; Halpern, A.; Kaminker, J.S.; Kennedy, C.; Mungall, C.J.; Sullivan, B.A.; Sutton, G.G.; *et al*. Heterochromatic sequences in a drosophila whole-genome shotgun assembly. *Genome Biol.* **2002**, doi:10.1186/gb-2002-3-12-research0085.
68. Misra, S.; Crosby, M.A.; Mungall, C.J.; Matthews, B.B.; Campbell, K.S.; Hradecky, P.; Huang, Y.; Kaminker, J.S.; Millburn, G.H.; Prochnik, S.E.; *et al*. Annotation of the drosophila melanogaster euchromatic genome: A systematic review. *Genome Biol.* **2002**, doi:10.1186/gb-2002-3-12-research0083.
69. Stapleton, M.; Carlson, J.; Brokstein, P.; Yu, C.; Champe, M.; George, R.; Guarin, H.; Kronmiller, B.; Pacleb, J.; Park, S.; *et al*. A drosophila full-length cdna resource. *Genome Biol.* **2002**, doi:10.1186/gb-2002-3-12-research0080.
70. Tomancak, P.; Beaton, A.; Weiszmann, R.; Kwan, E.; Shu, S.; Lewis, S.E.; Richards, S.; Ashburner, M.; Hartenstein, V.; Celniker, S.E.; *et al*. Systematic determination of patterns of gene expression during drosophila embryogenesis. *Genome Biol.* **2002**, doi:10.1186/gb-2002-3-12-research0088.
71. Kaminker, J.S.; Bergman, C.M.; Kronmiller, B.; Carlson, J.; Svirskas, R.; Patel, S.; Frise, E.; Wheeler, D.A.; Lewis, S.E.; Rubin, G.M.; *et al*. The transposable elements of the drosophila melanogaster euchromatin: A genomics perspective. *Genome Biol.* **2002**, doi:10.1186/gb-2002-3-12-research0084.
72. Ohler, U.; Liao, G.C.; Niemann, H.; Rubin, G.M. Computational analysis of core promoters in the drosophila genome. *Genome Biol.* **2002**, doi:10.1186/gb-2002-3-12-research0087.

73. Marygold, S.J.; Leyland, P.C.; Seal, R.L.; Goodman, J.L.; Thurmond, J.; Strelets, V.B.; Wilson, R.J.; the FlyBase consortium. Flybase: Improvements to the bibliography. *Nucleic Acids Res.* **2013**, *41*, D751–D757.
74. Daines, B.; Wang, H.; Wang, L.; Li, Y.; Han, Y.; Emmert, D.; Gelbart, W.; Wang, X.; Li, W.; Gibbs, R.; *et al*. The drosophila melanogaster transcriptome by paired-end RNA sequencing. *Genome Res.* **2011**, *21*, 315–324.
75. Mungall, C.J.; Misra, S.; Berman, B.P.; Carlson, J.; Frise, E.; Harris, N.; Marshall, B.; Shu, S.; Kaminker, J.S.; Prochnik, S.E.; *et al*. An integrated computational pipeline and database to support whole-genome sequence annotation. *Genome Biol.* **2002**, doi:10.1186/gb-2002-3-12-research0081.
76. Lewis, S.E.; Searle, S.M.; Harris, N.; Gibson, M.; Lyer, V.; Richter, J.; Wiel, C.; Bayraktaroglir, L.; Birney, E.; Crosby, M.A.; *et al*. Apollo: A sequence annotation editor. *Genome Biol.* **2002**, doi:10.1186/gb-2002-3-12-research0082.
77. Reese, M.G.; Kulp, D.; Tammana, H.; Haussler, D. Genie—Gene finding in drosophila melanogaster. *Genome Res.* **2000**, *10*, 529–538.
78. Wood, V.; Gwilliam, R.; Rajandream, M.A.; Lyne, M.; Lyne, R.; Stewart, A.; Sgouros, J.; Peat, N.; Hayles, J.; Baker, S.; *et al*. The genome sequence of schizosaccharomyces pombe. *Nature* **2002**, *415*, 871–880.
79. Stein, L.D.; Bao, Z.; Blasiar, D.; Blumenthal, T.; Brent, M.R.; Chen, N.; Chinwalla, A.; Clarke, L.; Clee, C.; Coghlan, A.; *et al*. The genome sequence of caenorhabditis briggsae: A platform for comparative genomics. *PLoS Biol.* **2003**, *1*, e45.
80. Richards, S.; Liu, Y.; Bettencourt, B.R.; Hradecky, P.; Letovsky, S.; Nielsen, R.; Thornton, K.; Hubisz, M.J.; Chen, R.; Meisel, R.P.; *et al*. Comparative genome sequencing of drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution. *Genome Res.* **2005**, *15*, 1–18.
81. Drosophila 12 Genomes Consortium; Clark, A.G.; Eisen, M.B.; Smith, D.R.; Bergman, C.M.; Oliver, B.; Markow, T.A.; Kaufman, T.C.; Kellis, M.; Gelbart, W.; *et al*. Evolution of genes and genomes on the drosophila phylogeny. *Nature* **2007**, *450*, 203–218.
82. Lin, M.F.; Carlson, J.W.; Crosby, M.A.; Matthews, B.B.; Yu, C.; Park, S.; Wan, K.H.; Schroeder, A.J.; Gramates, L.S.; St Pierre, S.E.; *et al*. Revisiting the protein-coding gene catalog of drosophila melanogaster using 12 fly genomes. *Genome Res.* **2007**, *17*, 1823–1836.
83. Stark, A.; Lin, M.F.; Kheradpour, P.; Pedersen, J.S.; Parts, L.; Carlson, J.W.; Crosby, M.A.; Rasmussen, M.D.; Roy, S.; Deoras, A.N.; *et al*. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature* **2007**, *450*, 219–232.
84. Majoros, W.H.; Ohler, U. Modeling the evolution of regulatory elements by simultaneous detection and alignment with phylogenetic pair hmms. *PLoS Comput. Biol.* **2010**, *6*, e1001037.
85. Arunachalam, M.; Jayasurya, K.; Tomancak, P.; Ohler, U. An alignment-free method to identify candidate orthologous enhancers in multiple drosophila genomes. *Bioinformatics* **2010**, *26*, 2109–2115.

86. Aerts, S.; Quan, X.J.; Claeys, A.; Naval Sanchez, M.; Tate, P.; Yan, J.; Hassan, B.A. Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in drosophila uncovers a regulatory basis for sensory specification. *PLoS Biol.* **2010**, *8*, e1000435.
87. Mackay, T.F. Quantitative trait loci in drosophila. *Nat. Rev. Genet.* **2001**, *2*, 11–20.
88. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J.P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367.
89. Mackay, T.F.; Richards, S.; Stone, E.A.; Barbadilla, A.; Ayroles, J.F.; Zhu, D.; Casillas, S.; Han, Y.; Magwire, M.M.; Cridland, J.M.; *et al*. The drosophila melanogaster genetic reference panel. *Nature* **2012**, *482*, 173–178.
90. King, E.G.; Merkes, C.M.; McNeil, C.L.; Hoofer, S.R.; Sen, S.; Broman, K.W.; Long, A.D.; Macdonald, S.J. Genetic dissection of a model complex trait using the drosophila synthetic population resource. *Genome Res.* **2012**, *22*, 1558–1566.
91. Ayroles, J.F.; Carbone, M.A.; Stone, E.A.; Jordan, K.W.; Lyman, R.F.; Magwire, M.M.; Rollmann, S.M.; Duncan, L.H.; Lawrence, F.; Anholt, R.R.; *et al*. Systems genetics of complex traits in drosophila melanogaster. *Nat. Genet.* **2009**, *41*, 299–307.
92. Massouras, A.; Waszak, S.M.; Albarca-Aguilera, M.; Hens, K.; Holcombe, W.; Ayroles, J.F.; Dermitzakis, E.T.; Stone, E.A.; Jensen, J.D.; Mackay, T.F.; *et al*. Genomic variation and its impact on gene expression in drosophila melanogaster. *PLoS Genet.* **2012**, *8*, e1003055.
93. Jordan, K.W.; Craver, K.L.; Magwire, M.M.; Cubilla, C.E.; Mackay, T.F.; Anholt, R.R. Genome-wide association for sensitivity to chronic oxidative stress in drosophila melanogaster. *PLoS One* **2012**, *7*, e38722.
94. Jumbo-Lucioni, P.; Bu, S.; Harbison, S.T.; Slaughter, J.C.; Mackay, T.F.; Moellering, D.R.; de Luca, M. Nuclear genomic control of naturally occurring variation in mitochondrial function in drosophila melanogaster. *BMC Genomics* **2012**, doi:10.1186/1471-2164-13-659.
95. Magwire, M.M.; Fabian, D.K.; Schweyen, H.; Cao, C.; Longdon, B.; Bayer, F.; Jiggins, F.M. Genome-wide association studies reveal a simple genetic basis of resistance to naturally coevolving viruses in drosophila melanogaster. *PLoS Genet.* **2012**, *8*, e1003057.
96. Harbison, S.T.; McCoy, L.J.; Mackay, T.F. Genome-wide association study of sleep in drosophila melanogaster. *BMC Genomics* **2013**, doi:10.1186/1471-2164-14-281.
97. King, E.G.; Macdonald, S.J.; Long, A.D. Properties and power of the drosophila synthetic population resource for the routine dissection of complex traits. *Genetics* **2012**, *191*, 935–949.
98. ENCODE Project Consortium. The encode (encyclopedia of DNA elements) project. *Science* **2004**, *306*, 636–640.
99. Gerstein, M.B.; Lu, Z.J.; van Nostrand, E.L.; Cheng, C.; Arshinoff, B.I.; Liu, T.; Yip, K.Y.; Robilotto, R.; Rechtsteiner, A.; Ikegami, K.; *et al*. Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science* **2010**, *330*, 1775–1787.
100. modENCODE Consortium; Roy, S.; Ernst, J.; Kharchenko, P.V.; Kheradpour, P.; Negre, N.; Eaton, M.L.; Landolin, J.M.; Bristow, C.A.; Ma, L.; *et al*. Identification of functional elements and regulatory circuits by drosophila modencode. *Science* **2010**, *330*, 1787–1797.

101. Riddle, N.C.; Minoda, A.; Kharchenko, P.V.; Alekseyenko, A.A.; Schwartz, Y.B.; Tolstorukov, M.Y.; Gorchakov, A.A.; Jaffe, J.D.; Kennedy, C.; Linder-Basso, D.; *et al*. Plasticity in patterns of histone modifications and chromosomal proteins in drosophila heterochromatin. *Genome Res.* **2011**, *21*, 147–163.
102. Kharchenko, P.V.; Alekseyenko, A.A.; Schwartz, Y.B.; Minoda, A.; Riddle, N.C.; Ernst, J.; Sabo, P.J.; Larschan, E.; Gorchakov, A.A.; Gu, T.; *et al*. Comprehensive analysis of the chromatin landscape in drosophila melanogaster. *Nature* **2011**, *471*, 480–485.
103. Nègre, N.; Brown, C.D.; Ma, L.; Bristow, C.A.; Miller, S.W.; Wagner, U.; Kheradpour, P.; Eaton, M.L.; Loriaux, P.; Sealfon, R.; *et al*. A cis-regulatory map of the drosophila genome. *Nature* **2011**, *471*, 527–531.
104. Spradling, A.C.; Rubin, G.M. Transposition of cloned p elements into drosophila germ line chromosomes. *Science* **1982**, *218*, 341–347.
105. Searles, L.L.; Jokerst, R.S.; Bingham, P.M.; Voelker, R.A.; Greenleaf, A.L. Molecular cloning of sequences from a drosophila RNA polymerase ii locus by p element transposon tagging. *Cell* **1982**, *31*, 585–592.
106. Rubin, G.M.; Spradling, A.C. Genetic transformation of drosophila with transposable element vectors. *Science* **1982**, *218*, 348–353.
107. Cooley, L.; Kelley, R.; Spradling, A. Insertional mutagenesis of the drosophila genome with single P elements. *Science* **1988**, *239*, 1121–1128.
108. Cooley, L.; Thompson, D.; Spradling, A.C. Constructing deletions with defined endpoints in drosophila. *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 3170–3173.
109. O'Kane, C.J.; Gehring, W.J. Detection *in situ* of genomic regulatory elements in drosophila. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 9123–9127.
110. Bellen, H.J.; Levis, R.W.; Liao, G.; He, Y.; Carlson, J.W.; Tsang, G.; Evans-Holm, M.; Hiesinger, P.R.; Schulze, K.L.; Rubin, G.M.; *et al*. The bdgp gene disruption project: Single transposon insertions associated with 40% of drosophila genes. *Genetics* **2004**, *167*, 761–781.
111. Bellen, H.J.; Levis, R.W.; He, Y.; Carlson, J.W.; Evans-Holm, M.; Bae, E.; Kim, J.; Metaxakis, A.; Savakis, C.; Schulze, K.L.; *et al*. The drosophila gene disruption project: Progress using transposons with distinctive site specificities. *Genetics* **2011**, *188*, 731–743.
112. Jacobson, J.W.; Medhora, M.M.; Hartl, D.L. Molecular structure of a somatically unstable transposable element in drosophila. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 8684–8688.
113. Garza, D.; Medhora, M.; Koga, A.; Hartl, D.L. Introduction of the transposable element mariner into the germline of drosophila melanogaster. *Genetics* **1991**, *128*, 303–310.
114. Franz, G.; Savakis, C. Minos, a new transposable element from drosophila hydei, is a member of the tc1-like family of transposons. *Nucleic Acids Res.* **1991**, *19*, 6646.
115. Loukeris, T.G.; Arca, B.; Livadaras, I.; Dialektaki, G.; Savakis, C. Introduction of the transposable element minos into the germ line of drosophila melanogaster. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 9485–9489.
116. Handler, A.M.; McCombs, S.D.; Fraser, M.J.; Saul, S.H. The lepidopteran transposon vector, piggybac, mediates germ-line transformation in the mediterranean fruit fly. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 7520–7525.

117. Handler, A.M.; Harrell, R.A., 2nd. Germline transformation of drosophila melanogaster with the piggybac transposon vector. *Insect Mol. Biol.* **1999**, *8*, 449–457.
118. Spradling, A.C.; Rubin, G.M. The effect of chromosomal position on the expression of the drosophila xanthine dehydrogenase gene. *Cell* **1983**, *34*, 47–57.
119. Levis, R.; Hazelrigg, T.; Rubin, G.M. Effects of genomic position on the expression of transduced copies of the white gene of drosophila. *Science* **1985**, *229*, 558–561.
120. Taillebourg, E.; Dura, J.M. A novel mechanism for p element homing in drosophila. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 6856–6861.
121. Cheng, Y.; Kwon, D.Y.; Arai, A.L.; Mucci, D.; Kassis, J.A. P-element homing is facilitated by engrailed polycomb-group response elements in drosophila melanogaster. *PLoS One* **2012**, *7*, e30437.
122. Thorpe, H.M.; Smith, M.C. *In vitro* site-specific integration of bacteriophage DNA catalyzed by a recombinase of the resolvase/invertase family. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 5505–5510.
123. Groth, A.C.; Fish, M.; Nusse, R.; Calos, M.P. Construction of transgenic drosophila by using the site-specific integrase from phage phic31. *Genetics* **2004**, *166*, 1775–1782.
124. Venken, K.J.; He, Y.; Hoskins, R.A.; Bellen, H.J. P[acman]: A bac transgenic platform for targeted insertion of large DNA fragments in d. Melanogaster. *Science* **2006**, *314*, 1747–1751.
125. Bischof, J.; Maeda, R.K.; Hediger, M.; Karch, F.; Basler, K. An optimized transgenesis system for drosophila using germ-line-specific phic31 integrases. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 3312–3317.
126. Markstein, M.; Pitsouli, C.; Villalta, C.; Celniker, S.E.; Perrimon, N. Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. *Nat. Genet.* **2008**, *40*, 476–483.
127. Brand, A.H.; Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **1993**, *118*, 401–415.
128. Manseau, L.; Baradaran, A.; Brower, D.; Budhu, A.; Elefant, F.; Phan, H.; Philp, A.V.; Yang, M.; Glover, D.; Kaiser, K.; *et al*. Gal4 enhancer traps expressed in the embryo, larval brain, imaginal discs, and ovary of drosophila. *Dev. Dyn.* **1997**, *209*, 310–322.
129. Hayashi, S.; Ito, K.; Sado, Y.; Taniguchi, M.; Akimoto, A.; Takeuchi, H.; Aigaki, T.; Matsuzaki, F.; Nakagoshi, H.; Tanimura, T.; *et al*. Getdb, a database compiling expression patterns and molecular locations of a collection of gal4 enhancer traps. *Genesis* **2002**, *34*, 58–61.
130. Horn, C.; Offen, N.; Nystedt, S.; Hacker, U.; Wimmer, E.A. Piggybac-based insertional mutagenesis and enhancer detection as a tool for functional insect genomics. *Genetics* **2003**, *163*, 647–661.
131. Jenett, A.; Rubin, G.M.; Ngo, T.T.; Shepherd, D.; Murphy, C.; Dionne, H.; Pfeiffer, B.D.; Cavallaro, A.; Hall, D.; Jeter, J.; *et al*. A gal4-driver line resource for drosophila neurobiology. *Cell Rep.* **2012**, *2*, 991–1001.

132. Stark, A.; Dickson, B.J. Vt (vienna tile) gal4 Driver Lines. Available online: http://stockcenter.vdrc.at/control/vtlibrary/ (accessed on 10 January 2014).
133. Suster, M.L.; Seugnet, L.; Bate, M.; Sokolowski, M.B. Refining gal4-driven transgene expression in drosophila with a gal80 enhancer-trap. *Genesis* **2004**, *39*, 240–245.
134. McGuire, S.E.; Le, P.T.; Osborn, A.J.; Matsumoto, K.; Davis, R.L. Spatiotemporal rescue of memory dysfunction in drosophila. *Science* **2003**, *302*, 1765–1768.
135. Lee, T.; Luo, L. Mosaic analysis with a repressible cell marker for studies of gene function in neuronal morphogenesis. *Neuron* **1999**, *22*, 451–461.
136. Han, D.D.; Stein, D.; Stevens, L.M. Investigating the function of follicular subpopulations during drosophila oogenesis through hormone-dependent enhancer-targeted cell ablation. *Development* **2000**, *127*, 573–583.
137. Osterwalder, T.; Yoon, K.S.; White, B.H.; Keshishian, H. A conditional tissue-specific transgene expression system using inducible gal4. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 12596–12601.
138. Roman, G.; Endo, K.; Zong, L.; Davis, R.L. P[switch], a system for spatial and temporal control of gene expression in drosophila melanogaster. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 12602–12607.
139. Stebbins, M.J.; Urlinger, S.; Byrne, G.; Bello, B.; Hillen, W.; Yin, J.C. Tetracycline-inducible systems for drosophila. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10775–10780.
140. Lai, S.L.; Lee, T. Genetic mosaic with dual binary transcriptional systems in drosophila. *Nat. Neurosci.* **2006**, *9*, 703–709.
141. Potter, C.J.; Tasic, B.; Russler, E.V.; Liang, L.; Luo, L. The q system: A repressible binary system for transgene expression, lineage tracing, and mosaic analysis. *Cell* **2010**, *141*, 536–548.
142. Yagi, R.; Mayer, F.; Basler, K. Refined lexa transactivators and their use in combination with the drosophila gal4 system. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 16166–16171.
143. Dang, D.T.; Perrimon, N. Use of a yeast site-specific recombinase to generate embryonic mosaics in drosophila. *Dev. Genet.* **1992**, *13*, 367–375.
144. Harrison, D.A.; Perrimon, N. Simple and efficient generation of marked clones in drosophila. *Curr. Biol.* **1993**, *3*, 424–433.
145. Golic, K.G. Site-specific recombination between homologous chromosomes in drosophila. *Science* **1991**, *252*, 958–961.
146. Yu, H.H.; Chen, C.H.; Shi, L.; Huang, Y.; Lee, T. Twin-spot marcm to reveal the developmental origin and identity of neurons. *Nat. Neurosci.* **2009**, *12*, 947–953.
147. Venken, K.J.; Schulze, K.L.; Haelterman, N.A.; Pan, H.; He, Y.; Evans-Holm, M.; Carlson, J.W.; Levis, R.W.; Spradling, A.C.; Hoskins, R.A.; *et al*. Mimic: A highly versatile transposon insertion resource for engineering drosophila melanogaster genes. *Nat. Methods* **2011**, *8*, 737–743.
148. Fire, A.; Xu, S.; Montgomery, M.K.; Kostas, S.A.; Driver, S.E.; Mello, C.C. Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans. *Nature* **1998**, *391*, 806–811.

149. Hannon, G.J. RNA interference. *Nature* **2002**, *418*, 244–251.
150. Kennerdell, J.R.; Carthew, R.W. Use of dsrna-mediated genetic interference to demonstrate that frizzled and frizzled 2 act in the wingless pathway. *Cell* **1998**, *95*, 1017–1026.
151. Misquitta, L.; Paterson, B.M. Targeted disruption of gene function in drosophila by RNA interference (RNA-i): A role for nautilus in embryonic somatic muscle formation. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 1451–1456.
152. Hammond, S.M.; Bernstein, E.; Beach, D.; Hannon, G.J. An RNA-directed nuclease mediates post-transcriptional gene silencing in drosophila cells. *Nature* **2000**, *404*, 293–296.
153. Kennerdell, J.R.; Carthew, R.W. Heritable gene silencing in drosophila using double-stranded RNA. *Nat. Biotechnol.* **2000**, *18*, 896–898.
154. Clemens, J.C.; Worby, C.A.; Simonson-Leff, N.; Muda, M.; Maehama, T.; Hemmings, B.A.; Dixon, J.E. Use of double-stranded RNA interference in drosophila cell lines to dissect signal transduction pathways. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 6499–6503.
155. Kiger, A.A.; Baum, B.; Jones, S.; Jones, M.R.; Coulson, A.; Echeverri, C.; Perrimon, N. A functional genomic analysis of cell morphology using RNA interference. *J. Biol.* **2003**, doi:10.1186/1475-4924-2-27.
156. Lum, L.; Yao, S.; Mozer, B.; Rovescalli, A.; von Kessler, D.; Nirenberg, M.; Beachy, P.A. Identification of hedgehog pathway components by RNAi in drosophila cultured cells. *Science* **2003**, *299*, 2039–2045.
157. Boutros, M.; Kiger, A.A.; Armknecht, S.; Kerr, K.; Hild, M.; Koch, B.; Haas, S.A.; Paro, R.; Perrimon, N.; Heidelberg Fly Array, C. Genome-wide RNAi analysis of growth and viability in drosophila cells. *Science* **2004**, *303*, 832–835.
158. Foley, E.; O'Farrell, P.H. Functional dissection of an innate immune response by a genome-wide RNAi screen. *PLoS Biol.* **2004**, *2*, E203.
159. Dietzl, G.; Chen, D.; Schnorrer, F.; Su, K.-C.; Barinova, Y.; Fellner, M.; Gasser, B.; Kinsey, K.; Oppel, S.; Scheiblauer, S.; *et al*. A genome-wide transgenic RNAi library for conditional gene inactivation in drosophila. *Nature* **2007**, *448*, 151–156.
160. Ueda, R. RNAi Fly—A Comprehensive RNAi-Mutant Fly Bank. Available online: http://www.
shigen.nig.ac.jp/fly/nigfly/about/aboutrnai.jsp/ (accessed on 13 January 2014).
161. Ni, J.-Q.; Liu, L.-P.; Binari, R.; Hardy, R.; Shim, H.-S.; Cavallaro, A.; Booker, M.; Pfeiffer, B.D.; Markstein, M.; Wang, H.; *et al*. A drosophila resource of transgenic RNAi lines for neurogenetics. *Genetics* **2009**, *182*, 1089–1100.
162. Schnorrer, F.; Schonbauer, C.; Langer, C.C.; Dietzl, G.; Novatchkova, M.; Schernhuber, K.; Fellner, M.; Azaryan, A.; Radolf, M.; Stark, A.; *et al*. Systematic genetic analysis of muscle morphogenesis and function in drosophila. *Nature* **2010**, *464*, 287–291.
163. Neely, G.G.; Kuba, K.; Cammarato, A.; Isobe, K.; Amann, S.; Zhang, L.; Murata, M.; Elmen, L.; Gupta, V.; Arora, S.; *et al*. A global *in vivo* drosophila RNAi screen identifies not3 as a conserved regulator of heart function. *Cell* **2010**, *141*, 142–153.

164. Pospisilik, J.A.; Schramek, D.; Schnidar, H.; Cronin, S.J.; Nehme, N.T.; Zhang, X.; Knauf, C.; Cani, P.D.; Aumayr, K.; Todoric, J.; *et al*. Drosophila genome-wide obesity screen reveals hedgehog as a determinant of brown *versus* white adipose cell fate. *Cell* **2010**, *140*, 148–160.
165. Neely, G.G.; Hess, A.; Costigan, M.; Keene, A.C.; Goulas, S.; Langeslag, M.; Griffin, R.S.; Belfer, I.; Dai, F.; Smith, S.B.; *et al*. A genome-wide drosophila screen for heat nociception identifies alpha2delta3 as an evolutionarily conserved pain gene. *Cell* **2010**, *143*, 628–638.
166. Ghosh, A.; Kling, T.; Snaidero, N.; Sampaio, J.L.; Shevchenko, A.; Gras, H.; Geurten, B.; Gopfert, M.C.; Schulz, J.B.; Voigt, A.; *et al*. A global *in vivo* drosophila RNAi screen identifies a key role of ceramide phosphoethanolamine for glial ensheathment of axons. *PLoS Genet.* **2013**, *9*, e1003980.
167. Czech, B.; Preall, J.B.; McGinn, J.; Hannon, G.J. A transcriptome-wide RNAi screen in the drosophila ovary reveals factors of the germline piRNA pathway. *Mol. Cell* **2013**, *50*, 749–761.
168. Kondo, S.; Booker, M.; Perrimon, N. Cross-species RNAi rescue platform in drosophila melanogaster. *Genetics* **2009**, *183*, 1165–1173.
169. Langer, C.C.H.; Ejsmont, R.K.; Schönbauer, C.; Schnorrer, F.; Tomancak, P. *In vivo* RNAi rescue in drosophila melanogaster with genomic transgenes from drosophila pseudoobscura. *PLoS One* **2010**, *5*, e8928.
170. Schulz, J.G.; David, G.; Hassan, B.A. A novel method for tissue-specific RNAi rescue in drosophila. *Nucleic Acids Res.* **2009**, *37*, e93.
171. Lawrence, P.A.; Bodmer, R.; Vincent, J.P. Segmental patterning of heart precursors in drosophila. *Development* **1995**, *121*, 4303–4308.
172. Murray, M.J.; Merritt, D.J.; Brand, A.H.; Whitington, P.M. *In vivo* dynamics of axon pathfinding in the drosophilia CNS: A time-lapse study of an identified motorneuron. *J. Neurobiol.* **1998**, *37*, 607–621.
173. Murphy, K.C. Use of bacteriophage lambda recombination functions to promote gene replacement in escherichia coli. *J. Bacteriol.* **1998**, *180*, 2063–2071.
174. Zhang, Y.; Buchholz, F.; Muyrers, J.P.; Stewart, A.F. A new logic for DNA engineering using recombination in escherichia coli. *Nat. Genet.* **1998**, *20*, 123–128.
175. Muyrers, J.P.; Zhang, Y.; Testa, G.; Stewart, A.F. Rapid modification of bacterial artificial chromosomes by et-recombination. *Nucleic Acids Res.* **1999**, *27*, 1555–1557.
176. Yu, D.; Ellis, H.M.; Lee, E.C.; Jenkins, N.A.; Copeland, N.G.; Court, D.L. An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 5978–5983.
177. Venken, K.J.T.; Carlson, J.W.; Schulze, K.L.; Pan, H.; He, Y.; Spokony, R.; Wan, K.H.; Koriabine, M.; de Jong, P.J.; White, K.P.; *et al*. Versatile P[acman] bac libraries for transgenesis studies in drosophila melanogaster. *Nat. Methods* **2009**, *6*, 431–434.
178. Ejsmont, R.K.; Sarov, M.; Winkler, S.; Lipinski, K.A.; Tomancak, P. A toolkit for high-throughput, cross-species gene engineering in drosophila. *Nat. Methods* **2009**, *6*, 435–437.

179. Horn, C.; Jaunich, B.; Wimmer, E.A. Highly sensitive, fluorescent transformation marker for drosophila transgenesis. *Dev. Genes Evol.* **2000**, *210*, 623–629.
180. Poser, I.; Sarov, M.; Hutchins, J.R.; Heriche, J.K.; Toyoda, Y.; Pozniakovsky, A.; Weigl, D.; Nitzsche, A.; Hegemann, B.; Bird, A.W.; *et al*. Bac transgeneomics: A high-throughput method for exploration of protein function in mammals. *Nat. Methods* **2008**, *5*, 409–415.
181. Wu, J.S.; Luo, L. A protocol for mosaic analysis with a repressible cell marker (Marcm) in drosophila. *Nat. Protoc.* **2006**, *1*, 2583–2589.
182. Rong, Y.S.; Golic, K.G. Gene targeting by homologous recombination in drosophila. *Science* **2000**, *288*, 2013–2018.
183. Gong, W.J.; Golic, K.G. Ends-out, or replacement, gene targeting in drosophila. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 2556–2561.
184. Choi, C.M.; Vilain, S.; Langen, M.; van Kelst, S.; de Geest, N.; Yan, J.; Verstreken, P.; Hassan, B.A. Conditional mutagenesis in drosophila. *Science* **2009**, doi:10.1126/ science.1168275.
185. Huang, J.; Zhou, W.; Dong, W.; Watson, A.M.; Hong, Y. From the cover: Directed, efficient, and versatile modifications of the drosophila genome by genomic engineering. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 8284–8289.
186. Gloor, G.B.; Nassif, N.A.; Johnson-Schlitz, D.M.; Preston, C.R.; Engels, W.R. Targeted gene replacement in drosophila via p element-induced gap repair. *Science* **1991**, *253*, 1110–1117.
187. Kim, Y.G.; Cha, J.; Chandrasegaran, S. Hybrid restriction enzymes: Zinc finger fusions to Fok I cleavage domain. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 1156–1160.
188. Bibikova, M.; Golic, M.; Golic, K.G.; Carroll, D. Targeted chromosomal cleavage and mutagenesis in drosophila using zinc-finger nucleases. *Genetics* **2002**, *161*, 1169–1175.
189. Auer, T.O.; Duroure, K.; de Cian, A.; Concordet, J.P.; Del Bene, F. Highly efficient crispr/cas9-mediated knock-in in zebrafish by homology-independent DNA repair. *Genome Res.* **2014**, *24*, 142–153.
190. Beumer, K.; Bhattacharyya, G.; Bibikova, M.; Trautman, J.K.; Carroll, D. Efficient gene targeting in drosophila with zinc-finger nucleases. *Genetics* **2006**, *172*, 2391–2403.
191. Maeder, M.L.; Thibodeau-Beganny, S.; Osiak, A.; Wright, D.A.; Anthony, R.M.; Eichtinger, M.; Jiang, T.; Foley, J.E.; Winfrey, R.J.; Townsend, J.A.; *et al*. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol. Cell* **2008**, *31*, 294–301.
192. Sander, J.D.; Dahlborg, E.J.; Goodwin, M.J.; Cade, L.; Zhang, F.; Cifuentes, D.; Curtin, S.J.; Blackburn, J.S.; Thibodeau-Beganny, S.; Qi, Y.; *et al*. Selection-free zinc-finger-nuclease engineering by context-dependent assembly (coda). *Nat. Methods* **2011**, *8*, 67–69.
193. Kim, S.; Lee, M.J.; Kim, H.; Kang, M.; Kim, J.S. Preassembled zinc-finger arrays for rapid construction of zfns. *Nat. Methods* **2011**, doi:10.1038/nmeth0111-7a.
194. Doyon, Y.; Vo, T.D.; Mendel, M.C.; Greenberg, S.G.; Wang, J.; Xia, D.F.; Miller, J.C.; Urnov, F.D.; Gregory, P.D.; Holmes, M.C. Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures. *Nat. Methods* **2011**, *8*, 74–79.

195. Mahfouz, M.M.; Li, L.; Shamimuzzaman, M.; Wibowo, A.; Fang, X.; Zhu, J.K. *De novo*-engineered transcription activator-like effector (tale) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 2623–2628.
196. Miller, J.C.; Tan, S.; Qiao, G.; Barlow, K.A.; Wang, J.; Xia, D.F.; Meng, X.; Paschon, D.E.; Leung, E.; Hinkley, S.J.; *et al*. A tale nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **2011**, *29*, 143–148.
197. Morbitzer, R.; Elsaesser, J.; Hausner, J.; Lahaye, T. Assembly of custom tale-type DNA binding domains by modular cloning. *Nucleic Acids Res.* **2011**, *39*, 5790–5799.
198. Moscou, M.J.; Bogdanove, A.J. A simple cipher governs DNA recognition by tal effectors. *Science* **2009**, doi:10.1126/science.1178817.
199. Brouns, S.J.; Jore, M.M.; Lundgren, M.; Westra, E.R.; Slijkhuis, R.J.; Snijders, A.P.; Dickman, M.J.; Makarova, K.S.; Koonin, E.V.; van der Oost, J. Small crispr rnas guide antiviral defense in prokaryotes. *Science* **2008**, *321*, 960–964.
200. Gasiunas, G.; Barrangou, R.; Horvath, P.; Siksnys, V. Cas9-crrna ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E2579–E2586.
201. Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J.A.; Charpentier, E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **2012**, *337*, 816–821.
202. Bassett, A.R.; Tibbit, C.; Ponting, C.P.; Liu, J.L. Highly efficient targeted mutagenesis of drosophila with the crispr/cas9 system. *Cell Rep.* **2013**, *4*, 220–228.
203. Gratz, S.J.; Cummings, A.M.; Nguyen, J.N.; Hamm, D.C.; Donohue, L.K.; Harrison, M.M.; Wildonger, J.; O'Connor-Giles, K.M. Genome engineering of drosophila with the crispr RNA-guided cas9 nuclease. *Genetics* **2013**, *194*, 1029–1035.
204. Yu, Z.; Ren, M.; Wang, Z.; Zhang, B.; Rong, Y.S.; Jiao, R.; Gao, G. Highly efficient genome modifications mediated by crispr/cas9 in drosophila. *Genetics* **2013**, *195*, 289–291.
205. Sebo, Z.L.; Lee, H.B.; Peng, Y.; Guo, Y. A simplified and efficient germline-specific crispr/cas9 system for drosophila genomic engineering. *Fly* **2014**, *8*, 52–57.
206. Ren, X.; Sun, J.; Housden, B.E.; Hu, Y.; Roesel, C.; Lin, S.; Liu, L.-P.; Yang, Z.; Mao, D.; Sun, L.; *et al*. Optimized gene editing technology for drosophila melanogaster using germ line-specific cas9. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 19012–19017.
207. Gratz, S.J.; Ukken, F.P.; Rubinstein, C.D.; Thiede, G.; Donohue, L.K.; Cummings, A.M.; O'Connor-Giles, K.M. Highly specific and efficient crispr/cas9-catalyzed homology-directed repair in drosophila. *Genetics* **2014**, doi:10.1534/genetics.113.160713.
208. Kondo, S.; Ueda, R. Highly improved gene targeting by germline-specific cas9 expression in drosophila. *Genetics* **2013**, *195*, 715–721.
209. Baena-Lopez, L.A.; Alexandre, C.; Mitchell, A.; Pasakarnis, L.; Vincent, J.P. Accelerated homologous recombination and subsequent genome modification in drosophila. *Development* **2013**, *140*, 4818–4825.

210. Del Valle Rodriguez, A.; Didiano, D.; Desplan, C. Power tools for gene expression and clonal analysis in drosophila. *Nat. Methods* **2012**, *9*, 47–55.
211. Venken, K.J.; Bellen, H.J. Genome-wide manipulations of drosophila melanogaster with transposons, flp recombinase, and phic31 integrase. *Methods Mol. Biol.* **2012**, *859*, 203–228.
212. Gratz, S.J.; Wildonger, J.; Harrison, M.M.; O'Connor-Giles, K.M. Crispr/cas9-mediated genome engineering and the promise of designer flies on demand. *Fly* **2013**, doi:10.4161/fly.26566.
213. Reiter, L.T.; Potocki, L.; Chien, S.; Gribskov, M.; Bier, E. A systematic analysis of human disease-associated gene sequences in drosophila melanogaster. *Genome Res.* **2001**, *11*, 1114–1125.
214. Chien, S.; Reiter, L.T.; Bier, E.; Gribskov, M. Homophila: Human disease gene cognates in drosophila. *Nucleic Acids Res.* **2002**, *30*, 149–151.
215. Richter, A.; Boch, J. Designer tales team up for highly efficient gene induction. *Nat. Methods* **2013**, *10*, 207–208.
216. Crocker, J.; Stern, D.L. Tale-mediated modulation of transcriptional enhancers *in vivo*. *Nat. Methods* **2013**, *10*, 762–767.
217. Mali, P.; Esvelt, K.M.; Church, G.M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **2013**, *10*, 957–963.
218. modENCODE Consortium Modencode Comparative Genomics Whitepaper. Available online: http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/modENCODE_ComparativeGenomics_WhitePaper.pdf (accessed on 16 January 2014).

Genome-Wide Analysis of Alpharetroviral Integration in Human Hematopoietic Stem/Progenitor Cells

Arianna Moiani, Julia Debora Suerth, Francesco Gandolfi, Ermanno Rizzi, Marco Severgnini, Gianluca De Bellis, Axel Schambach and Fulvio Mavilio

Abstract: Gene transfer vectors derived from gamma-retroviruses or lentiviruses are currently used for the gene therapy of genetic or acquired diseases. Retroviral vectors display a non-random integration pattern in the human genome, targeting either regulatory regions (gamma-retroviruses) or the transcribed portion of expressed genes (lentiviruses), and have the potential to deregulate gene expression at the transcriptional or post-transcriptional level. A recently developed alternative vector system derives from the avian sarcoma-leukosis alpha-retrovirus (ASLV) and shows favorable safety features compared to both gamma-retroviral and lentiviral vectors in preclinical models. We performed a high-throughput analysis of the integration pattern of self-inactivating (SIN) alpha-retroviral vectors in human $CD34^+$ hematopoietic stem/progenitor cells (HSPCs) and compared it to previously reported gamma-retroviral and lentiviral vectors integration profiles obtained in the same experimental setting. Compared to gamma-retroviral and lentiviral vectors, the SIN-ASLV vector maintains a preference for open chromatin regions, but shows no bias for transcriptional regulatory elements or transcription units, as defined by genomic annotations and epigenetic markers (H3K4me1 and H3K4me3 histone modifications). Importantly, SIN-ASLV integrations do not cluster in hot spots and target potentially dangerous genomic loci, such as the EVI2A/B, RUNX1 and LMO2 proto-oncogenes at a virtually random frequency. These characteristics predict a safer profile for ASLV-derived vectors for clinical applications.

Reprinted from *Genes*. Cite as: Moiani, A.; Suerth, J.D.; Gandolfi, F.; Rizzi, E.; Severgnini, M.; de Bellis, G.; Schambach, A.; Mavilio, F. Genome-Wide Analysis of Alpharetroviral Integration in Human Hematopoietic Stem/Progenitor Cells. *Genes* **2014**, *5*, 415-429.

1. Introduction

Transplantation of hematopoietic stem cells genetically modified by retroviral vectors has proven its clinical efficacy in a number of seminal clinical trials for the treatment of severe monogenic disorders [1–8]. However, some of these studies also showed the genotoxic risks associated with the insertion of foreign DNA in the human genome, which limit the clinical application of integrating vectors (reviewed in [9]). Several efforts have been made to improve the safety of retroviral vectors, leading to the design of safer constructs and the development of robust *in vitro* and *in vivo* genotoxic assays to predict the potential risk associated with their integration into the genome [10–12]. High-definition mapping of integration sites of vectors derived from the Moloney murine leukemia virus (MLV) and human immunodeficiency virus (HIV) in murine and human cells revealed non-random profiles with a strong tendency to target active regulatory regions for MLV-derived gamma-retroviral vectors [13,14] and transcribed regions for HIV-derived lentiviral vectors [15,16]. These integration patterns explain the relatively high risk to deregulate gene

expression at the transcriptional or post-transcriptional level observed in pre-clinical, as well as in clinical studies (reviewed in [9]).

Small-scale surveys of integration sites of vectors derived from alpha-retroviruses, such as the avian sarcoma-leukosis virus (ASLV), in different cell types indicated a more random pattern compared to other retroviruses, with a slight preference for transcription units, but no apparent preference for promoters and transcription start sites (TSSs) [17–20]. This potentially more favorable integration profile prompted the development of a replication-deficient, self-inactivating (SIN) ASLV-derived vector capable of efficiently transducing murine and human cells [21]. This vector was able to sustain long-term transgene expression in murine and human hematopoietic progenitors at levels comparable to those obtained with SIN-MLV and SIN-HIV vectors and to correct the X-linked chronic granulomatous disease (X-CGD) phenotype in a mouse model of the disease [20,22].

We and others previously reported that MLV, SIN-MLV and SIN-HIV integrations are highly clustered in the human genome, with cell-specific patterns that correlate with the transcriptional program and the epigenetic landscape of each cell type [14–16,19,23–26]. In this study, we report a high-definition analysis of the integration patterns of SIN-MLV, SIN-ASLV and SIN-HIV vectors in human $CD34^+$ hematopoietic stem/progenitor cells (HSPCs), which was carried out to evaluate their comparative genotoxic potential in a clinically relevant target cell. We show that the SIN-ASLV integration profile is close to random, with no preferential targeting of TSSs or transcribed genes compared to SIN-MLV and SIN-HIV. The SIN-ASLV vector does not target CpG islands, conserved non-coding regions (CNCs) or elements enriched in transcription factor binding sites (TFBS), is less frequently associated with epigenetically defined promoter and enhancer regions compared to SIN-MLV and is randomly associated with repetitive elements in the genome. Similarly, we observed no preference for transcribed regions compared to SIN-HIV. Heterochromatic regions are excluded by the integration pattern of all three vectors. Interestingly, the ASLV vector showed no apparent clustering in the genome and has no association with the typical integration hot spots observed for MLV- and HIV-based vectors. These results highlight a safer integration profile of alpha-retroviral vectors in human cells, supporting their development as a clinical gene transfer tool.

2. Experimental

2.1. Vectors and Cells

Human $CD34^+$ HSPCs were purified form umbilical cord blood, pre-stimulated for 48 h in serum-free Iscove's modified Dulbecco medium supplemented with 20% Fetal Calf Serum (FCS), 20 ng/mL human thrombopoietin, 100 ng/mL Flt-3 ligand, 20 ng/mL interleukin-6 and 100 ng/mL stem cell factor, as previously described [23]. HSPCs were transduced with the SIN-ASLV vector, pAlpha.SIN.EFS.EGFP.WPRE (noTATA), expressing GFP under the control of the elongation factor 1α promoter, pseudotyped in an amphotropic envelope by three-plasmid transfection in 293T cells, as previously described [20]. Cells were infected by 3 rounds of spinoculation (1500 rpm for

45 min) in the presence of 4 μg/mL polybrene. Transduction efficiency was evaluated by cytofluorimetric analysis of GFP expression 48 h after infection.

2.2. Amplification, Sequencing, and Analysis of Retroviral Integration Sites

Genomic DNA was extracted from a pool of 3.5×10^6 $CD34^+/GFP^+$ cells enriched by fluorescence-activated cell sorting, after a brief period in culture to dilute unintegrated vectors. 3'-LTR vector-genome junctions were amplified by LM-PCR adapted to the GS-FLX Genome Sequencer (Roche/454 Life Sciences) pyrosequencing platform, as previously described [14]. Raw sequence reads were processed by an automated bioinformatic pipeline that eliminated small and redundant sequences [14] and mapped on the University of California at Santa Cruz (UCSC) hg19 release of the human genome [14]. All UCSC RefSeq genes having their TSS at ±50 kb from an integration site were annotated as targets. Genomic features were annotated when their genomic coordinates overlapped for ≥1 nucleotide with a ±1 kb interval around each integration site. We used UCSC tracks for both CpG islands and conserved TFBSs, and the previously described genomic coordinates of 82,335 mammalian conserved non-coding sequences (CNCs) [27]. Raw sequences having a single or ambiguous match in the genome (the latter mapping in multiple genomic positions with a difference in the identity <2) were blasted on the UCSC RepeatMasker database. DNase I hypersensitive sites from publicly available data [28] were annotated when overlapping for at least 1 bp with a ±1-kb interval around an integration. Repetitive elements were annotated when directly targeted by each integration site. Sequences having multiple matches were collapsed and counted as one when matching in the same genomic positions and were univocally associated with the single type of repetitive element they targeted.

For the association of the integrations with epigenetically defined chromatin states, we used publicly available ChIP-Seq data (NIH Roadmap Epigenomics Mapping Consortium database) that we re-annotated in the UCSC hg19 release of the human genome. We analyzed the distribution of integration sites around histone modifications (H3K4me1, H3K4me3, H3K36me3, H3K27me3) using the seqMINER platform [29]. Previously generated SIN-MLV, SIN-HIV integrations and random control sequences datasets [14] were also re-annotated on the UCSC hg19 genome. For all pairwise comparisons, we applied a 2-sided Fisher's exact test. The threshold for statistical significance was set at a p-value < 0.01.

3. Results and Discussion

3.1. SIN-ASLV Vectors Exhibit an Almost Random Integration Profile in the Genome of Human $CD34^+$ HSPCs

To generate a high-definition alpha-retroviral integration profile in human HSPCs, we transduced umbilical cord blood-derived $CD34^+$ cells with a previously described SIN-ASLV vector carrying a GFP expression cassette under the control of the intron-less, 240-bp version of the elongation factor-1α (EFS) promoter [20]. Cells were transduced at 10% to 20% efficiency and were selected for GFP expression by cell sorting 10 days after infection, to dilute unintegrated vectors. Vector-genome junctions were amplified from genomic DNA by ligation-mediated

(LM)-PCR and pyrosequenced, as previously described [14]. Raw sequences (available at GenBank with the accession number SRR1282019) were processed by a previously described bioinformatic pipeline [14] and mapped on the UCSC hg19 release of the human genome, to obtain 8250 unique insertion sites. Two datasets of SIN-MLV (13,097) and SIN-HIV (31,827) vector integrations, previously generated in human umbilical cord blood-derived $CD34^+$ cells in comparable experimental conditions, and a set of in-silico generated normalized random sites (40,000) [14,26] were re-annotated on the hg19 genome and used for comparison. To identify differences in the integration preferences of SIN-ASLV compared to SIN-MLV and SIN-HIV in human HSPCs, we first analyzed the distribution of integration sites around RefSeq genes in the human genome: integration was annotated as TSS-proximal when occurring in an interval of ±2.5 kb from the TSS of any RefSeq gene, intragenic when occurring inside a RefSeq gene >2.5 kb from the TSS and intergenic in all other cases.

The high-definition profile of SIN-ASLV integration showed only a modest preference for TSSs (6.97% of the integration sites were annotated as TSS-proximal) compared to SIN-HIV and random sites (3.45% and 3.16%, respectively), which was significantly lower than that observed for the SIN-MLV vector (23.38%, $p < 0.01$). Similarly, SIN-ASLV showed only a slight tendency to integrate into genes (49.48% *vs.* 40.58% of random sites), significantly lower than that observed for SIN-HIV vectors (76.77%, $p < 0.01$). As a consequence, the frequency of SIN-ASLV integration outside transcription units was only slightly lower than random (43.55% *vs.* 56.26%) and significantly higher than those observed for the other two vectors (34.36% and 19.78%, respectively, $p < 0.01$) (Table 1). A plot of the relative distance of SIN-ASLV integration sites in an interval of ±50 kb from any TSS revealed a spread distribution with only a modest accumulation in the ±2.5 kb interval around TSS compared to the SIN-MLV vector. A higher definition map (100-bp intervals) showed the absence of integrations in the basal promoter region, most likely occupied by the RNA PolII basal transcriptional machinery. Integrations of the SIN-HIV vector were under-represented in a much wider interval of ±2.5-kb around the TSS (Figure 1).

Table 1. Integration distribution around RefSeq genes and genomic features in the genome of human hematopoietic stem/progenitor cells (HSPCs).

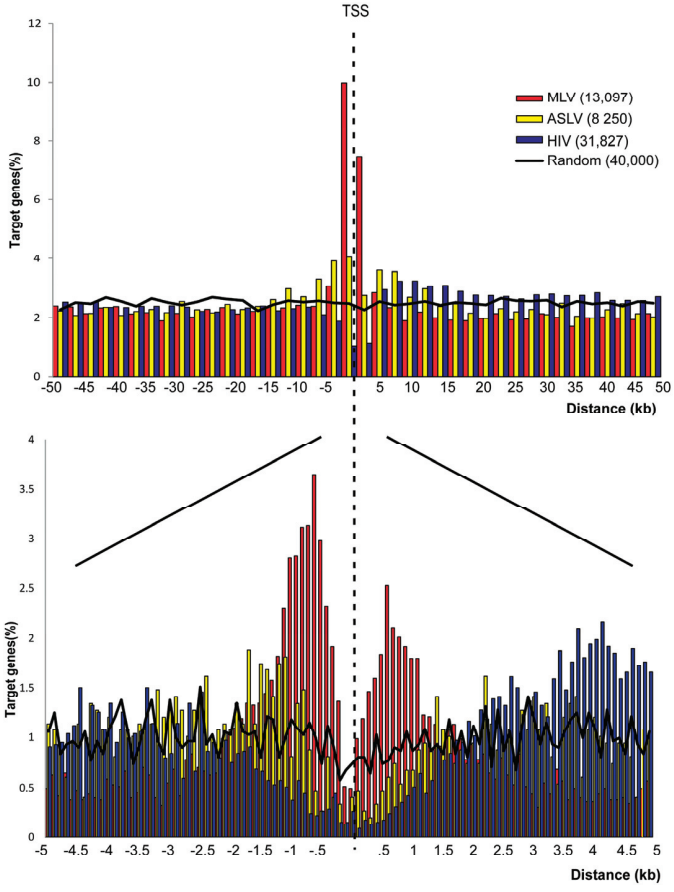
Vector	Intergenic (%)	TSS-proximal (%)	Intragenic (%)	CpG islands (%)	CNCs (%)	TFBS (%)	Total integrations
SIN-ASLV	43.55 *	6.97	49.48	2.84	5.49	55.70	8250
SIN-MLV	34.36 *	23.38 *	42.26	17.68 *	8.42	69.95 *	13,097
SIN-HIV	19.78 *	3.45	76.77 *	1.23	4.58	54.61	31,827
Random	56.26	3.16	40.58	1.76	6.05	51.01	40,000

Percentage of self-inactivating (SIN)-Moloney murine leukemia virus (MLV), SIN-avian sarcoma-leukosis alpha-retrovirus (ASLV) and SIN-HIV integrations and random sequences targeting intergenic, transcription start sites (TSS)-proximal and intragenic regions, regions annotated as CpG islands, conserved non-coding (CNC) regions and transcription factor binding sites (TFBS). For all the comparison with random sites, we applied a two-sided Fisher's exact test. * $p < 0.01$.

This analysis indicates that SIN-ASLV vector integrations have an almost random distribution in the human genome, with only a modest preference for genes and promoter regions compared to SIN-HIV and SIN-MLV vectors, suggesting entirely different modalities of target site selection.

We previously reported that SIN-MLV integrations are enriched around annotated CpG islands and conserved TFBSs and moderately enriched around mammalian, evolutionarily conserved non-coding sequences (CNCs) [14,25,26]. SIN-ASLV integrations were found associated with these genomic features at almost a random frequency, as observed for SIN-HIV integrations, and at a much lower frequency compared to SIN-MLV integrations (CpGs: 2.84% *vs.* 17.68%; TFBSs: 55.70% *vs.* 69.95%; CNCs: 5.49% *vs.* 8.42%, $p < 0.01$ in all cases) (Table 1), suggesting again that SIN-ASLV integrations have no obvious association with functional genomic elements.

Figure 1. Genomic distribution of SIN-MLV, SIN-ASLV and SIN-HIV integrations in human HSPCs. The distribution of the distance of SIN-MLV (red bars), SIN-ASLV (yellow bars) and SIN-HIV (blue bars) integration sites from the TSS of targeted genes at 2500-bp (**a**) or 50-bp (**b**) resolution. The percentage of genes targeted at each position is plotted on the y-axis. The black line indicates the distribution of random control sites.



We then looked at the tendency of the three types of retroviral vectors to target repetitive elements, by blasting both single- and multiple-match sequences to the UCSC RepeatMasker

database and by annotating repetitive elements directly targeted by each integration site. Interestingly, only SIN-ASLV integrations were associated with repetitive elements with an almost random frequency (51% *vs.* 50%), while both SIN-MLV and SIN-HIV integrations were significantly under-represented in repetitive regions (37% and 45%, respectively, $p < 0.01$) (Table 2). By looking at the different classes of repetitive elements, we found that all three vectors have a slightly higher preference to integrate in short interspersed nuclear elements (SINEs) compared to random controls (17% to 20% *vs.* 15%), probably as a consequence of the fact that SINEs are often located in transcribed regions and contain PolII promoters [30,31]. On the contrary, integrations in long interspersed nuclear elements (LINEs), long terminal repeats (LTRs) and other repetitive elements were under-represented or close to random (Table 2). Finally, integration in satellite elements was observed at a random frequency only for SIN-ASLV vectors (0.43% *vs.* 0.35%), while both SIN-MLV and SIN-HIV integrations were significantly under-represented in these regions (0.01% and 0.05%, respectively, $p < 0.01$) (Table 2).

Table 2. Integrations targeting repetitive elements in the genome.

Vector	Repetitive elements (%)	LINEs (%)	SINEs (%)	Satellites (%)	LTRs (%)	Others (%)
SIN-ASLV (8,899)	51.29	19.50	19.77	0.43	6.26	5.35
SIN-MLV (13,606)	37.75	10.96	17.26	0.01	5.09	4.42
SIN-HIV (32,964)	45.78	19.45	16.86	0.05	4.08	5.35
Random (40,000)	50.96	21.27	14.69	0.35	9.57	5.10

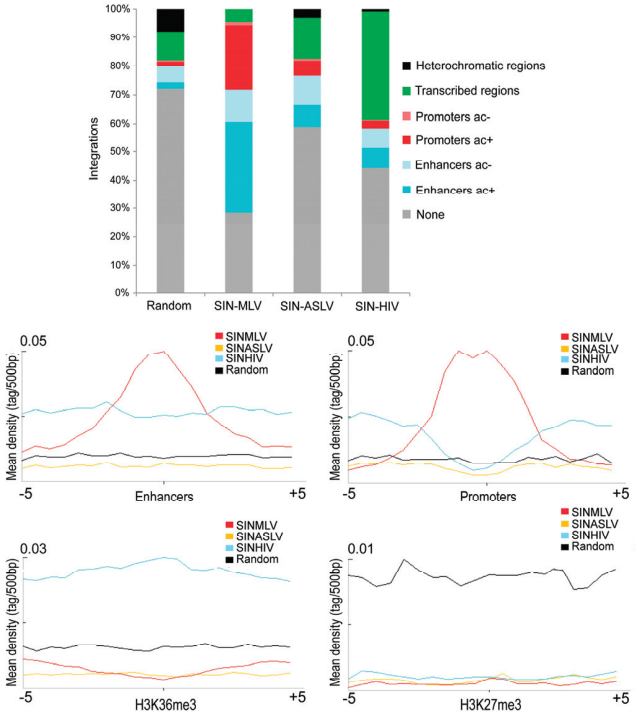
Percentage of SIN-MLV, SIN-ASLV, SIN-HIV integrations and random sequences targeting repetitive elements and the percentage targeting each specific element: LINEs, short interspersed nuclear elements (SINEs), satellites, LTRs and all the other elements.

Overall, these data indicate a remarkably random pattern of integration for the ASLV-derived vector, which shows none of the characteristic preferences of gamma-retroviruses and lentiviruses for genes and genetic elements associated with gene function and regulation.

3.2. SIN-ASLV Integration Is Not Associated with Epigenetically-Defined Functional Genomic Regions

Many studies have reported a strong correlation between MLV and HIV integration sites and distinct epigenetic markers in different cell types (reviewed in [9]). In human $CD34^+$ HSPCs, MLV integrations are strongly associated with histone modifications marking transcriptionally active PolII promoters and enhancers, while HIV integrations correlate with epigenetic markers of active PolII elongation within transcription units [14,32]. We therefore investigated the association of SIN-ASLV integrations with defined epigenetic markers of functional genomic elements. Taking advantage of publicly available ChIP-Seq data in the genome of human $CD34^+$ HSPCs, we analyzed the association of SIN-MLV, SIN-ASLV and SIN-HIV integrations with specific histone modifications defining active or poised PolII promoters, enhancers, transcribed regions and heterochromatin (H3K4me1, H3K4me3, H3K36me3, H3K27me3 and H3K27ac).

Figure 2. Association of vector integration sites with different epigenetically-defined chromatin states. (**a**) The percentage of integration sites associated with specific, epigenetically defined genomic regions for each vector type. Chromatin states are categorized on the basis of the combination of different epigenetic marks mapped by ChIP-seq in human HSPCs. Only integration sites that are unambiguously associated with one chromatin state were used for the analysis. (**b**) The mean densities of H3K4me1, H3K4me3, H3K36me3 and H3K27me3 ChIP-seq fragments in a 5-kb window around all SIN-MLV (red), SIN-ASLV (yellow) and SIN-HIV (light blue) integration sites and random sequences (black). ac: H3K27ac.



More than 60% and 70% of SIN-MLV and SIN-HIV integrations sites, respectively, were univocally associated with a defined chromatin state, compared to only 40% of the SIN-ASLV integration sites, a frequency very close to the 30% observed for random sequences. In particular, SIN-ASLV integrations were found around regulatory regions, *i.e.*, enhancer (H3K4me1^{+}) and promoters (H3K4me3^{+}), at a much lower frequency compared to SIN-MLV (10% *vs.* 37% in enhancers and 6% *vs.* 26% in promoters, respectively, $p < 0.01$), a tendency comparable to that observed for SIN-HIV (10% in enhancers and 4% in promoters) and slightly higher than that of the random sample (3% and 2%, respectively). Moreover, SIN-ASLV integrations were poorly associated with a marker of transcribed gene bodies (H3K36me3) compared to SIN-HIV (15% *vs.* 38%, $p < 0.01$). All three vectors were under-represented in heterochromatic regions marked by H3K27me3 compared to random sites, with the SIN-ASLV vector showing the highest association

(Figure 2A). This analysis is in agreement with the associations observed at the level of DNA sequence and genomic annotations, and confirms the preference of SIN-MLV and SIN-HIV vectors for, respectively, regulatory sequences and transcribed regions and an almost random integration pattern for SIN-ASLV. The modest bias observed for SIN-ASLV integrations in DNase I hypersensitive regions compared to the random sample (Table S1) can be explained by a certain tendency to integrate in "open" chromatin regions, as observed for most retroviruses [14,16,33,34].

The differences between SIN-ASLV and the other two vectors in targeting defined chromatin regions are highlighted by plotting the average integration densities of each vector type around each histone modification. Indeed, we clearly observe a peak of SIN-MLV integration sites in a ±2.5-kb interval from epigenetically-defined enhancers and promoters, while the distribution of the SIN-ASLV and SIN-HIV integrations around these elements is similar to that observed for random sequences (Figure 2B). The quasi-random association of ASLV integrations in regulatory element predicts a much lower genotoxic risk compared to MLV-derived vectors, whose tendency to target active regulatory elements is at the basis of their propensity to cause insertional deregulation of gene expression [9]. Most (>70%) of the genes targeted by all three vectors in HSPCs are actively expressed (Figure S1), an expected finding, considering that retroviral target site selection is highly favored by an open chromatin state [14,16,33,34]. However, the SIN-ASLV vector targets the transcribed portion of active genes at a much lower frequency compared to HIV and is devoid of splicing signals, thus predicting a much lower risk to interfere with gene regulation at the post-transcriptional level [22,35,36].

3.3. SIN-ASLV Vector Integrations Are Not Clustered in the Human Genome

The integration profile of MLV- and HIV-derived vectors in the human genome is characterized by heavy clustering into integration hot spots, where MLV forms narrow clusters overlapping active regulatory elements and HIV larger clusters targeting a subset of transcribed genes, in both cases in a cell-specific fashion [14–16,24,37]. On the contrary, the SIN-ASLV vector showed no significant clustering when we applied a statistical definition of clusters adjusted to the numerosity of the sample [14], which for the SIN-ASLV dataset was three integrations in 53,920 bp. By this threshold, we were able to identify only 484 clusters, a significantly lower frequency compared to the 1,415 and 2,724 identified for the SIN-MLV and SIN-HIV vectors, respectively ($p < 0.01$). Only 21% of all SIN-ASLV integrations are clustered, compared to 56% and 51% of SIN-MLV and SIN-HIV integrations, respectively ($p < 0.01$) (Table 3). Moreover, SIN-ASLV clusters are mostly made of few (three or four) integrations with few clusters containing up to nine integrations, while SIN-MLV and SIN-HIV clusters contain up to 37 and 122 integrations, respectively. From these data, it appears that, contrary to other retroviral vectors, SIN-ASLV integrations do not form hot spots of integrations in the human genome. Interestingly, when we looked at integration clusters at single genomic loci, we observed that the frequency of SIN-ASLV integrations at the typical MLV or HIV hot spots is very low and comparable to the frequency observed for random sequences. Figure 3 shows a comparison of the integration pattern of the three vectors in the NF1-EVI2A/B, RUNX1, LMO2 and PACS1 loci, four known hot spots for MLV or HIV integration. The same scenario is true for the EVI1/MDS1 (MECOM) locus and all other MLV or

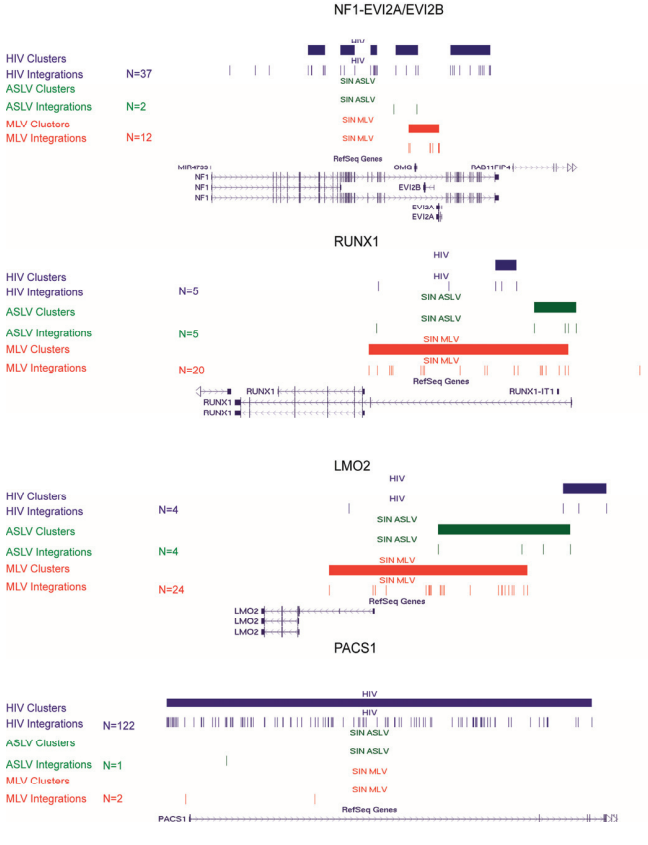
HIV integration hot spots (data not shown). The low frequency of SIN-ASLV integration in proto-oncogenes responsible for the severe adverse events observed in clinical gene therapy trials, such as LMO2, provide a further indication of its lower genotoxic profile.

Table 3. Clusters of integration sites in the genome of human HSPCs.

	SIN-MLV (13,097)	SIN-ASLV (8250)	SIN-HIV (31,827)
Clusters	1415	484	2724
Integrations in clusters (%)	56	21	51
Average cluster dimension	5.1	3.6	5.9

The number of SIN-MLV, SIN-ASLV and SIN-HIV clusters of integrations, the percentage of integrations in clusters and the average cluster dimension, calculated based on random sequences distribution in the genome. The threshold for cluster definition was defined at a *p*-value of <0.01 by a statistical algorithm that adjusts for the numerosity of the sample [14].

Figure 3. SIN-MLV, SIN-ASLV and SIN-HIV integration sites and clusters in CD34$^+$ HSPC-specific loci. Distribution of SIN-MLV (red), SIN-ASLV (green) and SIN-HIV (blue) integration clusters (horizontal solid bars) and integrations (vertical marks) in the NF1-EVI2A/B, RUNX1, LMO2 and PACS1 loci, as displayed by the UCSC Genome Browser.



Although the SIN-ASLV integration profile shows none of the features typical of MLV- and HIV-derived vectors, it is not completely random and shows a general preference for euchromatic regions. It is now know that both MLV and HIV pre-integration complexes (PICs) are targeted to chromatin by a tethering mechanism involving the interaction of the viral integrase with host cell factors: the LEDGF/p75 chromatin component interacts with the HIV integrase and directs its integration into transcribed gene bodies [38,39], while the MLV integrase appears to bind to bromodomain-containing BET proteins specifically associated with acetylated histones around TSSs and active regulatory elements [40–42]. Although it is likely that ASLV also may adopt a tethering mechanism to direct its integration in favorable genomic regions, the details are unknown. The integration preferences uncovered by our analysis predict an interaction with a broader range of host cell factors, which tether the PICs to open chromatin regions with unspecified or very subtle functional characteristics, thus leading to a more random profile characterized by the absence of hot spots.

4. Conclusions

Overcoming the genotoxic consequences of retroviral vector integration in the host cell genome is one of the major issues for the application of retroviral-based gene transfer in clinical trials. The strong preference to target TSSs, active regulatory elements or transcribed genes, together with the high frequency of clustering around hot spots, is a characteristics shared by all retroviral vectors currently used in clinical gene therapy. These characteristics are at the basis of the potential of retroviral insertion to deregulate gene expression at the transcriptional or post-transcriptional level, which has been observed to cause clonal expansion and contribute to neoplastic transformation in a number of cases (reviewed by [9]). Many efforts are being made to improve the safety of currently available retroviral vectors by removing the viral transcriptional control element and avoiding dominant, long-range acting enhancers in the transgene expression cassette. Retargeting vector integration has proven more difficult and was so far unsuccessful. No strategy is obviously perfect, and even a completely random integration machinery would not abolish the risk of inducing an insertional oncogenic mutation in the host cell genome.

Based on a genome-wide analysis of >8000 integration sites in human HSPCs, we show that a SIN-ASLV vector has a quasi-random integration pattern that privileges active chromatin regions, but is not associated with active regulatory elements, like MLV, or with transcribed genes, like HIV. More importantly, the SIN-ASLV vector showed no integration hot spots and no preferences for subsets of genes with a defined ontology or genes that were previously identified as being activated by retroviral insertion into tumors. Previous evaluations of ASLV-derived vectors in pre-clinical models proved its ability to sustain long-term transgene expression in murine and human hematopoietic progenitors and to correct the pathology in a mouse model of X-linked Chronic Granulomatous Disease (X-CGD), with no evidence of post-transcriptional interference [20,22]. Combined with the use of short-range or cell-specific transcriptional regulatory elements, an ASLV vector appears to offer a very safe profile and to be an ideal candidate for *ex vivo* gene therapy applications.

Acknowledgments

This work was supported by grants from the European Research Council and the Italian National Research Council (EPIGEN). Ermanno Rizzi is financially supported by the Italian Research Ministry grants “Futuro in Ricerca” RBFR08U07M_003 and RBFR126B8I_003.

Author Contributions

Conceived and designed the experiments: Arianna Moiani, Fulvio Mavilio. Performed the experiments: Arianna Moiani, Ermanno Rizzi, and Julia Debora Suerth. Analyzed the data: Arianna Moiani, Francesco Gandolfi, Marco Severgnini, and Gianluca De Bellis. Contributed reagents/materials/analysis tools: Julia Debora Suerth and Axel Schambach. Wrote the paper: Arianna Moiani and Fulvio Mavilio.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Hacein-Bey-Abina, S.; Le Deist, F.; Carlier, F.; Bouneaud, C.; Hue, C.; de Villartay, J.P.; Thrasher, A.J.; Wulffraat, N.; Sorensen, R.; Dupuis-Girod, S.; *et al.* Sustained correction of X-linked severe combined immunodeficiency by *ex vivo* gene therapy. *N. Engl. J. Med.* **2002**, *346*, 1185–1193.
2. Ott, M.G.; Schmidt, M.; Schwarzwaelder, K.; Stein, S.; Siler, U.; Koehl, U.; Glimm, H.; Kuhlcke, K.; Schilz, A.; Kunkel, H.; *et al.* Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRMD16 or SETBP1. *Nat. Med.* **2006**, *12*, 401–409.
3. Mavilio, F.; Pellegrini, G.; Ferrari, S.; di Nunzio, F.; di Iorio, E.; Recchia, A.; Maruggi, G.; Ferrari, G.; Provasi, E.; Bonini, C.; *et al.* Correction of junctional epidermolysis bullosa by transplantation of genetically modified epidermal stem cells. *Nat. Med.* **2006**, *12*, 1397–1402.
4. Aiuti, A.; Cattaneo, F.; Galimberti, S.; Benninghoff, U.; Cassani, B.; Callegaro, L.; Scaramuzza, S.; Andolfi, G.; Mirolo, M.; Brigida, I.; *et al.* Gene therapy for immunodeficiency due to adenosine deaminase deficiency. *N. Engl. J. Med.* **2009**, *360*, 447–458.
5. Cartier, N.; Hacein-Bey-Abina, S.; Bartholomae, C.C.; Veres, G.; Schmidt, M.; Kutschera, I.; Vidaud, M.; Abel, U.; Dal-Cortivo, L.; Caccavelli, L.; *et al.* Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science* **2009**, *326*, 818–823.
6. Boztug, K.; Schmidt, M.; Schwarzer, A.; Banerjee, P.P.; Diez, I.A.; Dewey, R.A.; Bohm, M.; Nowrouzi, A.; Ball, C.R.; Glimm, H.; *et al.* Stem-cell gene therapy for the Wiskott-Aldrich syndrome. *N. Engl. J. Med.* **2010**, *363*, 1918–1927.

7. Biffi, A.; Montini, E.; Lorioli, L.; Cesani, M.; Fumagalli, F.; Plati, T.; Baldoli, C.; Martino, S.; Calabria, A.; Canale, S.; *et al.* Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. *Science* **2013**, doi:10.1126/science.1233158.
8. Aiuti, A.; Biasco, L.; Scaramuzza, S.; Ferrua, F.; Cicalese, M.P.; Baricordi, C.; Dionisio, F.; Calabria, A.; Giannelli, S.; Castiello, M.C.; *et al.* Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. *Science* **2013**, doi:10.1126/science.1233151.
9. Cavazza, A.; Moiani, A.; Mavilio, F. Mechanisms of retroviral integration and mutagenesis. *Human Gene Ther.* **2013**, *24*, 119–131.
10. Montini, E.; Cesana, D.; Schmidt, M.; Sanvito, F.; Ponzoni, M.; Bartholomae, C.; Sergi Sergi, L.; Benedicenti, F.; Ambrosi, A.; di Serio, C.; *et al.* Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. *Nat. Biotechnol.* **2006**, *24*, 687–696.
11. Modlich, U.; Navarro, S.; Zychlinski, D.; Maetzig, T.; Knoess, S.; Brugman, M.H.; Schambach, A.; Charrier, S.; Galy, A.; Thrasher, A.J.; *et al.* Insertional transformation of hematopoietic cells by self-inactivating lentiviral and gammaretroviral vectors. *Mol. Ther.* **2009**, *17*, 1919–1928.
12. Montini, E.; Cesana, D.; Schmidt, M.; Sanvito, F.; Bartholomae, C.C.; Ranzani, M.; Benedicenti, F.; Sergi, L.S.; Ambrosi, A.; Ponzoni, M.; *et al.* The genotoxic potential of retroviral vectors is strongly modulated by vector design and integration site selection in a mouse model of hsc gene therapy. *J. Clin. Invest.* **2009**, *119*, 964–975.
13. Wu, X.; Li, Y.; Crise, B.; Burgess, S.M. Transcription start regions in the human genome are favored targets for mlv integration. *Science* **2003**, *300*, 1749–1751.
14. Cattoglio, C.; Pellin, D.; Rizzi, E.; Maruggi, G.; Corti, G.; Miselli, F.; Sartori, D.; Guffanti, A.; di Serio, C.; Ambrosi, A.; *et al.* High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. *Blood* **2010**, *116*, 5507–5517.
15. Schroder, A.R.; Shinn, P.; Chen, H.; Berry, C.; Ecker, J.R.; Bushman, F. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **2002**, *110*, 521–529.
16. Wang, G.P.; Ciuffi, A.; Leipzig, J.; Berry, C.C.; Bushman, F.D. HIV integration site selection: Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* **2007**, *17*, 1186–1194.
17. Barr, S.D.; Leipzig, J.; Shinn, P.; Ecker, J.R.; Bushman, F.D. Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. *J. Virol.* **2005**, *79*, 12035–12044.
18. Hu, J.; Renaud, G.; Gomes, T.J.; Ferris, A.; Hendrie, P.C.; Donahue, R.E.; Hughes, S.H.; Wolfsberg, T.G.; Russell, D.W.; Dunbar, C.E. Reduced genotoxicity of avian sarcoma leukosis virus vectors in rhesus long-term repopulating cells compared to standard murine retrovirus vectors. *Mol. Ther.* **2008**, *16*, 1617–1623.
19. Mitchell, R.S.; Beitzel, B.F.; Schroder, A.R.; Shinn, P.; Chen, H.; Berry, C.C.; Ecker, J.R.; Bushman, F.D. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2004**, *2*, e234.

20. Suerth, J.D.; Maetzig, T.; Brugman, M.H.; Heinz, N.; Appelt, J.U.; Kaufmann, K.B.; Schmidt, M.; Grez, M.; Modlich, U.; Baum, C.; *et al.* Alpharetroviral self-inactivating vectors: Long-term transgene expression in murine hematopoietic cells and low genotoxicity. *Mol. Ther.* **2012**, *20*, 1022–1032.
21. Suerth, J.D.; Maetzig, T.; Galla, M.; Baum, C.; Schambach, A. Self-inactivating alpharetroviral vectors with a split-packaging design. *J. Virol.* **2010**, *84*, 6626–6635.
22. Kaufmann, K.B.; Brendel, C.; Suerth, J.D.; Mueller-Kuller, U.; Chen-Wichmann, L.; Schwable, J.; Pahujani, S.; Kunkel, H.; Schambach, A.; Baum, C.; *et al.* Alpharetroviral vector-mediated gene therapy for X-CGD: Functional correction and lack of aberrant splicing. *Mol. Ther.* **2013**, *21*, 648–661.
23. Cattoglio, C.; Facchini, G.; Sartori, D.; Antonelli, A.; Miccio, A.; Cassani, B.; Schmidt, M.; von Kalle, C.; Howe, S.; Thrasher, A.J.; *et al.* Hot spots of retroviral integration in human $CD34^+$ hematopoietic cells. *Blood* **2007**, *110*, 1770–1778.
24. Cattoglio, C.; Maruggi, G.; Bartholomae, C.; Malani, N.; Pellin, D.; Cocchiarella, F.; Magnani, Z.; Ciceri, F.; Ambrosi, A.; von Kalle, C.; *et al.* High-definition mapping of retroviral integration sites defines the fate of allogeneic T cells after donor lymphocyte infusion. *PLoS One* **2010**, *5*, e15688.
25. Cavazza, A.; Cocchiarella, F.; Bartholomae, C.; Schmidt, M.; Pincelli, C.; Larcher, F.; Mavilio, F. Self-inactivating MLV vectors have a reduced genotoxic profile in human epidermal keratinocytes. *Gene Ther.* **2013**, doi:10.1038/gt.2013.18.
26. Moiani, A.; Miccio, A.; Rizzi, E.; Severgnini, M.; Pellin, D.; Suerth, J.D.; Baum, C.; de Bellis, G.; Mavilio, F. Deletion of the LTR enhancer/promoter has no impact on the integration profile of MLV vectors in human hematopoietic progenitors. *PLoS One* **2013**, *8*, e55721.
27. Kim, S.Y.; Pritchard, J.K. Adaptive evolution of conserved noncoding elements in mammals. *PLoS Genet.* **2007**, *3*, 1572–1586.
28. John, S.; Sabo, P.J.; Canfield, T.K.; Lee, K.; Vong, S.; Weaver, M.; Wang, H.; Vierstra, J.; Reynolds, A.P.; Thurman, R.E.; *et al.* Genome-scale mapping of DNase I hypersensitivity. *Curr. Protoc. Mol. Biol.* **2013**, doi:10.1002/0471142727.mb2127s103.
29. Ye, T.; Krebs, A.R.; Choukrallah, M.A.; Keime, C.; Plewniak, F.; Davidson, I.; Tora, L. Seqminer: An integrated chip-seq data interpretation platform. *Nucleic Acids Res.* **2010**, *39*, e35.
30. Ferrigno, O.; Virolle, T.; Djabari, Z.; Ortonne, J.P.; White, R.J.; Aberdam, D. Transposable b2 sine elements can provide mobile rna polymerase ii promoters. *Nat. Genet.* **2001**, *28*, 77–81.
31. Lunyak, V.V.; Prefontaine, G.G.; Nunez, E.; Cramer, T.; Ju, B.G.; Ohgi, K.A.; Hutt, K.; Roy, R.; Garcia-Diaz, A.; Zhu, X.; *et al.* Developmentally regulated activation of a SINE b2 repeat as a domain boundary in organogenesis. *Science* **2007**, *317*, 248–251.

32. Biasco, L.; Ambrosi, A.; Pellin, D.; Bartholomae, C.; Brigida, I.; Roncarolo, M.G.; di Serio, C.; von Kalle, C.; Schmidt, M.; Aiuti, A. Integration profile of retroviral vector in gene therapy treated patients is cell-specific according to gene expression and chromatin conformation of target cell. *EMBO Mol. Med.* **2011**, *3*, 89–101.
33. Berry, C.; Hannenhalli, S.; Leipzig, J.; Bushman, F.D. Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput. Biol.* **2006**, *2*, e157.
34. Lewinski, M.K.; Yamashita, M.; Emerman, M.; Ciuffi, A.; Marshall, H.; Crawford, G.; Collins, F.; Shinn, P.; Leipzig, J.; Hannenhalli, S.; *et al.* Retroviral DNA integration: Viral and cellular determinants of target-site selection. *PLoS Pathog.* **2006**, *2*, e60.
35. Cesana, D.; Sgualdino, J.; Rudilosso, L.; Merella, S.; Naldini, L.; Montini, E. Whole transcriptome characterization of aberrant splicing events induced by lentiviral vector integrations. *J. Clin. Invest.* **2012**, *122*, 1667–1676.
36. Moiani, A.; Paleari, Y.; Sartori, D.; Mezzadra, R.; Miccio, A.; Cattoglio, C.; Cocchiarella, F.; Lidonnici, M.R.; Ferrari, G.; Mavilio, F. Lentiviral vector integration in the human genome induces alternative splicing and generates aberrant transcripts. *J. Clin. Invest.* **2012**, *122*, 1653–1666.
37. Ambrosi, A.; Glad, I.K.; Pellin, D.; Cattoglio, C.; Mavilio, F.; di Serio, C.; Frigessi, A. Estimated comparative integration hotspots identify different behaviors of retroviral gene transfer vectors. *PLoS Comput. Biol.* **2011**, *7*, e1002292.
38. Ciuffi, A.; Llano, M.; Poeschla, E.; Hoffmann, C.; Leipzig, J.; Shinn, P.; Ecker, J.R.; Bushman, F. A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.* **2005**, *11*, 1287–1289.
39. Llano, M.; Vanegas, M.; Fregoso, O.; Saenz, D.; Chung, S.; Peretz, M.; Poeschla, E.M. LEDGF/p75 determines cellular trafficking of diverse lentiviral but not murine oncoretroviral integrase proteins and is a component of functional lentiviral preintegration complexes. *J. Virol.* **2004**, *78*, 9524–9537.
40. De Rijck, J.; de Kogel, C.; Demeulemeester, J.; Vets, S.; El Ashkar, S.; Malani, N.; Bushman, F.D.; Landuyt, B.; Husson, S.J.; Busschots, K.; *et al.* The BET family of proteins targets Moloney murine leukemia virus integration near transcription start sites. *Cell Rep.* **2013**, *5*, 886–894.
41. Gupta, S.S.; Maetzig, T.; Maertens, G.N.; Sharif, A.; Rothe, M.; Weidner-Glunde, M.; Galla, M.; Schambach, A.; Cherepanov, P.; Schulz, T.F. Bromo- and extraterminal domain chromatin regulators serve as cofactors for murine leukemia virus integration. *J. Virol.* **2013**, *87*, 12721–12736.
42. Sharma, A.; Larue, R.C.; Plumb, M.R.; Malani, N.; Male, F.; Slaughter, A.; Kessl, J.J.; Shkriabai, N.; Coward, E.; Aiyer, S.S.; *et al.* BET proteins promote efficient murine leukemia virus integration at transcription start sites. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 12036–12041.

Pharmacogenomics: Current State-of-the-Art

Daniel F. Carr, Ana Alfirevic and Munir Pirmohamed

Abstract: The completion of the human genome project 10 years ago was met with great optimism for improving drug therapy through personalized medicine approaches, with the anticipation that an era of genotype-guided patient prescribing was imminent. To some extent this has come to pass and a number of key pharmacogenomics markers of inter-individual drug response, for both safety and efficacy, have been identified and subsequently been adopted in clinical practice as pre-treatment genetic tests. However, the universal application of genetics in treatment guidance is still a long way off. This review will highlight important pharmacogenomic discoveries which have been facilitated by the human genome project and other milestone projects such as the International HapMap and 1000 genomes, and by the continued development of genotyping and sequencing technologies, including rapid point of care pre-treatment genetic testing. However, there are still many challenges to implementation for the many other reported biomarkers which continue to languish within the discovery phase. As technology advances over the next 10 years, and the costs fall, the field will see larger genetic data sets, including affordable whole genome sequences, which will, it is hoped, improve patient outcomes through better diagnostic, prognostic and predictive biomarkers.

Reprinted from *Genes*. Cite as: Carr, D.F.; Alfirevic, A.; Pirmohamed, M. Pharmacogenomics: Current State-of-the-Art. *Genes* **2014**, *5*, 430-443.

1. Introduction

The field of pharmacogenomics can trace its roots back to significantly earlier than the first draft publication human genome sequence in 2001 [1] and the subsequent completion in 2003. Indeed, the term "Pharmacogenetics" was first coined by Friedrich Vogel in 1959 [2], just 6 years after Watson and Crick's discovery of the structure of DNA [3].

Though significant progress has been made in the field since 2003, it could be argued that pharmacogenomics has failed to live up to expectations. A vast number of discoveries relating to genomic variability and drug response have been made in the last 10 years. The challenge remains to translate these findings into clinical practice for the benefit of the patient.

2. The International HapMap Project

The completion of the first phase of the International Hapmap project [4], a catalogue of common genetic variations within individuals of diverse ethnicities, in 2003, provided a rich data resource which enabled researchers to investigate the association of variants across the human genome with a wide range of clinical phenotypes. Indeed, the data derived from this allowed the creation of SNP arrays whereby researchers could analyze patient genotype for 100,000 s to millions of SNPs at a time. Thus, for the first time, unbiased genetic analysis of clinical phenotype/genotype associations was possible using platforms created for the Genome Wide Association Study (GWAS). More recent advances in genomics have seen the development of next generation sequencing methodologies which do not require *a priori* knowledge of genetic variation.

Table 1. Genetic biomarkers of (**A**) adverse drug reactions and (**B**) inter-individual variability of drug efficacy identified, or confirmed, from genome wide association studies identifying.

A	Year	Drug	Indication	Phenotype	Population	Associated Loci	SNP/Allele	Ref.
	2008	Simvastatin	Hypercholesterolaemia	Myopathy	Caucasian	SLCO1B1	rs4149056 (c.521T>C/*5)	[6]
	2008	Bisphosphonate	Multiple Myeloma	Osteonecrosis of the jaw	Spanish	CYP2C8	rs1934951	[7]
	2008	Ximelagatran	Anticoagulant	Hepatotoxicity	Caucasian	HLA-DRB1	*07 and *02	[8]
	2009	Flucloxacillin	Macrolide antibiotic	Hepatotoxicity	Caucasian	HLA-B	B*57:01	[9]
	2011	Carbamazepine	Epilepsy	Skin Rash/Hypersensitivity	Japanese Causasian	HLA-A	A*31:01	[10,11]
	2013	Allopurinol	Gout	SJS-TEN	Japanese	HLA-B	B*58:01	[12]
B	**Year**	**Drug**	**Indication**	**Phenotype**	**Population**	**Gene Loci**	**Allele**	**Ref.**
	2009	Clopidogrel	Antiplatelet		Amish	CYP2C19	*2	[13]
	2009	Pegylated Interferon	Hepatitis C		Caucasian	*IFNL3* (IL-28B)	rs12979860	[14]
	2009	Warfarin	Anticoagulant		Caucasian	CYP2C9 and VKORC1	*2,*3/c.-1639G>A	[15]

Mining of the National Human Genome Institute (NHGRI) genome wide association study (GWAS) catalogue [5] (accessed on 22nd April 2014) showed that a total of 1885 peer-reviewed articles have been published reporting GWAS findings. However, only 93 (4.9%) are related to inter-individual variability of drug response phenotypes (either safety or efficacy).

Despite the small numbers of studies to date, which have investigated common genetic variant associations with pharmacological phenotypes, a number of important genotype-phenotype associations have been identified in pharmacogenomics using GWAS approaches (Table 1). The interesting phenomenon, when compared with GWAS of complex diseases, is that the effect size for pharmacogenomics phenotypes on the whole seems to be larger than that observed for complex diseases, which allows for (a) smaller sample sizes to be studied, which is more time- and cost-efficient, and (b) some variants to be considered for clinical implementation in terms of use prior to drug prescription, which contrasts with complex diseases, where the relative risks identified are usually below 1.5 and have not been used clinically. Whilst common genetic variation has been shown, in a number of examples to be important factors determining inter-individual variability in drug response, the role of rare, or private, variants is unclear in pharmacogenomics phenotypes, and is an important area for further research.

3. The 1000 Genomes Project

The first findings from the 1000 genomes project were reported in 2010 [16]. This has provided researchers with a population scale map of rare variants to complement and enrich existing knowledge of common variants gained from the HapMap project. With the per-base cost of sequencing using "next generation sequencing (NGS)" platforms continuing to fall, it is likely that studies will further investigate the role of rare genetic variants in defining variation in drug response phenotypes.

A recent study reported the mapping of rare and common variants within 12 Cytochrome P450 genes, thought to be responsible for metabolizing 75% of prescribed drugs [17]. Using whole exome sequence data for 2203 African Americans and 4300 Caucasians, researchers were able to identify novel, potentially deleterious alleles in major drug metabolizing enzymes in 7.6%–11.7% of individuals. The power of NGS technologies to identify rare variants could allow for greater understanding of their contribution to inter-individual variability in drug responses where common variants explain a limited degree of variability. One such example where rare variants may enhance our understanding of variability is the case of warfarin dose prediction. Incorporating CYP2C9*2, *3 and VKORC1 polymorphisms along with clinical confounding factors into a dosing algorithm allows clinicians to predict ~60% [18] of dose variability. However, it is entirely plausible that incorporating the contribution of rare functional genetic variants into such algorithms may in the future allow for even greater accuracy in warfarin dose prediction. Indeed a number of small scale studies and case reports have already identified rare missense variants in the VKORC1 gene in warfarin resistant patients [19–22]. Another example is drug-induced torsades de pointes where at least 10% of cases of may be due to rare mutations in the congenital long QT syndrome genes [23]. At least 23% of Caucasian subjects with drug-induced torsades de pointes carry a variant within 22 congenital arrhythmia genes (which include the 13 congenital long QT syndrome

genes), compared with a background rate of 1.7% in 60 control subjects from the 1000 Genomes CEU data [24]. As greater numbers of NGS-based analyses in pharmacogenetic studies are undertaken, so the contribution of rare variants in other drug safety and efficacy phenotypes will be better understood.

4. Non-Coding RNAs

Large-scale, coordinated, international, research efforts, such as the ENCODE project [25] have expelled the myth that large regions of our human genome are "junk" DNA. Indeed, the presence of non-coding RNAs has altered the scientific community's perception of the central dogma of molecular biology. To date, only a very small number of studies have investigated the application of small non-coding RNA molecules as biomarkers of variable drug response.

However, studies have shown a potential utility for specific microRNAs as markers of both drug-induced liver injury (miR-122 and -192) [26] and severe skin reactions (miR-18a-5p) [27]. Though such biomarkers are still at the discovery stage, further clinical validation may see these and other non-coding RNAs enter clinical practice as early-stage pharmacogenomics predictors of adverse drug reactions.

5. Clinical Utilization of Pharmacogenomics

The biggest challenge that has faced the field of pharmacogenomics in the 10 years since the completion of the human genome project is clearly the application of genetic markers of variable drug response to decision-making in relation to prescribing. Indeed, as recently as 2011 it has been estimated that, of the >150,000 papers reporting claimed biomarkers, less than 100 has made it into clinical utility [28]. This of course refers to all biomarkers, not just pharmacogenomic biomarkers; whether pharmacogenomics biomarkers have been more successful is unclear, and would require formal analysis.

There are currently 121 Food and Drug Administation (FDA) drug labels referring to pharmacogenetic biomarkers of drug safety or efficacy [29]. Only a very small proportion of these drug labels mandate clinicians to test for pharmacogenetic markers (e.g., abacavir and *HLA-B*57:01*; and carbamazepine and HLA-*B*15:02* in Southern Asians). However, a large number of testing guideline position papers have been published by the Clinical Pharmacogenetics Implementation Consortium [30]. The aim of this collaborative effort is to help clinicians understand how genetic tests may be used to guide treatment decisions.

One reason why many pharmacogenetic biomarkers have failed to move from discovery to clinical implementation is that many genotype/phenotype associations fail to be independently replicated. It has been recognized that variability in phenotype definition could contribute to this, particularly in relation to adverse drug reactions. To this end a number of phenotype standardization efforts have been undertaken in recent years. These include drug-induced skin injury [31], liver injury [32] and Torsade de Pointes (long-QT syndrome) [33]. It is hoped that, with studies applying consistent phenotype definitions, future pharmacogenetic studies may identify replicable pharmacogenetic biomarkers that, with sufficient weight of evidence, could find

their way into clinical utility. Another reason for lack of replication is the inability to find replication sample sets, particularly where the phenotype is a rare adverse event. In such cases, functional genomic analyses may reduce false positives, and provide more confidence for implementation because of insights into the mechanism of effect of that biomarker.

6. The Future…

6.1. Point of Care Genetic Testing

As our understanding of the human genome has grown, so the technology with which we can analyze it has developed in terms of speed and fidelity. The polymerase chain reaction, first described by Kary Mullis in 1983 [34], allowed sensitive analysis of DNA and ultimately yielded a number of pharmacogenetic biomarkers in the pre-genome era based on prior knowledge of gene function (candidate gene studies). As with our understanding of the genome, PCR-based technology has advanced significantly and now allows for rapid and accurate genotype detection. In recent years a number of studies have investigated the potential utility of rapid point-of-care (POC) genetic testing.

Two key randomized control trials have recently utilized pre-treatment genetic testing, with POC devices, to guide the use of drugs used in cardiovascular medicine, the anticoagulant coumarin derivatives (e.g., warfarin) [35] and the antiplatelet drug clopidogrel [36]. The EU-PACT study [35] randomized patients to either genotype-guided or standard dosing. The genotype guided group, utilized molecular beacon technology to genotype for the *CYP2C9*2* and *CYP2C9*3* variants and a promoter polymorphism of *VKORC1* prior to the patient receiving warfarin. The results generated were then incorporated into a warfarin dose calculator, incorporating a number of key non-genetic factors affecting dose requirement. The warfarin trial found that patients randomized to pharmacogenetic-guided dosing spent a mean percentage time in the therapeutic range of 67.4% compared with 60.3% on the standard dosing protocol. In this example, the healthcare professional was presented with an individualized, genotype-guided dosing regimen for the patient based on not only clinical variables but also genetic factors. The technology used in the EU-PACT trial had a lead time of approximately 2 h to obtain the required genotypes.

In the case of the RAPID-GENE trial [36], patients were typed for the *CYP2C19*2* allele, which has been associated with a lack of efficacy of clopidogrel following percutaneous coronary intervention (PCI). Patients were randomized to standard (75 mg/day clopidogrel) or genotype-guided arms. In the genotype guided arm, patients carrying the *CYP2C19*2* allele were prescribed 10 mg prasugrel while wild-type patients were given the standard therapy of 75 mg/day clopidogrel. The trial demonstrated that genotype-guided dosing significantly reduced adverse events (platelet reactivity) compared to standard treatment. The genotyping technology utilized for RAPID-GENE, the Spartan RX CYP2C19, is able to produce genotypes in less than 60 min. In August 2013, the Spartan RX CYP2C19 device became the first on-demand rapid genetic testing device for prescribing guidance to gain Food and Drug Administration (FDA) approval. However, it is important to note that it is not approved as a POC device and must be operated within a clinical laboratory environment.

POC Genotype technologies continue to advance and molecular biology methodologies, such as SmartAmp 2 [37] are reducing the time to obtain a genotype further. It is already possible to determine genotypes for pharmacogenomics-related variants such as those associated with warfarin dose requirement (*VKORC1* and *CYP2C9*) in 30–40 min [37,38]. With addition of microfluidics technology to devices [39,40], the size of the equipment for genotyping will decrease and subsequently, the portability will also increase. It is the ease of use and interpretation of results which are the key attributes for POC genetic test which will facilitate their adoption by health-care professionals. In future years, it is likely that many more devices and tests will be applied to clinical care and obtain regulatory approval.

6.2. Companion Diagnostics

For a number of years now the notion of "one size fits all" in both drug prescribing and drug development has seemed an outdated concept. Indeed, the pharmaceutical industry has re-focused its efforts away from the previous block buster drug model to develop more "niche-busting" products which are licensed alongside a companion diagnostic assay (Table 2). This allows for drugs which may be largely considered ineffective in the wider population to be targeted to a subset of patients likely to respond well to the treatment. To date, the vast majority of these have been developed in the oncology field but it is likely that many more examples of population stratification using genomics methodologies for targeted treatment will emerge for other indications. An added advantage of undertaking targeted therapies in cancer using companion diagnostics has been the rapid approvals obtained from the FDA, despite pivotal efficacy trials testing smaller numbers than is usual in non-stratified trials [41].

While the majority of companion diagnostics products to date have focused on variation in pharmacodynamic factors for stratification, there is also a growing interest in individualizing drug doses based on pharmacokinetic variability. For example, dose escalation of tivantinib, a non-small cell lung cancer therapy, is based on stratification for the *CYP2C19* genotype [42]. This is consistent with European Medicines Agency guidance on pharmacogenetic effects on drug pharmacokinetics [43]. Thus, in the future, it is likely that regulatory approval for a new drug could be dependent on dose stratification based on the underlying metabolizer phenotypes.

6.3. Pre-Emptive Genotyping

The possibility of pre-emptying genotyping is being explored in the U.S., for example, through the Vanderbilt Personalized Medicines Program, where patients are genotyped on a 184-variant platform [44]. The theory behind this approach is that if a genotype is available to the physician in the electronic medical record, at the point of prescribing, they will be able to make a rational decision as to the choice and/or dose of the drug. It also highlights an important issue in relation to the evidence for clinical implementation: we cannot possibly undertake randomised controlled trials or prospective studies for every genomic variant that is identified, and other methodologies for evaluating the clinical utility of a genomic biomarker will have to be utilized. It is also important to note that such information is already used by physicians with respect to

non-pharmacogenomic tests, for example renal function tests allow clinicians to reduce the dose of a drug that is renally excreted [45]. Whether and how such pre-prescription genotyping will be used by prescribers, and whether it leads to improvement in clinical outcomes, will require careful evaluation over the next few years. Indeed it can be seen as a prelude to a time when whole genome sequencing is so cheap that is undertaken routinely with the data being available within electronic medical records.

6.4. Personal Genomes and Clinical Applications

The per-base cost of sequencing has fallen exponentially since 2003 and has for many years exceeded the trajectory of Moores Law (the number of transistors on integrated circuits approximately doubles every 2 years) [46]. With this is mind it is feasible to imagine in the not too distant future that whole genome sequencing will be a cost effective option for healthcare providers. However, with the >3 billion base pairs confirmed in 2003 by the human genome project and the complexity of interpreting the role of genetic variation in inter-individual drug response, it is likely the challenge for the next 10 years will be in producing the tools with which to interpret this data and provide meaningful outputs that can be utilized by healthcare professionals.

6.5. Ethical Considerations

A key ethical issue with regard to the implementation of personalised medicine relates to the fact that it may lead to health inequalities, within and between countries. Given the costs related to developing, manufacturing and obtaining approval for genetic tests, in addition to defining the role of ethnic genetic differences, it is easy to see how resource poor countries and communities could miss out on the benefits of personalised medicine advances in the future. This may be offset by the advances in genomics technologies and the vast reduction in costs of implementing them that has taken place over the last 10 years. Another issue which is becoming increasingly important with the use of next generation sequencing technologies is whether patients should be informed about incidental findings. An analysis of 1000 participants' exomes showed that the frequency of actionable (*i.e.*, pathogenic or likely pathogenic single nucleotide variants) incidental findings was 3.4% in European Caucasians and 1.2% in Africans [53]. Some guidelines have been produced [54], but there is still controversy [54,55], and the debate will no doubt become more intense as more people have their genomes sequenced, and more variants are classified as being actionable.

6.6. Educating Stakeholders

In order for pharmacogenomics, as one of the technologies that is important for personalised medicine, to realize wider uptake into healthcare provision, there needs to be greater awareness and education. This applies not only to healthcare professional but to patients who are the ultimate stakeholder in pharmacogenomics, and stand to gain the most if we can improve predictability of how patients will respond to drugs.

Table 2. *In vitro* companion diagnostic kits currently licensed by the Food and Drug Administration (FDA).

Biomarker	Prevalance	Indication	Drug	Assay Kit	Technology	Manufacturer
HER2 gene amplification	22.2% [47]	Breast cancer	trastuzumab (Herceptin)	Inform	FISH	Ventana Medical Systems
				PathVysion #	FISH	Abbott Molecular Inc.
				SPOT-Light HER2	CISH	Life Technologies
				InSite HER2	IHC	Biogenex Laboratories, Inc.
				HercepTest	IHC	Dako Denmark
				HER2 PharmDx	CISH	Dako Denmark
				HER2 PharmDx *	FISH	Dako Denmark
				PATHWAY Her2	IHC	Ventana Medical Systems
				Bond Oracle	IHC	Leica Biosystems
EGFR protein expression	60%–80% [48]	Colorectal cancer	cetuximab (Erbitux) panitumumab (Vectibix)	EGFR pharmDx	IHC	Dako North America
c-Kit protein expression	100% [49]	Gastrointestinal stromal tumours	imatinib (Gleevac)	c-kit pharmDx	IHC	Dako North America
ALK gene rearrangement	7.5% [50]	Non-small cell lung cancer	crizotinib (Xalkori)	VYSIS	FISH	Abbott Molecular Inc.
BRAF p.V600E mutation	75.4% [51]	Melanoma	vemurafenib (Zelboraf)	Cobas 4800	Real-time PCR	Roche Molecular Systems, Inc
KRAS mutation	30%–60% [52]	Colorectal cancer	cetuximb (Erbitux)	therascreen	Real-time PCR	Qiagen

* Indicated for assessment of breast cancer patients considered for pertuzumab (Perjeta) or # cyclophosphamide, doxorubicin, 5-FU treatment.

Clinicians need to be made aware of the availability of genetic tests relating to treatment decisions, and how to interpret them. Lack of familiarity with genetic tests may be one reason for the poor uptake into clinical practice [45]. The lack of training is amongst the most common reasons cited as a barrier to pharmacogenomic implementation [56].

6.7. Regulatory Environment

There are of course many other issues which need to be tackled in order to facilitate the implementation of pharmacogenomic testing into clinical practice—many of these also apply to other biomarker strategies that are important for personalised medicine [41]. Amongst this is the regulatory environment, which has a big influence on the diagnostics industry. With the development of companion diagnostics, it will be important for there to be streamlined procedures which allow for simultaneous approval of the drug and diagnostic. There are differences between the FDA and European Medicines Agency for the development of companion diagnostics, with the latter being less stringent, but likely to adopt similar procedures in a drive to ensure there is global harmonization of the regulatory procedures needed for approval [57]. For the diagnostics industry, the requirement for a clinical utility study may become prohibitive in terms of cost, unless there are clear pathways for protection of intellectual property and re-imbursement.

7. Conclusions

Since the completion of the human genome, there has been steady, albeit slow, progress in the identification and implementation of biomarkers into clinical practice. This progress is likely to continue, and hopefully accelerate as our ability to interrogate the human genome becomes more cost- and time-efficient, and we start embracing, and intelligently interpreting, different sources of data to define the clinical validity and utility of biomarkers. Outcomes research will thus become particularly important, and will depend on having access to curated electronic healthcare databases where patients can be followed longitudinally from the time of having a biomarker assessed to the time a drug is prescribed, and forward into the future to define the clinical outcome of the patient (in comparison to relevant *a priori* defined controls).

Acknowledgments

MP is a NIHR Senior Investigator. We thank the UK Department of Health (NHS Chair of Pharmacogenetics), Wolfson Foundation, EU-FP7 and the MRC (MRC Centre for Drug Safety Science) for grant funding.

Author Contributions

All authors contributed equally.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; *et al.* Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921.
2. Vogel, F. Moderne probleme der humangenetik. *Ergeb. Inn. Med. Kinderheilkd* **1959**, *12*, 52–125.
3. Watson, J.D.; Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **1953**, *171*, 737–738.
4. International HapMap Consortium. The international hapmap project. *Nature* **2003**, *426*, 789–796.
5. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J.P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367.
6. Link, E.; Parish, S.; Armitage, J.; Bowman, L.; Heath, S.; Matsuda, F.; Gut, I.; Lathrop, M.; Collins, R. Slco1b1 variants and statin-induced myopathy—A genomewide study. *New Engl. J. Med.* **2008**, *359*, 789–799.
7. Sarasquete, M.E.; Garcia-Sanz, R.; Marin, L.; Alcoceba, M.; Chillon, M.C.; Balanzategui, A.; Santamaria, C.; Rosinol, L.; de la Rubia, J.; Hernandez, M.T.; *et al.* Bisphosphonate-related osteonecrosis of the jaw is associated with polymorphisms of the cytochrome p450 cyp2c8 in multiple myeloma: A genome-wide single nucleotide polymorphism analysis. *Blood* **2008**, *112*, 2709–2712.
8. Kindmark, A.; Jawaid, A.; Harbron, C.G.; Barratt, B.J.; Bengtsson, O.F.; Andersson, T.B.; Carlsson, S.; Cederbrant, K.E.; Gibson, N.J.; Armstrong, M.; *et al.* Genome-wide pharmacogenetic investigation of a hepatic adverse event without clinical signs of immunopathology suggests an underlying immune pathogenesis. *Pharmacogenomics J.* **2008**, *8*, 186–195.
9. Daly, A.K.; Donaldson, P.T.; Bhatnagar, P.; Shen, Y.; Pe'er, I.; Floratos, A.; Daly, M.J.; Goldstein, D.B.; John, S.; Nelson, M.R.; *et al.* Hla-b*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat. Genet.* **2009**, *41*, 816–819.
10. McCormack, M.; Alfirevic, A.; Bourgeois, S.; Farrell, J.J.; Kasperaviciute, D.; Carrington, M.; Sills, G.J.; Marson, T.; Jia, X.; de Bakker, P.I.; *et al.* Hla-a*3101 and carbamazepine-induced hypersensitivity reactions in europeans. *New Engl. J. Med.* **2011**, *364*, 1134–1143.
11. Ozeki, T.; Mushiroda, T.; Yowang, A.; Takahashi, A.; Kubo, M.; Shirakata, Y.; Ikezawa, Z.; Iijima, M.; Shiohara, T.; Hashimoto, K.; *et al.* Genome-wide association study identifies hla-a*3101 allele as a genetic risk factor for carbamazepine-induced cutaneous adverse drug reactions in japanese population. *Hum. Mol. Gen.* **2011**, *20*, 1034–1041.
12. Tohkin, M.; Kaniwa, N.; Saito, Y.; Sugiyama, E.; Kurose, K.; Nishikawa, J.; Hasegawa, R.; Aihara, M.; Matsunaga, K.; Abe, M.; *et al.* A whole-genome association study of major determinants for allopurinol-related stevens-johnson syndrome and toxic epidermal necrolysis in japanese patients. *Pharmacogenomics J.* **2013**, *13*, 60–69.

13. Shuldiner, A.R.; O'Connell, J.R.; Bliden, K.P.; Gandhi, A.; Ryan, K.; Horenstein, R.B.; Damcott, C.M.; Pakyz, R.; Tantry, U.S.; Gibson, Q.; *et al.* Association of cytochrome p450 2c19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. *JAMA* **2009**, *302*, 849–857.
14. Ge, D.; Fellay, J.; Thompson, A.J.; Simon, J.S.; Shianna, K.V.; Urban, T.J.; Heinzen, E.L.; Qiu, P.; Bertelsen, A.H.; Muir, A.J.; *et al.* Genetic variation in il28b predicts hepatitis c treatment-induced viral clearance. *Nature* **2009**, *461*, 399–401.
15. Takeuchi, F.; McGinnis, R.; Bourgeois, S.; Barnes, C.; Eriksson, N.; Soranzo, N.; Whittaker, P.; Ranganath, V.; Kumanduri, V.; McLaren, W.; *et al.* A genome-wide association study confirms vkorc1, cyp2c9, and cyp4f2 as principal genetic determinants of warfarin dose. *PLoS Genet.* **2009**, *5*, e1000433.
16. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **2010**, *467*, 1061–1073.
17. Gordon, A.S.; Tabor, H.K.; Johnson, A.D.; Snively, B.M.; Assimes, T.L.; Auer, P.L.; Ioannidis, J.P.; Peters, U.; Robinson, J.G.; Sucheston, L.E.; *et al.* Quantifying rare, deleterious variation in 12 human cytochrome p450 drug-metabolism genes in a large-scale exome dataset. *Hum. Mol. Gen.* **2014**. 23, 1957–1963.
18. Wadelius, M.; Chen, L.Y.; Lindh, J.D.; Eriksson, N.; Ghori, M.J.; Bumpstead, S.; Holm, L.; McGinnis, R.; Rane, A.; Deloukas, P. The largest prospective warfarin-treated cohort supports genetic forecasting. *Blood* **2009**, *113*, 784–792.
19. Ainle, F.N.; Mumford, A.; Tallon, E.; McCarthy, D.; Murphy, K. A vitamin k epoxide reductase complex subunit 1 mutation in an irish patient with warfarin resistance. *Irish J. Med. Sci.* **2008**, *177*, 159–161.
20. Harrington, D.J.; Gorska, R.; Wheeler, R.; Davidson, S.; Murden, S.; Morse, C.; Shearer, M.J.; Mumford, A.D. Pharmacodynamic resistance to warfarin is associated with nucleotide substitutions in vkorc1. *J. Thrombosis Haemost.* **2008**, *6*, 1663–1670.
21. Harrington, D.J.; Underwood, S.; Morse, C.; Shearer, M.J.; Tuddenham, E.G.; Mumford, A.D. Pharmacodynamic resistance to warfarin associated with a val66met substitution in vitamin k epoxide reductase complex subunit 1. *Thromb. Haemost.* **2005**, *93*, 23–26.
22. Rost, S.; Fregin, A.; Ivaskevicius, V.; Conzelmann, E.; Hortnagel, K.; Pelz, H.J.; Lappegard, K.; Seifried, E.; Scharrer, I.; Tuddenham, E.G.; *et al.* Mutations in vkorc1 cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature* **2004**, *427*, 537–541.
23. Behr, E.R.; Roden, D. Drug-induced arrhythmia: Pharmacogenomic prescribing? *Eur. Heart J.* **2013**, *34*, 89–95.
24. Ramirez, A.H.; Shaffer, C.M.; Delaney, J.T.; Sexton, D.P.; Levy, S.E.; Rieder, M.J.; Nickerson, D.A.; George, A.L., Jr.; Roden, D.M. Novel rare variants in congenital cardiac arrhythmia genes are frequent in drug-induced torsades de pointes. *Pharmacogenomics J.* **2013**, *13*, 325–329.
25. Consortium, E.P.; Bernstein, B.E.; Birney, E.; Dunham, I.; Green, E.D.; Gunter, C.; Snyder, M. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74.

26. Starkey Lewis, P.J.; Dear, J.; Platt, V.; Simpson, K.J.; Craig, D.G.; Antoine, D.J.; French, N.S.; Dhaun, N.; Webb, D.J.; Costello, E.M.; *et al.* Circulating micrornas as potential markers of human drug-induced liver injury. *Hepatology* **2011**, *54*, 1767–1776.
27. Ichihara, A.; Wang, Z.; Jinnin, M.; Izuno, Y.; Shimozono, N.; Yamane, K.; Fujisawa, A.; Moriya, C.; Fukushima, S.; Inoue, Y.; *et al.* Upregulation of mir-18a-5p contributes to epidermal necrolysis in severe drug eruptions. *J. Allergy Clin. Immunol.* **2013**, *133*, 1065–1074.
28. Poste, G. Bring on the biomarkers. *Nature* **2011**, *469*, 156–157.
29. US Food and Drug Administration. Table of Pharmacogenomic Biomarkers in Drug Labeling. Available online: http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm/ (6 November 2013).
30. Relling, M.V.; Klein, T.E. Cpic: Clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. *Clin. Pharmacol. Ther.* **2011**, *89*, 464–467.
31. Pirmohamed, M.; Friedmann, P.S.; Molokhia, M.; Loke, Y.K.; Smith, C.; Phillips, E.; La Grenade, L.; Carleton, B.; Papaluca-Amati, M.; Demoly, P.; *et al.* Phenotype standardization for immune-mediated drug-induced skin injury. *Clin. Pharmacol. Ther.* **2011**, *89*, 896–901.
32. Pirmohamed, M.; Aithal, G.P.; Behr, E.; Daly, A.; Roden, D. The phenotype standardization project: Improving pharmacogenetic studies of serious adverse drug reactions. *Clin. Pharmacol. Ther.* **2011**, *89*, 784–785.
33. Behr, E.R.; January, C.; Schulze-Bahr, E.; Grace, A.A.; Kaab, S.; Fiszman, M.; Gathers, S.; Buckman, S.; Youssef, A.; Pirmohamed, M.; *et al.* The international serious adverse events consortium (isaec) phenotype standardization project for drug-induced torsades de pointes. *Eur. Heart J.* **2012**. *34*, 1958–1963
34. Saiki, R.K.; Gelfand, D.H.; Stoffel, S.; Scharf, S.J.; Higuchi, R.; Horn, G.T.; Mullis, K.B.; Erlich, H.A. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **1988**, *239*, 487–491.
35. Pirmohamed, M.; Burnside, G.; Eriksson, N.; Jorgensen, A.L.; Toh, C.H.; Nicholson, T.; Kesteven, P.; Christersson, C.; Wahlstrom, B.; Stafberg, C.; *et al.* A randomized trial of genotype-guided dosing of warfarin. *New Engl. J. Med.* **2013**, *369*, 2294–2303.
36. Roberts, J.D.; Wells, G.A.; Le May, M.R.; Labinaz, M.; Glover, C.; Froeschl, M.; Dick, A.; Marquis, J.F.; O'Brien, E.; Goncalves, S.; *et al.* Point-of-care genetic testing for personalisation of antiplatelet treatment (rapid gene): A prospective, randomised, proof-of-concept trial. *Lancet* **2012**, *379*, 1705–1711.
37. Lezhava, A.; Ishidao, T.; Ishizu, Y.; Naito, K.; Hanami, T.; Katayama, A.; Kogo, Y.; Soma, T.; Ikeda, S.; Murakami, K.; *et al.* Exciton primer-mediated snp detection in smartamp2 reactions. *Hum. Mut.* **2010**, *31*, 208–217.
38. Aomori, T.; Yamamoto, K.; Oguchi-Katayama, A.; Kawai, Y.; Ishidao, T.; Mitani, Y.; Kogo, Y.; Lezhava, A.; Fujita, Y.; Obayashi, K.; *et al.* Rapid single-nucleotide polymorphism detection of cytochrome p450 (cyp2c9) and vitamin k epoxide reductase (vkorc1) genes for the warfarin dose adjustment by the smart-amplification process version 2. *Clin. Chem.* **2009**, *55*, 804–812.

39. Burn, J. Company profile: Quantumdx group limited. *Pharmacogenomics* **2013**, *14*, 1011–1015.
40. Stedtfeld, R.D.; Tourlousse, D.M.; Seyrig, G.; Stedtfeld, T.M.; Kronlein, M.; Price, S.; Ahmad, F.; Gulari, E.; Tiedje, J.M.; Hashsham, S.A. Gene-z: A device for point of care genetic testing using a smartphone. *Lab. Chip* **2012**, *12*, 1454–1462.
41. *Realising the potential of stratified medicine*; Academy of Medical Sciences: London, UK, 2013.
42. Yamamoto, N.; Murakami, H.; Hayashi, H.; Fujisaka, Y.; Hirashima, T.; Takeda, K.; Satouchi, M.; Miyoshi, K.; Akinaga, S.; Takahashi, T.; *et al.* Cyp2c19 genotype-based phase i studies of a c-met inhibitor tivantinib in combination with erlotinib, in advanced/metastatic non-small cell lung cancer. *Br. J. Cancer* **2013**, *109*, 2803–2809.
43. European Medicines Agency. Guideline on the Use of Pharmacogenetic Methodologies in the Pharmacokinetic Evaluation of Medicinal Products. Available online: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/02/WC500121954.pdf (accessed on 30 October 2012).
44. Pulley, J.M.; Denny, J.C.; Peterson, J.F.; Bernard, G.R.; Vnencak-Jones, C.L.; Ramirez, A.H.; Delaney, J.T.; Bowton, E.; Brothers, K.; Johnson, K.; *et al.* Operational implementation of prospective genotyping for personalized medicine: The design of the vanderbilt predict project. *Clin. Pharmacol. Ther.* **2012**, *92*, 87–95.
45. Pirmohamed, M. Pharmacogenetics: Past, present and future. *Drug Discov. Today* **2011**, *16*, 852–861.
46. Moore, G.E. Cramming more components onto integrated circuits. *Electronics* **1965**, *38*, 114–117.
47. Ross, J.S.; Slodkowska, E.A.; Symmans, W.F.; Pusztai, L.; Ravdin, P.M.; Hortobagyi, G.N. The her-2 receptor and breast cancer: Ten years of targeted anti-her-2 therapy and personalized medicine. *Oncologist* **2009**, *14*, 320–368.
48. Goldstein, N.S.; Armin, M. Epidermal growth factor receptor immunohistochemical reactivity in patients with american joint committee on cancer stage iv colon adenocarcinoma: Implications for a standardized scoring system. *Cancer* **2001**, *92*, 1331–1346.
49. Went, P.T.; Dirnhofer, S.; Bundi, M.; Mirlacher, M.; Schraml, P.; Mangialaio, S.; Dimitrijevic, S.; Kononen, J.; Lugli, A.; Simon, R.; *et al.* Prevalence of kit expression in human tumors. *J. Clin. Oncol.* **2004**, *22*, 4514–4522.
50. Martelli, M.P.; Sozzi, G.; Hernandez, L.; Pettirossi, V.; Navarro, A.; Conte, D.; Gasparini, P.; Perrone, F.; Modena, P.; Pastorino, U.; *et al.* Eml4-alk rearrangement in non-small cell lung cancer and non-tumor lung tissues. *Am. J. Pathol.* **2009**, *174*, 661–670.
51. Greaves, W.O.; Verma, S.; Patel, K.P.; Davies, M.A.; Barkoh, B.A.; Galbincea, J.M.; Yao, H.; Lazar, A.; Aldape, K.D.; Medeiros, L.J., *et al.* Frequency and spectrum of braf mutations in a retrospective, single-institution study of 1112 cases of melanoma. *J. Mol. Diagn.* **2012**, *15*, 220–226.

52. Brink, M.; de Goeij, A.F.; Weijenberg, M.P.; Roemen, G.M.; Lentjes, M.H.; Pachen, M.M.; Smits, K.M.; de Bruine, A.P.; Goldbohm, R.A.; van den Brandt, P.A. K-ras oncogene mutations in sporadic colorectal cancer in the netherlands cohort study. *Carcinogenesis* **2003**, *24*, 703–710.
53. Dorschner, M.O.; Amendola, L.M.; Turner, E.H.; Robertson, P.D.; Shirts, B.H.; Gallego, C.J.; Bennett, R.L.; Jones, K.L.; Tokita, M.J.; Bennett, J.T., *et al*. Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am. J. Hum. Genet.* **2013**, *93*, 631–640.
54. Green, R.C.; Berg, J.S.; Grody, W.W.; Kalia, S.S.; Korf, B.R.; Martin, C.L.; McGuire, A.L.; Nussbaum, R.L.; O'Daniel, J.M.; Ormond, K.E., *et al*. Acmg recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med*. **2013**, *15*, 565–574.
55. Kocarnik, J.M.; Fullerton, S.M. Returning pleiotropic results from genetic testing to patients and research participants. *JAMA* **2014**, *311*, 795–796.
56. Patel, H.N.; Ursan, I.D.; Zueger, P.M.; Cavallari, L.H.; Pickard, A.S. Stakeholder views on pharmacogenomic testing. *Pharmacotherapy* **2014**, *34*, 151–165.
57. Senderowicz, A.M.; Pfaff, O. Similarities and differences in the oncology drug approval process between fda and european union with emphasis on in vitro companion diagnostics. *Clin. Cancer Res.* **2014**, *20*, 1445–1452.

Changing Histopathological Diagnostics by Genome-Based Tumor Classification

Michael Kloth and Reinhard Buettner

Abstract: Traditionally, tumors are classified by histopathological criteria, *i.e.*, based on their specific morphological appearances. Consequently, current therapeutic decisions in oncology are strongly influenced by histology rather than underlying molecular or genomic aberrations. The increase of information on molecular changes however, enabled by the Human Genome Project and the International Cancer Genome Consortium as well as the manifold advances in molecular biology and high-throughput sequencing techniques, inaugurated the integration of genomic information into disease classification. Furthermore, in some cases it became evident that former classifications needed major revision and adaption. Such adaptations are often required by understanding the pathogenesis of a disease from a specific molecular alteration, using this molecular driver for targeted and highly effective therapies. Altogether, reclassifications should lead to higher information content of the underlying diagnoses, reflecting their molecular pathogenesis and resulting in optimized and individual therapeutic decisions. The objective of this article is to summarize some particularly important examples of genome-based classification approaches and associated therapeutic concepts. In addition to reviewing disease specific markers, we focus on potentially therapeutic or predictive markers and the relevance of molecular diagnostics in disease monitoring.

Reprinted from *Genes*. Cite as: Kloth, M.; Buettner, R. Changing Histopathological Diagnostics by Genome-Based Tumor Classification. *Genes* **2014**, *5*, 444-459.

1. Approaches to a Genome-Based Tumor Classification

1.1. The 2008 WHO Classification of Hematological Malignancies

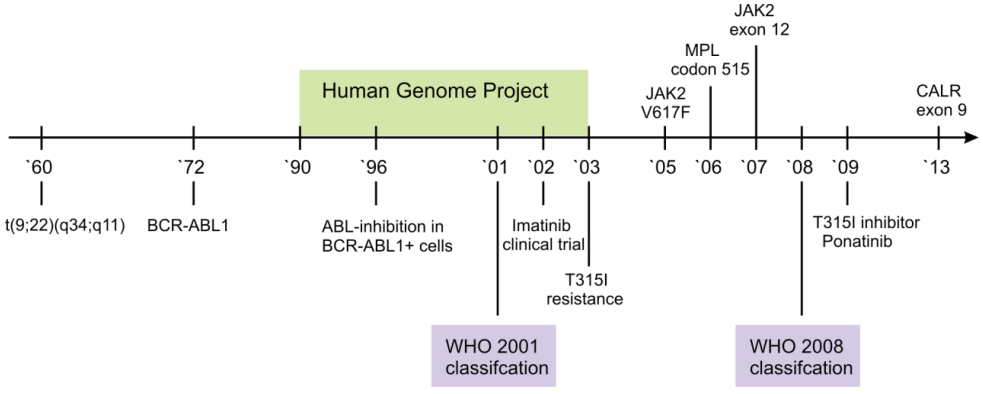
The 2008 WHO classification of chronic myeloid malignancies is at present the most evolved approach to a taxonomy considering defined molecular aberrations [1]. The malignancies that are included are now classified into five categories:

(1) Acute myeloid leukemia (AML) and related precursor neoplasms;
(2) Myelodysplastic syndromes (MDS);
(3) Myeloproliferative neoplasms (MPN);
(4) Myelodysplastic/Myeloproliferative neoplasms (MDS/MPN);
(5) Myeloid and lymphoid neoplasms with eosinophilia and abnormalities of *PDGFRA*, *PDGFRB*, or *FGFR1*.

The integration of histology and genetics is particularly visible in the category of myeloproliferative neoplasms (MPN). Classification of specific entities into MPN is dependent on presence or absence of *BCR-ABL1*, the disease-causing translocation in CML [2]. The first description of the associated karyotype t(9;22)(q34;q11), according to an abnormally short

chromosome 22, was published as early as 1960 and is widely known as the Philadelphia (Ph) chromosome [3]. Due to the high specificity of *BCR-ABL1,* its detection is mandatory for diagnosis of CML and is further underscored by the influence on therapy with the small molecule tyrosine kinase inhibitor (TKI) Imatinib [4]. Nevertheless, the remarkable journey, from the genomic aberration to the specific therapy, required approximately forty years and started long before the elucidation of the human genome (Figure 1).

Figure 1. Timeline of the elucidation of genomic alterations in myeloproliferative neoplasms. Major breakthroughs in the understanding of the Ph+ neoplasm CML are depicted below the line, those in the understanding of Ph- neoplasms above. Note the significant impact of the human genome project on the elucidation on Ph- specific genomic alterations.



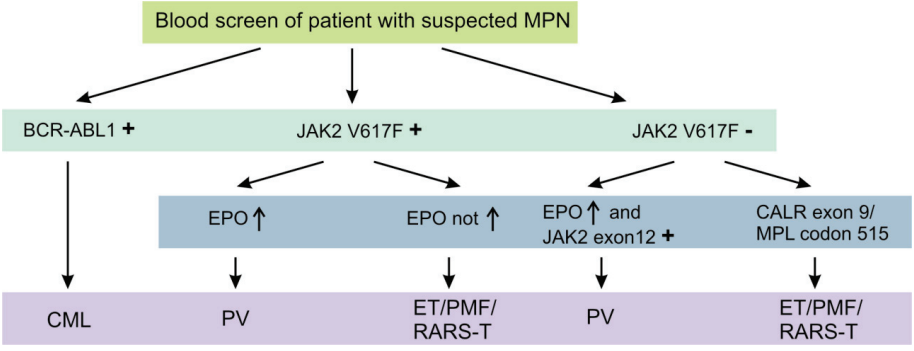
Further myeloproliferative neoplasms are characterized by the absence of *BCR-ABL1*, including the three Ph-negative classic MPNs—polycythemia vera (PV), essential thrombocythemia (ET) and progressive myelofibrosis (PMF). Despite the absence of *BCR-ABL1*, it has consecutively been shown that these MPNs are themselves also characterized by additional recurrent aberrations [5,6] (Figure 1).

The *JAK2 V617F* mutation is detectable in approximately 95% of all PV patients and 50% of both ET and PMF patients [7–9]. It is further detected in patients with refractory anemia with ring sideroblasts and thrombocytosis (RARS-T), but in less than 5% of patients with acute myeloid leukemia (AML) or myeloid dysplastic syndrome (MDS) and not in solid tumors [10,11]. Although, the diagnostic process is initially dominated by peripheral blood cell count and serum erythropoetin (EPO) levels, *JAK2 V617F* or less common *JAK2 Exon 12* mutations [12] confirm the diagnosis of a suspected PV without the need of a bone marrow biopsy [13]. Beside the two most important genetic aberrations in the diagnostic algorithm of MPNs, namely *BCR-ABL1* and *JAK2 V617F*, several other potentially helpful recurrent aberrations are known. These include the presence of recurrent mutations in *CALR exon 9* [14] and *MPL codon 515* [15], essentially occurring in *JAK2 V617F* negative cases (Figure 2).

Another important example of the integration of molecular aberrations in the 2008 WHO classification of hematological malignancies is the newly introduced group of myeloid and

lymphoid neoplasms with eosinophilia and abnormalities of *PDGFRA*, *PDGFRB*, or *FGFR1*. This reclassification highlights the consideration of the targetable alterations *FIP1L1-PDGFRA* or *PDGFRB*-rearrangements and those harboring *FGFR1*-rearrangements, indicating response or resistance to Imatinib [2,16].

Figure 2. Diagnostic algorithm of classic myeloproliferative neoplasms using specific molecular aberrations. Detection of the molecular aberrations depicted above is highly suggestive for the suspected myeloproliferative disorder. Nevertheless, at least in the case of absence of these specific aberrations, a bone marrow biopsy should be performed.



1.2. Lung Cancer as a Paradigm: Advances in the Molecular Characterization of Solid Malignancies

The ongoing comprehensive characterizations of solid tumors, such as those conducted by The Cancer Genome Atlas (TCGA) will significantly impact upcoming tumor classifications. This also includes lung cancer, the leading cause of cancer death worldwide [17] and an example for substantial advances by genome-based therapy approaches [18]. In general, NSCLC is subclassified into adenocarcinoma, squamous cell carcinoma and large cell carcinoma. Future classifications need to characterize clinically relevant subtypes, instead of the traditional distinction of non-small cell lung cancers (NSCLC) and small-cell lung cancer (SSLC). Concepts for a reclassification of lung adenocarcinoma were suggested, particularly with respect to reclassify large cell carcinoma based on genomic aberrations into adenocarcinoma, squamous cell carcinoma and large cell neuroendocrine carcinoma, respectively. Recommendations for (immuno)histological diagnostic work-up, and also for determining specific molecular aberrations in lung adenocarcinoma subtypes have been published [19].

Major rationales for changes in adenocarcinoma histologic variants are that the invasive mucinous type shows frequent mutations in *KRAS* and hardly any in *EGFR*, whereas non-mucinous adenocarcinoma of predominant lepidic subtype is characterized by frequent mutations in *EGFR* and fewer in *KRAS* [20]. In this context, it is worth noting and described in more detail below, that clinical studies with the TKIs Gefitinib and Erlotinib showed significantly improved survival in patients suffering from lung adenocarcinoma harboring mutations in the kinase domain of

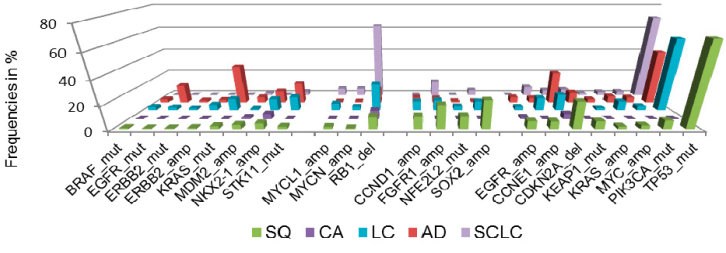
EGFR [21,22]. Besides the mutations in *EGFR* and *KRAS*, the targetable translocation *EML4-ALK* recurrently occurs in lung adenocarcinoma [23] and is most frequent in tumours with mucinous signet-ring appearance. Clinical trials using the tyrosine kinase inhibitors Crizotinib and Ceritinib have now shown improved progression-free survival in those patients [24,25].

In contrast, squamous cell carcinoma of the lung is commonly associated with mutations in the *NFE2L2/KEAP*-axis [26] and often harbors targetable alterations of *FGFR1* [27] and in fewer cases mutations in *DDR2*. Interestingly, DDR2-transformed cell lines maintain SRC phosphorylation and are sensitive to Dasatinib [28], proved by the response in squamous cell lung cancer patients [29].

Beyond the combined loss of *RB1 and TP53* in neuroendocrine pulmonary tumors [30], it was shown that low- and intermediate-grade pulmonary cacinoids harbor recurrent mutations in chromatin remodeling complexes [31], whereas the high-grade neuroendocrine tumor, small cell lung cancer, is associated with sequential changes, including the deletion of *PTEN* [32].

Additionally, we assessed cancer genome alterations linked to histomorphological and immunohistochemical features, considering high therapeutic relevance and improved patient outcome [33] (Figure 3). By this approach, we devised a genomic-based prediction model of lung cancer subtypes. This model shows that the majority of large cell cancers could be reassigned to adenocarcinoma, squamous cell carcinoma or small-cell lung cancer. By the combined analysis of immunohistochemical, genomic and clinical features it becomes further evident that personalized approaches significantly improve the outcome of patients with advanced lung cancer and other solid cancers.

Figure 3. Frequencies of significant genomic alterations in histological subgroups of lung cancer. Colors of histological subtypes are encoded as follows: green—squamous cell lung cancer (SQ), purple—carcinoid tumor, light blue—large cell lung cancer (LC), red—adenocarcinoma of the lung (AD), dark blue—small cell lung cancer (SCLC). Data adapted from [33].



Beyond the principally well-characterized situation in lung cancer, we already know of highly recurrent genomic alterations in many other solid malignancies. Prostate cancer is the most common cancer in men and is characterized by fewer mutations in typical cancer genes, when compared to other solid cancers [34–36]. However, genomic alterations in androgen signaling, the rearrangement of *ETS* transcription factors, especially the fusion of *TMPRSS2-ERG* [37], as well as the deletion of *PTEN* [38] are known to be highly recurrent in primary cancers [39], early-onset cancers [40] and castration-resistant prostate cancers [41]. Beside the outstanding therapeutic role of

androgen deprivation, recent efforts investigate a mechanistic rationale of PARP inhibition in *ETS*-rearranged prostate cancers [42].

1.3. EWS and the Importance of Translocations in the Diagnostic Workup of Mesenchymal Malignancies

Tumors ascribed to the family of Ewing's sarcomas or primitive neuroectodermal tumors (PNET) are the second most common bone tumors in children, but can also arise from any other tissue. The recent molecular understanding of these aggressive tumors has greatly advanced new therapeutic approaches. However, whereas chemotherapy improved the survival rate from 10% to 70%–80% in localized disease, the survival of patients with distant metastases is still poor [43]. The WHO classification of Ewing sarcoma (ES) as a single entity is underlined by its cytogenetic signature t(11;12)(q24;q12), according to the translocation *EWSR1-FLI1* in approximately 85% of patients. Although, *EWSR1* translocations can be found in many other mesenchymal tumors, almost all of the remaining ES cases are characterized by further ES-specific *EWSR1* translocations [44,45]. This also includes the second most common cytogenetic aberration t(21;22)(q22;q12), corresponding to *EWSR1-ERG* [46] (Table 1).

Table 1. Fusion partners of *EWSR1*-rearrangements in different soft tissue tumors. Overlapping rearrangements in different histopathological entities are in bold. Adapted from [44,45].

Histological type	Translocation	EWS-rearrangements
Ewing's Sarcoma	t(11;22)(q24;q12)	EWSR1-FLI1
	t(21;22)(q22;q12)	EWSR1-ERG
	t(7;22)(q22;q12)	EWSR1-ETV1
	t(17;22)(q21;q12)	EWSR1-ETV4
	t(2;22)(q36;q12)	EWSR1-FEV
	inv(22)(q12q12)	EWSR1-PATZ1
	t(2;22)(q31;q12)	EWSR1-SP3
	t(20;22)(q13;q12)	EWSR1-NFATC2
	t(4;22)(q31;12)	EWSR1-SMARCA5
	t(17;22)(q12;q12)	EWSR1-E1AF
	inv(22)(q21;12)	EWSR1-ZSG
Angiomatoid fibrous histiocytoma	t(12;22)(q13;q12)	**EWSR1-ATF1**
	t(2;22)(q33;q12)	**EWSR1-CREB1**
Clear cell sarcoma	t(12;22)(q13;q12)	**EWSR1-ATF1**
	t(2;22)(q33;q12)	**EWSR1-CREB1**
Malignant gastrointestinal neuroectodermal tumor	t(12;22)(q13;q12)	**EWSR1-ATF1**
	t(2;22)(q33;q12)	**EWSR1-CREB1**
Myoepithelial tumor of soft tissue and bone	t(1;22)(q23;q12)	EWSR1-PBX1
	t(19;22)(q13;q12)	EWSR1-ZNF444
	t(6;22)(p21;q12)	EWSR1-POU5F1
Extraskeletal myxoid chondrosarcoma	t(9;22)(q22;q12)	EWSR1-NR4A3
Myxoid liposarcoma	t(12;22)(q13;q12)	EWSR1-DD1T3

Intriguingly, the well-characterized molecular situation of ES is in contrast to the fact that the cell of origin of the small round cell appearing tumor is not known. A further complication of the histological diagnosis is caused by atypical ES, including large/epithelioid/clear/spindle cell ES, vascular-like ES, adamantinoma-like ES, ES with neuroectodermal features, synovial sarcoma-like PNET and sclerosing PNET [47]. Beyond this difficult histopathological classification, sarcomas may be genetically classified from a near-diploid karyotpe to a more complex genomic instability. The first is characterized by highly recurrent translocations, the latter by numerical and structural abnormalities affecting multiple chromosomes [48]. In the case of small round cell appearing sarcoma, recently, new subtypes were defined by recurrent translocations of BCOR-CCNB3 [49] and CIC-DUX4 [50].

Furthermore, we already know several diagnostically relevant genomic rearrangements in soft tissue malignancies [51]. The recurrently occurring reciprocal translocation t(X;18) is characteristic of synovial sarcoma, and leads to the potentially therapeutic relevant SSX-S18 protein [52]. The translocation *FUS-CHOP* is detected in myxoid liposarcoma [53] and recurrent amplifications of the E3-Ubiqutitin ligase *MDM2*, as well as *CDK4*, in well differentiated and dedifferentiated liposarcoma [54,55]. Interestingly, ongoing efforts investigate the reactivation of p53 by MDM2-p53 interaction inhibitors [56].

1.4. Cancer of Unknown Primary Origin (CUP)

Carcinoma of an unknown primary origin (CUP) is descriptive of a metastatic cancer without an identifiable primary tumor site. CUP accounts for 3%–5% of all cancer diagnoses and is usually characterized by an aggressive metastatic growth and a challenging clinical presentation. In theory, CUP could be considered as a unique biological entity, or in the opposite view, as a group of different entities. The classification of CUP is essentially based on the prognostic outcome, thereby distinguishing favorable and unfavorable carcinomas [57–59] (Table 2).

Table 2. Classification of cancers of unknown primary origin.

Clinically favorable CUP	Clinically unfavorable CUP
Extragonadal germ-cell cancer	Metastatic adenocarcinoma
Peritoneal papillary adenocarcinoma	Non papillary malignant ascites
Adenocarcinoma in axillary lymph nodes	Multiple cerebral metastases
Cervical squamous-cell carcinoma	Squamous-cell carcinoma of the abdominopelvic cavity
Neuroendocrine carcinoma	Lytic bone metastases
Blastic bone metastases and PSA elevation	

This classification predominantly depends on the morphological and immunohistochemical appearance. At present, the specimen is investigated by the use of specific antibodies (investigated epitopes in parentheses):

(1) Identification of the cancer type:
Carcinoma (CK AE1/3), mesothelioma (Calretinin, BerEP4), sarcoma (Vimentin), lymphoma (LCA), melanoma (HMB-45, MITF, S100);

(2) Identification of the subtype:
Adenocarcinoma (CK7, CK20), squamous cell carcinoma (CK5/6, p40, p63), hepatocellular carcinoma (Hepar1), renal cell carcinoma (RCC, PAX8, CA9), urothelial carcinoma (GATA3, S100P, Uroplakin), thyroid cancer (hTG, TTF1), neuroendocrine cancer (CD56, Synaptophysin, ChromoA), germ-cell tumor (PLAP);

(3) Identification of the origin:
Lung (TTF1, NapsinA), colorectal cancer (CDX2, CK20), breast (ER, PR), pancreas (CDX2, CK7, CK20), ovary (Ca125, ER, WT1), prostate (PSA, PSAP, AR).

However, only 30% of all cases studied can be assigned to a primary tumor and even fewer benefit from a change in therapy regimen. Over the last decade, the molecular characterization by gene-expression profiling in various tumor types has led to the development of several gene signature assays with identification rates of a putative tissue of origin in up to 90% of cases. Furthermore, it was shown that CUP can be treated along actionable genomic alterations and recent whole exome sequencing revealed the existence of known recurrent mutations [60]. These include therapeutically relevant mutations in *PIK3CA*, *MET*, *FGFR3*, *IDH1* as well as several others and it has been already demonstrated that targeted therapies can significantly influence a more favorable outcome in CUP patients [61]. As another variation on this theme, a drug-sensitizing genome alteration in one tumor type may not confer drug susceptibility in another histology, as has been observed in the case of *BRAF* mutations that confer *MEK* and *BRAF* dependency in melanomas [62,63] but not in colorectal carcinomas resulting from *EGFR* activation [64]. These interactions highlight the need for classifications integrating cancer genome alterations, but also histomorphological and immunohistochemical features.

2. Monitoring of Malignancies

Molecular monitoring of CML is the most advanced routinely used surveillance strategy and reflects the need for standardization and quality controls of diagnostic tests. A main objective is the identification of patients, which have a worse response or resistance to therapy with tyrosine kinase inhibitors. The quantification of *BCR-ABL1* transcripts in peripheral blood is thereby of outstanding value in the early phase after initial drug administration, corresponding to the molecular response rate (MMR) upon TKI therapy [65–67]. The MMR is calculated in relation to a control gene (e.g., *BCR* or *ABL1*) and was standardized between laboratories using the international scale (IS) [68]. Essentially, the WHO has undertaken extensive efforts to simplify and standardize the assay by providing reference reagents. The importance of such efforts is reflected by a comparable poor overall survival of patients with an early MMR >10%, in turn leading to universally valid changes in the associated NCCN and ELN guidelines [66].

Despite significant progress in therapy of lung adenocarcinoma, all patients with *EGFR* mutations and *ALK* or *ROS1* translocations receiving specific tyrosine kinase inhibitors will ultimately experience relapse. Recent work highlights the potential of noninvasive detection and monitoring of resistance mutations in free circulating plasma DNA of lung cancer patients. The most prominent example is the known resistance mechanisms mediated by T790M in *EGFR* [69].

Beside the potential of directly targeting cancers harboring T790M by new tyrosine kinase inhibitors [70], it is important to note that T790M cells proliferate more slowly, thereby enabling resensitizing of tumors to primarily used TKIs after temporary withdrawal of the drug. Since it is difficult in clinical practice to undertake repeat biopsies at multiple metastatic sites, noninvasive targeted sequencing techniques may further enable additional therapeutic strategies [71]. It thereby becomes evident that new diagnostic techniques will not only lead to major influences in treatment and monitoring guidelines, but will influence therapeutic guidelines and also the classification of tumors. Especially, the opportunity of non-invasive testing seems to be an attractive and emerging field in diagnostic and therapeutic concepts, as further delineated by testing for *TMPRSS2-ERG* in the case of suspicion of prostate cancer [72].

3. Clinical Success of Targeted Therapeutic Approaches Based on Molecular Biomarkers

3.1. BCR-ABL1

As depicted above, *BCR-ABL1* is the driving lesion in CML, leading to the development of the first small molecule TKI, Imatinib [4,73]. Several multicenter studies confirmed that the overall survival of advanced-stage cancer patients in a chronic disease phase climbed from approximately 50% before 2002 up to approximately 90% after introduction of Imatinib in 2002 [74]. The success of the targeted approach in CML patients is further underscored by ongoing studies investigating the discontinuation of Imatinib [75] and the groundbreaking question of a potential cure [76]. Despite these promising developments, several further therapeutic possibilities for second line therapies are underway. These include the use of second generation TKIs in case of therapeutic failure or intolerance, e.g., Dasatinib, Nilotinib, Bosutinib and Ponatinib, the latter particularly used in case of a secondary resistance by T315I mutation [75] (Figure 1).

3.2. BRAF

The most common melanoma mutation in *BRAF exon 15*, the activating mutation V600E, leads to response rates of more than 50% of all patients treated with the specific TKI Vemurafenib. Further, therapy with Vemurafenib is associated with a relative reduction in death of 63% when compared to standard therapeutic regime with Dacarbazine [62]. In contrast to the activating biology of *BRAF V600E*, ongoing clinical trials investigate the potency of Dasatinib in tumors harboring inactivating *BRAF exon 11* mutations (NCT01514864 clinicaltrials.gov).

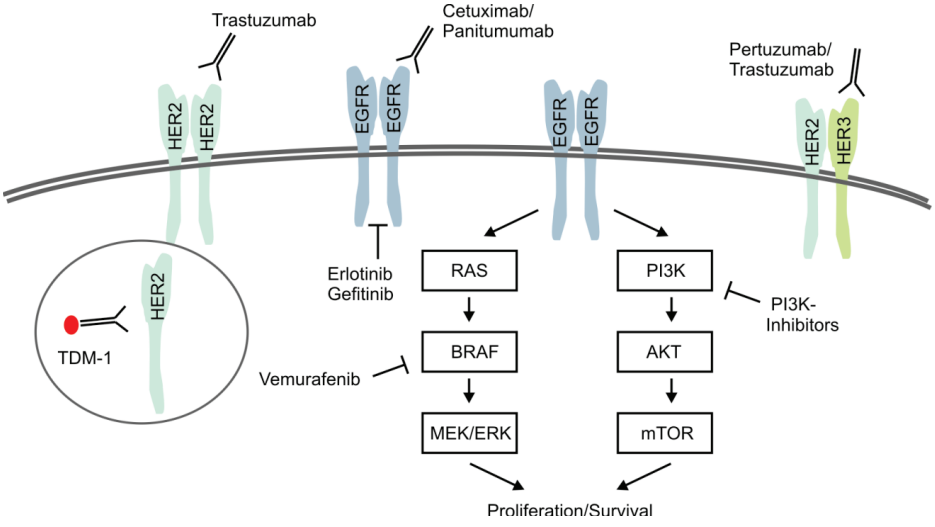
3.3. EGFR-Family and KRAS Mutational Status

Genetic aberrations in members of the *EGFR*-family are well known for targeted therapies, including HER2- and EGFR-targeted inhibition of downstream signaling cascades. *HER2/ERBB2* is primary known as being amplified and activated in breast cancer causing high recurrence rates and increased mortality in approximately 15% of all patients [77]. Patients with amplification of *HER2/ERBB2* treated with the monoclonal antibody Trastuzumab in combination with chemotherapy showed improved outcome in several studies [78]. Beyond Trastuzumab several ongoing studies

are investigating further drug therapies targeting the HER2-axis. These include the combined inhibition of HER2/HER3-heterodimerization and activation by Trastuzumab/Pertuzumab [79] and the use of the covalent immunoconjugate Trastuzumab-Emtansine (T-DM1) [80] (Figure 4). Moreover, several trials evaluate the therapeutic significance of small molecule inhibition in *HER2*-positive breast cancer, e.g., Lapatinib, Afatinib, Pazopanib and Neratinib [81].

Comparable to the situation in *HER2*-amplified breast cancer, substantial progress has been made by the introduction of EGFR-targeted therapies in the treatment of lung cancer and colorectal cancer. These efforts become evident by comparing the median overall survival of lung adenocarcinoma patients under standard therapeutic regimes of approximately 12 months with approximately 2 years under EGFR-targeted therapy with Erlotinib or Gefitinib in *EGFR*-mutated cancers [82]. As depicted above, further efforts are made to overcome primary and secondary therapeutic resistance by next generation TKIs [70]. Similar positive achievements were made for treatment of metastatic colorectal cancer (mCRC) by an improvement of survival from 12 months with fluoruracil monotherapy up to approximately 2 years with EGFR/VEGFR-targeted therapy combined with chemotherapy [83].

Figure 4. Predictive Biomarkers for Targeted and Selective Therapies. Signaling of EGFR-family receptors is characterized by homo-/heterodimerization and subsequent activation of the targetable downstream signaling pathways RAS/RAF and PI3K/AKT. Present therapeutic approaches focus on the inhibition of ligand-dependent activation, dimerization and receptor tyrosine kinases. Immunoconjugates, e.g., T-DM1, specifically deliver chemotherapeutic agents by the process of receptor internalization. As described in the text in more detail, ongoing efforts investigate the effectiveness of combined or dual approaches.



Beyond that, it recently became evident that we need to predict therapeutic response to cetuximab/panitumumab in mCRC not only by *KRAS* mutational status, but also by *NRAS*

mutational status [84], highlighting the increasing importance of mutations in downstream or interacting pathways. As depicted above, it becomes also clear that combined approaches, like the inhibition of the EGFR/BRAF-axis in *BRAF V600E* mutated colorectal cancers [64], could be used to overcome primary resistance in histological subtypes.

4. Conclusions

The purpose of integrating pathogenetic and molecular information into disease classification systems, exemplified by the 2008 WHO classification of hematological malignancies, reflects the high clinical relevance for predicting therapy outcome and prognosis. The human genome project and emerging technologies in the last decade have led to fundamental pathogenetic breakthroughs, which substantially improved the translation into clinical practice and individual therapeutic possibilities. Altogether, this data underlines the significant influence of cancer genomics and the substantial increase in genomic information on the process of defining tumor entities and effective and selective treatment approaches.

Author Contributions

Wrote the paper: MK and RB.

Conflicts of Interest

MK declares no conflict of interest. RB is serving on Scientific Advisory Boards and has received lecture honoraria from AstraZeneca, Qiagen, Roche, Lilly, MerckSerono, Pfizer and is a co-founder and co-owner of Targos Molecular Pathology, Inc. RB received grant support from the Deutsche Forschungsgemeinschaft, Deutsche Krebshilfe and the Bundesministerium für Bildung und Forschung (BMBF).

References

1. Vardiman, J.W.; Thiele, J.; Arber, D.A.; Brunning, R.D.; Borowitz, M.J.; Porwit, A.; Harris, N.L.; le Beau, M.M.; Hellstrom-Lindberg, E.; Tefferi, A.; *et al*. The 2008 revision of the world health organization (who) classification of myeloid neoplasms and acute leukemia: Rationale and important changes. *Blood* **2009**, *114*, 937–951.
2. Cross, N.C.; Reiter, A. Fibroblast growth factor receptor and platelet-derived growth factor receptor abnormalities in eosinophilic myeloproliferative disorders. *Acta Haematol.* **2008**, *119*, 199–206.
3. Nowell, P.C.; Hungerford, D.A. Chromosome studies on normal and leukemic human leukocytes. *J. Natl. Cancer Inst.* **1960**, *25*, 85–109.
4. Kantarjian, H.; Sawyers, C.; Hochhaus, A.; Guilhot, F.; Schiffer, C.; Gambacorti-Passerini, C.; Niederwieser, D.; Resta, D.; Capdeville, R.; Zoellner, U.; *et al*. Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. *N. Engl. J. Med.* **2002**, *346*, 645–652.

5. Tefferi, A.; Skoda, R.; Vardiman, J.W. Myeloproliferative neoplasms: Contemporary diagnosis using histology and genetics. *Nat. Rev. Clin. Oncol.* **2009**, *6*, 627–637.
6. Vannucchi, A.M.; Guglielmelli, P.; Tefferi, A. Advances in understanding and management of myeloproliferative neoplasms. *CA Cancer J. Clin.* **2009**, *59*, 171–191.
7. Levine, R.L.; Wadleigh, M.; Cools, J.; Ebert, B.L.; Wernig, G.; Huntly, B.J.; Boggon, T.J.; Wlodarska, I.; Clark, J.J.; Moore, S.; *et al*. Activating mutation in the tyrosine kinase jak2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell* **2005**, *7*, 387–397.
8. Kralovics, R.; Passamonti, F.; Buser, A.S.; Teo, S.S.; Tiedt, R.; Passweg, J.R.; Tichelli, A.; Cazzola, M.; Skoda, R.C. A gain-of-function mutation of jak2 in myeloproliferative disorders. *N. Engl. J. Med.* **2005**, *352*, 1779–1790.
9. James, C.; Ugo, V.; Le Couedic, J.P.; Staerk, J.; Delhommeau, F.; Lacout, C.; Garcon, L.; Raslova, H.; Berger, R.; Bennaceur-Griscelli, A.; *et al*. A unique clonal jak2 mutation leading to constitutive signalling causes polycythaemia vera. *Nature* **2005**, *434*, 1144–1148.
10. Szpurka, H.; Tiu, R.; Murugesan, G.; Aboudola, S.; Hsi, E.D.; Theil, K.S.; Sekeres, M.A.; Maciejewski, J.P. Refractory anemia with ringed sideroblasts associated with marked thrombocytosis (rars-t), another myeloproliferative condition characterized by jak2 v617f mutation. *Blood* **2006**, *108*, 2173–2181.
11. Nishii, K.; Nanbu, R.; Lorenzo, V.F.; Monma, F.; Kato, K.; Ryuu, H.; Katayama, N. Expression of the jak2 v617f mutation is not found in de novo aml and mds but is detected in mds-derived leukemia of megakaryoblastic nature. *Leukemia* **2007**, *21*, 1337–1338.
12. Scott, L.M.; Tong, W.; Levine, R.L.; Scott, M.A.; Beer, P.A.; Stratton, M.R.; Futreal, P.A.; Erber, W.N.; McMullin, M.F.; Harrison, C.N.; *et al*. Jak2 exon 12 mutations in polycythemia vera and idiopathic erythrocytosis. *N. Engl. J. Med.* **2007**, *356*, 459–468.
13. Tefferi, A.; Vainchenker, W. Myeloproliferative neoplasms: Molecular pathophysiology, essential clinical understanding, and treatment strategies. *J. Clin. Oncol.* **2011**, *29*, 573–582.
14. Klampfl, T.; Gisslinger, H.; Harutyunyan, A.S.; Nivarthi, H.; Rumi, E.; Milosevic, J.D.; Them, N.C.; Berg, T.; Gisslinger, B.; Pietra, D.; *et al*. Somatic mutations of calreticulin in myeloproliferative neoplasms. *N. Engl. J. Med.* **2013**, *369*, 2379–2390.
15. Pardanani, A.D.; Levine, R.L.; Lasho, T.; Pikman, Y.; Mesa, R.A.; Wadleigh, M.; Steensma, D.P.; Elliott, M.A.; Wolanskyj, A.P.; Hogan, W.J.; *et al*. Mpl515 mutations in myeloproliferative and other myeloid disorders: A study of 1182 patients. *Blood* **2006**, *108*, 3472–3476.
16. Gotlib, J. World health organization-defined eosinophilic disorders: 2011 update on diagnosis, risk stratification, and management. *Am. J. Hematol.* **2011**, *86*, 677–688.
17. Siegel, R.; Naishadham, D.; Jemal, A. Cancer statistics, 2013. *CA Cancer J. Clin.* **2013**, *63*, 11–30.
18. Buettner, R.; Wolf, J.; Thomas, R.K. Lessons learned from lung cancer genomics: The emerging concept of individualized diagnostics and treatment. *J. Clin. Oncol.* **2013**, *31*, 1858–1865.

19. Travis, W.D.; Brambilla, E.; Noguchi, M.; Nicholson, A.G.; Geisinger, K.R.; Yatabe, Y.; Beer, D.G.; Powell, C.A.; Riely, G.J.; van Schil, P.E.; *et al*. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *J. Thorac. Oncol.* **2011**, *6*, 244–285.
20. Hata, A.; Katakami, N.; Fujita, S.; Kaji, R.; Imai, Y.; Takahashi, Y.; Nishimura, T.; Tomii, K.; Ishihara, K. Frequency of egfr and kras mutations in japanese patients with lung adenocarcinoma with features of the mucinous subtype of bronchioloalveolar carcinoma. *J. Thorac. Oncol.* **2010**, *5*, 1197–1200.
21. Roberts, P.J.; Stinchcombe, T.E. Kras mutation: Should we test for it, and does it matter? *J. Clin. Oncol.* **2013**, *31*, 1112–1121.
22. Rosell, R.; Carcereny, E.; Gervais, R.; Vergnenegre, A.; Massuti, B.; Felip, E.; Palmero, R.; Garcia-Gomez, R.; Pallares, C.; Sanchez, J.M.; *et al*. Erlotinib *versus* standard chemotherapy as first-line treatment for european patients with advanced egfr mutation-positive non-small-cell lung cancer (eurtac): A multicentre, open-label, randomised phase 3 trial. *Lancet Oncol.* **2012**, *13*, 239–246.
23. Kwak, E.L.; Bang, Y.J.; Camidge, D.R.; Shaw, A.T.; Solomon, B.; Maki, R.G.; Ou, S.H.; Dezube, B.J.; Janne, P.A.; Costa, D.B.; *et al*. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N. Engl. J. Med.* **2010**, *363*, 1693–1703.
24. Doebele, R.C. Targeted therapies: Time to shift the burden of proof for oncogene-positive cancer? *Nat. Rev. Clin. Oncol.* **2013**, *10*, 492–493.
25. Shaw, A.T.; Kim, D.W.; Mehra, R.; Tan, D.S.; Felip, E.; Chow, L.Q.; Camidge, D.R.; Vansteenkiste, J.; Sharma, S.; de Pas, T.; *et al*. Ceritinib in alk-rearranged non-small-cell lung cancer. *N. Engl. J. Med.* **2014**, *370*, 1189–1197.
26. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **2012**, *489*, 519–525.
27. Weiss, J.; Sos, M.L.; Seidel, D.; Peifer, M.; Zander, T.; Heuckmann, J.M.; Ullrich, R.T.; Menon, R.; Maier, S.; Soltermann, A.; *et al*. Frequent and focal fgfr1 amplification associates with therapeutically tractable fgfr1 dependency in squamous cell lung cancer. *Sci. Transl. Med.* **2010**, *2*, 62ra93.
28. Hammerman, P.S.; Sos, M.L.; Ramos, A.H.; Xu, C.; Dutt, A.; Zhou, W.; Brace, L.E.; Woods, B.A.; Lin, W.; Zhang, J.; *et al*. Mutations in the ddr2 kinase gene identify a novel therapeutic target in squamous cell lung cancer. *Cancer Discov.* **2011**, *1*, 78–89.
29. Pitini, V.; Arrigo, C.; di Mirto, C.; Mondello, P.; Altavilla, G. Response to dasatinib in a patient with sqcc of the lung harboring a discoid-receptor-2 and synchronous chronic myelogenous leukemia. *Lung Cancer* **2013**, *82*, 171–172.
30. Peifer, M.; Fernandez-Cuesta, L.; Sos, M.L.; George, J.; Seidel, D.; Kasper, L.H.; Plenker, D.; Leenders, F.; Sun, R.; Zander, T.; *et al*. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat. Genet.* **2012**, *44*, 1104–1110.
31. Fernandez-Cuesta, L.; Peifer, M.; Lu, X.; Sun, R.; Ozretic, L.; Seidel, D.; Zander, T.; Leenders, F.; George, J.; Muller, C.; *et al*. Frequent mutations in chromatin-remodelling genes in pulmonary carcinoids. *Nat. Commun.* **2014**, *5*, doi:10.1038/ncomms4518.

32. McFadden, D.G.; Papagiannakopoulos, T.; Taylor-Weiner, A.; Stewart, C.; Carter, S.L.; Cibulskis, K.; Bhutkar, A.; McKenna, A.; Dooley, A.; Vernon, A.; *et al*. Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. *Cell* **2014**, *156*, 1298–1311.
33. The Clinical Lung Cancer Genome Project (CLCGP); Network Genomic Medicine (NGM). A genomics-based classification of human lung tumors. *Sci. Transl. Med.* **2013**, *5*, 209ra153.
34. Taylor, B.S.; Schultz, N.; Hieronymus, H.; Gopalan, A.; Xiao, Y.; Carver, B.S.; Arora, V.K.; Kaushik, P.; Cerami, E.; Reva, B.; *et al*. Integrative genomic profiling of human prostate cancer. *Cancer Cell* **2010**, *18*, 11–22.
35. Barbieri, C.E.; Baca, S.C.; Lawrence, M.S.; Demichelis, F.; Blattner, M.; Theurillat, J.P.; White, T.A.; Stojanov, P.; van Allen, E.; Stransky, N.; *et al*. Exome sequencing identifies recurrent spop, foxa1 and med12 mutations in prostate cancer. *Nat. Genet.* **2012**, *44*, 685–689.
36. Baca, S.C.; Prandi, D.; Lawrence, M.S.; Mosquera, J.M.; Romanel, A.; Drier, Y.; Park, K.; Kitabayashi, N.; MacDonald, T.Y.; Ghandi, M.; *et al*. Punctuated evolution of prostate cancer genomes. *Cell* **2013**, *153*, 666–677.
37. Tomlins, S.A.; Rhodes, D.R.; Perner, S.; Dhanasekaran, S.M.; Mehra, R.; Sun, X.W.; Varambally, S.; Cao, X.; Tchinda, J.; Kuefer, R.; *et al*. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *Science* **2005**, *310*, 644–648.
38. Li, J.; Yen, C.; Liaw, D.; Podsypanina, K.; Bose, S.; Wang, S.I.; Puc, J.; Miliaresis, C.; Rodgers, L.; McCombie, R.; *et al*. Pten, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **1997**, *275*, 1943–1947.
39. Berger, M.F.; Lawrence, M.S.; Demichelis, F.; Drier, Y.; Cibulskis, K.; Sivachenko, A.Y.; Sboner, A.; Esgueva, R.; Pflueger, D.; Sougnez, C.; *et al*. The genomic complexity of primary human prostate cancer. *Nature* **2011**, *470*, 214–220.
40. Weischenfeldt, J.; Simon, R.; Feuerbach, L.; Schlangen, K.; Weichenhan, D.; Minner, S.; Wuttig, D.; Warnatz, H.J.; Stehr, H.; Rausch, T.; *et al*. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell* **2013**, *23*, 159–170.
41. Grasso, C.S.; Wu, Y.M.; Robinson, D.R.; Cao, X.; Dhanasekaran, S.M.; Khan, A.P.; Quist, M.J.; Jing, X.; Lonigro, R.J.; Brenner, J.C.; *et al*. The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **2012**, *487*, 239–243.
42. Brenner, J.C.; Ateeq, B.; Li, Y.; Yocum, A.K.; Cao, Q.; Asangani, I.A.; Patel, S.; Wang, X.; Liang, H.; Yu, J.; *et al*. Mechanistic rationale for inhibition of poly(adp-ribose) polymerase in ets gene fusion-positive prostate cancer. *Cancer Cell* **2011**, *19*, 664–678.
43. Ross, K.A.; Smyth, N.A.; Murawski, C.D.; Kennedy, J.G. The biology of ewing sarcoma. *ISRN Oncol.* **2013**, *2013*, doi:10.1155/2013/759725.
44. Riggi, N.; Cironi, L.; Suva, M.L.; Stamenkovic, I. Sarcomas: Genetics, signalling, and cellular origins. Part 1: The fellowship of tet. *J. Pathol.* **2007**, *213*, 4–20.
45. Demicco, E.G. Sarcoma diagnosis in the age of molecular pathology. *Adv. Anat. Pathol.* **2013**, *20*, 264–274.

46. Wang, W.L.; Patel, N.R.; Caragea, M.; Hogendoorn, P.C.; Lopez-Terrada, D.; Hornick, J.L.; Lazar, A.J. Expression of erg, an ets family transcription factor, identifies erg-rearranged ewing sarcoma. *Mod. Pathol.* **2012**, *25*, 1378–1383.
47. Machado, I.; Noguera, R.; Mateos, E.A.; Calabuig-Farinas, S.; Lopez, F.I.; Martinez, A.; Navarro, S.; Llombart-Bosch, A. The many faces of atypical ewing's sarcoma. A true entity mimicking sarcomas, carcinomas and lymphomas. *Virchows Arch.* **2011**, *458*, 281–290.
48. Taylor, B.S.; Barretina, J.; Maki, R.G.; Antonescu, C.R.; Singer, S.; Ladanyi, M. Advances in sarcoma genomics and new therapeutic targets. *Nat. Rev. Cancer* **2011**, *11*, 541–557.
49. Pierron, G.; Tirode, F.; Lucchesi, C.; Reynaud, S.; Ballet, S.; Cohen-Gogo, S.; Perrin, V.; Coindre, J.M.; Delattre, O. A new subtype of bone sarcoma defined by bcor-ccnb3 gene fusion. *Nat. Genet.* **2012**, *44*, 461–466.
50. Choi, E.Y.; Thomas, D.G.; McHugh, J.B.; Patel, R.M.; Roulston, D.; Schuetze, S.M.; Chugh, R.; Biermann, J.S.; Lucas, D.R. Undifferentiated small round cell sarcoma with t(4;19)(q35;q13.1) cic-dux4 fusion: A novel highly aggressive soft tissue tumor with distinctive histopathology. *Am. J. Surg. Pathol.* **2013**, *37*, 1379–1386.
51. Tanas, M.R.; Goldblum, J.R. Fluorescence *in situ* hybridization in the diagnosis of soft tissue neoplasms: A review. *Adv. Anat. Pathol.* **2009**, *16*, 383–391.
52. Trautmann, M.; Sievers, E.; Aretz, S.; Kindler, D.; Michels, S.; Friedrichs, N.; Renner, M.; Kirfel, J.; Steiner, S.; Huss, S.; *et al*. Ss18-ssx fusion protein-induced wnt/beta-catenin signaling is a therapeutic target in synovial sarcoma. *Oncogene* **2013**, doi:10.1038/onc.2013.443.
53. Shing, D.C.; McMullan, D.J.; Roberts, P.; Smith, K.; Chin, S.F.; Nicholson, J.; Tillman, R.M.; Ramani, P.; Cullinane, C.; Coleman, N. Fus/erg gene fusions in ewing's tumors. *Cancer Res.* **2003**, *63*, 4568–4576.
54. Leach, F.S.; Tokino, T.; Meltzer, P.; Burrell, M.; Oliner, J.D.; Smith, S.; Hill, D.E.; Sidransky, D.; Kinzler, K.W.; Vogelstein, B. P53 mutation and mdm2 amplification in human soft tissue sarcomas. *Cancer Res.* **1993**, *53*, 2231–2234.
55. Pilotti, S.; Della Torre, G.; Lavarino, C.; Sozzi, G.; Minoletti, F.; Vergani, B.; Azzarelli, A.; Rilke, F.; Pierotti, M.A. Molecular abnormalities in liposarcoma: Role of mdm2 and cdk4-containing amplicons at 12q13–22. *J. Pathol.* **1998**, *185*, 188–190.
56. Shangary, S.; Wang, S. Small-molecule inhibitors of the mdm2-p53 protein-protein interaction to reactivate p53 function: A novel approach for cancer therapy. *Annu. Rev. Pharmacol. Toxicol.* **2009**, *49*, 223–241.
57. Pavlidis, N.; Pentheroudakis, G. Cancer of unknown primary site. *Lancet* **2012**, *379*, 1428–1435.
58. Massard, C.; Loriot, Y.; Fizazi, K. Carcinomas of an unknown primary origin—Diagnosis and treatment. *Nat. Rev. Clin. Oncol.* **2011**, *8*, 701–710.
59. Stella, G.M.; Senetta, R.; Cassenti, A.; Ronco, M.; Cassoni, P. Cancers of unknown primary origin: Current perspectives and future therapeutic strategies. *J. Transl. Med.* **2012**, *10*, 12.
60. Tothill, R.W.; Li, J.; Mileshkin, L.; Doig, K.; Siganakis, T.; Cowin, P.; Fellowes, A.; Semple, T.; Fox, S.; Byron, K.; *et al*. Massively-parallel sequencing assists the diagnosis and guided treatment of cancers of unknown primary. *J. Pathol.* **2013**, *231*, 413–423.

61. Tan, D.S.; Montoya, J.; Ng, Q.S.; Chan, K.S.; Lynette, O.; Sakktee Krisna, S.; Takano, A.; Lim, W.T.; Tan, E.H.; Lim, K.H. Molecular profiling for druggable genetic abnormalities in carcinoma of unknown primary. *J. Clin. Oncol.* **2013**, *31*, e237–e239.
62. Chapman, P.B.; Hauschild, A.; Robert, C.; Haanen, J.B.; Ascierto, P.; Larkin, J.; Dummer, R.; Garbe, C.; Testori, A.; Maio, M.; *et al*. Improved survival with vemurafenib in melanoma with braf v600e mutation. *N. Engl. J. Med.* **2011**, *364*, 2507–2516.
63. Flaherty, K.T.; Puzanov, I.; Kim, K.B.; Ribas, A.; McArthur, G.A.; Sosman, J.A.; O'Dwyer, P.J.; Lee, R.J.; Grippo, J.F.; Nolop, K.; *et al*. Inhibition of mutated, activated braf in metastatic melanoma. *N. Engl. J. Med.* **2010**, *363*, 809–819.
64. Prahallad, A.; Sun, C.; Huang, S.; di Nicolantonio, F.; Salazar, R.; Zecchin, D.; Beijersbergen, R.L.; Bardelli, A.; Bernards, R. Unresponsiveness of colon cancer to braf(v600e) inhibition through feedback activation of egfr. *Nature* **2012**, *483*, 100–103.
65. Hanfstein, B.; Muller, M.C.; Hehlmann, R.; Erben, P.; Lauseker, M.; Fabarius, A.; Schnittger, S.; Haferlach, C.; Gohring, G.; Proetel, U.; *et al*. Early molecular and cytogenetic response is predictive for long-term progression-free and overall survival in chronic myeloid leukemia (cml). *Leukemia* **2012**, *26*, 2096–2102.
66. Oehler, V.G. Update on current monitoring recommendations in chronic myeloid leukemia: Practical points for clinical practice. *Hematol. Am. Soc. Hematol. Educ. Program* **2013**, *2013*, 176–183.
67. Merx, K.; Muller, M.C.; Kreil, S.; Lahaye, T.; Paschka, P.; Schoch, C.; Weisser, A.; Kuhn, C.; Berger, U.; Gschaidmeier, H.; *et al*. Early reduction of bcr-abl mrna transcript levels predicts cytogenetic response in chronic phase cml patients treated with imatinib after failure of interferon alpha. *Leukemia* **2002**, *16*, 1579–1583.
68. Cross, N.C.; White, H.E.; Muller, M.C.; Saglio, G.; Hochhaus, A. Standardized definitions of molecular response in chronic myeloid leukemia. *Leukemia* **2012**, *26*, 2172–2175.
69. Maheswaran, S.; Sequist, L.V.; Nagrath, S.; Ulkus, L.; Brannigan, B.; Collura, C.V.; Inserra, E.; Diederichs, S.; Iafrate, A.J.; Bell, D.W.; *et al*. Detection of mutations in egfr in circulating lung-cancer cells. *N. Engl. J. Med.* **2008**, *359*, 366–377.
70. Walter, A.O.; Sjin, R.T.; Haringsma, H.J.; Ohashi, K.; Sun, J.; Lee, K.; Dubrovskiy, A.; Labenski, M.; Zhu, Z.; Wang, Z.; *et al*. Discovery of a mutant-selective covalent inhibitor of egfr that overcomes t790m-mediated resistance in nsclc. *Cancer Discov.* **2013**, *3*, 1404–1415.
71. Forshew, T.; Murtaza, M.; Parkinson, C.; Gale, D.; Tsui, D.W.; Kaper, F.; Dawson, S.J.; Piskorz, A.M.; Jimenez-Linan, M.; Bentley, D.; *et al*. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **2012**, *4*, 136ra168.
72. Tomlins, S.A.; Aubin, S.M.; Siddiqui, J.; Lonigro, R.J.; Sefton-Miller, L.; Miick, S.; Williamsen, S.; Hodge, P.; Meinke, J.; Blase, A.; *et al*. Urine tmprss2:Erg fusion transcript stratifies prostate cancer risk in men with elevated serum psa. *Sci. Transl. Med.* **2011**, *3*, 94ra72.

73. Druker, B.J.; Tamura, S.; Buchdunger, E.; Ohno, S.; Segal, G.M.; Fanning, S.; Zimmermann, J.; Lydon, N.B. Effects of a selective inhibitor of the abl tyrosine kinase on the growth of bcr-abl positive cells. *Nat. Med.* **1996**, *2*, 561–566.
74. Kantarjian, H.; O'Brien, S.; Jabbour, E.; Garcia-Manero, G.; Quintas-Cardama, A.; Shan, J.; Rios, M.B.; Ravandi, F.; Faderl, S.; Kadia, T.; *et al*. Improved survival in chronic myeloid leukemia since the introduction of imatinib therapy: A single-institution historical experience. *Blood* **2012**, *119*, 1981–1987.
75. Baccarani, M.; Deininger, M.W.; Rosti, G.; Hochhaus, A.; Soverini, S.; Apperley, J.F.; Cervantes, F.; Clark, R.E.; Cortes, J.E.; Guilhot, F.; *et al*. European leukemianet recommendations for the management of chronic myeloid leukemia: 2013. *Blood* **2013**, *122*, 872–884.
76. Mahon, F.X. Is going for cure in chronic myeloid leukemia possible and justifiable? *Hematol. Am. Soc. Hematol. Educ. Program* **2012**, *2012*, 122–128.
77. Ross, J.S.; Slodkowska, E.A.; Symmans, W.F.; Pusztai, L.; Ravdin, P.M.; Hortobagyi, G.N. The her-2 receptor and breast cancer: Ten years of targeted anti-her-2 therapy and personalized medicine. *Oncologist* **2009**, *14*, 320–368.
78. Arteaga, C.L.; Sliwkowski, M.X.; Osborne, C.K.; Perez, E.A.; Puglisi, F.; Gianni, L. Treatment of her2-positive breast cancer: Current status and future perspectives. *Nat. Rev. Clin. Oncol.* **2012**, *9*, 16–32.
79. Swain, S.M.; Kim, S.B.; Cortes, J.; Ro, J.; Semiglazov, V.; Campone, M.; Ciruelos, E.; Ferrero, J.M.; Schneeweiss, A.; Knott, A.; *et al*. Pertuzumab, trastuzumab, and docetaxel for her2-positive metastatic breast cancer (cleopatra study): Overall survival results from a randomised, double-blind, placebo-controlled, phase 3 study. *Lancet Oncol.* **2013**, *14*, 461–471.
80. Hurvitz, S.A.; Dirix, L.; Kocsis, J.; Bianchi, G.V.; Lu, J.; Vinholes, J.; Guardino, E.; Song, C.; Tong, B.; Ng, V.; *et al*. Phase ii randomized study of trastuzumab emtansine *versus* trastuzumab plus docetaxel in patients with human epidermal growth factor receptor 2-positive metastatic breast cancer. *J. Clin. Oncol.* **2013**, *31*, 1157–1163.
81. Incorvati, J.A.; Shah, S.; Mu, Y.; Lu, J. Targeted therapy for her2 positive breast cancer. *J. Hematol. Oncol.* **2013**, *6*, 38.
82. Pao, W.; Chmielecki, J. Rational, biologically based treatment of egfr-mutant non-small-cell lung cancer. *Nat. Rev. Cancer* **2010**, *10*, 760–774.
83. Van Cutsem, E.; Kohne, C.H.; Hitre, E.; Zaluski, J.; Chang Chien, C.R.; Makhson, A.; D'Haens, G.; Pinter, T.; Lim, R.; Bodoky, G.; *et al*. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *N. Engl. J. Med.* **2009**, *360*, 1408–1417.
84. Douillard, J.Y.; Oliner, K.S.; Siena, S.; Tabernero, J.; Burkes, R.; Barugel, M.; Humblet, Y.; Bodoky, G.; Cunningham, D.; Jassem, J.; *et al*. Panitumumab-folfox4 treatment and ras mutations in colorectal cancer. *N. Engl. J. Med.* **2013**, *369*, 1023–1034.

GWAS to Sequencing: Divergence in Study Design and Analysis

Christopher Ryan King and Dan L. Nicolae

Abstract: The success of genome-wide association studies (GWAS) in uncovering genetic risk factors for complex traits has generated great promise for the complete data generated by sequencing. The bumpy transition from GWAS to whole-exome or whole-genome association studies (WGAS) based on sequencing investigations has highlighted important differences in analysis and interpretation. We show how the loss in power due to the allele frequency spectrum targeted by sequencing is difficult to compensate for with realistic effect sizes and point to study designs that may help. We discuss several issues in interpreting the results, including a special case of the winner's curse. Extrapolation and prediction using rare SNPs is complex, because of the selective ascertainment of SNPs in case-control studies and the low amount of information at each SNP, and naive procedures are biased under the alternative. We also discuss the challenges in tuning gene-based tests and accounting for multiple testing when genes have very different sets of SNPs. The examples we emphasize in this paper highlight the difficult road we must travel for a two-letter switch.

Reprinted from *Genes*. Cite as: King, C.R.; Nicolae, D.L. GWAS to Sequencing: Divergence in Study Design and Analysis. *Genes* **2014**, *5*, 460–476.

1. Introduction

The Human Genome Project has paved the way to the data revolution in complex disease genetics, by permitting the development of databases of genetic variation, such as HapMap [1], and machinery for producing genome-wide data, such as genotyping arrays and high-throughput sequencing technologies. Our understanding of the genetic risk factors for complex traits has evolved from a few loci discovered with positional cloning approaches in the 1990s to thousands of replicated associations from genome-wide association studies (GWAS), available to the public, as well as scientists in the NHGRI catalog [2]. Interest has shifted recently to discovering disease association with data from whole-genome or whole-exome sequencing studies, and so far, these have had limited success. GWAS has delivered on their early promise to speed up the search for disease genes, and there are bold predictions about what sequencing can achieve [3] on the way to the era of personalized medicine. Sequencing could offer a complete picture of genetic variation—from SNPs to Copy Number Variants (CNVs) and insertions-deletions—for the subjects in the study and for future patients and has led to successful discoveries in Mendelian diseases. So far, sequencing has had limited success for complex diseases, mostly in candidate gene studies. Whole-genome and whole-exome sequencing investigations have only demonstrated the complicated architecture of common traits, sometimes indirectly through a lack of findings in single-SNP low-frequency analyses.

The success of GWAS and of the corresponding analytical tools leads naturally to an investigation of what is different between the two strategies. The goal of this paper is to compare some of the divergent aspects of GWAS and sequencing studies with the hope of guiding future sequencing investigations. We focus on two key distinctions. First, we look at consequences that follow from investigating SNPs with low minor allele frequency (MAF), including the ability to detect novel SNPs. It is important to reiterate that GWAS analyses cover, directly (through genotyping or imputation) or indirectly (through linkage disequilibrium), most of the common variants in the studied populations. This implies that the goal of sequence-based studies is to detect association with low frequency and rare variants. Even though sequencing studies can be used to investigate high MAF SNPs, we ignore their role, since traditional genotyping is dramatically more cost effective. Furthermore, we do not discuss the fact that sequencing studies permit the investigation of structural variation, an important characteristic for diseases, such as autism, where these variants play an important role. In Section 2, we develop a simple analytical formula for the power of a burden test and use it to illustrate the factors affecting power with a contrast to GWAS and scenarios for improving them. In Section 3, we illustrate two novel problems with estimation and prediction using sequencing data. First, we show that case-control studies, which add rare SNPs into a super-SNP or test the distribution of case- and control-private SNPs, can be misleading if analyzed naively. Second, we show that the optimal prediction for previously observed and novel rare SNPs can be strikingly different.

Second, we turn to issues surrounding the use of gene-based tests. In Section 4, we discuss the difficulty of selecting and tuning gene-based test statistics and contrast this to the case in GWAS. We show the alternative hypothesis, which would recommend that a particular procedure can be quite unstable even with seemingly irrelevant details of a gene. We do not recommend a particular testing procedure, but highlight concerns guiding the tuning parameter selection. Finally, in Section 5, we highlight the sharp distinctions between multiple-testing-adjustment strategies for GWAS and gene-based tests.

2. Power of Sequencing *versus* GWAS

The relatively minor number of associations with rare variants seems surprising to many, but was predicted by prior knowledge on the genetics of complex phenotypes. For example, the lack of major linkage loci for diseases, like type 2 diabetes [4], suggests that there are no genes with many rare variants with very large effects. Given this lack of observed associations, it is useful to investigate the relative contributions of factors driving power. We will illustrate with a burden-style test for which an analytical power calculation is straightforward.

The goal here is not to calculate power nor to find realistic sample sizes for genetic association studies with rare variants. Existing software (e.g., [5]) can perform such calculations. Our aim is to use simple analytical calculations to gain insight into what drives power and what are possible strategies for designing optimal investigations. A comparison to GWAS will illustrate the challenges ahead of us. One important set of shared assumptions for GWAS and WGAS is that of the

unconfoundedness of associations. Recent work has suggested that approaches to adjusting for population structure, which work well in GWAS, may not in WGAS [6–8]. However, the literature on this topic is rapidly evolving, and we will set this problem aside for purposes of discussion.

Assume a balanced design with n cases and n controls. It can be shown (see Appendix A for the assumptions used in the derivation of this) that the non-centrality parameter (NCP) for burden tests [9,10] can be approximated by:

$$\sqrt{n}\,\frac{k_1}{\sqrt{k}}\,\frac{E_M}{\sqrt{V_M + E_M - E_M^2}}\,(\gamma - 1) \tag{1}$$

where the test is done on a set of k SNPs, out of which, k_1 are associated with a common odds-ratio (OR) of γ, and $E_M,\ V_M$ are the mean and variance of the minor allele frequency (MAF) for the SNPs in the set. This formula works for single SNP analyses, as well, with $k_1/\sqrt{k} = 1$ and the term about frequency replaced by the corresponding function of MAF. Note that the power of the test is approximately linear in the NCP in the interesting range of moderate values.

All the terms, except the one containing elements of the MAF distribution, are easy to calculate and interpret. The MAF term can be approximated using 1000 Genomes Project data and calculations conditional on an SNP being polymorphic in a study. For 5000 cases and 5000 controls of European descent, and filtering to SNPs with MAF $< 1\%$, that term is close to 0.046, and the non-centrality parameter when $k = 100$, $k_1 = 10$ and γ = 3 is approximately 6.47. Those settings yield a power of 87% at the genome-wide 5×10^{-8} significance level. We will discuss the four terms in Formula (1) and contrast the results between GWAS and sequencing.

Sample size: The simplest way to double the NCP is to increase the sample size by a factor of four. This requires the least amount of innovation, but takes a huge effort and expense, especially when using existing cohorts, since ascertaining and phenotyping additional samples comparable with existing data is very difficult. As is common with many GWAS meta-analyses, a cost-effective increase in the sample size requires the use of ancestry-diverse populations. Additional diversity increases heterogeneity and will affect power to a larger degree than in GWAS, both because the effective MAF decreases (many rare alleles are population-specific) and because a similarly defined set of SNPs (e.g., all exonic SNPs in a given gene) will have different elements in different populations, with powerful tests requiring the presence of functional/causal variants in each (sub)population. We also anticipate that cryptic gene-environment interactions (GxE) provides a substantial amount of heterogeneity in effect sizes. GxE has been long known to exist for some complex traits (e.g., for a review in psychiatric phenotypes, see [11]); given how difficult is it to anticipate relevant modifiers, measure them accurately and statistically detect them [12], it seems likely that unknown environmental modifiers are not uniformly distributed across populations. We will not expand on the difficulties inherent to adjusting for structure in diverse samples, but note that this is much more challenging in sequencing, since rare SNPs can be specific for relatively recent and small-scale demographic events [9,10].

Sparsity of signals and variant annotation: The next term in Formula (1) has to do with the number of associated SNPs relative to investigated SNPs, which we call the sparsity of the signal. Figure 1 shows the impact of sparsity on sample sizes needed to design powerful association studies. For

GWAS ($k_1 = k = 1$), the sparsity term is equal to one for associated SNPs. In sequencing studies, it is possible to increase the power by reducing the number of non-associated SNPs (for sets where k is large compared to k_1). The annotation of SNPs through functional status, eQTL (expression quantitative trait loci) studies, ENCODE, prior data, *etc.*, will allow more useful definitions for the analyzed sets by excluding SNPs with a low *a priori* likelihood of being associated. This is a fruitful area of current research and one that is implicit in some study designs, such as exome sequencing.

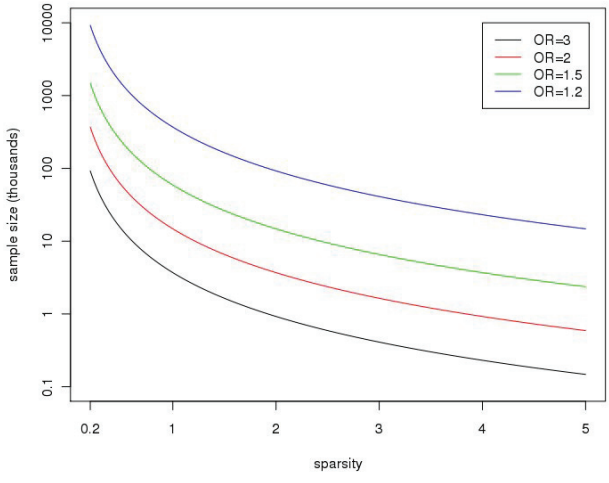


Figure 1. The plot shows the sample sizes (on the y-axis, in thousands) needed to achieve 80% power at the 10^{-6} significance level as a function of "sparsity", $k_1/\sqrt{k}$ (on the x-axis), with k and k_1 as defined in the text. It is assumed for these calculations that the k SNPs are independent (no linkage disequilibrium), with the minor allele frequency (MAF) sampled from a beta distribution with parameters selected to match allele frequencies from the CEU of the 1000 Genomes Project, $B(0.14,0.73)$; the distribution is truncated at 0.01 (so SNPs have MAF $< 1\%$) and only polymorphic SNPs when sequencing 10,000 subjects are selected. Calculations are based on the NCP in Equation (1). OR, odds ratio.

The MAF distribution: The dominant term in the denominator of Formula (1) is given by the mean MAF, so a simple approximation to the third component of NCP is $\sqrt{E_M}$. This is the term in NCP that explains most of the difference between sequence association and GWAS. The corresponding term for a single-SNP test with a risk allele frequency of 0.2 is approximately 10 times larger than our 1000 Genomes-based estimate. In order to have comparable power between sequencing and GWAS, this loss would have to be balanced by the other terms (sample size, sparsity and effect size). There is no easy strategy to increase this term in unrelated individuals, but one available route is to shift the study design to families or isolated populations, where alleles which are rare in the larger population are locally common.

Phenotyping/environment: We can also increase power by analyzing datasets with a larger effect size; this corresponds to the last term in Formula (1), $(\gamma - 1)$. This can be done using stratified analysis: by analyzing sub-phenotypes and/or by accounting for environment (when GxE is present). This is a common issue for GWAS and sequencing, and we illustrate the impact of stratification on the effect size using a single SNP as the unit of analysis. Let us assume that γ_T is the mean effect corresponding to the cases in the most at risk strata (with the rest of cases being "controls" with respect to the variants in the set under investigation). Let α be the proportion of relevant cases, and let p be the control MAF. It follows that in the full set of cases (relevant and irrelevant), the MAF is approximated by $\alpha p\gamma_T+(1-\alpha)p$, and the corresponding effect size is $\gamma-1 \approx \alpha(\gamma_T-1)$. Therefore, if phenotyping or sub-setting by environment allows one to find the relevant cases, analyzing a smaller sample size (of αn) leads to an increase of $1/\alpha$ in the fourth term of NCP and to a $1/\sqrt{\alpha}$ overall increase in NCP.

3. Prediction Using Rare and Novel SNPs: A Different Winner's Curse

Aside from association discovery, one of the major goals of GWAS is to estimate the effect sizes of SNPs on traits, which can be used for the prediction of unrealized phenotypes on newly sequenced individuals. For example, SNP genotyping platforms have recently been used for risk and pharmacogenomic prediction by several companies, such as 23andMe, Life Technologies, and Pathway Genomics. Prediction using SNPs discovered in a sequencing study can be performed analogously to GWAS, as long as adequate data has been gathered. One major difference between GWAS and sequencing is that newly-sequenced individuals will regularly carry novel SNPs in disease-associated genes, and most discovered SNPs will have too little information for accurate per-SNP estimates [13,14]. Quantitative estimates of personal risk based on sequencing association studies will therefore require an evidence-based estimate of the effect of previously unobserved and seldomly observed rare SNPs. Given that mutations in the gene in question have already been associated with disease and that harmful SNPs are thought to be more likely to be rare [15–20], ignoring these SNPs (setting the effect to zero) is unlikely to be accurate. Naively, we could estimate a "rare SNP effect" based on the rare SNPs observed in previous sequencing studies and apply that estimate to new SNPs and known rare SNPs alike. We illustrate two problems with no analog in GWAS that occur when rare SNPs are lumped into a super-SNP for estimation or prediction. The major results are that: (1) rare alleles in a sequencing study can cumulatively have a substantial per-allele OR, which depends on disease prevalence, even if log odds-ratios (lORs) are centered at zero; (2) the prediction of new samples based on that OR is substantially inaccurate.

First, unlike GWAS, prediction with new SNPs depends non-trivially on the variability of rare SNP effects. With GWAS, previous data will give the investigator an estimate of the effect of each SNP; a plug-in prediction can be formed using these estimates: $logit(\hat{Y}_i) = G_i\hat{\beta} + \alpha$, where $logit$ is the logistic function, $\hat{Y}_i$ is the predicted probability for person i, G_i is the vector of genotypes for that person, α is an intercept, which depends on disease frequency, and β are SNP lORs. The naive plug-in prediction is not quite correct due to the uncertainty in SNP effects; however, the

inaccuracy with GWAS-based estimates tends to be negligible for reasons discussed below. In contrast, the effect of a rare SNP in an associated gene is not precisely known, and the impact of that uncertainty on prediction is substantial. For example, even if SNPs in an associated gene are as likely to be risk-decreasing as risk-increasing, the correct prediction in the context of a rare disease for a newly sequenced individual is that carrying a novel SNP increases their odds of being affected. Qualitatively, the uncertainty in SNP effects makes one less confident in the plug-in estimate and pushes the best estimate from the raw prevalence closer to 50:50. To give a numerical example, if the population of lORs for novel SNPs is Gaussian, with a mean of zero and standard deviation of one, and the disease frequency is 1%, then the marginal OR for carrying an allele (*versus* no minor alleles) is 1.9.

This is a well-known phenomenon from the literature comparing marginal and conditional random effects [21–23]; derivation of the effect size and additional explanation is offered in the Appendix. A useful formula for the risk associated with carrying new SNPs can be derived under the assumption that their lORs are Gaussian distributed with mean μ and standard deviation σ along with standard logistic regression assumptions. Define c as a constant related to the disease prevalence (the threshold in the Appendix $c = -log(\frac{p}{1-p}) \approx -log(p)$) and g_i as the number of alleles in that individual, then:

$$logit\left(Pr\{Y_i = 1|g_i, \mu, \sigma\}\right) \approx \frac{-c + \mu g_i}{\sqrt{1 + \nu^2\sigma^2 g_i}} \tag{2}$$

where $\nu \approx 0.625$. When the mean lOR is zero and the standard deviation is not large (it is almost assuredly less than one in areas without overwhelming evidence for linkage), the lOR for having an SNP (*versus* no SNPs) is approximately $\frac{c\nu^2\sigma^2}{2\sqrt{1+\nu^2\sigma^2}}$, which increases sharply with the standard deviation of lORs and scales with the negative log of disease prevalence. This is the lOR that we estimate when regressing the outcome on the number of SNPs carried and that we would use for prediction, absent other information about the effects of particular SNPs.

A related result occurs for GWAS-based plug-in estimates, the details of which depend on the choice of statistical estimators used for effect estimates and the pattern of linkage disequilibrium. For any estimate of an SNP's lOR for which a central limit theorem applies, σ^2 in Equation (2) can be replaced with the square of the standard error of the estimate, which will usually scale as the inverse of the sample size and the MAF. Given the relatively low cost of GWAS data acquisition and the need to overcome the burden of genome-wide multiple testing, we are accustomed to gathering enough data for precise estimates. For example, the expected standard error of a lOR of zero with a MAF of 0.3 and 3000 cases and 3000 controls is 0.06. If the standard deviation of the population of novel SNP lORs is 0.5, then the marginal lOR for a new SNP is 70 times bigger than the previously observed SNP.

The above effect is observable regardless of the sample size and MAF of SNPs used in the calculation. One might expect that since case control-based estimates of ORs are consistent for prospective associations, that this effect would be corrected by empirically estimating the per-allele OR and using that for future data; however, there is a unique twist for the group of SNPs with MAFs, such that they are reasonably likely to be monomorphic in the original study. The observed lOR for

all rare SNPs together does estimate the marginal effect of future rare SNPs, but that prediction breaks down when stratified by whether or not the SNP was observed as polymorphic in the case-control study. Case-control designs are somewhat more efficient for discovering rare risk-increasing SNPs compared to risk-decreasing SNPs [20,24–27], so as a group, previously observed SNPs are more harmful in future samples than newly observed SNPs.

In Figure 2, we plot how the probability of an SNP being discovered (the minor allele is observed in at least one participant) in a case-control study depends on both the OR and the MAF when the MAF is low compared to the sample size. The absolute probability of a SNP appearing at least once in the study increases markedly with OR in this range of MAF. Intuitively, compared to a population sample, a risk-increasing SNP's greater frequency among cases more than makes up for the decline in its frequency among controls. As a result, the observed odds of a rare SNP appearing in a case are inflated, and the finite pool of remaining SNPs at that MAF contains a preponderance of protective and small effects.

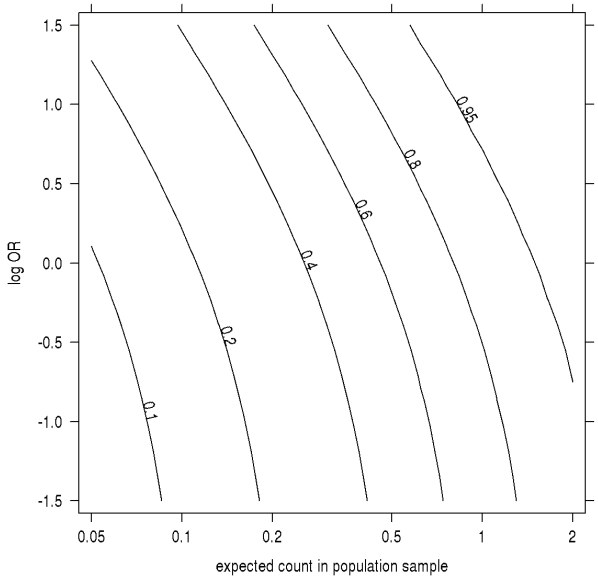


Figure 2. Sampling probability by MAF, log odds-ratio. The contour plot has on the x-axis the allelic expected count in a population sample the same size as the control group (sample sizes times MAF) and, on the y-axis, the log-odds ratio. Contours are the absolute probability of being sampled in a case-control study of 100 cases and 100 controls when prevalence equals 1%.

This is similar to the first problem discussed above, except that we have selectively observed SNPs based on their true odds ratio. Figure 3 shows that when the lORs of SNPs in a gene are assumed to come from a population with high variance, the odds of an SNP appearing only in cases and the expected lOR of observed SNPs varies with MAF and substantially favors an increase in risk. This

is not a Bayesian argument; it relies only on the rate at which SNPs appear in the sample, even for fixed SNP effects. To give a numerical example, with a sample size of 100, a prevalence of 5%, 125 SNPs with a MAF of 0.002 and lORs drawn from a standard normal, an average of 36% of the SNPs are discovered in the original sequencing study with an average per-allele estimated lOR of 0.33. In a new replication or prediction sample, the average per-allele lOR based on previously discovered SNPs is 0.66, but the per-allele lOR of new SNPs is only 0.05. The numerical result depends heavily on a number of parameters; we have deferred a detailed exploration of the phenomenon to another work [28]. The longer report is available for download, and re-demonstration of the importance of each factor is beyond the scope of this paper. However, there are three notable features to which we wish to briefly draw attention: (1) the tail behavior of SNP lORs is very influential; large ORs enrich even rare SNPs into the population of cases; (2) the effect occurs at a MAF around the minimum observable in a study; regardless of the size of the original dataset observed, rare SNPs are unrepresentative of future rare SNPs; (3) the effect vanishes under the null hypothesis that no SNPs in a gene are associated.

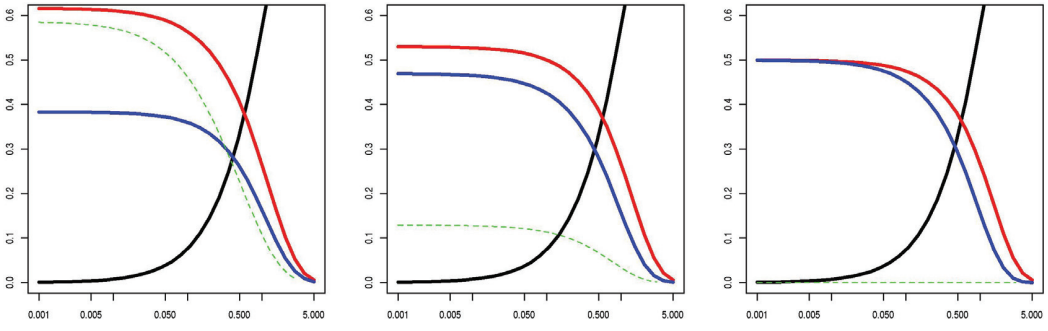


Figure 3. Observed data probabilities by MAF. X-axis N·MAF. The y-axis shows the probability of each special data type conditional on the SNP being polymorphic: occurring only in cases (red), once in controls and zero times in cases (blue) and all other (black). Green = expected log-odds-ratio (OR) of sampled SNPs (same numeric scale). The log-ORs are assumed to be distributed left: $N(0, 1)$; center: $N(0, 0.5^2)$; right: $N(0, 0.25^2)$; other settings are as in Figure 2.

4. Implicit and Explicit Models in Association Studies

In contrast to GWAS, genetics practitioners with sequencing data are currently faced with a dizzying selection of methods to test for an association between genotype and phenotype, each of which has tuning parameters. In GWAS, a simple allelic test is the overwhelmingly most commonly used test. The additive allelic model performs well regardless of the true risk model when linkage disequilibrium between a tested marker and causal allele is imperfect [29]. While some authors have suggested GWAS schemes that incorporate prior knowledge and more complex risk models through explicit Bayesian calculations or alpha-spending procedures (see also Section 5), the dominant

technique in the literature is to report SNP-level evidence and the allelic effect from a particular dataset. While the commonly used methods make usual regression-type assumptions about the distribution of the trait, the effects of confounders and covariates and the measurement error of SNPs, they make minimal assumptions about the effects of other SNPs or how effect size varies with SNP- or gene-level features.

Because of the small amount of information at each rare SNP, all sequencing association tests of which we are aware pool information in some way across SNPs, which are regarded as belonging to a unit (gene) or being "similar," and some pool information across genes that are "similar." These techniques have tuning parameters appropriate under a particular alternative hypothesis and that may suffer a substantial loss of power under other alternatives. A full comparison of proposed tests for sequencing data is beyond the scope of this article; however, we will discuss a few of the most common tests. In this section, we will discuss the role and meaning of some of these tuning parameters. Ignoring these tuning parameters as if the investigator were still using relatively assumption-free GWAS techniques is unlikely to work well, and the importance of these analytic decisions represents a substantial divergence from GWAS.

One extreme of this approach is to try to swap tuning parameters for explicit models and assumptions. We have advocated multi-level modeling of effect sizes or lORs using SNP- and gene-level features as predictors, with stated assumptions, such as the functional form of associations, the linearity and additivity of associations, distributional requirements and exchangeability between SNPs, where required [30].

However, most proposed tests are not model-based summaries. In some cases, we can gain insight into these tests by constructing a map from the tuning parameters to a genetic model, which would imply those as optimal in some way. For example, $C(\alpha)$ and diagonal kernel SKAT [5] can be derived from an explicit model with weights on the $j-th$ SNP, $w_j \propto E[lOR_j^2]$ [31], and therefore, any scheme of weights as a function of MAF can be understood in terms of the implied variance of SNP effects. The addition of correlation structures to SNP effects also follows a simple model-based logic; for example, if effects are expected to substantially go the same direction (such as a group of loss-of-function SNPs), one can balance between the burden-type and variance-component-type test [5]. Similarly, several authors [32,33] have pointed out that optimal weights for burden-type tests are proportional to per-SNP lORs. Figure 4 shows curves for three MAF-depended weights proposed in the literature [34]. Notably, the implied lOR curve depends heavily on the upper limit of MAF included in the pooling procedure.

However, the tuning parameters of some proposed tests are more challenging. For example, SKAT with the Gaussian kernel does not map to a meaningful model of SNP effects, but is suspected to work reasonably well under several alternatives and detects some non-linear effects and epistatic interactions [35–37]. The kernel itself is a tuning parameter for SKAT and related methods; kernel-based techniques are sensitive to the choice of kernel and, aside from a few special cases [35,38], are difficult to choose between *a priori*. Additionally, while most kernels can up- or down-weight SNPs, transforming prior information into a calibrated SNP "similarity" or "distance" measure is a task for which we have little guidance. While SKAT and related tests can be motivated by

a simple variance components model, they are not automatically robust to non-Gaussian SNP effects, such as a mixture of causal and non-causal alleles [39,40]. This is not to suggest that the many variants of kernel-based tests are poorly applied tools or that they perform poorly compared to other tests, just to highlight the fundamental difficulty of interpreting and guiding the key analytic decision.

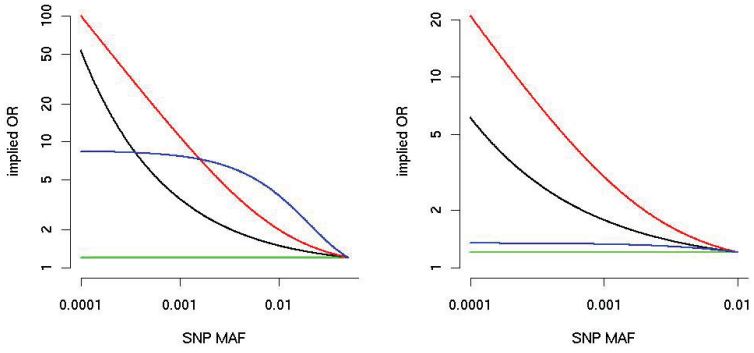


Figure 4. Implied alternative OR (on the y-axis, logarithmic scale) as a function of MAF (x-axis) for three burden weighting schemes. The black line corresponds to the Madsen–Browning weight [10]; the red line corresponds to the attributable risk weight [41], and the blue line corresponds to the default in SKAT, Beta(25,1) [42]; the green line is for equal weights. For the **left** panel, the MAF is truncated at 5%, and for the **right** panel at 1%. We assume that the OR of the SNP with the largest MAF is 1.2.

There are numerous specific deviations from linear Gaussian SNP effects, which hypothetically should influence the tuning parameter selection. In the implicit model tests described above, these issues are difficult to address in planning and power analysis. When considering sequencing data as potential negative evidence in replication studies, each has to be explored on a case-by-case basis. Explicit-model methods have the advantage of facilitating graphical model checks (for an example, see [30]), posterior-predictive diagnostics [43] and prior-data conflict summaries [44,45]. The price of these checks is a relatively high computational burden, stricter distributional assumptions, additional investigator effort eliciting the model (and suitable priors for Bayesian methods [46]) and an unclear definition of a "good enough" model.

5. On Multiplicity

The common strategy used in GWAS for ranking and follow-up of new discoveries is to focus on the SNPs with the most significant *p*-values. As we discussed in Section 2, association tests of common single SNPs yield *p*-values that reflect a combination of sample size, effect size and MAF; however, for the range of MAFs in GWAS, rankings based on *p*-values correlate well with those based on effect size. Adjustment for multiple testing is usually done independent of any information on SNPs, and in a Bayesian framework, this corresponds to an equal prior probability of each SNP

being associated. There have been approaches developed to incorporate prior knowledge, such as stratified false discovery control [47] and weighted Bonferroni criteria [48]. Although there are several sources of information for these procedures in GWAS (such as effect on expression, effect on related phenotypes, position relative to gene elements, MAF), formal methods have not been used extensively for several reasons. In general, the available information is difficult to translate to the right scale, and there is low prior confidence that information on tag-SNPs is useful, since the causal SNP is unobserved.

Sequencing-based association studies are even more challenging, because there is more variability in the units of analysis than in GWAS. Gene units vary enormously in the number of SNPs, linkage disequilibrium (LD) pattern, the plausible ratio of causal SNPs, MAF spectrum and annotations. For example, what is more likely to be associated, a gene with two non-synonymous SNPs or a gene with ten non-synonymous SNPs? A gene with ten singletons (variants with only one observed copy of the non-reference allele) *versus* a gene with 10 total minor alleles with varying MAF? A set of ten non-synonymous SNPs *versus* a set of ten intronic SNPs? Furthermore the calculation of optimal weights will include an interplay of subject matter knowledge (e.g., assumptions on the effect sizes for different annotations) and the choice of statistical methods (e.g., some methods will accommodate signal sparsity well).

The complicated assessment of prior probabilities for a set of SNPs is one of the issues in using *p*-values for ranking genes and for deciding on efficient follow-up studies. *p*-values might be a poor proxy for the probability of replication, especially when the signals come from the very rare alleles that might not appear in the subjects used for replication. *p*-values contain little information on strategies for functional validation, because they do not inform on the best variants to be investigated. Relying only on *p*-values for decision-making has a bigger impact in sequencing studies than in GWAS, and we hope that developing better and more diverse measures of significance will become a more active area of research.

6. Discussion

Many people have been surprised by the lack of substantial findings from the recent studies on rare variants performed with whole-genome or whole-exome sequencing and from platforms, such as the exome chip. The reality is that for complex traits, there was little prior evidence in favor of genetic models that would give such studies high power (with multiple rare variants with a large effect per unit of study). The whole literature of the recent past, which is too extensive to be cited here, on investigating low frequency variants using imputation from population-based sequencing shows that large effect SNPs are uncommon for the diseases where they exist. This advocates for the development of more efficient strategies than the brute force sequencing of large, poorly phenotyped cohorts. The detailed annotation of variants should improve the sparsity of signals in the units of analysis, and careful phenotyping and incorporation of environmental factors should lead to the discovery of larger effects.

Much of the analytical effort on the association with sequencing data has been put into the development of novel testing tools. We argue in this paper that it is equally important to focus on other aspects of the process, from the design of the study to the interpretation of results. Furthermore, hypothesis tests and multiplicity adjustments should fit into the paradigm of a careful design that we set out above; model-based tests should incorporate the complexity that we expect without resorting to black boxes or poorly characterized weights. We should also be on guard for excessive parsimony; lumping together rare SNPs into a super-SNP creates a variable with properties that depend on the sampling scheme, minor allele frequency distribution and effects on phenotype distribution in complex ways.

Acknowledgments

The research was supported in part by the NIH grants, U01DK085501, P50MH094267 and R01MH101820. Dr. King was supported by National Institute of General Medical Sciences (NIGMS) T32GM007281 and F30HL103105. We are grateful to Nancy J. Cox and Hae Kyung Im for helpful discussions.

Author Contributions

Both authors conceived and designed the study, performed analytical calculations and simulations, and wrote the paper.

Appendix

A. The Derivation of the Power Formula

The following assumptions are used for the calculation of power: (1) there are n cases and m controls; (2) the association test is performed on a set of k SNPs, out of which, k_1 are associated, and the calculations are done conditional on k and k_1, ignoring the variability in those numbers that is associated with sequencing; (3) MAFs are sampled from a distribution with mean E_M and variance V_M; (4) for simplicity, we assume that the effect sizes of the associated SNPs are independent of MAF; (5) the SNPs are in linkage equilibrium; (6) all associations are with the rare allele; and (7) the effect sizes are sampled from a distribution with a mean odds ration equal to γ.

The association method used for illustration is the "burden" test, where, for each individual, we calculate a score based on the genotypes for the k SNPs. Let G_j denote the number of rare alleles at the j-th SNP, and let w_j be a fixed prespecified weight (it does not depend on the observed data; this is needed to simplify the analytical calculations). For each subject, we calculate:

$$S = \sum_{j=1}^{k} w_j G_j$$

then correlate this with the trait for the detection of association. For case-control studies, this can be done using a two-sample *t*-test. Note that power for a two sample normal test is governed by the non-centrality parameter,

$$\sqrt{\frac{nm}{n+m}}\,\frac{\mu_1-\mu_2}{\sigma}$$

with classical notation (and assuming equal variance). Note that for large k and n, the burden test will be close to a two sample normal test, and the mean and variance of the scores are the key determinants of power. The approximation is fairly accurate, even for small k, as long as n remains large.

If we denote with P_j the MAF for the j-th SNP, we have that:

$$\mathrm{E}(S) = \mathrm{E}\left[\mathrm{E}(S|P)\right] = \mathrm{E}\left(\sum_{j=1}^{k} w_j 2P_j\right) = \sum_{j=1}^{k} 2w_j E_M = 2\bar{w}kE_M$$

where $\bar{w}$ is the average weight. Similarly,

$$\mathrm{Var}(S) = \mathrm{Var}\left[\mathrm{E}(S|P)\right] + \mathrm{E}\left[\mathrm{Var}(S|P)\right] = \mathrm{Var}\left(2\sum_{j=1}^{k} w_j P_j\right) + \mathrm{E}\left(2\sum_{j=1}^{k} w_j^2 P_j(1-P_j)\right) =$$

$$4V_M \sum_{j=1}^{k} w_j^2 + 2\left(\sum_{j=1}^{k} w_j^2\right)(E_M - V_M - E_M^2) = 2k\overline{w^2}\left[V_m + E_M - E_M^2\right]$$

In the case of equal weights ($w_j = 1$),

$$\mathrm{E}(S) = 2kE_M, \quad \mathrm{Var}(S) = 2k(V_M + E_M - E_M^2)$$

For a rare associated SNP, its MAF is approximated by the product of the MAF in controls and the odds ratio. Because MAF and the odds ratios are independent (Assumption 4), we obtain that the mean MAF is approximated by $E_M\gamma$. This leads to the following mean score in cases (assuming equal weights),

$$\mathrm{E}(S) \approx 2(k-k_1)E_M + 2k_1E_M\gamma = 2kE_M + 2k_1E_M(\gamma - 1)$$

One can similarly derive a formula for $\mathrm{Var}(S)$ in the cases.

Assuming that the variances of the scores are not greatly different in cases and controls (valid with mild assumptions), the non-centrality parameter is approximated by:

$$\sqrt{\frac{2nm}{n+m}}\,\frac{k_1}{\sqrt{k}}\,\frac{E_M}{\sqrt{V_M + E_M - E_M^2}}\,(\gamma - 1)$$

B. Marginal Effects with Uncertainty

The marginal effect of a SNP whose true log odds ratio comes from a known distribution is easy to quantitatively analyze when using a model of the binary disease outcome as a dichotomized latent liability plus an SNP effect. That is, one can recast a traditional logistic regression model as a model where each individual has an unobserved quantitative trait, and individuals whose quantitative

trait is greater than some threshold demonstrate a positive binary trait. Effects of covariates (such as SNPs) add or subtract to the unobserved quantitative trait; when the liability has a logistic distribution (slightly heavier tailed than a Gaussian distribution), the effects on the latent scale are the same as lORs. Consider a logistic model shown in Figure A1; the plotted curve is the density of the latent liability and the dotted line the threshold above which individuals are affected by disease and below which they are unaffected if the base rate of the disease is 5%. The area under the curve to the right of the threshold are affected individuals and to the left of the threshold unaffected. The blue area under the curve is the fraction of individuals in the population who, with a moderately risk-increasing SNP, would cross the liability threshold and become affected by the disease. The smaller red area is the individuals who, with a risk-decreasing SNP of the same magnitude, lOR would cross the liability threshold in the other direction and cease to be affected.

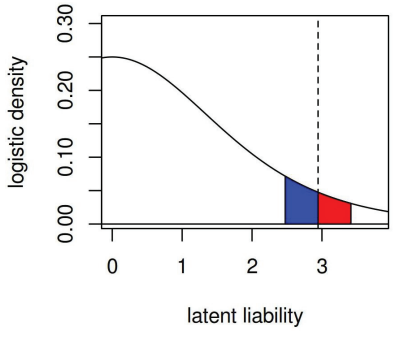


Figure A1. Density of latent trait before SNP effects. The dotted line indicates the case threshold. The blue area corresponds to controls that become cases if possessing an SNP with OR = 1.6. The red area indicates cases that become controls if possessing an SNP with lOR = 1/1.6.

When G_i is a vector of genotypes for person i, β are SNP lORs Gaussian distributed with mean μ and standard deviation σ, c is the threshold above, $I()$ the indicator function and X_i the latent liability, then we can write:

$$Y_i|g_i, \mu, \sigma = I(X_i + G_i\beta > c) \tag{A1}$$

One can approximate a logistic variable by a Gaussian scaled by 1.6, yielding:

$$Y_i|g_i, \mu, \sigma = I(Z_i\sqrt{1.6^2 + \sigma^2 g_i} + \mu g_i > c) \tag{A2}$$

for Z_i, a standard normal. One can then re-apply the normal-logistic approximation:

$$Y_i|g_i, \mu, \sigma = I(X_i^* + \frac{\mu g_i - c}{\sqrt{1 + \sigma^2 g_i/1.6^2}} > 0) \tag{A3}$$

where X^* is again logistic distributed, returning to a usual form for logistic regression.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **2005**, *437*, 1299–1320.
2. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J.P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367.
3. Burn, J. Should we sequence everyone's genome? Yes. *BMJ* **2013**, *346*, doi:10.1136/bmj.f3133.
4. Guan, W.; Pluzhnikov, A.; Cox, N.J.; Boehnke, M.; International Type 2 Diabetes Linkage Analysis Consortium. Meta-analysis of 23 type 2 diabetes linkage studies from the International Type 2 Diabetes Linkage Analysis Consortium. *Hum. Hered.* **2008**, *66*, 35–49.
5. Lee, S.; Wu, M.C.; Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **2012**, *13*, 762–775.
6. Mathieson, I.; McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **2012**, *44*, 243–246.
7. Liu, Q.; Nicolae, D.L.; Chen, L.S. Marbled inflation from population structure in gene-based association studies with rare variants. *Genet. Epidemiol.* **2013**, *37*, 286–292.
8. Babron, M.C.; de Tayrac, M.; Rutledge, D.N.; Zeggini, E.; GÃl'nin, E. Rare and Low Frequency Variant Stratification in the UK Population: Description and Impact on Association Tests. *PLoS One* **2012**, *7*, e46519.
9. Li, B.; Leal, S. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *Am. J. Hum. Genet.* **2008**, *83*, 311–321.
10. Madsen, B.E.; Browning, S.R. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet.* **2009**, *5*, e1000384.
11. Caspi, A.; Moffitt, T.E. Gene-environment interactions in psychiatry: Joining forces with neuroscience. *Nat. Rev. Neurosci.* **2006**, *7*, 583–590.
12. Hunter, D.J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **2005**, *6*, 287–298.
13. Coventry, A.; Bull-Otterson, L.M.; Liu, X.; Clark, A.G.; Maxwell, T.J.; Crosby, J.; Hixson, J.E.; Rea, T.J.; Muzny, D.M.; Lewis, L.R.; *et al*. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* **2010**, *1*, 131.
14. Keinan, A.; Clark, A.G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **2012**, *336*, 740–743.
15. Pritchard, J.K.; Cox, N.J. The allelic architecture of human disease genes: Common disease-common variant... or not? *Hum. Mol. Genet.* **2002**, *11*, 2417–2423.

16. Pritchard, J.K. Are Rare Variants Responsible for Susceptibility to Complex Diseases? *Am. J. Hum. Genet.* **2001**, *69*, 124–137.
17. Manolio, T.A.; Collins, F.S.; Cox, N.J.; Goldstein, D.B.; Hindorff, L.A.; Hunter, D.J.; McCarthy, M.I.; Ramos, E.M.; Cardon, L.R.; Chakravarti, A.; *et al*. Finding the missing heritability of complex diseases. *Nature* **2009**, *461*, 747–753.
18. Eyre-Walker, A. Evolution in Health and Medicine Sackler Colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 1752–1756.
19. Gorlov, I.P.; Gorlova, O.Y.; Sunyaev, S.R.; Spitz, M.R.; Amos, C.I. Shifting Paradigm of Association Studies: Value of Rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **2008**, *82*, 100–112.
20. Li, B.; Leal, S.M. Discovery of Rare Variants via Sequencing: Implications for the Design of Complex Trait Association Studies. *PLoS Genet.* **2009**, *5*, e1000481.
21. Zeger, S.L.; Liang, K.; Albert, P.S. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics* **1988**, *44*, 1049–1060.
22. Neuhaus, J.M.; Kalbfleisch, J.D.; Hauck, W.W. A Comparison of cluster-specific and population-averaged Approaches for Analyzing Correlated Binary Data. *Int. Stat. Rev. Rev. Int. Stat.* **1991**, *59*, 25–35.
23. Subramanian, S.V.; O'Malley, A.J. Modeling neighborhood effects: The futility of comparing mixed and marginal approaches. *Epidemiology* **2010**, *21*, 475–478; discussion 479–481.
24. Longmate, J.A.; Larson, G.P.; Krontiris, T.G.; Sommer, S.S. Three Ways of Combining Genotyping and Resequencing in Case-Control Association Studies. *PLoS One* **2010**, *5*, e14318.
25. Curtin, K.; Iles, M.M.; Camp, N.J. Identifying rarer genetic variants for common complex diseases: Diseased *versus* neutral discovery panels. *Ann. Hum. Genet.* **2009**, *73*, 54–60.
26. Edwards, T.L.; Song, Z.; Li, C. Enriching Targeted Sequencing Experiments for Rare Disease Alleles. *Bioinformatics* **2011**, *27*, 2112–2118.
27. Yang, F.; Thomas, D.C. Two-Stage Design of Sequencing Studies for Testing Association with Rare Variants. *Hum. Hered.* **2011**, *71*, 209–220.
28. King, C.R.; Rathouz, P.J.; Nicolae, D.L. Generalizing from sequencing studies. *arXiv* **2013**, arXiv:1312.7714.
29. Clayton, D.; Chapman, J.; Cooper, J. Use of unphased multilocus genotype data in indirect association studies. *Genet. Epidemiol.* **2004**, *27*, 415–428.
30. King, C.R.; Rathouz, P.J.; Nicolae, D.L. An Evolutionary Framework for Association Testing in Resequencing Studies. *PLoS Genet.* **2010**, *6*, e1001202.
31. Zelterman, D.; Chen, C. Homogeneity Tests Against Central-Mixture Alternatives. *J. Am. Stat. Assoc.* **1988**, *83*, 179–182.
32. Morris, A.P.; Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **2010**, *34*, 188–193.

33. Price, A.L.; Kryukov, G.V.; de Bakker, P.I.; Purcell, S.M.; Staples, J.; Wei, L.; Sunyaev, S.R. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *Am. J. Hum. Genet.* **2010**, *86*, 832–838.
34. Bansal, V.; Libiger, O.; Torkamani, A.; Schork, N.J. An application and empirical comparison of statistical analysis methods for associating rare variants to a complex phenotype. In Proceedings of the Pacific Symposium on Biocomputing, Kohala Coast, HI, USA, 3–7 January 2011; pp. 76–87.
35. Schaid, D.J. Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations. *Hum. Hered.* **2010**, *70*, 109–131.
36. Schaid, D.J. Genomic Similarity and Kernel Methods II: Methods for Genomic Information. *Hum. Hered.* **2010**, *70*, 132–140.
37. Pan, W. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet. Epidemiol.* **2011**, *35*, 211–216.
38. Hofmann, T. Kernel methods in machine learning. *Ann. Stat.* **2008**, *36*, 1171–1220.
39. Ladouceur, M.; Dastani, Z.; Aulchenko, Y.S.; Greenwood, C.M.T.; Richards, J.B. The Empirical Power of Rare Variant Association Methods: Results from Sanger Sequencing in 1,998 Individuals. *PLoS Genet.* **2012**, *8*, e1002496.
40. Xu, C.; Ladouceur, M.; Dastani, Z.; Richards, J.B.; Ciampi, A.; Greenwood, C.M.T. Multiple Regression Methods Show Great Potential for Rare Variant Association Tests. *PLoS One* **2012**, *7*, e41694.
41. Sul, J.H.; Han, B.; He, D.; Eskin, E. An Optimal Weighted Aggregated Association Test for Identification of Rare Variants Involved in Common Diseases. *Genetics* **2011**, *188*, 181–188.
42. Wu, M.; Lee, S.; Cai, T.; Li, Y.; Boehnke, M.; Lin, X. Rare-variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am. J. Hum. Genet.* **2011**, *89*, 82–93.
43. Meng, X.L. Posterior Predictive *p*-Values. *Ann. Stat.* **1994**, *22*, 1142–1160.
44. Bayarri, M.J.; Castellanos, M.E. Bayesian Checking of the Second Levels of Hierarchical Models. *Stat. Sci.* **2007**, *22*, 322–343.
45. Gelman, A. Comment: Bayesian Checking of the Second Levels of Hierarchical Models. *Stat. Sci.* **2007**, *22*, 349–352.
46. Yi, N.; Zhi, D. Bayesian analysis of rare variants in genetic association studies. *Genet. Epidemiol.* **2011**, *35*, 57–69.
47. Sun, L.; Craiu, R.V.; Paterson, A.D.; Bull, S.B. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genet. Epidemiol.* **2006**, *30*, 519–530.
48. Roeder, K.; Wasserman, L. Genome-Wide Significance Levels and Weighted Hypothesis Testing. *Stat. Sci.* **2009**, *24*, 398–413.

Imprinted Genes and the Environment: Links to the Toxic Metals Arsenic, Cadmium and Lead

Lisa Smeester, Andrew E. Yosim, Monica D. Nye, Cathrine Hoyo, Susan K. Murphy and Rebecca C. Fry

Abstract: Imprinted genes defy rules of Mendelian genetics with their expression tied to the parent from whom each allele was inherited. They are known to play a role in various diseases/disorders including fetal growth disruption, lower birth weight, obesity, and cancer. There is increasing interest in understanding their influence on environmentally-induced disease. The environment can be thought of broadly as including chemicals present in air, water and soil, as well as food. According to the Agency for Toxic Substances and Disease Registry (ATSDR), some of the highest ranking environmental chemicals of concern include metals/metalloids such as arsenic, cadmium, and lead. The complex relationships between toxic metal exposure, imprinted gene regulation/expression and health outcomes are understudied. Herein we examine trends in imprinted gene biology, including an assessment of the imprinted genes and their known functional roles in the cell, particularly as they relate to toxic metals exposure and disease. The data highlight that many of the imprinted genes have known associations to developmental diseases and are enriched for their role in the TP53 and AhR pathways. Assessment of the promoter regions of the imprinted genes resulted in the identification of an enrichment of binding sites for two transcription factor families, namely the zinc finger family II and PLAG transcription factors. Taken together these data contribute insight into the complex relationships between toxic metals in the environment and imprinted gene biology.

Reprinted from *Genes*. Cite as: Smeester, L.; Yosim, A.E.; Nye, M.D.; Hoyo, C.; Murphy, S.K.; Fry, R.C. Imprinted Genes and the Environment: Links to the Toxic Metals Arsenic, Cadmium and Lead. *Genes* **2014**, *5*, 477-496.

1. Introduction

There is heightened interest in understanding the role of epigenetic mechanisms in cell signaling regulation and disease. This is particularly the case when attempting to discern the etiology of disease where a cause is hitherto unknown. Recent studies suggest that disease can be influenced by the environment via epigenetic mechanisms [1–3], a seemingly Lamarkian notion discordant with the tenets set forth by Mendel's work. Yet the advent of modern epigenetics as a distinct field of study is much more storied than simply pitting Lamark's theory of "soft inheritance" against Mendel's firm genetic basis of heredity. While controversial to this day, some of the earliest experiments that suggested non-Mendelian inheritance, and possibly the first indication of parent-of-origin phenomena, came from Kammerer's midwife toad experiments [4] which pointed to the ability of a modified environment to modulate heritable shifts in mating.

However, it was not until the early 1940s, the same time Huxley's Modern Synthesis sought an interdisciplinary approach to inheritance and evolution [5], that the term epigenetics was first introduced by developmental biologist Conrad Waddington [6]. Describing mechanisms that

determine cell fate and differentiation during development, Waddington's epigenetics arose as a conceptual means to describe how a complex network of genes and gene-environment interactions brought about phenotype in an evolutionary context [7].

It was more recently though, with Holliday linking the term epigenetics with the ability of DNA methylation to modulate gene activity, that the definition of epigenetics began to shift to being inclusive of any mechanism with the ability to modulate gene activity without a change in DNA sequence [8]. These epigenetic modifications to an individual's genome include, but are not limited to, three commonly studied mechanisms: DNA methylation, histone modification, and non-coding RNA expression [9].

While such marks are generally stable and heritable in nature, there is the potential for epigenetic modifications to be reversible, as seen with their innate reversibility during critical stages of fetal development, as epigenetic tags are added and removed [10–12]. Such changes to the epigenetic landscape are vital in the process of normal development [13], yet imprinted genes are protected from this process [14].

Changes to the epigenome can also be induced by the environment resulting in abnormal physiologic changes [3,15]. Stable epigenetic modifications are implicated in adult onset disease such as cancer, neurodevelopmental/neurodegenerative disorders, and autoimmune disorders [16] among others, and also have the potential to be trans-generational in nature [17,18]. However, it should be noted that the stability of such epigenetic alterations and their link to later life health outcomes is still under active debate [19,20].

Current research has shown that epigenetic modifications can be induced by exposure to environmental contaminants, such as toxic metals [21–23]. Such modifications have been linked to later life health outcomes including cancers, heart disease, kidney disease, and various neurological conditions [20,24,25]. While cadmium and lead are metals, arsenic is a metalloid, with shared properties of both metals and non-metals. For the purposes of this article, all three elements will hereby be referred to as "metals." While these metals are among the most studied, many other metals have demonstrated toxicity and are associated with epigenetic alterations including nickel and chromium [21]. In addition, new research is currently investigating various under-studied metals for toxicity and associated epigenetic alterations including tungsten and cobalt [26,27].

While research continues into the stability of epigenetic alterations associated with environmental contaminants and their associations with negative health endpoints, many have theorized that such reversibility may provide the opportunity for therapeutic targets for disease prevention following environmental exposure [24,28].

1.1. What is Genomic Imprinting?

Evidence that parental genomes are not equivalent was first described in mouse models by both the Surani and McGrath groups in the early 1980s [29,30]. The researchers attempted to generate viable embryos using only maternal or paternal chromosomes. They found that normal development required genetic material from both parents; maternal and paternal genomes were not interchangeable, indicating the first experiments to demonstrate mammalian imprinting. While the exact mechanism of these imprinting phenomena was unknown at that time, they hypothesized a

process that operated pre-fertilization, yet impacted post-fertilization expression. To date, researchers have used a variety of strategies including genome-wide studies, gene-specific experiments, and transcriptome analysis to determine which human or mammalian genes are imprinted [31–33]. Imprinted genes are vulnerable to genetic and epigenetic perturbation and have been tied to adverse health outcomes. As imprinted genes are monoallelically expressed with one of the copies of the gene silenced in a parent-of-origin dependent manner, only one copy is functional. As a result, mutations or epigenetic alterations on one allele that would normally have minimal impact for a biallelically expressed gene may lead to detrimental consequences for an imprinted gene.

As it is a critical part of the epigenome, the inheritance and manifestation of traits associated with imprinted genes is regulated through epigenetic marks. The term "imprintome" was first coined to describe a set of "*cis*-acting imprint regulatory elements" [34]. This term refers to the mechanisms needed for modifying expression including DNA methylation and histone modification, which are two such mechanisms that are well established as being required for the appropriate maintenance of imprinted gene expression [35]. The imprintome is vulnerable to the environment and potentially modified by a host of environmental chemicals and contaminants [36].

Many imprinted genes are grouped in clusters and possess imprinting control regions (ICRs) or a central control region [37]. These ICRs, as well as other regulatory regions associated with imprinted genes, are referred to as differentially methylated regions (DMRs) and display ~50% methylation, where one of the parental alleles is methylated and the other unmethylated in a manner based on parent of origin. These DMRs represent discrete DNA elements that carry a heritable epigenetic mark that distinguishes the parental alleles.

1.2. Evolution of Imprinted Genes: A Fight between the Parental Chromosomes

Perhaps the most widely accepted hypothesis for evolutionary underpinnings of the origins of genomic imprinting is the parental conflict theory or the "battle of the sexes" [38,39]. Central to this hypothesis is the struggle to control maternal resources during fetal development, with paternal genes favoring increased use of maternal resources in order to promote the fittest possible offspring, and to divert resources from offspring of other males. Contextually, this hypothesis posits that imprinting arose during early mammalian evolution, where females were able to simultaneously gestate offspring from multiple males. The basis then for the desired growth and resource extraction of the offspring carrying the male's DNA is an attempt to out-compete offspring from other males. In contrast, maternal genes will suppress fetal growth to ensure equal reproductive success among all her offspring. Supporting this argument, many paternally expressed genes are growth promoting and metabolism-related, whereas maternally expressed genes tend to be growth limiting. The placenta, which can serve to regulate nutrients and growth for the developing fetus, is thought to play a pivotal role in this maternal-paternal fight over resources and control of fetal growth [40]. Within the placenta, a large number of the known imprinted genes are expressed, and genomic imprinting has been confirmed in all placental mammals studied thus far [41].

1.3. Imprinted Genes and Their Relationship to Human Health

Imprinted genes have been associated with various human adverse outcomes including diabetes, cancer, developmental disorders, behavioral disorders, and reproductive diseases [42,43]. Prader-Willi and Angelman syndromes were the first disorders to suggest an imprinted mechanism. Though each syndrome manifests differently, it was shown they both arise from deletions in the same region of chromosome fifteen [44]; Prader-Willi results from the loss of a cluster of paternally expressed genes, whereas Angelman syndrome results from the loss of maternal expression within the 15q11-q13 region.

Imprinted genes have also been associated with altered cellular growth resulting in cancer [45,46]. Imprinted genes are particularly vulnerable because they are functionally haploid, thus any epigenetic or genetic perturbations may have a greater impact. For example, epigenetically-induced silencing of the active allele of an imprinted tumor suppressor gene could result in complete loss of expression which in turn would influence cell growth or proliferation [47]. Conversely, epigenetic alterations can also result in activation of the otherwise silent copy of an imprinted growth-promoting gene, contributing to loss of growth regulation. Both of these types of alterations are referred to as "loss of imprinting" (LOI). In fact, LOI has been found across a broad spectrum of tumors and is one of the most common alterations in cancer [46]. As examples, in cancer, the active copy of tumor suppressor cyclin-dependent kinase inhibitor 1C (*CDKN1C*) is frequently aberrantly silenced and the silent copy of the growth promoting insulin-like growth factor II (*IGF2*) gene is often inappropriately activated [48].

1.4. Links between Imprinted Genes and Toxic Environmental Metals

Environmental contaminants are currently estimated to be responsible for almost five million deaths and over eighty million Disability-Adjusted Life Years (DALYs) globally [49], and are thought to be involved in 13% to 37% of the global disease burden [50]. Toxic metals represent some of the highest priority contaminants as determined by the Agency for Toxic Substances and Disease Registry (ATSDR) [51].

A number of studies have observed links between exposure to toxic metals and epigenetic events tied to imprinted genes. For example, quantitative analysis was conducted on multiple imprinted gene DMRs in peripheral blood from individuals followed as part of the Cincinnati Lead Study [52]. The researchers identified early childhood lead exposure was associated with hypomethylation of the gene pleiomorphic adenoma gene-like 1 (*PLAGL1*) [53]. Prenatal lead exposure has also been linked to decreases in global DNA methylation in cord blood [54]. In addition, lead has been shown to disrupt global DNA methylation patterns in embryonic stem cells [55].

Similarly, exposure to arsenic and cadmium has been observed to alter the methylation of both experimentally validated and predicted imprinted genes. Specifically, in adults exposed to inorganic arsenic, the known imprinted gene anoctamin 1, calcium activated chloride channel (*ANO1*) and predicted imprinted gene forkhead box F1 (*FOXF1*) shows increased promoter methylation in leukocytes [56]. Conversely, in a separate study, the imprinted gene insulin (*INS*)

exhibits decreased promoter methylation as a result of arsenic exposure [57]. In a cohort of mother-newborn pairs, the putative imprinted gene zic family member 1 (*ZIC1*), as well as *ANO1*, are differentially methylated in leukocytes, where cadmium exposure was associated with hypermethylation of *ZIC1* in mothers and *ANO1* in newborns [58].

Tobacco is a common source of cadmium exposure [59,60]. Tobacco smoke exposure in utero has been associated with altered DNA methylation [61,62]. These studies showed that among smoking mothers, ten imprinted genes including aryl-hydrocarbon receptor repressor (AHRR), growth factor independent 1 transcription repressor (GFI1), and cytochrome P450, family 1, subfamily A, polypeptide 1 (CYP1A1) were hypomethylated in newborn cord blood [61]. Additionally, the imprinted gene *IGF2* was found to be hypermethylated in cord blood of newborns born to smoking mothers [62]. Infant gender was correlated with differential methylation, as males exhibited smoking related methylation changes at *IGF2* while female newborns did not [62]. While the mechanisms of such sex-associated methylation patterning are unknown, other studies have shown similar sex-differentiated changes in methylation [63,64]. Further research is needed to understand gender specific epigenetic alterations, including imprinting, and may help to inform both sex-linked susceptibility to disease, as well as potentially being predictive of severity and/or prognosis. Additionally, research into the biological basis underlying the roll an individual's sex plays in disease development may afford the ability to develop targeted therapies based on differential responses.

1.5. Study Aim

In the present study, we set out to analyze imprinted genes for their involvement in shared biological pathways and to determine their known interactions with arsenic, cadmium, and lead. This analysis included an assessment of: (i) known relationships between the proteins encoded by the imprinted genes; (ii) common functionality of the proteins encoded by the imprinted genes in the cell; and (iii) common transcription factor-based regulatory regions present in the promoter regions of the imprinted genes.

2. Methods

Imprinted Gene List and Network, Pathway, and Functional Enrichment Analysis

We analyzed imprinted genes derived from a publically available database from GeneImprint [65] that were filtered for experimentally validated and computationally predicted imprinted genes. The genes within the GeneImprint database are classified as predicted based upon chromosomal location [66]. From those, individuals have confirmed some to be imprinted based on actual experiments with cDNA showing parent-of-origin monoallelic expression from humans or other methodologies. Specifically, of the 197 genes, 90 were experimentally confirmed. These two imprinted gene sets were further analyzed for known metals relationships using the comparative toxicogenomics database (CTD) resulting in two additional gene lists with $n = 43$, and $n = 14$, respectively. Toxic metals were prioritized based on their 2011 ATSDR rankings [51]. Metal and imprinted gene relationships were determined using the CTD [67]. The CTD is a public resource

that synthesizes current scientific literature on interactions between chemicals, genes, proteins, and the diseases associated with each. However, it should be noted that while the CTD is a useful tool that may be utilized to query a centralized public repository for known gene-metal interactions, the database may be limited in use for emerging findings, as there may be a delay between recent publications and inclusion in the database.

In order to identify biological pathways enriched within the imprinted gene sets, the four gene lists were analyzed for enrichment using Ingenuity Pathway Analysis (IPA) [68]. Ingenuity allows for the mapping of genes, proteins, and their corresponding regulatory networks as a useful tool for the identification of molecular pathways in disease. As a secondary method, for verification, The Database for Annotation, Visualization and Integrated Discovery (DAVIDv6.7) was also used to analyze the four gene lists.

As a method to identify transcription factor binding site enrichment within the imprintome, the imprinted gene set(s) were analyzed using the Genomatix Matinspector module (Genomatix Software Inc., Ann Arbor, MI, USA) [69]. The analysis was used to examine the four gene sets for common regulatory sequences and/or known regulation by common transcription factors. Where multiple promoter regions were possible for a given gene, a single promoter region was selected to maximize the number of experimentally verified 5' complete transcripts. The promoter regions were analyzed with the additional search criteria of 1000 base pairs upstream, and 50 base pairs downstream relative to the transcription start site. The genes were analyzed with a minimum core and matrix similarity of 1.00, the highest level of sensitivity possible. The *p*-value generated is the probability to obtain an equal or greater number of sequences with a match in a randomly drawn sample of the same size as the input sequence set. The lower this probability the higher is the importance of the observed common transcription factor.

3. Results and Discussion

3.1. Enriched Biological Trends within the Imprinted Gene Set

We analyzed imprinted genes derived from a publically available database that were filtered for experimentally validated and computationally predicted imprinted genes. Of the 197 genes, 90 were experimentally confirmed. These gene lists can be found in Table S1. These two imprinted gene sets were further analyzed for metals relationships using the CTD. Specifically, these associations include cellular perturbations such as altered mRNA and/or protein expression as well as epigenetic modifications (e.g., DNA methylation). The analysis identified known metals interactions for $n = 43$ predicted and experimentally validated imprinted genes and $n = 14$ experimentally validated imprinted genes with known metals interactions (Table S2).

The resulting four imprinted gene lists (both computationally predicted and validated) and the metals-associated genes were analyzed using IPA for known molecular interactions and for enrichment of diseases and canonical pathways. Many of the genes are involved in gene expression, embryonic development, organismal development, gastrointestinal disease, endocrine system disorders, hereditary disorders, cell morphology, cellular development, cell growth and proliferation, cell death and survival (Table 1). The results of IPA analysis were supported by

analysis performed using DAVID which found similar enrichment of several of the functional pathways such as embryonic development and gene regulation (Table 1).

In addition, enrichment analyses for canonical pathways were performed using IPA. The imprinted gene set is enriched for two canonical pathways, namely the TP53 and aryl-hydrocarbon receptor (AhR) signalling pathways (Table 1). A total of five genes were associated with TP53 including cyclin-dependent kinase 4 (*CDK4*), *PLAGL1*, retinoblastoma 1 (*RB1*), tumor protein p73 (*TP73*), and Wilms tumor 1 (*WT1*) (enrichment p value = 0.00107). A total of five genes were also associated with AhR including aldehyde dehydrogenase 1 family, member L1 (*ALDH1L1*), *CDK4*, cytochrome P450, family 1, subfamily B, polypeptide 1 (*CYP1B1*), *RB1*, and *TP73* (enrichment p value < 0.01). The following three genes were common to both pathways: *CDK4*, *RB1*, and *TP73*. This shared gene set perhaps is not surprising, as imprinted genes play a large role in development, and these pathways are fundamental in regulating development, signaling, and cellular responses to stress. However, it should be noted that the observation of the links between the imprinted gene set and these two critical biological pathways has not been previously reported.

Table 1. Summary of enriched biological processes/functions of the imprinted gene set.

Networks	***p*-value**
Embryonic Development, Organismal Development, Gene Expression *	1×10^{-39}
Embryonic Organ Development **	1.7×10^{-12}
Gene Expression, Developmental Disorder, Endocrine System Disorders *	1×10^{-39}
Diseases and Disorders	**Average *p*-value**
Developmental Disorder *	0.001
Endocrine System Disorders *	0.001
Organismal Injury and Abnormalities *	0.002
Gastrointestinal Disease *	0.002
Hereditary Disorder *	0.006
Prader-Willi Syndrome **	0.02
Beckwith-Wiedemann Syndrome **	0.0007
Molecular and Cellular Functions	**Average *p*-value**
Gene Expression *	<0.001
Transcription Regulation **	8.2×10^{-8}
Cell Morphology *	0.002
Cell Morphogenesis Involved in Differentiation	0.005
Cellular Development *	0.003
Development-associated Proteins **	1.8×10^{-10}
Cell Death and Survival *	0.003
Cell Signaling *	0.003
Canonical Pathways	***p*-value**
TP53 Signaling *	0.001
Aryl Hydrocarbon Receptor Signaling *	0.006

(*) Ingenuity Pathway Analysis (IPA) results; (**) Database for Annotation, Visualization and Integrated Discovery (DAVID) results.

Enrichment analyses for canonical pathways were also performed for the experimentally validated imprinted gene set (n = 90), where the TP53 signalling pathway was enriched ($p \leq 0.001$) represented by the following four genes *PLAGL1*, *RB1*, *TP73*, and *WT1*. Imprinted genes involved

in the AhR signalling pathway were also present including *RB1* and *TP73*, and the pathway was marginally significant ($p = 0.09$). Further enrichment analysis was performed on the two imprinted gene sets filtered for known association with our prioritized metals; those genes that are predicted or experimentally validated ($n = 43$), and just those experimentally validated ($n = 14$). The TP53 signaling pathway was also enriched in both the predicted/validated and experimentally validated gene sets ($p < 5.7 \times 10^{-5}$, and $p < 0.0023$, respectively). Enrichment of the AhR signaling pathway was significant ($p < 0.00023$) represented by the genes *CDK4*, *CYP1B1*, *RB1*, and *TP73* in the computationally predicted/experimentally validated gene set, but not in the set of only experimentally validated metal-associated imprinted genes.

3.1.1. TP53 Signaling-Associated Imprinted Genes

The TP53 tumor suppressor protein is known as the guardian of the genome. It is a key transcriptional regulator that responds to a variety of cellular stresses including damage induced by various environmental contaminants. It serves to control key cellular processes such as DNA repair, cell-cycle progression, angiogenesis, and apoptosis pathways critical for influencing apoptosis or cell-cycle arrest [70]. In addition to its critical roles in DNA repair pathways, TP53 can act as a transcriptional regulator. Mutations within the *TP53* gene are responsible for Li-Fraumeni syndrome [71] and loss in functionality of the tumor suppressor is thought to be a contributing factor in the majority of cancer cases [72,73].

Of the five imprinted genes found to be associated with TP53, TP73 has been shown to increase activation of phosphorylated TP53 in mouse embryo fibroblasts [74] and PLAGL1 was shown to act as a transcriptional co-activator and enhance the activity of TP53 in both human carcinoma $P53^{+/-}$ and HeLa cells [75]. In *in vitro* models of rat kidney and human osteoblast-like cells, binding of WT1 to TP53 has been shown to stabilize TP53, as well as inhibit TP53-mediated apoptosis [76]. Further, TP53 has been shown to modulate CDK4/RB1 through CDKN1A in human colon cancer cells [77]. TP53 can increase the expression of CDKN1A, which in turn decreases phosphorylation of RB1 [78]. Additionally, CDKN1A may also decrease phosphorylation of RB1 via CDK2-Cyclin D1 complex [79].

3.1.2. AhR Signaling-Associated Imprinted Genes

The aryl-hydrocarbon receptor is a transcription factor involved in cell cycle regulation, and an initiator of biological responses to xenobiotics. AhR has been shown to regulate enzymes such as cytochrome P450 and other xenobiotic metabolizing enzyme genes including putative imprinted gene *CYP1B1* [80]. In addition to xenobiotic metabolism, the AhR pathway plays a key role in organismal development processes [81]. In vertebrates, AhR is important for cellular proliferation and differentiation as well as many developmental pathways [82]. In addition to its role in mediating xenobiotics, the AhR pathway contributes to gene regulation and carcinogenesis [83].

Of the five genes found to be associated with AhR, CDK4 is involved in regulation of RB1 through phosphorylation [84]. It has been shown that RB1 increases activation of the AhR-Arnt complex in hepatoma cells [85] and this heterodimer complex is known to regulate CYP1B1 [86] and

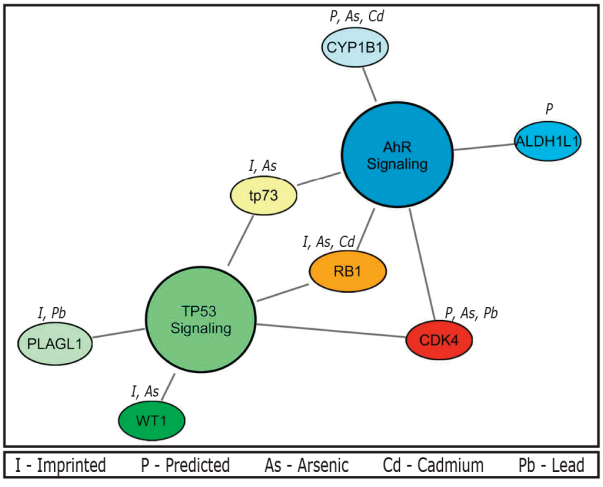
ALDH1 [87]. AhR has also been shown to decrease activation of phosphorylated E2F1 in tandem with RB1 [88,89]. This is worth noting, as expression of *TP73*, as well as many cyclins, is regulated by E2F1 [90,91].

3.1.3. Relationships between Imprinted Genes and Toxic Metals

Using data collected from the CTD, the seven unique genes within the TP53 and AhR pathways were detailed according to their known relationships to metals (Figure 1). The current body of literature pertaining to these relationships includes associations of the metals with altered gene regulation. Of the three genes (*CDK4*, *RB1*, and *TP73*) common to both pathways, upregulation of *CDK4* is observed upon exposure to both arsenic in a glioma cell line [92] and lead measured in peripheral blood [93]. Similarly, exposure to inorganic arsenic is associated with increased TP73 protein expression and activation of the downstream TP73-TP53 pathway in leukemia cells [94]. Conversely, there is decreased expression of *RB1* from arsenic [95] and cadmium exposure [96].

Of the genes specific to the TP53 pathway, exposure to inorganic arsenic is associated with decreased expression of the *WT1* gene in leukemia cell lines [97]. To date, the only known interaction between *PLAGL1* and environmental metals is hypomethylation associated with lead exposure [53]. Importantly, changes to signaling within the TP53 pathway by environmental contaminants such as inorganic arsenic and cadmium can impact the balance between apoptosis and proliferation in epithelial cancer cell lines [98,99].

Figure 1. Top canonical pathways and their relationships to toxic metals. The aryl-hydrocarbon receptor (AhR) (p = 0.001) and TP53 (p = 0.007) networks display known interactions between pathway genes and priority metals (arsenic, cadmium, and lead). Imprinted status is also noted; abbreviations are shown in the figure legend.



Gene-metal interactions that disrupt the AhR pathway can disturb xenobiotic metabolism and increase an individual's susceptibility to a range of negative health outcomes. Studies have shown

decreased expression of *CYP1B1* in response to metals such as arsenic [100] and cadmium in acute promyelocytic leukemia cell lines [101]. Interactions between *ALDH1L1* and environmental metals have not been studied. It should be mentioned that the analyses for gene-metal interactions were performed using the CTD database, which may not be fully comprehensive due to querying limitations and time delays between published findings and inclusion in the database. Tools such as CTD or IPA's canonical pathway analysis may be prone to inherent selection bias. Resources such as these that index interactions from published research may disproportionately display enrichment for a particular outcome or pathway as a consequence of overrepresentation of that particular interaction within the literature. Nevertheless, the alteration of these putative imprinted genes within the AhR pathway by inorganic arsenic, cadmium, and possibly other toxic metals may then serve as a potential mechanism of later life health outcomes including cancers and susceptibility to exogenous chemicals.

3.2. Sequence Specific Patterns of TF Elements in the Imprinted Genes

An analytical method was used to explore the four separate gene sets including experimentally validated and predicted imprinted genes (n = 197), experimentally validated imprinted genes (n = 90), experimentally validated and predicted imprinted genes associated with metals enrichment (n = 43), and experimentally validated imprinted genes associated with metals enrichment (n = 14) for common regulatory sequences and/or known regulation by common transcription factors. These results demonstrated that two transcription factor families were identified across all four gene lists, namely the family containing TF2B (Transcription Factor II B) and the PLAG family (Table 2).

TF2B is a transcription factor that mediates interactions between RNA polymerase II and promoter regions [102]. The other significantly enriched transcription factor family was the PLAG family. The PLAG family contains the transcription factors pleiomorphic adenoma gene 1 (PLAG1), pleiomorphic adenoma gene-like 2 (PLAGL2), and Pleiomorphic Adenoma Gene-Like 1 (PLAGL1, also known as ZAC1) encoded by the imprinted gene *PLAGL1*.

Interestingly, PLAGL1 is a zinc finger protein transcription factor that has been implicated as a regulatory hub in an "imprinted gene network" (IGN) controlling embryonic growth and cell proliferation, and proposed as a regulator of expression of other imprinted genes including *IGF2*, *H19*, and *CDKN1C* [103,104]. The data here expand upon this notion and in fact support that many, specifically 133 of the 171 analyzed imprinted genes (*i.e.*, 77%–78%) have binding sites for PLAG transcription factors, thus greatly expanding the current list of potential imprinted gene targets of PLAG.

Likewise, when only the metal-associated imprinted genes (n = 14) were analyzed, the most significantly enriched transcription factor families were once again the families of *TF2B* ($p < 2.62 \times 10^{-4}$) and *PLAG* ($p < 4.9 \times 10^{-3}$). TF2B motifs were enriched in 7 of 14 (50%) metal associated gene sequences.

Table 2. Enriched transcription factors amongst the imprinted gene sets.

Gene set	Transcription factor families	Transcription factors	Genes	*p*-value	Representative consensus sequence
Experimentally Validated & Predicted IGs (*n* = 197)	TF2B	GTF2B	93/171 (54%)	3.79E-45	ccgCGCC [1]
	PLAG	PLAG1 PLAGL1 PLAGL2	133/171 (78%)	2.10E-08	gaGGGGgcgggggggggggggg [2]
Experimentally Validated IGs (*n* = 90)	TF2B	GTF2B	38/71 (54%)	4.72E-19	ccgCGCC
	PLAG	PLAG1 PLAGL1 PLAGL2	55/71 (77%)	3.48E-04	gaGGGGgcgggggggggggggg
Experimentally Validated & Predicted IGs Metals-associated (*n* = 43)	TF2B	GTF2B	21/43 (49%)	1.09E-10	ccgCGCC
	PLAG	PLAG1 PLAGL1 PLAGL2	35/43 (81%)	1.31E-04	gaGGGGgcgggggggggggggg
Experimentally Validated IGs Metals-associated (*n* = 14)	TF2B	GTF2B	7/14 (50%)	2.62E-04	ccgCGCC
	PLAG	PLAG1 PLAGL1 PLAGL2	13/14 (93%)	4.90E-03	gaGGGGgcgggggggggggggg

[1] Consensus sequence based on most conserved nucleotide at each position for 210 sequences; [2] Consensus sequence based on most conserved nucleotide at each position for 337 sequences.

Transcription factors are known regulators of gene expression but they also serve an important role in regulating access to the DNA within genes. This access has consequences for both transcription and DNA methylation patterning. The binding of transcription factors to specific sites within the promoter region may protect CpG islands from methylation [105]. Recent work has found that areas with high transcription factor binding tend to have lower methylation [106].

Expanding upon previous work linking transcription factor occupancy with footprints left by DNase 1 [107,108], the Stamatoyannopoulos group studied numerous cell and tissue lines and found such footprint occupancy represented a viable quantitative measure of transcription factor occupancy and such occupancy afforded protection from DNA methylation [109].

In this context, our laboratory has recently hypothesized that in the case of cadmium and likely other environmental contaminants as well, distinct methylation patterns may represent "environmental footprints" or indicators of transcription factor occupancy during times of DNA methylation [58]. Based on the results from the present study, it can be hypothesized that specific

transcription factor families may impact occupancy related to imprinted gene regulatory regions and subsequently their potential for DNA methylation. The identification of these regulators of the methylation "footprints" may serve as important biomarkers of environmental exposure, and may help to support a mechanistic link between toxic metals exposure, imprinted gene alterations, and later life health outcomes.

4. Conclusions and Future Research Directions

Using a survey of current experimentally validated and predicted imprinted genes, we set out to examine shared functionality and pathway enrichment within the imprinted gene set. Notably, we found: (i) common functionality among proteins encoded by the imprinted gene set; (ii) several of the putative and known imprinted genes play a role in the TP53 and AhR pathways and many of these imprinted genes have been shown to interact with toxic environmental metals specifically in their ability to modify, and in certain cases disrupt, apoptosis, cell cycle arrest, ligand metabolism, DNA repair, and may promote the development of certain cancers; and (iii) common transcription factor-based regulatory regions for TF2B and PLAG present in the promoter regions of the experimentally validated and predicted imprinted genes.

As a result of their impact of access to DNA, transcription factors may contribute to specific DNA methylation patterning upon exposure to metals. Our lab has recently hypothesized that distinct methylation patterns due to exposure to environmental contaminants may represent "environmental footprints" of transcription factor occupancy during DNA methylating events. Based on our results, it can be hypothesized that specific transcription factor families such as TF2B and PLAG may impact the occupancy and subsequent methylation of the imprinted gene set. Additional research is needed to understand the mechanisms linking metals exposure, transcription factor occupancy and the imprintome. It is also worth noting that we prioritized the current research on cadmium, lead, and arsenic. While these metals represent some of the most studied, the lack of inclusion the emerging roll additional metals play in terms of impact on the epigenome is not a reflection of their toxicity, or relationship to genetic imprinting. Furthermore, metals are only one class of the broad range of environmental contaminants that have been shown, or may be associated with genetic imprinting and other epigenetic alterations.

Our novel finding that *CDK4*, *RB1*, and *TP73* are associated with two biologically critical pathways, TP53 and AhR may have implications for understanding the biological mechanisms of how these imprinted genes may ultimately be associated with negative health outcomes. Importantly, as more databases are populated with published work on imprinted genes and their interactions with toxic metals, other statistically significant pathways may be associated with the imprintome, or subsets of imprinted genes, and may provide further mechanistic links between genetic imprinting and disease.

Furthermore, our current findings have implications for understanding perturbed biological pathways that may provide insight into early life exposures and later life health consequences. In addition, it is worth noting, that while the idea is still under debate, further studies are needed to determine the extent of trans-generational inheritability of specific patterns of methylation and their contributions to disease. Combined with the dysregulation of imprinted genes by toxic-metals, it

may be possible to link ancestral exposure to environmental contaminants with current patterns of disease.

Based on our findings, further research is recommended to investigate the biological consequences of the imprinted gene set and its relationship to transcription factor occupancy. Specifically, transcription factor knockdown experiments or controlled toxicological experiments may help to elucidate the relationship between imprinted genes, transcription factor occupancy, environmental exposures, and associated health consequences.

Acknowledgments

This research was supported in part by the National Institute of Environmental Health Sciences (NIEHS) (ES019315, ES010126, ES005948, ES016772, ES022831), the National Cancer Institute (CA057726), the National Institute of Diabetes and Digestive and Kidney Diseases (DK085173) and by the USEPA (RD-83543701). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the USEPA.

Author Contributions

Rebecca C. Fry conceived of and oversaw the study; Lisa Smeester and Andrew E. Yosim performed analyses; all authors contributed to the writing of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Hou, L.; Zhang, X.; Wang, D.; Baccarelli, A. Environmental chemical exposures and human epigenetics. *Int. J. Epidemiol.* **2012**, *41*, 79–105.
2. Gluckman, P.D.; Hanson, M.A.; Cooper, C.; Thornburg, K.L. Effect of in utero and early-life conditions on adult health and disease. *N. Engl. J. Med.* **2008**, *359*, 61–73.
3. Jirtle, R.L.; Skinner, M.K. Environmental epigenomics and disease susceptibility. *Nat. Rev. Genet.* **2007**, *8*, 253–262.
4. Vargas, A.O. Did paul kammerer discover epigenetic inheritance? A modern look at the controversial midwife toad experiments. *J. Exp. Zool. B Mol. Dev. Evol.* **2009**, *312*, 667–678.
5. Lamm, E. Review of: Julian Huxley, evolution: The modern synthesis—The definitive edition, with a new forward by Massimo Pigliucci and Gerd B. Müller. MIT Press, 2010. *Integr. Psych. Behav.* **2011**, *45*, 154–159.
6. Jablonka, E.; Lamb, M.J. The changing concept of epigenetics. *Ann. N. Y. Acad. Sci.* **2002**, *981*, 82–96.
7. Haig, D. The (dual) origin of epigenetics. *Cold Spring Harb. Symp. Quant. Biol.* **2004**, *69*, 67–70.
8. Holliday, R. The inheritance of epigenetic defects. *Science* **1987**, *238*, 163–170.

9. Jaenisch, R.; Bird, A. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nat. Genet.* **2003**, *33*, 245–254.
10. Li, E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat. Rev. Genet.* **2002**, *3*, 662–673.
11. Dean, W.; Santos, F.; Reik, W. Epigenetic reprogramming in early mammalian development and following somatic nuclear transfer. *Semin. Cell Dev. Biol.* **2003**, *14*, 93–100.
12. Kiefer, J.C. Epigenetics in development. *Dev. Dyn.* **2007**, *236*, 1144–1156.
13. Reik, W.; Dean, W.; Walter, J. Epigenetic reprogramming in mammalian development. *Science* **2001**, *293*, 1089–1093.
14. Delaval, K.; Feil, R. Epigenetic regulation of mammalian genomic imprinting. *Curr. Opin. Genet. Dev.* **2004**, *14*, 188–195.
15. Feinberg, A.P. Phenotypic plasticity and the epigenetics of human disease. *Nature* **2007**, *447*, 433–440.
16. Portela, A.; Esteller, M. Epigenetic modifications and human disease. *Nat. Biotechnol.* **2010**, *28*, 1057–1068.
17. Jablonka, E.; Raz, G. Transgenerational epigenetic inheritance: Prevalence, mechanisms, and implications for the study of heredity and evolution. *Q. Rev. Biol.* **2009**, *84*, 131–176.
18. Skinner, M.K.; Manikkam, M.; Guerrero-Bosagna, C. Epigenetic transgenerational actions of environmental factors in disease etiology. *Trends Endocrinol. Metab.* **2010**, *21*, 214–222.
19. Feinberg, A.P.; Tycko, B. The history of cancer epigenetics. *Nat. Rev. Cancer* **2004**, *4*, 143–153.
20. Perera, F.; Herbstman, J. Prenatal environmental exposures, epigenetics, and disease. *Reprod. Toxicol.* **2011**, *31*, 363–373.
21. Salnikow, K.; Zhitkovich, A. Genetic and epigenetic mechanisms in metal carcinogenesis and cocarcinogenesis: Nickel, arsenic, and chromium. *Chem. Res. Toxicol.* **2008**, *21*, 28–44.
22. Baccarelli, A.; Bollati, V. Epigenetics and environmental chemicals. *Curr. Opin. Pediatr.* **2009**, *21*, 243–251.
23. Cheng, T.F.; Choudhuri, S.; Muldoon-Jacobs, K. Epigenetic targets of some toxicologically relevant metals: A review of the literature. *J. Appl. Toxicol.* **2012**, *32*, 643–653.
24. Egger, G.; Liang, G.; Aparicio, A.; Jones, P.A. Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **2004**, *429*, 457–463.
25. Aguilera, O.; Fernandez, A.F.; Munoz, A.; Fraga, M.F. Epigenetics and environment: A complex relationship. *J. Appl. Physiol.* **2010**, *109*, 243–251.
26. Verma, R.; Xu, X.; Jaiswal, M.K.; Olsen, C.; Mears, D.; Caretti, G.; Galdzicki, Z. *In vitro* profiling of epigenetic modifications underlying heavy metal toxicity of tungsten-alloy and its components. *Toxicol. Appl. Pharmacol.* **2011**, *253*, 178–187.
27. Li, Q.; Ke, Q.; Costa, M. Alterations of histone modifications by cobalt compounds. *Carcinogenesis* **2009**, *30*, 1243–1251.
28. Kelly, T.K.; de Carvalho, D.D.; Jones, P.A. Epigenetic modifications as therapeutic targets. *Nat. Biotechnol.* **2010**, *28*, 1069–1078.
29. Moore, G.E.; Oakey, R. The role of imprinted genes in humans. *Genome Biol.* **2011**, *12*, 106.

30. Munshi, A.; Duvvuri, S. Genomic imprinting—The story of the other half and the conflicts of silencing. *J. Genet. Genomics* **2007**, *34*, 93–103.
31. Babak, T.; Deveale, B.; Armour, C.; Raymond, C.; Cleary, M.A.; van der Kooy, D.; Johnson, J.M.; Lim, L.P. Global survey of genomic imprinting by transcriptome sequencing. *Curr. Biol.* **2008**, *18*, 1735–1741.
32. Wang, X.; Sun, Q.; McGrath, S.D.; Mardis, E.R.; Soloway, P.D.; Clark, A.G. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One* **2008**, *3*, e3839.
33. Bartolomei, M.S.; Ferguson-Smith, A.C. Mammalian genomic imprinting. *Cold Spring Harb. Perspect Biol.* **2011**, *3*, a002592.
34. Jirtle, R.L. Epigenome: The program for human health and disease. *Epigenomics* **2009**, *1*, 13–16.
35. Skaar, D.A.; Li, Y.; Bernal, A.J.; Hoyo, C.; Murphy, S.K.; Jirtle, R.L. The human imprintome: Regulatory mechanisms, methods of ascertainment, and roles in disease susceptibility. *ILAR J.* **2012**, *53*, 341–358.
36. Murphy, S.; Hoyo, C. Sculpting our future: Environmental nudging of the imprintome. In *Environmental Epigenomics in Health and Disease*; Jirtle, R.L., Tyson, F.L., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 51–73.
37. Barlow, D.P. Genomic imprinting: A mammalian epigenetic discovery model. *Annu. Rev. Genet.* **2011**, *45*, 379–403.
38. Moore, T.; Haig, D. Genomic imprinting in mammalian development: A parental tug-of-war. *Trends Genet.* **1991**, *7*, 45–49.
39. Kinoshita, T.; Ikeda, Y.; Ishikawa, R. Genomic imprinting: A balance between antagonistic roles of parental chromosomes. *Semin. Cell Dev. Biol.* **2008**, *19*, 574–579.
40. Frost, J.M.; Moore, G.E. The importance of imprinting in the human placenta. *PLoS Genet.* **2010**, *6*, e1001015.
41. Renfree, M.B.; Suzuki, S.; Kaneko-Ishino, T. The origin and evolution of genomic imprinting and viviparity in mammals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2013**, doi: 10.1098/rstb.2012.0151.
42. Úbeda, F.; Wilkins, J. Imprinted genes and human disease: An evolutionary perspective. In *Genomic Imprinting*; Wilkins, J., Ed.; Springer: New York, NY, USA, 2008; Volume 626, pp. 101–115.
43. Wilkins, J.F.; Ubeda, F. Diseases associated with genomic imprinting. *Prog. Mol. Biol. Transl. Sci.* **2011**, *101*, 401–445.
44. Knoll, J.H.M.; Nicholls, R.D.; Magenis, R.E.; Graham, J.M.; Lalande, M.; Latt, S.A.; Opitz, J.M.; Reynolds, J.F. Angelman and prader-willi syndromes share a common chromosome 15 deletion but differ in parental origin of the deletion. *Am. J. Med. Genet.* **1989**, *32*, 285–290.
45. Monk, D. Deciphering the cancer imprintome. *Brief. Funct. Genomics* **2010**, *9*, 329–339.
46. Jelinic, P.; Shaw, P. Loss of imprinting and cancer. *J. Pathol.* **2007**, *211*, 261–268.
47. Jones, P.A.; Baylin, S.B. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* **2002**, *3*, 415–428.

48. Leick, M.B.; Shoff, C.J.; Wang, E.C.; Congress, J.L.; Gallicano, G.I. Loss of imprinting of IGF2 and the epigenetic progenitor model of cancer. *Am. J. Stem Cells* **2012**, *1*, 59–74.
49. Pruss-Ustun, A.; Vickers, C.; Haefliger, P.; Bertollini, R. Knowns and unknowns on burden of disease due to chemicals: A systematic review. *Environ. Health* **2011**, *10*, 9.
50. Pruss-Ustun, A.; Bonjour, S.; Corvalan, C. The impact of the environment on health by country: A meta-synthesis. *Environ. Health* **2008**, *7*, 7.
51. Agency for toxic substances and disease registry (ATSDR). Available online: http://www.atsdr.cdc.gov/ (accessed on 28 Feburary 2014).
52. Murphy, S.K.; Hoyo, C. *Sculpting Our Future: Environmental Nudging of the Imprintome*; Jirtle, R.L., Tyson, F.L., Eds.; Springer: New York, NY, USA, 2013; pp. 51–73.
53. Hoyo, C. Department of Biological Sciences, Center for Human Health and Environment, Campus Box 7633, NC State University, Raleigh, NC 27695, USA. Unpublished work, 2014.
54. Pilsner, J.R.; Hu, H.; Ettinger, A.; Sanchez, B.N.; Wright, R.O.; Cantonwine, D.; Lazarus, A.; Lamadrid-Figueroa, H.; Mercado-Garcia, A.; Tellez-Rojo, M.M.; *et al.* Influence of prenatal lead exposure on genomic methylation of cord blood DNA. *Environ. Health Perspect.* **2009**, *117*, 1466–1471.
55. Senut, M.-C.; Sen, A.; Cingolani, P.; Shaik, A.; Land, S.J.; Ruden, D.M. Lead exposure disrupts global DNA methylation in human embryonic stem cells and alters their neuronal differentiation. *Toxicol. Sci.* **2014**, 142–161.
56. Smeester, L.; Rager, J.E.; Bailey, K.A.; Guan, X.; Smith, N.; Garcia-Vargas, G.; del Razo, L.M.; Drobna, Z.; Kelkar, H.; Styblo, M.; *et al.* Epigenetic changes in individuals with arsenicosis. *Chem. Res. Toxicol.* **2011**, *24*, 165–167.
57. Bailey, K.A.; Wu, M.C.; Ward, W.O.; Smeester, L.; Rager, J.E.; García-Vargas, G.; del Razo, L.M.; Drobná, Z.; Stýblo, M.; Fry, R.C. Arsenic and the epigenome: Interindividual differences in arsenic metabolism related to distinct patterns of DNA methylation. *J. Biochem. Mol. Toxicol.* **2013**, *27*, 106–115.
58. Sanders, A.P.; Smeester, L.; Rojas, D.; Debussycher, T.; Wu, M.C.; Wright, F.A.; Zhou, Y.H.; Laine, J.E.; Rager, J.E.; Swamy, G.K.; *et al.* Cadmium exposure and the epigenome: Exposure-associated patterns of DNA methylation in leukocytes from mother-baby pairs. *Epigenetics* **2013**, doi:10.4161/epi.26798.
59. Agency for Toxic Substances and Disease Registry (ATSDR). *Toxicological Profile for Cadmium*; ATSDR: Atlanta, GA, USA, 2012.
60. International Agency for Research on Cancer (IARC). Cadmium and cadmium compounds. IARC: Lyon, France, 2012.
61. Joubert, B.R.; Haberg, S.E.; Nilsen, R.M.; Wang, X.; Vollset, S.E.; Murphy, S.K.; Huang, Z.; Hoyo, C.; Midttun, O.; Cupul-Uicab, L.A.; *et al.* 450k epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.* **2012**, *120*, 1425–1431.
62. Murphy, S.K.; Adigun, A.; Huang, Z.; Overcash, F.; Wang, F.; Jirtle, R.L.; Schildkraut, J.M.; Murtha, A.P.; Iversen, E.S.; Hoyo, C. Gender-specific methylation differences in relation to prenatal exposure to cigarette smoke. *Gene* **2012**, *494*, 36–43.

63. Tobi, E.W.; Lumey, L.H.; Talens, R.P.; Kremer, D.; Putter, H.; Stein, A.D.; Slagboom, P.E.; Heijmans, B.T. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum. Mol. Genet.* **2009**, *18*, 4046–4053.
64. Kundakovic, M.; Gudsnuk, K.; Franks, B.; Madrid, J.; Miller, R.L.; Perera, F.P.; Champagne, F.A. Sex-specific epigenetic disruption and behavioral changes following low-dose in utero bisphenol a exposure. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 9956–9961.
65. Genomic Imprinting Database. Available online: http://www.geneimprint.org/ (accessed on 19 Feburary 2014).
66. Luedi, P.P.; Dietrich, F.S.; Weidman, J.R.; Bosko, J.M.; Jirtle, R.L.; Hartemink, A.J. Computational and experimental identification of novel human imprinted genes. *Genome Res.* **2007**, *17*, 1723–1730.
67. Comparative Toxicogenomics Database. Available online: http://ctdbase.org/ (accessed on 19 March 2014).
68. Ingenuity Pathway Analysis (IPA). Available online: http://www.ingenuity.com/ (Accessed on 18 March 2014).
69. Genomatix Software Suite v3.1. Available online: http://www.genomatix.de/ (Accessed on 18 March 2014).
70. Levine, A.J. P53, the cellular gatekeeper for growth and division. *Cell* **1997**, *88*, 323–331.
71. Birch, J.M. Li-fraumeni syndrome. *Eur. J. Cancer* **1994**, *30A*, 1935–1941.
72. Soussi, T.; Legros, Y.; Lubin, R.; Ory, K.; Schlichtholz, B. Multifactorial analysis of p53 alteration in human cancer: A review. *Int. J. Cancer* **1994**, *57*, 1–9.
73. Royds, J.A.; Iacopetta, B. P53 and disease: When the guardian angel fails. *Cell Death Differ.* **2006**, *13*, 1017–1026.
74. Flores, E.R.; Tsai, K.Y.; Crowley, D.; Sengupta, S.; Yang, A.; McKeon, F.; Jacks, T. P63 and p73 are required for p53-dependent apoptosis in response to DNA damage. *Nature* **2002**, *416*, 560–564.
75. Huang, S.M.; Schonthal, A.H.; Stallcup, M.R. Enhancement of p53-dependent gene activation by the transcriptional coactivator ZAC1. *Oncogene* **2001**, *20*, 2134–2143.
76. Maheswaran, S.; Englert, C.; Bennett, P.; Heinrich, G.; Haber, D.A. The WT1 gene product stabilizes p53 and inhibits p53-mediated apoptosis. *Genes Dev.* **1995**, *9*, 2143–2156.
77. Draus, J.M.; Elliott, M.J.; Atienza, C., Jr.; Stilwell, A.; Wong, S.L.; Dong, Y.; Yang, H.; McMasters, K.M. P53 gene transfer does not enhance E2F-1-mediated apoptosis in human colon cancer cells. *Exp. Mol. Med.* **2001**, *33*, 209–219.
78. Shav-Tal, Y.; Zipori, D. The role of activin a in regulation of hemopoiesis. *Stem Cells* **2002**, *20*, 493–500.
79. Adams, P.D.; Sellers, W.R.; Sharma, S.K.; Wu, A.D.; Nalin, C.M.; Kaelin, W.G., Jr. Identification of a cyclin-cdk2 recognition motif present in substrates and p21-like cyclin-dependent kinase inhibitors. *Mol. Cell Biol.* **1996**, *16*, 6623–6633.
80. Nebert, D.W.; Roe, A.L.; Dieter, M.Z.; Solis, W.A.; Yang, Y.; Dalton, T.P. Role of the aromatic hydrocarbon receptor and [*Ah*] gene battery in the oxidative stress response, cell cycle control, and apoptosis. *Biochem. Pharmacol.* **2000**, *59*, 65–85.

81. Hankinson, O. The aryl hydrocarbon receptor complex. *Annu. Rev. Pharmacol. Toxicol.* **1995**, *35*, 307–340.
82. Tijet, N.; Boutros, P.C.; Moffat, I.D.; Okey, A.B.; Tuomisto, J.; Pohjanvirta, R. Aryl hydrocarbon receptor regulates distinct dioxin-dependent and dioxin-independent gene batteries. *Mol. Pharmacol.* **2006**, *69*, 140–153.
83. Barouki, R.; Coumoul, X.; Fernandez-Salguero, P.M. The Aryl hydrocarbon receptor, more than a xenobiotic-interacting protein. *FEBS Lett.* **2007**, *581*, 3608–3615.
84. Gao, N.; Flynn, D.C.; Zhang, Z.; Zhong, X.S.; Walker, V.; Liu, K.J.; Shi, X.; Jiang, B.H. G1 cell cycle progression and the expression of G1 cyclins are regulated by Pi3k/Akt/Mtor/P70s6k1 signaling in human ovarian cancer cells. *Am. J. Physiol. Cell Physiol.* **2004**, *287*, C281–C291.
85. Huang, G.; Elferink, C.J. Multiple mechanisms are involved in ah receptor-mediated cell cycle arrest. *Mol. Pharmacol.* **2005**, *67*, 88–96.
86. Vrzal, R.; Ulrichova, J.; Dvorak, Z. Aromatic hydrocarbon receptor status in the metabolism of xenobiotics under normal and pathophysiological conditions. *Biomed. Pap. Med. Fac. Univ. Palacky Olomouc Czech Repub.* **2004**, *148*, 3–10.
87. Lindros, K.O.; Oinonen, T.; Kettunen, E.; Sippel, H.; Muro-Lupori, C.; Koivusalo, M. Aryl hydrocarbon receptor-associated genes in rat liver: Regional coinduction of aldehyde dehydrogenase 3 and glutathione transferase YA. *Biochem. Pharmacol.* **1998**, *55*, 413–421.
88. Marlowe, J.L.; Puga, A. Aryl hydrocarbon receptor, cell cycle regulation, toxicity, and tumorigenesis. *J. Cell Biochem.* **2005**, *96*, 1174–1184.
89. Puga, A.; Barnes, S.J.; Dalton, T.P.; Chang, C.; Knudsen, E.S.; Maier, M.A. Aromatic hydrocarbon receptor interaction with the retinoblastoma protein potentiates repression of E2f-dependent transcription and cell cycle arrest. *J. Biol. Chem.* **2000**, *275*, 2943–2950.
90. Urist, M.; Tanaka, T.; Poyurovsky, M.V.; Prives, C. P73 induction after DNA damage is regulated by checkpoint kinases Chk1 and Chk2. *Genes Dev.* **2004**, *18*, 3041–3054.
91. Murray, A.W. Recycling the cell cycle: Cyclins revisited. *Cell* **2004**, *116*, 221–234.
92. Zhao, S.; Zhang, J.; Zhang, X.; Dong, X.; Sun, X. Arsenic trioxide induces different gene expression profiles of genes related to growth and apoptosis in glioma cells dependent on the p53 status. *Mol. Biol. Rep.* **2008**, *35*, 421–429.
93. Tian, Y.; Green, P.; Stamova, B.; Hertz-Picciotto, I.; Pessah, I.; Hansen, R.; Yang, X.; Gregg, J.; Ashwood, P.; Jickling, G.; *et al.* Correlations of gene expression with blood lead levels in children with autism compared to typically developing controls. *Neurotox Res.* **2011**, *19*, 1–13.
94. Lunghi, P.; Costanzo, A.; Levrero, M.; Bonati, A. Treatment with arsenic trioxide (Ato) and Mek1 inhibitor activates the p73-p53aip1 apoptotic pathway in leukemia cells. *Blood* **2004**, *104*, 519–525.
95. Wu, X.; Shi, J.; Wu, Y.; Tao, Y.; Hou, J.; Meng, X.; Hu, X.; Han, Y.; Jiang, W.; Tang, S.; *et al.* Arsenic trioxide-mediated growth inhibition of myeloma cells is associated with an extrinsic or intrinsic signaling pathway through activation of trail or trail receptor 2. *Cancer Biol. Ther.* **2010**, *10*, 1201–1214.

96. Choi, Y.-J.; Yin, H.-Q.; Suh, H.-R.; Lee, Y.-J.; Park, S.-R.; Lee, B.-H. Involvement of e2f1 transcriptional activity in cadmium-induced cell-cycle arrest at G1 in human lung fibroblasts. *Environ. Mol. Mutagen.* **2011**, *52*, 145–152.
97. Glienke, W.C.; Chow, K.U.; Bauer, N.; Bergmann, L. Down-regulation of Wt1 expression in leukemia cell lines as part of apoptotic effect in arsenic treatment using two compounds. *Leuk. Lymphoma* **2006**, *47*, 1629–1638.
98. Sandoval, M.; Morales, M.; Tapia, R.; del Carmen Alarcón, L.; Sordo, M.; Ostrosky-Wegman, P.; Ortega, A.; López-Bayghen, E. P53 response to arsenic exposure in epithelial cells: Protein kinase B/Akt involvement. *Toxicol. Sci.* **2007**, *99*, 126–140.
99. Meplan, C.; Mann, K.; Hainaut, P. Cadmium induces conformational modifications of wild-type p53 and suppresses p53 response to DNA damage in cultured cells. *J. Biol. Chem.* **1999**, *274*, 31663–31670.
100. Zheng, P.-Z.; Wang, K.-K.; Zhang, Q.-Y.; Huang, Q.-H.; Du, Y.-Z.; Zhang, Q.-H.; Xiao, D.-K.; Shen, S.-H.; Imbeaud, S.; Eveno, E.; *et al.* Systems analysis of transcriptome and proteome in retinoic acid/arsenic trioxide-induced cell differentiation/apoptosis of promyelocytic leukemia. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 7653–7658.
101. Andrew, A.S.; Warren, A.J.; Barchowsky, A.; Temple, K.A.; Klei, L.; Soucy, N.V.; O'Hara, K.A.; Hamilton, J.W. Genomic and proteomic profiling of responses to toxic metals in human lung cells. *Environ. Health Perspect.* **2003**, *111*, 825–835.
102. Roeder, R.G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **1996**, *21*, 327–335.
103. Lui, J.C.; Finkielstain, G.P.; Barnes, K.M.; Baron, J. An imprinted gene network that controls mammalian somatic growth is down-regulated during postnatal growth deceleration in multiple organs. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **2008**, *295*, R189–R196.
104. Varrault, A.; Gueydan, C.; Delalbre, A.; Bellmann, A.; Houssami, S.; Aknin, C.; Severac, D.; Chotard, L.; Kahli, M.; le Digarcher, A.; *et al.* Zac1 regulates an imprinted gene network critically involved in the control of embryonic growth. *Dev. Cell* **2006**, *11*, 711–722.
105. Brandeis, M.; Frank, D.; Keshet, I.; Siegfried, Z.; Mendelsohn, M.; Nemes, A.; Temper, V.; Razin, A.; Cedar, H. SP1 elements protect a CPG island from *de novo* methylation. *Nature* **1994**, *371*, 435–438.
106. Feldmann, A.; Ivanek, R.; Murr, R.; Gaidatzis, D.; Burger, L.; Schubeler, D. Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet.* **2013**, *9*, e1003994.
107. Galas, D.J.; Schmitz, A. Dnase footprinting: A simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **1978**, *5*, 3157–3170.
108. Hesselberth, J.R.; Chen, X.; Zhang, Z.; Sabo, P.J.; Sandstrom, R.; Reynolds, A.P.; Thurman, R.E.; Neph, S.; Kuehn, M.S.; Noble, W.S.; *et al.* Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods* **2009**, *6*, 283–289.
109. Neph, S.; Vierstra, J.; Stergachis, A.B.; Reynolds, A.P.; Haugen, E.; Vernot, B.; Thurman, R.E.; John, S.; Sandstrom, R.; Johnson, A.K.; *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **2012**, *489*, 83–90.

The Impact of the Human Genome Project on Complex Disease

Jessica N. Cooke Bailey, Margaret A. Pericak-Vance and Jonathan L. Haines

Abstract: In the decade that has passed since the initial release of the Human Genome, numerous advancements in science and technology within and beyond genetics and genomics have been encouraged and enhanced by the availability of this vast and remarkable data resource. Progress in understanding three common, complex diseases: age-related macular degeneration (AMD), Alzheimer's disease (AD), and multiple sclerosis (MS), are three exemplars of the incredible impact on the elucidation of the genetic architecture of disease. The approaches used in these diseases have been successfully applied to numerous other complex diseases. For example, the heritability of AMD was confirmed upon the release of the first genome-wide association study (GWAS) along with confirmatory reports that supported the findings of that state-of-the art method, thus setting the foundation for future GWAS in other heritable diseases. Following this seminal discovery and applying it to other diseases including AD and MS, the genetic knowledge of AD expanded far beyond the well-known *APOE* locus and now includes more than 20 loci. MS genetics saw a similar increase beyond the *HLA* loci and now has more than 100 known risk loci. Ongoing and future efforts will seek to define the remaining heritability of these diseases; the next decade could very well hold the key to attaining this goal.

Reprinted from *Genes*. Cite as: Bailey, J.N.C.; Pericak-Vance, M.A.; Haines, J.L. The Impact of the Human Genome Project on Complex Disease. *Genes* **2014**, *5*, 518-535.

1. Introduction

In celebration of the 10th anniversary of the completion of the Human Genome Project, it is pertinent to take a step back and reflect on the progress that has been made in genetic and genomic research over the past decade by exploring the knowledge gleaned from the extensive wealth of information provided by the Human Genome Project (HGP). Herein we provide a concise historical overview of three signature human diseases that have strong but complex genetic etiologies: age-related macular degeneration (AMD), multiple sclerosis (MS), and Alzheimer's disease (AD). The significant progress in defining the genetic architecture of these diseases, beginning with the pre-genome-wide association study (GWAS)-era and concluding with the current state of each, and what lies ahead for these complex diseases reflects the great progress that has been made in general in the study of multifactorial diseases, and provides a brief glimpse at what we can hope the next decade of genomic research will provide.

2. Age-Related Macular Degeneration (AMD) and the First Genome-Wide Association Study

Age-related macular degeneration an ocular neurodegenerative disease that results primarily in loss of central vision, is a major cause of visual impairment and blindness in elderly populations worldwide. Although there was at one time substantial controversy over the strength of the genetic

effects in AMD, genetic and epidemiological research, established that there is a significant genetic component to AMD, estimated to be 45%–70% [1]. This was supported by twin studies that reported higher incidence of disease in monozygotic *versus* dizygotic twins [1–5] and family studies in which risk for developing AMD between first degree relatives ranges from 2–3 [3,6,7]. This knowledge encouraged the application of increasingly sophisticated genomic techniques to elucidate the genetic etiology of AMD susceptibility and pathogenesis. Prior to major genetic breakthroughs such as the completion of the HGP, it was well established that inflammatory and immunologic mediators contribute to AMD (e.g., [8–13]). However, this knowledge did not lead to identification of any confirmed genetic loci for AMD. Following the trends at that time in applying the available statistical genetic techniques, numerous genetic linkage studies using multiplex families and (primarily) affected sibships were attempted [14–22]. Notably, The *ABCA4* (*ABCR*) locus on chromosome 1p21, identified for its involvement in autosomal recessive Stargardt disease retinopathy [23–27], was one of the first loci identified as involved in AMD, though not all reports have been consistent [23–31]. While linkage studies continued to provide suggestive evidence of a role of genetics, they did not find any definitive locus for AMD. In a large meta-analysis of most of these genetic linkage studies, several chromosomal regions were identified as highly likely to harbor AMD genes, most convincingly including chromosome 1q23.3–q32 and 10q26 [32].

With the continuing evolution of HGP resources, in particular the identification of very large numbers of single nucleotide polymorphisms (SNPs) [33,34] multiple new experimental designs for identifying AMD loci were employed. SNPs provided several advantages over the then prevalent microsatellite markers; the two most important were the high density of SNPs across the genome, and their much higher fidelity in genotyping. The culmination of the efforts of four independent studies using four complementary study designs was a convergence on the discovery of the association between AMD and the gene encoding the complement factor H protein (*CFH*), located on chromosome 1q32. In one of the first reported genome-wide association studies (GWAS), Klein *et al.* [35] screened 96 AMD cases and 50 non-AMD controls to evaluate variants associated with AMD. The GWAS method implements a hypothesis-free approach in which a large number of SNPs are genotyped across the genome and evaluated for association with disease. This particular study evaluated 116,204 successfully genotyped SNPs and detected association between AMD and an intronic SNP in *CFH* ($p < 10^{-7}$). Linkage disequilibrium analysis and localization using resequencing in this region led to the discovery of a nonsynonymous SNP in exon 9 of *CFH*; this SNP, rs1061170, causes the substitution of a histidine for a tyrosine at amino acid 402 (Y402H).

Independently and concurrently with the Klein *et al.* study [35], Haines *et al.* [36] also identified the association with the Y402H variant, but by implementing a purely locational genomic approach. By focusing on and extending slightly beyond the 24 Mb region implicated by linkage studies of AMD [14,15,22] they identified a five-SNP haplotype spanning a 261 kb region surrounding the Regulators of Complement Activation (RCA) gene cluster by genotyping only 61 SNPs in two independent datasets. Both affected and unaffected individuals homozygous for the risk haplotype were sequenced for the genes residing in this haplotype. Their hypothesis was that having controlled for the locus specific genetic background (e.g., the haplotype), frequency

differences for variants between cases and controls would identify the causal variation. Scanning the coding region of *CFH* in those individuals, Y402H was by far the most significantly different of the 11 detected variants. Follow-up genotyping in the original datasets confirmed that the Y402H variant was significantly associated with risk for AMD and that a surprisingly high proportion of the genetic variation in AMD could be attributed to the Y402H variant.

Implementing yet another independent, concurrent, and complimentary approach to localize AMD-causing variants, Edwards *et al.* also identified the Y402H variant using a fine-mapping approach focused on this same general region on chromosome 1 [37]. This study centered efforts on 86 SNPs located in coding sequences encompassing the RCA locus in a case-control sample. The most significant of the 29 associated variants located in the RCA was again rs1061170 (Y402H) in *CFH*. Replication analysis evaluating this and 13 additional SNPs typed in a second case-control sample confirmed the association of Y402H with AMD. Further analysis established that that C (risk)-allele carrying individuals accounted for approximately half of cases.

Hageman and colleagues also confirmed the Y402H variant, applying yet a fourth genetic analysis method [38]. They applied prior biological knowledge of the involvement of *CFH* (also called *HF1*) in membranoproliferative glomerulonephritis type II (MPGNII), a disease in which patients develop ocular drusen nearly identical to those found in AMD patients. The genetic lesion for MPGNII resides in the same chromosome 1q31–32 region that was also implicated in linkage studies of AMD [14,15,22]. Evaluating two samples of unrelated individuals for AMD-associated variation in *CFH*, this group also detected evidence for association between AMD and the Y402H variant.

These four studies simultaneously reported the role of variation in a chromosome 1 region that had previously been highlighted in AMD linkage studies [14,15,22]; identifying this major genetic determinant of AMD, something that even a year earlier was thought not to exist, was a major landmark in genetics of complex disease. These results, while obviously important for AMD research, provided the first validation of the GWAS approach. Up to that point, hundreds of papers had been written about the potential of the GWAS study design, but very little had been published on actual implementation. Dr. Elias Zherhouni, then Director of NIH, highlighted these studies as a major breakthrough in health research [39]. This very strong validation imparted the necessary confidence in GWAS to invigorate its application to numerous other diseases. Over the past nine years more than 2000 GWAS studies that have been published [40].

Since the initial discovery of the Y402H *CFH* variant, substantial progress has been made in understanding the genetics of AMD. This includes the localization of the strongest single genetic effect in AMD on chromosome 10q26 (through positional localization approach) to the region containing *ARMS2* and *HTRA1* (e.g., [41–47]), though there is still controversy whether either one or both of these genes contains the causal variant (e.g. [46–49]).

In the *CFH* region, in addition to the high-effect Y402H variant, a deletion of *CFHR1* and *CFHR3* was detected and determine to be protective for AMD [50] (e.g., [51–53]). Various additional studies focused on the potential role of additional complement mediators in AMD. Gold *et al.* explored additional alternative complement pathway activators beyond CFH and determined that variants in complement factor b (*BF/CFB*) and *C2* are highly protective against AMD [54]. A

coding variant in *C3* was also determined to be associated with AMD [55–57]. A variant upstream of the *CFI* gene was also determined to influence AMD risk [58]. A variant in the *CFD* gene was associated with AMD but replicated almost solely in females [59]. More recently discovered AMD-associated loci have been detect in/near genes *ADAMTS9*, *B3GALTL*, *CETP*, *COL8A1-FILIP1L*, *IER3-DDR1*, *LIPC*, *RAD51B*, *SLC16A8*, *TGFBR1* and *TIMP3* [60–62].

GWAS studies in AMD now include over 1 million markers [61,63–65]. Though the traditional GWAS approach has been incredibly informative for many diseases, a great deal of the genetic proportion of many of these diseases remains to be fully elucidated even after applying straightforward and complex GWAS methods [66]. Approaches to enhance the detection of genetic variation associated with disease have necessarily expanded beyond the traditional GWAS to broaden the range of discovery and increase the power of detection; one technique that aids in this process is imputing variants—using known genetic information from a reference sample. Obviously increasing the sample size of genetic studies of complex diseases is crucial to accelerate the identification of disease-specific variants. Expanding the number of testable variants is now a more attainable goal using imputation, a technique that can significantly increase the number of tested variants beyond those interrogated by a GWAS through informing genotypes of untyped SNPs [67,68]. Combining known genotypes at GWAS-interrogated SNPs with available sequence data from a reference panel and inferring untyped SNPs in the dataset based on haplotype frequencies allows for the inference of numerous SNPs with varying degrees of confidence and accuracy. This method increases the power of GWAS by increasing the number of SNPs that can be tested, it can also lead to more efficient identification of causal variants and/or SNPs in high linkage disequilibrium with a causal variant [67,68]. Imputation has been implemented in several studies of AMD to enhance the ability to detect associated variants [61,63,69]. For example, the most recent publication from the AMD Gene Consortium reports seven novel variants that were detected using imputed data in addition to confirming 12 previously identified variants [61].

An additional method to utilize genome-wide data beyond the traditional association analysis is to perform pathway enrichment analyses. The goal of such analyses is identify biological relationships between associated genetic signals and pathways of interest in a particular disease. Pathway analysis can be performed by a comprehensive review of GWAS results to assess overrepresentation of SNPs meeting a specific threshold that occur within biological pathways [70]. These enhance GWAS by evaluating potentially biologically relevant signals that might otherwise be overlooked because of the numerous false-positive results that occur in large GWAS studies [70]. These have the potential to highlight otherwise undetected small and/or interactive effects that are important to evaluate in addition to and in the context of the overall genome-wide results. Using the INRICH (Interval-based Enrichment Analysis Tool for Genome Wide Association Studies) pathway analysis tool [71] to evaluate overall results, the AMD Gene Consortium not only confirmed previously implicated AMD pathways, but also determined additional pathways of interest in the most recent publication which detected enrichment of complement and atherosclerotic pathway-encoding genes as well as genes involved in pathways of collagen and extracellular region, complement and coagulation cascades, lipoprotein metabolism, and regulation of apoptosis [61].

The impressive impact that genetic information can have on our understanding of disease pathophysiology is highlighted in the recent publication by Yang *et al.* in which they report that *ARMS2/HTRA1* risk alleles contribute to AMD pathogenesis by decreasing the defense capabilities of superoxide dismutase 2 (SOD2) and thereby cause the retinal pigment epithelium to be more susceptibility to oxidative damage [72]. Having an explanation for the role the variants have in the disease is crucial to further elucidating disease mechanisms both genetically and physiologically. Additionally, genetic studies implicate VEGF as having a role in AMD and current AMD treatment and clinical trials utilize this information for treatment of neovascular AMD (reviewed in [73]), thus highlighting the utility of genetic data for clinical impact.

3. Alzheimer's Disease

Alzheimer's disease (AD) is a genetically heterogeneous neurologic disorder that is the leading cause of dementia among the elderly. It is characterized by the progressive loss of cognitive ability beyond what is normally associated with aging. AD is a complex disease that is influenced by both environmental and genetic mediators, the most significant of which is age [74,75]. The heritability of AD is estimated between 60%–80%. Before 1985, there was very significant debate about whether or not genetics played any role in AD (e.g., [76–80]). However, in 1987, using some of the earliest technologies employing genomic markers, a locus for the rare early onset AD (EOAD) was identified [81], and in 1991 the responsible variation in the *APP* gene was located [41].

Expansions of genomic marker sets, developed through early HGP efforts, were used to further identify two additional early onset genes in the early 1990's [82–86]. Simultaneously and independently, the emerging technologies of genomic markers and genetic linkage analysis were applied to the far more common late onset Alzheimer's disease (LOAD), which accounts for 99% of AD cases [87]. Using these techniques, Pericak-Vance *et al.* identified a locus on chromosome 19 near the gene encoding apolipoprotein E (*APOE*) [88,89], which was at that time thought to only be involved in cardiovascular disease. This locus has three distinct alleles: ∈2, ∈3, and ∈4. Corder *et al.* characterized a dose-dependent association between the *APOE*-∈4 allele and an increased risk of LOAD [90]. Mutations in the EOAD genes are causal, with very high penetrance, and opened avenues for exploring the pathophysiology of AD. However, in aggregate they explain less than 1% of AD. In contrast, *APOE* explains at least 25% of AD. A year after determining the role of the ∈4 variant in LOAD susceptibility, it became apparent that the ∈2 allele provided an independently protective effect on LOAD [91].

The *APOE* finding was pivotal for two reasons. Within the AD research community, it provided a new avenue and a completely different view of the genetic etiology of AD. More generally, however, this was one of the very first examples of how the emerging technologies of the HGP could be successfully applied to diseases lacking a simple Mendelian inheritance pattern, *i.e.*, what are commonly called complex diseases. The finding of the *APOE*-∈2 allele protective effect was also one of the first examples of different alleles carrying different effects on a complex disease, a pivotal moment in AD research and broadly in the field of genetics.

Innumerable attempts to identify additional genomic variations modulating the risk of LOAD followed these groundbreaking *APOE* discoveries, using the increasingly dense set of known

variations and emerging sequencing techniques (cataloged in Alzgene.org). These efforts were primarily applied to specific genes of interest; that is, employing a focused candidate gene approach. Although there were numerous reports of significant associations, no consensus arose that any of these were true effects. It was not until GWAS became a viable approach [92–94], and multiple datasets were combined, that additional LOAD loci become visible and confirmed [95–97]. The most recent efforts by the Alzheimer Disease Genetics Consortium (ADGC) and the International Genomics of Alzheimer's Project (IGAP) have greatly increased the number of known loci associated with LOAD. In the 2011 Naj *et al.* report, a three-stage design (discovery stage 1, replication stages 2–3) was utilized; this analysis evaluated >18,000 cases and >29,000 controls using both joint- and meta-analysis approaches and novel genome-wide significant hits were detected at SNPs in *MS4A4A*, *CD2AP*, *EPHA1* and *CD33* [96]. In Lambert *et al.* 2013, the IGAP reported an additional eleven novel LOAD susceptibility loci after analyzing genotyped and imputed data in a two-stage meta-analysis of >25,000 cases and >48,000 controls [95]. There are now over 20 loci identified that influence LOAD [95]. Importantly, using the pathway approach, the amyloid precursor protein and tau pathways are confirmed by this most recent large GWAS in addition to the newly implicated hippocampal synaptic function, cytoskeletal function and axonal transport, regulation of gene expression and post-translational modification of proteins, and microglial and myeloid cell function pathways [95].

4. Multiple Sclerosis

Multiple sclerosis (MS) is a common cause of neurological disability involving inflammatory demyelination of the central nervous system [98–101]. There is ample evidence that MS has a strong genetic component, but like so many other complex diseases, non-genetic influences are also important (e.g., [99,102–104]). MS is also a complex, heterogeneous disease in which significant efforts to unravel the role of genetics have been made. Unlike both AMD and LOAD, the first and strongest genetic effect in MS was identified well before the HGP was undertaken. Because MS is an autoimmune disease, it was strongly suspected that the major histocompatibility locus (*MHC*) would be involved. More specifically, there was a focus on the human leukocyte antigen loci on chromosome 6. In the early 1970's the *HLA* loci could be genotyped using blood antigen reactions, allowing assignment of genotypes without directly examining the DNA. Through a number of efforts (e.g., [99,103–109]) a strong risk association with the *HLA-DR* locus, and specifically the 15*01 allele was identified.

Despite this auspicious beginning, identifying additional MS loci languished. As with the other complex diseases, genetic linkage analysis was applied to multiplex MS families, with varying results. Some early genetic linkage studies confirmed the role of HLA [110], while others did not [111]. Additional studies, using the increasingly dense DNA marker sets and larger datasets, ultimately demonstrated and confirmed that the *HLA* locus was the single largest genetic effect, and that any other MS loci would have at most modest individual effects [109,110,112–116]. These studies did highlight several other possible loci, but did not have the resolution to identify specific associated genetic variations [115,116].

Finally, in 2007, nearly 30 years after the initial association finding, a second locus for MS was identified [117,118]. Gregory *et al.* employed a genomic convergence approach that integrated data from genetic linkage studies, genetic association studies, model system gene expression data, and *in vitro* functional data to narrow in on a specific locus and a functional polymorphism in the interleukin-7 receptor α chain (*IL7R*) [117]. Independently, the International Multiple Sclerosis Genetics Consortium (IMSGC) published results from one of the first large-scale GWAS studies, using 334,923 GWAS SNPs. The IMSGC used a hybrid study design that included a family-based study of 931 family trios and an independent dataset set of cases and controls [118]. These analyses confirmed the role of genetic variation in *IL7RA* and also highlighted variations in *IL2RA*.

These results also had broad implications for the field of MS resarch. The IMSGC GWAS was still one of the first such studies done with a well-powered dataset and demonstrated that family-based and case-control GWAS approaches were both useful methods for exploring genetic information. In addition, like AMD, the convergence of independent approaches (GWAS and gene-targeted methods) further validated that GWAS could identify relevant associated loci. Subsequent studies with much larger datasets [96,119,120] have now identified over 100 total loci associated with MS.

Efforts in MS have shown substantial increases in the number of independent loci associated with this disease. The most recent IMSGC study evaluated in two stages more than 80,000 individuals of European ancestry [119]. This analysis expanded the known MS loci by 48, raising the total number of discrete MS-associated loci to 103. In addition, the IMSGC interrogated specific genomic regions and hypotheses using a custom array, the Immunochip. The group efficiently utilized extensive amounts of data by assembling multiple studies and utilizing imputation methods and then applying conditional and joint analysis methods [119]. Such methods are becoming more common in the efforts to expand the power to detect common variation in many multifactorial diseases. The strongest of the novel hits from this analysis implicates a SNP in the region between *BCL10* and *DDAH1* [119]; *BCL10* is an activator of nuclear factor (NF)-κB signaling which is involved in gene expression control of inflammation, immunity, cell proliferation and apoptosis and has been explored as a clinical target for MS [73,119,121].

Pathway analysis in MS has also proven useful. The IMSGC study additionally sought to evaluate the Gene Ontology (GO) processes of the associated variants using the MetaCore ([122]); their results indicated, as expected, that most variants fall in or near genes with immune function [119]. Another recent endeavor to evaluate pathways involved in MS utilized results from eight MS GWAS datasets and prioritized genes in the cell adhesion molecule (CAM) biological pathway with the Cytoscape software [120,123]. Their findings highlighted five networks that were associated with susceptibility to MS—again supporting the utility of expanding beyond traditional case-control association analyses of GWAS data and encouraging the use of multiple datasets to determine enrichment of signals that might otherwise not have been detected using a traditional GWAS approach [120].

5. Conclusions and Future Directions

Since the completion of the HGP and shortly after the first GWAS, thousands more GWAS have been reported [40]; these have brought forth great progress in numerous diseases that previously were only hypothesized to have a genetic component. Large-scale collaborative efforts have raised the number of known AMD, AD, and MS loci to 19 [61], 20 [95], and 103 [119], respectively. Efforts to increase sample size have been successful, as evidenced by the largest and most recently reported analyses of AMD [61], AD [95] and MS [92], which each evaluated >74,000 individuals; however, other techniques are necessary to evaluate and explore as data becomes increasingly large and complex. Whole exome and whole genome sequencing are more recent approaches to generating genetic data that allow investigation far beyond the capabilities of the GWAS, and their utility is just starting to take shape in studies of many complex common diseases, including those mentioned herein. These will enable the study of rare and low frequency variants, which have been implicated as a potential source of missing heritability in many genetic diseases [66]. Analysis of data from exome arrays, designed to jointly interrogate data relevant to association studies of common variants and sequencing studies of rare variants, will improve genetic analysis of disease by providing greater coverage of known susceptibility loci and enhancing the likelihood for discovery of novel disease loci.

For each of these diseases, the fact that there is a genetic component is irrefutable; this knowledge has, over the past decade, most certainly been confirmed and expounded upon with the completion of the human genome sequence. As genetic knowledge continues to grow, and as clinical phenotyping techniques improve, further genetic variation influencing AMD, AD, and MS will likely be detectable and, hopefully, their roles in these diseases will be more clearly defined [124]. We can anticipate that as our understanding of genetic etiology of these diseases grows, future studies will further explore rare variations contributing to disease, the role of copy number variants, and the genetics of these diseases in non-European populations. Additionally, the role of currently undetermined environmental factors and their interactions with genetic variants must continue to be elucidated. The global objective of prior and ongoing studies is certainly to improve the current comprehension of new and existing disease loci in order that the biology of these diseases can be fully explicated in the hope of attaining improved strategies for disease treatment and prevention in the future.

The last decade has doubtlessly ushered in dramatic advances in the amount of shared data available to genetic researchers. Resources such as the NHLBI Grand Opportunity Exome Sequencing Project [125], the 1000Genomes [126], ENCODE [127], and the International HapMap Project [128,129] provide seemingly limitless amounts of data—all geared toward further understanding the intricacies of the human genome and how alterations of it influence human variation. We have provided a brief genetic history of three diseases that are exemplars of developing approaches to apply the incredible resources of the HGP. Such progress will most certainly continue to improve and exponentially increase in the next decade to facilitate a greater understanding of these and other complex diseases, as well as usher in the realization of personalized medicine.

Acknowledgments

This work was supported by NIH T32 EY007157 (Jessica N. Cooke Bailey), Alzheimer's Disease Genetics Consortium, funded by NIA grant U01AG032984 (Margaret A. Pericak-Vance, Jonathan L. Haines), BrightFocus Foundation grant A2011048 (Margaret A. Pericak-Vance), DOD grant W81XWH-12-1-0013 (Margaret A. Pericak-Vance, Jonathan L. Haines), NEI grants 7R01EY012118 (Margaret A. Pericak-Vance, Jonathan L. Haines), 1R01EY022310 (Jonathan L. Haines, Margaret A. Pericak-Vance), 1R01EY023164 (Margaret A. Pericak-Vance), and 1R01EY020928 (Margaret A. Pericak-Vance), NIA grants 1R01AG027944 (Margaret A. Pericak-Vance, Jonathan L. Haines) and R01AG19085 (Margaret A. Pericak-Vance, Jonathan L. Haines) and NINDS grant R01NS032830 (Margaret A. Pericak-Vance, Jonathan L. Haines).

Author Contributions

Jessica N. Cooke Bailey wrote and edited the manuscript. Margaret A. Pericak-Vance conceived the idea for the manuscript and reviewed and edited the manuscript. Jonathan L. Haines conceived the idea for the manuscript, wrote, reviewed and edited the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Seddon, J.M.; Cote, J.; Page, W.F.; Aggen, S.H.; Neale, M.C. The US twin study of age-related macular degeneration: Relative roles of genetic and environmental influences. *Arch. Ophthalmol.* **2005**, *123*, 321–327.
2. Hammond, C.J.; Webster, A.R.; Snieder, H.; Bird, A.C.; Gilbert, C.E.; Spector, T.D. Genetic influence on early age-related maculopathy: A twin study. *Ophthalmology* **2002**, *109*, 730–736.
3. Klaver, C.C.; Wolfs, R.C.; Assink, J.J.; van Duijn, C.M.; Hofman, A.; de Jong, P.T. Genetic risk of age-related maculopathy. Population-based familial aggregation study. *Arch. Ophthalmol.* **1998**, *116*, 1646–1651.
4. Klein, M.L.; Mauldin, W.M.; Stoumbos, V.D. Heredity and age-related macular degeneration. Observations in monozygotic twins. *Arch. Ophthalmol.* **1994**, *112*, 932–937.
5. Meyers, S.M. A twin study on age-related macular degeneration. *Trans. Am. Ophthalmol. Soc.* **1994**, *92*, 775–843.
6. Heiba, I.M.; Elston, R.C.; Klein, B.E.; Klein, R. Sibling correlations and segregation analysis of age-related maculopathy: The beaver dam eye study. *Genet. Epidemiol.* **1994**, *11*, 51–67.
7. Seddon, J.M.; Cote, J.; Davis, N.; Rosner, B. Progression of age-related macular degeneration: Association with body mass index, waist circumference, and waist-hip ratio. *Arch. Ophthalmol.* **2003**, *121*, 785–792.

8. Anderson, D.H.; Mullins, R.F.; Hageman, G.S.; Johnson, L.V. A role for local inflammation in the formation of drusen in the aging eye. *Am. J. Ophthalmol.* **2002**, *134*, 411–431.
9. Anderson, D.H.; Radeke, M.J.; Gallo, N.B.; Chapin, E.A.; Johnson, P.T.; Curletti, C.R.; Hancox, L.S.; Hu, J.; Ebright, J.N.; Malek, G.; *et al.* The pivotal role of the complement system in aging and age-related macular degeneration: Hypothesis re-visited. *Prog. Retin. Eye Res.* **2010**, *29*, 95–112.
10. Ding, X.; Patel, M.; Chan, C.C. Molecular pathology of age-related macular degeneration. *Prog. Retin. Eye Res.* **2009**, *28*, 1–18.
11. Patel, M.; Chan, C.C. Immunopathological aspects of age-related macular degeneration. *Semin. Immunopathol.* **2008**, *30*, 97–110.
12. Penfold, P.L.; Killingsworth, M.C.; Sarks, S.H. Senile macular degeneration: The involvement of immunocompetent cells. *Graefes Arch. Clin. Exp. Ophthalmol.* **1985**, *223*, 69–76.
13. Tuo, J.; Grob, S.; Zhang, K.; Chan, C.C. Genetics of immunological and inflammatory components in age-related macular degeneration. *Ocul. Immunol. Inflamm.* **2012**, *20*, 27–36.
14. Abecasis, G.R.; Yashar, B.M.; Zhao, Y.; Ghiasvand, N.M.; Zareparsi, S.; Branham, K.E.; Reddick, A.C.; Trager, E.H.; Yoshida, S.; Bahling, J.; *et al.* Age-related macular degeneration: A high-resolution genome scan for susceptibility loci in a population enriched for late-stage disease. *Am. J. Hum. Genet.* **2004**, *74*, 482–494.
15. Iyengar, S.K.; Song, D.; Klein, B.E.; Klein, R.; Schick, J.H.; Humphrey, J.; Millard, C.; Liptak, R.; Russo, K.; Jun, G.; *et al.* Dissection of genomewide-scan data in extended families reveals a major locus and oligogenic susceptibility for age-related macular degeneration. *Am. J. Hum. Genet.* **2004**, *74*, 20–39.
16. Klein, M.L.; Schultz, D.W.; Edwards, A.; Matise, T.C.; Rust, K.; Berselli, C.B.; Trzupek, K.; Weleber, R.G.; Ott, J.; Wirtz, M.K.; *et al.* Age-related macular degeneration. Clinical features in a large family and linkage to chromosome 1q. *Arch. Ophthalmol.* **1998**, *116*, 1082–1088.
17. Majewski, J.; Schultz, D.W.; Weleber, R.G.; Schain, M.B.; Edwards, A.O.; Matise, T.C.; Acott, T.S.; Ott, J.; Klein, M.L. Age-related macular degeneration—A genome scan in extended families. *Am. J. Hum. Genet.* **2003**, *73*, 540–550.
18. Seddon, J.M.; Santangelo, S.L.; Book, K.; Chong, S.; Cote, J. A genomewide scan for age-related macular degeneration provides evidence for linkage to several chromosomal regions. *Am. J. Hum. Genet.* **2003**, *73*, 780–790.
19. Tuo, J.; Bojanowski, C.M.; Chan, C.C. Genetic factors of age-related macular degeneration. *Prog. Retin. Eye Res.* **2004**, *23*, 229–249.
20. Weeks, D.E.; Conley, Y.P.; Mah, T.S.; Paul, T.O.; Morse, L.; Ngo-Chang, J.; Dailey, J.P.; Ferrell, R.E.; Gorin, M.B. A full genome scan for age-related maculopathy. *Hum. Mol. Genet.* **2000**, *9*, 1329–1349.
21. Weeks, D.E.; Conley, Y.P.; Tsai, H.J.; Mah, T.S.; Rosenfeld, P.J.; Paul, T.O.; Eller, A.W.; Morse, L.S.; Dailey, J.P.; Ferrell, R.E.; *et al.* Age-related maculopathy: An expanded genome-wide scan with evidence of susceptibility loci within the 1q31 and 17q25 regions. *Am. J. Ophthalmol.* **2001**, *132*, 682–692.

22. Weeks, D.E.; Conley, Y.P.; Tsai, H.J.; Mah, T.S.; Schmidt, S.; Postel, E.A.; Agarwal, A.; Haines, J.L.; Pericak-Vance, M.A.; Rosenfeld, P.J.; *et al.* Age-related maculopathy: A genomewide scan with continued evidence of susceptibility loci within the 1q31, 10q26, and 17q25 regions. *Am. J. Hum. Genet.* **2004**, *75*, 174–189.
23. Allikmets, R. A photoreceptor cell-specific ATP-binding transporter gene (ABCR) is mutated in recessive Stargardt macular dystrophy. *Nat. Genet.* **1997**, *17*, 122.
24. Allikmets, R. Further evidence for an association of ABCR alleles with age-related macular degeneration. The International ABCR Screening Consortium. *Am. J. Hum. Genet.* **2000**, *67*, 487–491.
25. Allikmets, R.; Shroyer, N.F.; Singh, N.; Seddon, J.M.; Lewis, R.A.; Bernstein, P.S.; Peiffer, A.; Zabriskie, N.A.; Li, Y.; Hutchinson, A.; *et al.* Mutation of the Stargardt disease gene (ABCR) in age-related macular degeneration. *Science* **1997**, *277*, 1805–1807.
26. Meyers, S.M.; Greene, T.; Gutman, F.A. A twin study of age-related macular degeneration. *Am. J. Ophthalmol.* **1995**, *120*, 757–766.
27. Shroyer, N.F.; Lewis, R.A.; Yatsenko, A.N.; Wensel, T.G.; Lupski, J.R. Cosegregation and functional analysis of mutant ABCR (ABCA4) alleles in families that manifest both Stargardt disease and age-related macular degeneration. *Hum. Mol. Genet.* **2001**, *10*, 2671–2678.
28. Fritsche, L.G.; Fleckenstein, M.; Fiebig, B.S.; Schmitz-Valckenberg, S.; Bindewald-Wittich, A.; Keilhauer, C.N.; Renner, A.B.; Mackensen, F.; Mossner, A.; Pauleikhoff, D.; *et al.* A subgroup of age-related macular degeneration is associated with mono-allelic sequence variants in the *ABCA4* gene. *Invest. Ophthalmol. Vis. Sci.* **2012**, *53*, 2112–2118.
29. Guymer, R.H.; Heon, E.; Lotery, A.J.; Munier, F.L.; Schorderet, D.F.; Baird, P.N.; McNeil, R.J.; Haines, H.; Sheffield, V.C.; Stone, E.M. Variation of codons 1961 and 2177 of the Stargardt disease gene is not associated with age-related macular degeneration. *Arch. Ophthalmol.* **2001**, *119*, 745–751.
30. Rivera, A.; White, K.; Stohr, H.; Steiner, K.; Hemmrich, N.; Grimm, T.; Jurklies, B.; Lorenz, B.; Scholl, H.P.; Apfelstedt-Sylla, E.; *et al.* A comprehensive survey of sequence variation in the ABCA4 (ABCR) gene in Stargardt disease and age-related macular degeneration. *Am. J. Hum. Genet.* **2000**, *67*, 800–813.
31. Webster, A.R.; Heon, E.; Lotery, A.J.; Vandenburgh, K.; Casavant, T.L.; Oh, K.T.; Beck, G.; Fishman, G.A.; Lam, B.L.; Levin, A.; *et al.* An analysis of allelic variation in the ABCA4 gene. *Invest. Ophthalmol. Vis. Sci.* **2001**, *42*, 1179–1189.
32. Fisher, S.A.; Abecasis, G.R.; Yashar, B.M.; Zareparsi, S.; Swaroop, A.; Iyengar, S.K.; Klein, B.E.; Klein, R.; Lee, K.E.; Majewski, J.; *et al.* Meta-analysis of genome scans of age-related macular degeneration. *Hum. Mol. Genet.* **2005**, *14*, 2257–2264.
33. dbSNP. Available online: http://www.ncbi.nlm.nih.gov/SNP/ (accessed on 7 March 2014).
34. Thorisson, G.A.; Stein, L.D. The SNP Consortium website: Past, present and future. *Nucleic Acids Res.* **2003**, *31*, 124–127.
35. Klein, R.J.; Zeiss, C.; Chew, E.Y.; Tsai J.Y.; Sackler R.S.; Haynes, C.; Henning, A.K.; SanGiovanni J.P.; Mane, S.M.; Mayne, S.T.; *et al*. Complement factor H polymorphism in age-related macular degeneration. *Science* **2005**, *308*, 385–389.

36. Haines, J.L.; Hauser, M.A.; Schmidt, S.; Scott, W.K.; Olson, L.M.; Gallins, P.; Spencer, K.L.; Kwan, S.Y.; Noureddine, M.; Gilbert, J.R.; *et al.* Complement factor H variant increases the risk of age-related macular degeneration. *Science* **2005**, *308*, 419–421.
37. Edwards, A.O.; Ritter, R., III; Abel, K.J.; Manning, A.; Panhuysen, C.; Farrer, L.A. Complement factor H polymorphism and age-related macular degeneration. *Science* **2005**, *308*, 421–424.
38. Hageman, G.S.; Anderson, D.H.; Johnson, L.V.; Hancox, L.S.; Taiber, A.J.; Hardisty, L.I.; Hageman, J.L.; Stockman, H.A.; Borchardt, J.D.; Gehrs, K.M.; *et al.* A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 7227–7232.
39. Zerhouni, E. House subcommittee of labor-HHS-Education appropriations. Available online: http://legislative.csancer.gov/files/appropriations-2006-04-06.pdf (accessed on 1 March 2014).
40. Welter, D.; Macarthur, J.; Morales, J.; Burdett, T.; Hall, P.; Junkins, H.; Klemm, A.; Flicek, P.; Manolio, T.; Hindorff, L.; *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **2014**, *42*, D1001–D1006.
41. Goate, A.; Chartier-Harlin, M.C.; Mullan, M.; Brown, J.; Crawford, F.; Fidani, L.; Giuffra, L.; Haynes, A.; Irving, N.; James, L.; *et al.* Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **1991**, *349*, 704–706.
42. Jakobsdottir, J.; Conley, Y.P.; Weeks, D.E.; Mah, T.S.; Ferrell, R.E.; Gorin, M.B. Susceptibility genes for age-related maculopathy on chromosome 10q26 29. *Am. J. Hum. Genet.* **2005**, *77*, 389–407.
43. Schmidt, S.; Hauser, M.A.; Scott, W.K.; Postel, E.A.; Agarwal, A.; Gallins, P.; Wong, F.; Chen, Y.S.; Spencer, K.; Schnetz-Boutaud, N.; *et al.* Cigarette smoking strongly modifies the association of LOC387715 and age-related macular degeneration. *Am. J. Hum. Genet.* **2006**, *78*, 852–864.
44. Schwartz, S.G.; Agarwal, A.; Kovach, J.L.; Gallins, P.J.; Cade, W.; Postel, E.A.; Wang, G.; Ayala-Haedo, J.; Spencer, K.M.; Haines, J.L.; *et al.* The ARMS2 A69S variant and bilateral advanced age-related macular degeneration. *Retina* **2012**, *32*, 1486–1491.
45. Shuler, R.K., Jr.; Hauser, M.A.; Caldwell, J.; Gallins, P.; Schmidt, S.; Scott, W.K.; Agarwal, A.; Haines, J.L.; Pericak-Vance, M.A.; Postel, E.A. Neovascular age-related macular degeneration and its association with LOC387715 and complement factor H polymorphism. *Arch. Ophthalmol.* **2007**, *125*, 63–67.
46. Wang, G. Chromosome 10q26 locus and age-related macular degeneration: A progress update. *Exp. Eye Res.* **2014**, *119*, 1–7.
47. Wang, G.; Dubovy, S.R.; Kovach, J.L.; Schwartz, S.G.; Agarwal, A.; Scott, W.K.; Haines, J.L.; Pericak-Vance, M.A. Variants at chromosome 10q26 locus and the expression of HTRA1 in the retina. *Exp. Eye Res.* **2013**, *112*, 102–105.
48. Dewan, A.; Liu, M.; Hartman, S.; Zhang, S.S.; Liu, D.T.; Zhao, C.; Tam, P.O.; Chan, W.M.; Lam, D.S.; Snyder, M.; *et al.* HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* **2006**, *314*, 989–992.

49. Yang, Z.; Camp, N.J.; Sun, H.; Tong, Z.; Gibbs, D.; Cameron, D.J.; Chen, H.; Zhao, Y.; Pearson, E.; Li, X.; *et al.* A variant of the HTRA1 gene increases susceptibility to age-related macular degeneration. *Science* **2006**, *314*, 992–993.
50. Hughes, A.E.; Orr, N.; Esfandiary, H.; Diaz-Torres, M.; Goodship, T.; Chakravarthy, U. A common CFH haplotype, with deletion of CFHR1 and CFHR3, is associated with lower risk of age-related macular degeneration. *Nat. Genet.* **2006**, *38*, 1173–1177.
51. Fritsche, L.G.; Lauer, N.; Hartmann, A.; Stippa, S.; Keilhauer, C.N.; Oppermann, M.; Pandey, M.K.; Kohl, J.; Zipfel, P.F.; Weber, B.H.; *et al.* An imbalance of human complement regulatory proteins CFHR1, CFHR3 and factor H influences risk for age-related macular degeneration (AMD). *Hum. Mol. Genet.* **2010**, *19*, 4694–4704.
52. Sawitzke, J.; Im, K.M.; Kostiha, B.; Dean, M.; Gold, B. Association assessment of copy number polymorphism and risk of age-related macular degeneration. *Ophthalmology* **2011**, *118*, 2442–2446.
53. Spencer, K.L.; Hauser, M.A.; Olson, L.M.; Schmidt, S.; Scott, W.K.; Gallins, P.; Agarwal, A.; Postel, E.A.; Pericak-Vance, M.A.; Haines, J.L. Protective effect of complement factor B and complement component 2 variants in age-related macular degeneration. *Hum. Mol. Genet.* **2007**, *16*, 1986–1992.
54. Gold, B.; Merriam, J.E.; Zernant, J.; Hancox, L.S.; Taiber, A.J.; Gehrs, K.; Cramer, K.; Neel, J.; Bergeron, J.; Barile, G.R.; *et al.* Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nat. Genet.* **2006**, *38*, 458–462.
55. Maller, J.B.; Fagerness, J.A.; Reynolds, R.C.; Neale, B.M.; Daly, M.J.; Seddon, J.M. Variation in complement factor 3 is associated with risk of age-related macular degeneration. *Nat. Genet.* **2007**, *39*, 1200–1201.
56. Spencer, K.L.; Olson, L.M.; Anderson, B.M.; Schnetz-Boutaud, N.; Scott, W.K.; Gallins, P.; Agarwal, A.; Postel, E.A.; Pericak-Vance, M.A.; Haines, J.L. C3 R102G polymorphism increases risk of age-related macular degeneration. *Hum. Mol. Genet.* **2008**, *17*, 1821–1824.
57. Yates, J.R.; Sepp, T.; Matharu, B.K.; Khan, J.C.; Thurlby, D.A.; Shahid, H.; Clayton, D.G.; Hayward, C.; Morgan, J.; Wright, A.F.; *et al.* Complement C3 variant and the risk of age-related macular degeneration. *N. Engl. J. Med.* **2007**, *357*, 553–561.
58. Fagerness, J.A.; Maller, J.B.; Neale, B.M.; Reynolds, R.C.; Daly, M.J.; Seddon, J.M. Variation near complement factor I is associated with risk of advanced AMD. *Eur. J. Hum. Genet.* **2009**, *17*, 100–104.
59. Stanton, C.M.; Yates, J.R.; den Hollander, A.I.; Seddon, J.M.; Swaroop, A.; Stambolian, D.; Fauser, S.; Hoyng, C.; Yu, Y.; Atsuhiro, K.; *et al.* Complement factor D in age-related macular degeneration. *Invest. Ophthalmol. Vis. Sci.* **2011**, *52*, 8828–8834.
60. Chen, W.; Stambolian, D.; Edwards, A.O.; Branham, K.E.; Othman, M.; Jakobsdottir, J.; Tosakulwong, N.; Pericak-Vance, M.A.; Campochiaro, P.A.; Klein, M.L.; *et al.* Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 7401–7406.

61. Fritsche, L.G.; Chen, W.; Schu, M.; Yaspan, B.L.; Yu, Y.; Thorleifsson, G.; Zack, D.J.; Arakawa, S.; Cipriani, V.; Ripke, S.; *et al.* Seven new loci associated with age-related macular degeneration. *Nat. Genet.* **2013**, *45*, 433–439.
62. Neale, B.M.; Fagerness, J.; Reynolds, R.; Sobrin, L.; Parker, M.; Raychaudhuri, S.; Tan, P.L.; Oh, E.C.; Merriam, J.E.; Souied, E.; *et al.* Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (LIPC). *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 7395–7400.
63. Helgason, H.; Sulem, P.; Duvvari, M.R.; Luo, H.; Thorleifsson, G.; Stefansson, H.; Jonsdottir, I.; Masson, G.; Gudbjartsson, D.F.; Walters, G.B.; *et al.* A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. *Nat. Genet.* **2013**, *45*, 1371–1374.
64. Ryu, E.; Fridley, B.L.; Tosakulwong, N.; Bailey, K.R.; Edwards, A.O. Genome-wide association analyses of genetic, phenotypic, and environmental risks in the age-related eye disease study. *Mol. Vis.* **2010**, *16*, 2811–2821.
65. Scheetz, T.E.; Fingert, J.H.; Wang, K.; Kuehn, M.H.; Knudtson, K.L.; Alward, W.L.; Boldt, H.C.; Russell, S.R.; Folk, J.C.; Casavant, T.L.; *et al.* A genome-wide association study for primary open angle glaucoma and macular degeneration reveals novel Loci. *PLoS One* **2013**, *8*, e58657.
66. Manolio, T.A.; Collins, F.S.; Cox, N.J.; Goldstein, D.B.; Hindorff, L.A.; Hunter, D.J.; McCarthy, M.I.; Ramos, E.M.; Cardon, L.R.; Chakravarti, A.; *et al.* Finding the missing heritability of complex diseases. *Nature* **2009**, *461*, 747–753.
67. Li, Y.; Willer, C.; Sanna, S.; Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **2009**, *10*, 387–406.
68. Marchini, J.; Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **2010**, *11*, 499–511.
69. Seddon, J.M.; Yu, Y.; Miller, E.C.; Reynolds, R.; Tan, P.L.; Gowrisankar, S.; Goldstein, J.I.; Triebwasser, M.; Anderson, H.E.; Zerbib, J.; *et al.* Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration. *Nat. Genet.* **2013**, *45*, 1366–1370.
70. Yaspan, B.L.; Bush, W.S.; Torstenson, E.S.; Ma, D.; Pericak-Vance, M.A.; Ritchie, M.D.; Sutcliffe, J.S.; Haines, J.L. Genetic analysis of biological pathway data through genomic randomization. *Hum. Genet.* **2011**, *129*, 563–571.
71. Lee, P.H.; O'Dushlaine, C.; Thomas, B.; Purcell, S.M. INRICH: Interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* **2012**, *28*, 1797–1799.
72. Yang, J.; Li, Y.; Chan, L.; Tsai, Y.T.; Wu, W.H.; Nguyen, H.V.; Hsu, C.W.; Li, X.; Brown, L.M.; Egli, D.; *et al.* Validation of genome-wide association study (GWAS)-identified disease risk alleles with patient-specific stem cell lines. *Hum. Mol. Genet.* **2014**, *23*, 3445–3455.
73. Van Lookeren, C.M.; Le Couter, J.; Yaspan, B.L.; Ye, W. Mechanisms of age-related macular degeneration and therapeutic opportunities. *J. Pathol.* **2014**, *232*, 151–164.

74. Herrup, K. Reimagining Alzheimer's disease—An age-based hypothesis. *J. Neurosci.* **2010**, *30*, 16755–16762.
75. Querfurth, H.W.; LaFerla, F.M. Alzheimer's disease. *N. Engl. J. Med.* **2010**, *362*, 329–344.
76. Breitner, J.C.; Folstein, M.F. Familial Alzheimer Dementia: A prevalent disorder with specific clinical features. *Psychol. Med.* **1984**, *14*, 63–80.
77. Breitner, J.C.; Folstein, M.F. Familial nature of Alzheimer's disease. *N. Engl. J. Med.* **1984**, *311*, 192.
78. Folstein, M. Alzheimer's disease: Challenge to psychiatry. *Hosp. Community Psychiatry* **1984**, *35*, 111.
79. Pericak-Vance, M.A.; Haines, J.L. Genetic susceptibility to Alzheimer disease. *Trends Genet.* **1995**, *11*, 504–508.
80. Powell, D.; Folstein, M.F. Pedigree study of familial Alzheimer disease. *J. Neurogenet.* **1984**, *1*, 189–197.
81. St George-Hyslop, P.H.; Tanzi, R.E.; Polinsky, R.J.; Haines, J.L.; Nee, L.; Watkins, P.C.; Myers, R.H.; Feldman, R.G.; Pollen, D.; Drachman, D.; *et al.* The genetic defect causing familial Alzheimer's disease maps on chromosome 21. *Science* **1987**, *235*, 885–890.
82. Levy-Lahad, E.; Lahad, A.; Wijsman, E.M.; Bird, T.D.; Schellenberg, G.D. Apolipoprotein E genotypes and age of onset in early-onset familial Alzheimer's disease. *Ann. Neurol.* **1995**, *38*, 678–680.
83. Levy-Lahad, E.; Wasco, W.; Poorkaj, P.; Romano, D.M.; Oshima, J.; Pettingell, W.H.; Yu, C.E.; Jondro, P.D.; Schmidt, S.D.; Wang, K.; *et al.* Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* **1995**, *269*, 973–977.
84. Levy-Lahad, E.; Wijsman, E.M.; Nemens, E.; Anderson, L.; Goddard, K.A.; Weber, J.L.; Bird, T.D.; Schellenberg, G.D. A familial Alzheimer's disease locus on chromosome 1. *Science* **1995**, *269*, 970–973.
85. Sherrington, R.; Rogaev, E.I.; Liang, Y.; Rogaeva, E.A.; Levesque, G.; Ikeda, M.; Chi, H.; Lin, C.; Li, G.; Holman, K.; *et al.* Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **1995**, *375*, 754–760.
86. St George-Hyslop, P.; Haines, J.; Rogaev, E.; Mortilla, M.; Vaula, G.; Pericak-Vance, M.; Foncin, J.F.; Montesi, M.; Bruni, A.; Sorbi, S.; *et al.* Genetic evidence for a novel familial Alzheimer's disease locus on chromosome 14. *Nat. Genet.* **1992**, *2*, 330–334.
87. Ridge, P.G.; Mukherjee, S.; Crane, P.K.; Kauwe, J.S. Alzheimer's disease: Analyzing the missing heritability. *PLoS One* **2013**, *8*, e79771.
88. Pericak-Vance, M.A.; Bebout, J.L.; Gaskell, P.C., Jr.; Yamaoka, L.H.; Hung, W.Y.; Alberts, M.J.; Walker, A.P.; Bartlett, R.J.; Haynes, C.A.; Welsh, K.A.; *et al.* Linkage studies in familial Alzheimer disease: Evidence for chromosome 19 linkage. *Am. J. Hum. Genet.* **1991**, *48*, 1034–1050.
89. Pericak-Vance, M.A.; Yamaoka, L.H.; Haynes, C.S.; Speer, M.C.; Haines, J.L.; Gaskell, P.C.; Hung, W.Y.; Clark, C.M.; Heyman, A.L.; Trofatter, J.A.; *et al.* Genetic linkage studies in Alzheimer's disease families. *Exp. Neurol.* **1988**, *102*, 271–279.

90. Corder, E.H.; Saunders, A.M.; Strittmatter, W.J.; Schmechel, D.E.; Gaskell, P.C.; Small, G.W.; Roses, A.D.; Haines, J.L.; Pericak-Vance, M.A. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **1993**, *261*, 921–923.
91. Corder, E.H.; Saunders, A.M.; Risch, N.J.; Strittmatter, W.J.; Schmechel, D.E.; Gaskell, P.C., Jr.; Rimmler, J.B.; Locke, P.A.; Conneally, P.M.; Schmader, K.E.; *et al.* Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nat. Genet.* **1994**, *7*, 180–184.
92. Beecham, G.W.; Martin, E.R.; Li, Y.J.; Slifer, M.A.; Gilbert, J.R.; Haines, J.L.; Pericak-Vance, M.A. Genome-wide association study implicates a chromosome 12 risk locus for late-onset Alzheimer disease. *Am. J. Hum. Genet.* **2009**, *84*, 35–43.
93. Coon, K.D.; Myers, A.J.; Craig, D.W.; Webster, J.A.; Pearson, J.V.; Lince, D.H.; Zismann, V.L.; Beach, T.G.; Leung, D.; Bryden, L.; *et al.* A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J. Clin. Psychiatry*. **2007**, *68*, 613–618.
94. Li, H.; Wetten, S.; Li, L.; St Jean, P.L.; Upmanyu, R.; Surh, L.; Hosford, D.; Barnes, M.R.; Briley, J.D.; Borrie, M.; *et al.* Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch. Neurol.* **2008**, *65*, 45–53.
95. Lambert, J.C.; Ibrahim-Verbaas, C.A.; Harold, D.; Naj, A.C.; Sims, R.; Bellenguez, C.; Jun, G.; Destefano, A.L.; Bis, J.C.; Beecham, G.W.; *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **2013**, *45*, 1452–1458.
96. Naj, A.C.; Jun, G.; Beecham, G.W.; Wang, L.S.; Vardarajan, B.N.; Buros, J.; Gallins, P.J.; Buxbaum, J.D.; Jarvik, G.P.; Crane, P.K.; *et al.* Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat. Genet.* **2011**, *43*, 436–441.
97. Seshadri, S.; Fitzpatrick, A.L.; Ikram, M.A.; Destefano, A.L.; Gudnason, V.; Boada, M.; Bis, J.C.; Smith, A.V.; Carassquillo, M.M.; Lambert, J.C.; *et al.* Genome-wide analysis of genetic loci associated with Alzheimer disease. *JAMA* **2010**, *303*, 1832–1840.
98. Gourraud, P.A.; Sdika, M.; Khankhanian, P.; Henry, R.G.; Beheshtian, A.; Matthews, P.M.; Hauser, S.L.; Oksenberg, J.R.; Pelletier, D.; Baranzini, S.E. A genome-wide association study of brain lesion distribution in multiple sclerosis. *Brain* **2013**, *136*, 1012–1024.
99. Hauser, S.L.; Chan, J.R.; Oksenberg, J.R. Multiple sclerosis: Prospects and promise. *Ann. Neurol.* **2013**, *74*, 317–327.
100. Nylander, A.; Hafler, D.A. Multiple sclerosis. *J. Clin. Invest*. **2012**, *122*, 1180–1188.
101. Oksenberg, J.R. Decoding multiple sclerosis: An update on genomics and future directions. *Expert. Rev. Neurother.* **2013**, *13*, 11–19.
102. Compston, A.; Coles, A. Multiple sclerosis. *Lancet* **2002**, *359*, 1221–1231.
103. Sadovnick, A.D.; Ebers, G.C. Epidemiology of multiple sclerosis: A critical overview. *Can. J. Neurol. Sci.* **1993**, *20*, 17–29.
104. Sawcer, S.; Compston, A. Multiple sclerosis: Light at the end of the tunnel. *Eur. J. Hum. Genet.* **2006**, *14*, 257–258.

105. Compston, A.; Sawcer, S. Genetic analysis of multiple sclerosis. *Curr. Neurol. Neurosci. Rep.* **2002**, *2*, 259–266.
106. Sadovnick, A.D. Familial recurrence risks and inheritance of multiple sclerosis. *Curr. Opin. Neurol. Neurosurg.* **1993**, *6*, 189–194.
107. Sadovnick, A.D.; Armstrong, H.; Rice, G.P.; Bulman, D.; Hashimoto, L.; Paty, D.W.; Hashimoto, S.A.; Warren, S.; Hader, W.; Murray, T.J.; *et al.* A population-based study of multiple sclerosis in twins: Update. *Ann. Neurol.* **1993**, *33*, 281–285.
108. Sadovnick, A.D.; Yee, I.M.; Guimond, C.; Reis, J.; Dyment, D.A.; Ebers, G.C. Age of onset in concordant twins and other relative pairs with multiple sclerosis. *Am. J. Epidemiol.* **2009**, *170*, 289–296.
109. Sawcer, S.; Ban, M.; Maranian, M.; Yeo, T.W.; Compston, A.; Kirby, A.; Daly, M.J.; de Jager, P.L.; Walsh, E.; Lander, E.S.; *et al.* A high-density screen for linkage in multiple sclerosis. *Am. J. Hum. Genet.* **2005**, *77*, 454–467.
110. Haines, J.L.; Ter-Minassian, M.; Bazyk, A.; Gusella, J.F.; Kim, D.J.; Terwedow, H.; Pericak-Vance, M.A.; Rimmler, J.B.; Haynes, C.S.; Roses, A.D.; *et al.* A complete genomic screen for multiple sclerosis underscores a role for the major histocompatability complex. The Multiple Sclerosis Genetics Group. *Nat. Genet.* **1996**, *13*, 469–471.
111. Ebers, G.C.; Kukay, K.; Bulman, D.E.; Sadovnick, A.D.; Rice, G.; Anderson, C.; Armstrong, H.; Cousin, K.; Bell, R.B.; Hader, W.; *et al.* A full genome search in multiple sclerosis. *Nat. Genet.* **1996**, *13*, 472–476.
112. Haines, J.L.; Bradford, Y.; Garcia, M.E.; Reed, A.D.; Neumeister, E.; Pericak-Vance, M.A.; Rimmler, J.B.; Menold, M.M.; Martin, E.R.; Oksenberg, J.R.; *et al.* Multiple susceptibility loci for multiple sclerosis. *Hum. Mol. Genet.* **2002**, *11*, 2251–2256.
113. Haines, J.L.; Terwedow, H.A.; Burgess, K.; Pericak-Vance, M.A.; Rimmler, J.B.; Martin, E.R.; Oksenberg, J.R.; Lincoln, R.; Zhang, D.Y.; Banatao, D.R.; *et al.* Linkage of the MHC to familial multiple sclerosis suggests genetic heterogeneity. The Multiple Sclerosis Genetics Group. *Hum. Mol. Genet.* **1998**, *7*, 1229–1234.
114. Kenealy, S.J.; Herrel, L.A.; Bradford, Y.; Schnetz-Boutaud, N.; Oksenberg, J.R.; Hauser, S.L.; Barcellos, L.F.; Schmidt, S.; Gregory, S.G.; Pericak-Vance, M.A.; *et al.* Examination of seven candidate regions for multiple sclerosis: Strong evidence of linkage to chromosome 1q44. *Genes Immun.* **2006**, *7*, 73–76.
115. McCauley, J.L.; Zuvich, R.L.; Bradford, Y.; Kenealy, S.J.; Schnetz-Boutaud, N.; Gregory, S.G.; Hauser, S.L.; Oksenberg, J.R.; Mortlock, D.P.; Pericak-Vance, M.A.; *et al.* Follow-up examination of linkage and association to chromosome 1q43 in multiple sclerosis. *Genes Immun.* **2009**, *10*, 624–630.
116. Pericak-Vance, M.A.; Rimmler, J.B.; Martin, E.R.; Haines, J.L.; Garcia, M.E.; Oksenberg, J.R.; Barcellos, L.F.; Lincoln, R.; Goodkin, D.E.; Hauser, S.L. Linkage and association analysis of chromosome 19q13 in multiple sclerosis. *Neurogenetics* **2001**, *3*, 195–201.
117. Gregory, S.G.; Schmidt, S.; Seth, P.; Oksenberg, J.R.; Hart, J.; Prokop, A.; Caillier, S.J.; Ban, M.; Goris, A.; Barcellos, L.F.; *et al.* Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. *Nat. Genet.* **2007**, *39*, 1083–1091.

118. Hafler, D.A.; Compston, A.; Sawcer, S.; Lander, E.S.; Daly, M.J.; de Jager, P.L.; de Bakker, P.I.; Gabriel, S.B.; Mirel, D.B.; Ivinson, A.J.; *et al.* Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* **2007**, *357*, 851–862.
119. Beecham, A.H.; Patsopoulos, N.A.; Xifara, D.K.; Davis, M.F.; Kemppinen, A.; Cotsapas, C.; Shah, T.S.; Spencer, C.; Booth, D.; Goris, A.; *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **2013**, *45*, 1353–1360.
120. Damotte, V.; Guillot-Noel, L.; Patsopoulos, N.A.; Madireddy, L.; El, B.M.; Ban, M.; Baranzini, S.; Barcellos, L.; Beecham, G.; Beecham, A.; *et al.* A gene pathway analysis highlights the role of cellular adhesion molecules in multiple sclerosis susceptibility. *Genes Immun.* **2014**, *15*, 126–132.
121. Yan, J.; Greer, J.M. NF-kappa B, a potential therapeutic target for the treatment of multiple sclerosis. *CNS Neurol. Disord. Drug Targets* **2008**, *7*, 536–557.
122. Metacore. Available online: http://thomsonreuters.com/metacore/ (accessed on 7 March 2014).
123. Cytoscape. Available online: http://www.cytoscape.org/ (accessed on 7 March 2014).
124. Ratnapriya, R.; Chew, E.Y. Age-related macular degeneration-clinical review and genetics update. *Clin. Genet.* **2013**, *84*, 160–166.
125. ESP. Available online: https://esp.gs.washington.edu/drupal/ (accessed on 7 March 2014).
126. 1000Genomes. Available online: http://www.1000genomes.org/ (accessed on 7 March 2014).
127. ENCODE. Available online: https://genome.ucsc.edu/ENCODE/ (accessed on 7 March 2014).
128. International HapMap Consortium. A haplotype map of the human genome. *Nature* **2005**, *437*, 1299–1320.
129. HapMap. Available online: http://hapmap.ncbi.nlm.nih.gov/ (accessed on 7 March 2014).

From Genotype to Functional Phenotype: Unraveling the Metabolomic Features of Colorectal Cancer

Oliver F. Bathe and Farshad Farshidfar

Abstract: Much effort in recent years has been expended in defining the genomic and epigenetic alterations that characterize colorectal adenocarcinoma and its subtypes. However, little is known about the functional ramifications related to various subtypes. Metabolomics, the study of small molecule intermediates in disease, provides a snapshot of the functional phenotype of colorectal cancer. Data, thus far, have characterized some of the metabolic perturbations that accompany colorectal cancer. However, further studies will be required to identify biologically meaningful metabolic subsets, including those corresponding to specific genetic aberrations. Moreover, further studies are necessary to distinguish changes due to tumor and the host response to tumor.

Reprinted from *Genes*. Cite as: Bathe, O.F.; Farshidfar, F. From Genotype to Functional Phenotype: Unraveling the Metabolomic Features of Colorectal Cancer. *Genes* **2014**, *5*, 536-560.

1. Introduction

Colorectal cancer (CRC) is the second leading cause of cancer death in the Western world. CRC invades locally to involve successive layers of the colon or rectum, then spreads via the lymphatic system to regional lymph nodes and/or metastasizes hematogenously to involve distant organs, such as the liver or lungs. Distant metastatic disease is present in about 25% of individuals. The features of the tumor that describe this behavior are reflected in the TNM (Tumor, Lymph Node, Metastasis) staging classification of tumor, where higher degrees of disease portend a worse prognosis. Other clinical and pathologic features that are well known to reflect biological behavior include the presence of obstruction or perforation; degree of differentiation; and presence of lymph or vascular invasion.

In recent years, it has become apparent that CRC has variable biologic behavior, which is not always reflected by its clinicopathological features. Rather, its growth properties, its tendency to metastasize, and its susceptibility to treatments are a function of its molecular properties. For example, colorectal cancers with lymph node involvement more frequently have *BRAF* and *KRAS* mutations [1,2], high levels of CCR7 [3], low levels of thymidylate gene expression [4], as well as *p16INK4A* promoter methylation [1]. The recognition that the molecular properties of CRC dictate biological behavior has led various investigators to subclassify CRC based on its "molecular signature" at the transcriptomic level [5–8]. Even more recently, coordinated efforts have been made to obtain highly detailed analysis of the CRC genome, as well as the downstream transcriptomic and epigenomic events [9]. As a result of these efforts, it is now feasible to refine the classification of CRC, based on molecular pathogenesis.

2. Molecular Subclassification of CRC: From Genomics to Phenotype

Most investigators classify sporadic CRC according to molecular pathway leading to its pathogenesis (Figure 1): chromosomal instability (CIN), microsatellite instability (MSI) and CpG island methylator phenotype (CIMP). Others have proposed a molecular classification system in which groups of CRC are defined according only to MSI and CIMP status in conjunction with clinicopathological features (Figure 2) [10,11]. More recently, based on data generated from The Cancer Genome Atlas Project, CRC has been designated as hypermutated or non-hypermutated, based on mutation rates (Figure 3).

Figure 1. Classification of CRC by pathogenic pathway. The classical pathway involves a progressive accumulation of mutations due to chromosomal instability as an adenoma develops into adenocarcinoma. Serrated polyps, which are thought to develop from hyperplastic polyps, are generated due to microsatellite instability and/or high levels of CpG island methylation. The adenocarcinomas that emerge from that pathway have distinct clinical and pathological features.

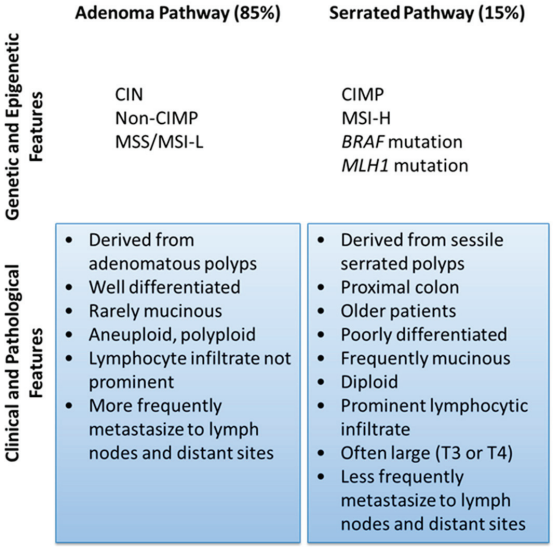


Figure 2. Relationship of CIMP expression phenotype and microsatellite instability. There is significant overlap between MSI-H tumors and CIMP-H tumors, although microsatellite stable (MSS) tumors and tumors with a low level of MSI (MSI-L) may have high levels of CpG island methylation.

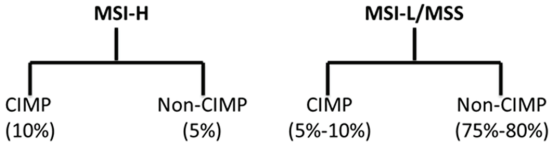
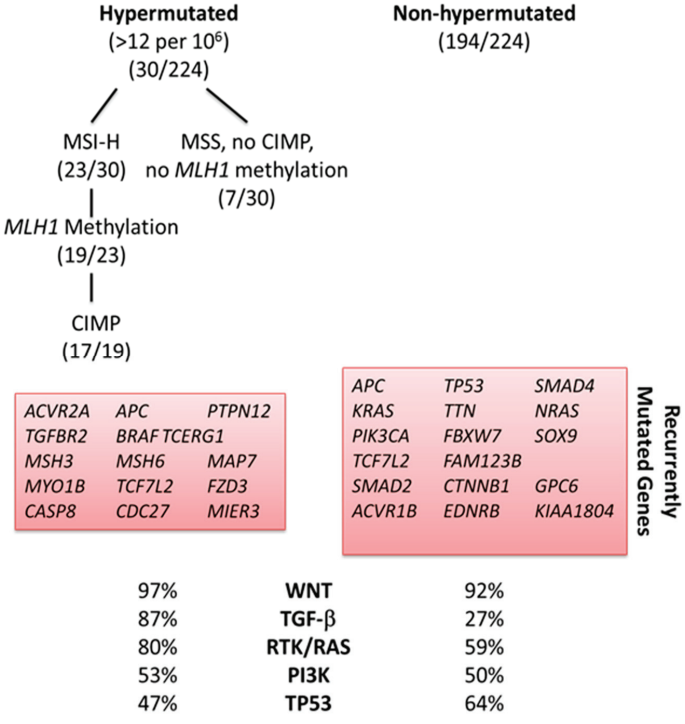


Figure 3. CRC subgroups identified by analysis of molecular data compiled from 224 tumors analyzed by The Cancer Genome Atlas Project [9]. Hypermutated CRC is highly enriched for hypermethylation, CIMP expression phenotype and *BRAF* mutations; it most frequently occurs in the proximal colon.



CRC due to CIN represents 80%–85% of sporadic cases. There are imbalances in chromosome number and loss of heterozygosity, as well as accumulation of mutations in tumor suppressor genes and oncogenes that activate pathways critical for CRC initiation and progression. In that pathway, CRC arises from adenomas.

Tumors with a high degree of MSI (MSI-H; about 15% of CRC) are characterized by frequent microsatellite length mutations. MSI occurs due to deficiencies in the mismatch repair (MMR) system, which recognizes and repairs nucleotide mismatches. Most sporadic MSI-H CRCs are caused by hypermethylation (epigenetic silencing) of the mismatch-repair gene *MLH1*. This silencing typically occurs in tumors of the CIMP phenotype. There is substantial overlap between MSI-H cancers and cancers containing a high degree of CIMP (Figure 2).

CIMP represents a specific type of epigenomic stability that is characterized by widespread promoter CpG island methylation and epigenetic gene silencing including tumor suppressor genes. CRCs with a high degree of CIMP are associated with older age, female gender, proximal tumor location, poor tumor differentiation, *BRAF* mutation, wild-type TP53, and high levels of global DNA methylation [12–14]. CIMP tumors have a distinct mRNA expression profile [9]. CIMP is also significantly associated with mucinous or signet ring cell morphology, as well as a marked peritumoral lymphocytic reaction, features that are also associated with MSI-H tumors [15,16].

CIMP and MSI-H tumors are thought to arise via the serrated adenoma pathway (*i.e.*, derived from sessile serrated adenomas, with progressive dysplasia) [16,17].

Importantly, the genomic and epigenomic subclass of CRC has clinical significance (and, therefore, biological significance). CIN is associated with a worse prognosis independent of stage and type of therapy [18]. MSI-H tumors have a better prognosis than microsatellite stable (MSS) CRC [19,20]. Hypermethylation is more common in cancers of the proximal colon, and most hypermutated CRCs originate in the proximal colon [9]. The effect of CIMP on prognosis is controversial, and analysis has been complicated by the high degree of overlap with MSI. MSI-H tumors containing a high level of CIMP are less prone to lymph node or distant metastasis [21]. However, in tumors that are not MSI-H, CIMP appears to be associated with worse survivals [22,23]. There are also numerous studies on patient outcomes as a function of individual molecular events (*BRAF* mutation, *KRAS* mutation, *TP53* mutation, *etc.*). However, given the numerous possible interactions with other genetic and molecular events, the findings from such studies should be interpreted with caution. As an example of this, among MSI-H cancers, *TGFBR2* mutations are associated with better survival [24], and this association with improved survival is even more pronounced in the presence of *BAX* mutations [25]. Finally, there is evidence that there is a link between genomic and epigenomic subclass and chemosensitivity [26]. Together, these observations demonstrate that the molecular features of CRC have biological consequences that translate to clinically significant outcomes. To this end, subclassification schemes will provide a good initial framework on which to study the disease.

However, subclassification based on genomic and epigenomic features has limitations, particularly when applying that information to an individual. That is, given the large number of combinations of molecular aberrations that are possible for any subclass of CRC, it is clear that there is substantial heterogeneity even within each subclass—even at the genomic level, let alone at the transcriptomic and proteomic levels. For example, while *BRAF* mutation frequency is particularly high in CIMP-H tumors, it may also occur (albeit infrequently) in CIMP-low tumors [10]. Indeed, there are no known mutations that are pathognomonic for any particular subtype of CRC. Therefore, if therapeutic strategies and decisions are to be derived for any individual, more work is required to define the biological phenotype by studying downstream pathophysiologies.

3. Functional Genomics: Defining the Biological Impact of the CRC Genome

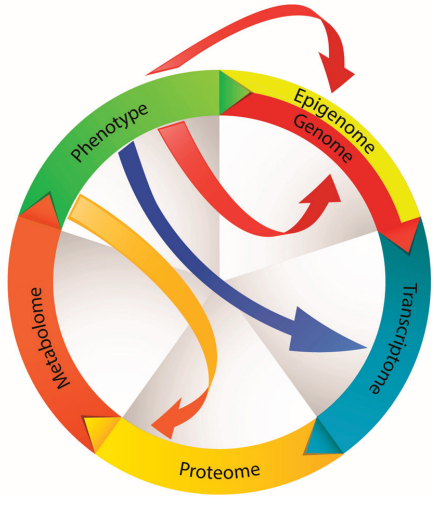
While cataloging the molecular features of CRC subtypes is important, what is most relevant from a clinical standpoint is how particular molecular features translate to differences in tumor biology. Which molecular features predispose to metastasis to various sites in the body, enhance tumor growth, induce angiogenesis, confer susceptibility to certain drugs, most affect the general health of the host (the patient), and are associated with the best (or worst) prognosis?

In general, it is well recognized that the fixed structure and sequence comprising the genome does not closely predict function or phenotype, as the information encoded by the genome is subject to modifications through a multitude of mechanisms and downstream events (Figure 4). There are multiple examples relevant to CRC. Hypermethylation of promoter-associated CpG islands is frequently seen in CRC and leads to transcriptional silencing; hypomethylation of CpG

islands outside of promoter regions is also a hallmark of CRC [27]. Long non-coding RNAs that overlap the 5' and 3' termini of genes may regulate the function of one or more genes [28–30]. MicroRNAs (miRNA) function to regulate expression at the post-transcriptional level, and numerous miRNAs are seen to be dysregulated in CRC [31–35]. Pseudogenes further complicate the interpretation of genomic and transcriptomic information. Pseudogenes resemble real genes, but contain premature stop codons and mutations that preclude their translation into functional proteins. Pseudogene-derived transcripts may act as a decoy to functionally significant miRNAs. For example, transcripts corresponding to the pseudogenes *PTENP1* and *KRAS1P* act as a decoy for miRNAs targeting *PTEN* and *KRAS* [36,37], which are known to be important in CRC. Protein translation is further regulated by a number of mechanisms that may vary in CRC [38,39]. Post-translational modifications, which may be altered in CRC, and differential expression of protein isoforms can further affect tumor biology or the host response to CRC [40].

In CRC, the genes and pathways that are particularly important in the initiation and progression of CRC include the WNT, MAPK, phosphatidylinositol-3-kinase (PI3K), TGF-β and p53 signaling pathways [9]. What is becoming increasingly clear is that there is a multitude of genetic and molecular events that can lead to the dysregulation of these signaling pathways in CRC. Moreover, the degree of dysregulation in each of these pathways may vary considerably between individuals, and the biological consequences of those variations are not currently predictable.

Figure 4. CRC is derived from an accumulation of genomic and epigenomic alterations. Alterations at the transcriptional level also occur due to the influence of regulatory RNAs (e.g., long noncoding RNA, miRNA, pseudogenes). Post-translational regulation and post-translational modifications further contribute to functional perturbations in tumor cells. These sequential and synchronous events contribute to the phenotype. Phenotype can further modify the genotype as well as any of the downstream events. According to this model, the metabolome represents the closest molecular representation of phenotype.



One example of how diverse (but related) mechanisms could lead to the same biological manifestation is in activation of the insulin-like growth factor (IGF)-PI3K pathway. It has recently been identified that about 7% of CRCs have an amplification of insulin-like growth factor 2 (*IGF2*) [9]. IGF2 overexpression may play an important role in the promotion of CRC [41,42]. In 15% of tumors without *IGF2* amplification, *IGF2* gene expression was also overexpressed. *IGF2* amplification or overexpression was associated with genomic events known to activate the PI3K pathway. We and others have also observed that CRC is often associated with high levels of expression of IGF1R (the receptor for IGF2) [43]. The *IRS2* gene, which encodes a protein that links IGF1R with PI3K, is frequently gained in CRC, which may also enhance the activity of this pathway [9]. Activation of the IGF1R-PI3K pathway is associated with multiple functional alterations in tumor cells, including changes in lipid metabolism and inflammatory events. Given the diversity of mechanisms by which this pathway (and any pathway) could be affected, it would be impractical to dissect all of the individual molecular events contributing to the function of that pathway. Rather, it may be more practical to dissect the phenotype so that therapeutic efforts could be directed at that particular imbalance.

Phenotype is therefore a product of preliminary molecular instructions (the genome) amended by a number of sequential and synchronous molecular events at the transcriptional and translational levels. Genomic information is, therefore, least reflective of phenotype and, as more downstream events are taken into account (especially in sum), molecular features more closely predict phenotype. Using this as a framework, it would be expected that proteomic and metabolomic profiles most closely reflect the phenotype and functional state of a cell (Figure 4). As an extension of this functional genomic model, phenotype can further modify any of the preceding molecular processes (including genotype, in cells susceptible to mutation), by affecting the conditions of the intracellular or extracellular microenvironment.

4. Metabolism: A Terminal Function Reflecting Phenotype

While it is possible to identify and measure some of the terminal elements of the proteome, protein function is significantly modified by the abundance of other proteins that may not be simultaneously measured. For example, soluble receptors and binding proteins may compete with functional receptors; ligand function can be modified by competing ligands or inhibitory proteins; and protein fragments can have significant biological effects. Since our knowledge of protein function is far from complete and since it is not yet possible to measure every single protein and protein fragment, one cannot make accurate inferences on phenotype just by evaluating the proteome.

Perhaps the best reflection of tumor phenotype (besides measurement of specific biological functions) is the metabolome. Any biological function is dependent on metabolic functions. Any alterations in metabolism and bioenergetics will alter the efficiency of downstream biological functions. Any perturbations in metabolism are identifiable by simultaneously measuring the abundance of co-related metabolites. Importantly, modifications of any of the constituents of the metabolome are measurable. Constituents of the metabolome—similar to proteins—are subject to some modifications such hydroxylation or amination. Derivatization of functionally known

metabolites is traceable by mass spectrometry methods and flux analysis. Moreover, because of the co-relationship of various metabolites, it is possible to extrapolate functional consequences of a metabolomic state despite such modifications.

Metabolic phenotype is dynamic; it is not static or even stable. There are a number of sources for the dynamic nature of the phenotype in tumor. First, while we often consider a tumor to be of a particular genotype, it must be appreciated that any given tumor is actually comprised of cells with substantial molecular heterogeneity [44]. Second, tumor phenotype is sculpted by the host immune response, where susceptible cells are eliminated and resistant cells survive the immune response, immunoediting [45,46]. Third, treatments, such as chemotherapy, alter phenotype by selective mechanisms, based on the pharmacology of any particular pharmacologic agent [44,47,48]. Fourth, the metabolic and inflammatory milieu within the tumor microenvironment may affect the function and phenotype of tumor cells, irrespective of genotype [49,50]. Finally, these same microenvironmental factors can predispose tumor cells to further mutational events [51]. The metabolic state of any tumor is, therefore, a product of all of these ever-changing influences at any given time.

4.1. Disordered Metabolism Is a Hallmark of Cancer

It has recently been recognized that one of the hallmarks of a cancer cell is the reprogramming of energy metabolism [52]. Cancer consists of rapidly proliferating cells, and large amounts of adenosine triphosphate (ATP) and substrates are required to support this proliferation. Sufficient ATP production is possible due to adaptations in metabolic pathways, including processes of carbohydrate, protein, lipid and nucleotide synthesis, which sustain the high metabolic demand.

The classic example of metabolic reprogramming is the Warburg Effect [53]. In normal cells, in the presence of sufficient oxygen, glucose is processed through oxidative phosphorylation, which is the most efficient means of generating ATP. Glycolysis only becomes a primary means to metabolize glucose in hypoxic conditions. However, in cancer cells, glycolysis is the dominant pathway for glucose metabolism, and this is independent of oxygen supply. The advantage to the tumor cell is that this is a much more rapid means of ATP production, which is necessary to support rapid cellular proliferation. The increased glucose processing in cancer cells forms the basis of ^{18}F-fluorodeoxyglucose positron emission tomography (FDG-PET), which is used to detect and monitor tumors including CRC [54,55].

Tumor cells also have other characteristic features of metabolic reprogramming, each of which function to support the rapidly expanding biomass within tumor. Glutamine uptake is enhanced to replenish the tricarboxylic acid (TCA) cycle; glutaminolysis also contributes to the production of acetyl-coenzyme A for subsequent lipid biosynthesis; and there is increased fatty acid and lipid synthesis, which sustains synthesis of cell membranes and lipid derivatives.

The altered metabolism of malignancy influences other hallmark functions of cancer such as proliferation, apoptosis and inflammation. A high rate of cell proliferation is supported by high levels of ATP and substrate. Defective mitochondrial morphology and function can affect susceptibility to apoptosis [56,57]. The metabolic milieu within tumor can produce an immunosuppressive microenvironment. For example, tryptophan depletion due to overexpression

of indoleamine 2,3-dioxygenase can suppress T cell responses against tumor [58–60]. Finally, alterations in fatty acid metabolism can result in a proinflammatory state, which is known to deleteriously affect tumor biology in CRC [61–63]. Therefore, not only is altered metabolism in itself a hallmark of cancer; it also supports the other disordered functions that characterize malignancy.

4.2. Genomic and Molecular Events Influencing Metabolism in CRC

While the genetic and molecular pathogenesis of CRC is being delineated in ever greater detail, the effects of the genetic and transcriptional events that characterize CRC are relatively poorly understood. That genotype affects metabolism is apparent in the features of FDG-PET, which varies with *KRAS* mutation status and HIF-1 expression [55]. Table 1 summarizes some of the known metabolic effects of genetic and epigenetic features that frequently accompany CRC. In a number of instances, important proteins can be dysregulated as a result of modulatory events at multiple levels (Figure 5).

KRAS mutations frequently accompany CRC. Tumors with KRAS mutations express high levels of GLUT1 (glucose transporter-1), providing the ability for enhanced glucose uptake and glycolysis, enabling survival in low glucose conditions [64]. KRAS protein expression and activation can be further modified by a number of mechanisms, including post-translational modification (Figure 5).

HIF-1 overexpression also frequently accompanies CRC [69]. HIF-1 transcription factor activates numerous target genes (reviewed in [70]). Not only is HIF-1 transcription factor a pivotal regulator of oxygen homeostasis; it also encourages glycolysis, contributes to the metabolism of nucleotides and iron, and exerts additional effects on cellular bioenergetics through its mitogenic effects. HIF-1 regulates genes involved in glucose metabolism, encouraging conversion of glucose to lactate. Specifically, HIF-1 increases expression of glucose transporters (GLUT1 and GLUT3) [71] and increases transcription of hexokinase-2 [71]. Direct activation of pyruvate dehydrogenase kinase 1 (PDK1) by HIF-1 leads to inactivation of pyruvate dehydrogenase (PDH), a key enzyme in TCA cycle. This inhibition of mitochondrial biogenesis results in shunting of pyruvate to lactate [72]. HIF-1 also alters the expression of pyruvate kinase (PK), the enzyme that catalyzes the last step of glycolysis. In a number of cancers, including CRC [73], the M2 isoform of (normally found in embryonic tissue) is the predominant PK isoform, and this is encouraged by HIF-1. PKM2 enhances aerobic glycolysis, which represents a selective growth advantage to tumor cells [74].

Figure 5. Examples of predicted metabolic effects of dysregulated proteins associated with CRC. The proteins are dysregulated as a consequence of modulatory molecular events at multiple levels. Alterations in metabolic function (and ultimately phenotype) result from the closely connected functional networks' response to these upstream signals. (GDP: guanine diphosphate; GTP: guanine triphosphate; GEF: guanine nucleotide-exchange factor; GAP: GTPase-activating protein; PPP: pentose phosphate pathway) [65–68].

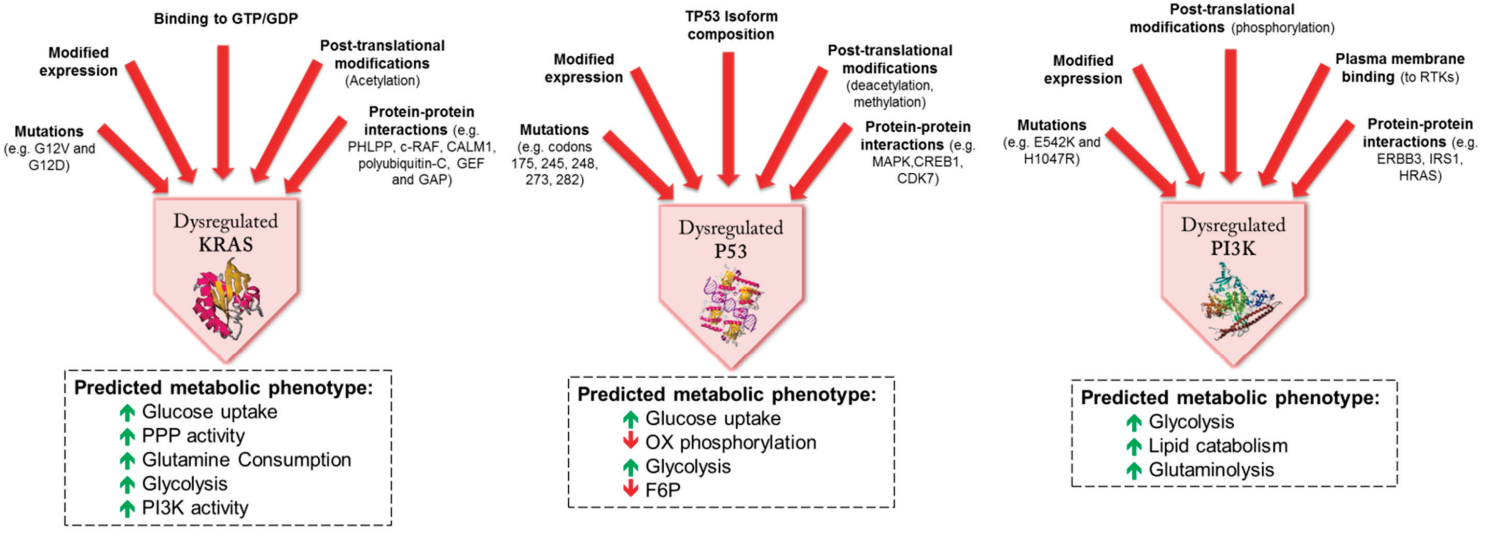


Table 1. Examples of genetic and epigenetic alterations in colorectal cancer (CRC) that have known or potential metabolic consequences.

Gene	Protein Product	Mechanism of Change in Function	Metabolic Effect
TGFBR2	TGF-beta receptor type-2	Inactivating mutation, overexpression	Activation of MAPK/ERK and TGF-β-SMAD pathway; inactivation leads to increased proliferation and decreased apoptosis
TP53	Tumor protein p53	Inactivating mutation or SNP in tumor suppressor	Inhibition of glucose transporters, inhibition of insulin receptor, activation of TCA cycle and oxidative phosphorylation
KRAS	GTPase kras	Activating mutation	Increased glucose uptake. Increased glycolysis, activation of PI3K pathway
PI3KCA	Phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha	Activating mutation	Increased lipid metabolism, growth-factor independence, increased glycolysis and glutaminolysis
SMAD4	Mothers against decapentaplegic homolog 4	Inactivating mutation	TGF-β signaling
TCF7L2	Transcription factor 7-like 2	Activating mutation	Increased Wnt signaling, increased glycolysis and lactate production
SMAD2	Mothers against decapentaplegic homolog 2	Inactivating mutation in tumor suppressor	TGF-β signaling
CTNNB1	Catenin beta-1	Activating mutation	Wnt signaling pathway
SOX9	SRY (sex determining region Y)-box 9	Mutation or Overexpression of transcription factor	Wnt signaling pathway, inactivation of insulin signaling, anti-proliferation
SOX9	SRY (sex determining region Y)-box 9	Mutation or Overexpression of transcription factor	Wnt signaling pathway, inactivation of insulin signaling, anti-proliferation
ACVR1B	Activin receptor type-1B	Mutation, Overexpression	Activation of TGF-β signaling
EDNRB	Endothelin B receptor	Mutation, hypermethylation, Overexpression	Response to peptide hormonal stimuli
FASN	Fatty acid synthase	Overexpression	Production of fatty acids from Acetyl-CoA
PTGS2 (COX2)	Prostaglandin G/H synthase 2	Overexpression	Modulated by HIF-2α, inducing TGF-β pathway
E-Cadherin (CDH1)	Cadherin 1, type 1, E-cadherin	Mutation, Overexpression	Activates Wnt signaling and lipid metabolism pathway
CDKN2A (p16-INK4a)	Cyclin dependent kinase inhibitor 2A	Mutation, deletion, Methylation	Leads to mitochondrial dysfunction and impaired phosphorylative oxidation, increased glycolysis

Table 1. *Cont.*

Gene	Protein Product	Mechanism of Change in Function	Metabolic Effect
THBS1/TSP1	Thrombospondin 1	Methylation	Regulator of TGF-β signaling, increased inflammation in adipose tissue
SDH	Succinate dehydrogenase complex, subunit B, iron sulfur	Underexpression (mechanism unclear)	Enzyme for TCA cycle, phosphorylative oxidation activity, decreased glucose uptake
PTEN	Phosphatase and tensin homolog	Inactivating mutation in tumor suppressor	Suppressor of PI3K/Akt pathway. Inactivation leads to increased glycolysis, lipogenesis and glycogenesis.
HIF-1α	Hypoxia-inducible factor 1-alpha	Overexpression and molecular stabilization	Activates glycolysis, deactivates TCA cycle and phosphorylative oxidation

TP53 mutations frequently occur in CRC. TP53 inhibits transcription of glucose transporters (GLUT1 and GLUT4) and the insulin receptor, reducing glucose uptake by cells. Glucose uptake is further inhibited by activation of NF-κB, which inhibits GLUT3 expression. TP53 also activates TP53-inducible glycolysis and apoptosis regulator (TIGAR), which catalyzes the conversion of fructose-2,6-bisphosphate to fructose-6-phosphate, an important substrate for the pentose phosphate pathway [75]. Accumulation of fructose-6-phosphate also inhibits glycolysis and encourages gluconeogenesis. TIGAR also activates ɣH2AX complex, which plays an important role in histone methylation of many genes, some of which may have metabolic functions such as *PTEN* [76]. Finally, TP53 activates transcription of *SCO2* (cytochrome C oxidase assembly gene), which encourages oxidative phosphorylation through regulation of complex IV of the mitochondrial respiratory chain [75]. Loss of TP53 function, which results from *TP53* mutations in CRC, therefore, results in enhanced cellular glucose uptake, accelerated glycolysis, as well as reduced oxidative phosphorylation. At the protein level, p53 isoform composition, post-translational deacetylation and interactions with other proteins may further modify p53 activity (Figure 5).

In CRC, *PTEN* inactivation occurs through mixed genetic and epigenetic mechanisms [77], and it is also controlled by post-translational modifications [78]. *PTEN* is a tumor suppressor gene that functions to antagonize the PI3K/AKT/mammalian target of rapamycin (mTOR) pathway. *PTEN* loss of function results in deregulation of PI3K signaling, leading to constitutively activated AKT. Constitutive activation of AKT results in metabolic changes that are characteristic of the Warburg effect [79], as well as lipogenesis [80].

Alterations in TGF-β signaling are frequent in CRC. Binding of TGF-β initiates downstream signaling involving phosphorylation of SMADs. *TGFBR2* mutations are among the recurrently mutated genes in hypermutated CRCs [9,81]. Mutations in *SMAD2*, *SMAD3*, and *SMAD4* occur primarily in tumors of mucinous histology [82]. These alterations typically result in suppression of the TGF-β antiproliferative effect [83,84].

4.3. Metabolomic Studies Related to CRC

Recently, the field of metabolomics has emerged as a means to more comprehensively study the contribution of tumor on the overall metabolic milieu. The metabolome can be evaluated in a multiplexed fashion using two primary technologies: nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS). Currently no single analytical technology is capable of detecting all metabolites in a biological sample and, since metabolites detectable in one analytical modality are not necessarily detectable in the other, these modalities are complimentary.

There has been considerable interest in the metabolomic analysis of CRC. Several groups have shown that metabolomic profiling of colon mucosa could distinguish between normal and malignant tissues using high resolution magic angle spinning (HR-MAS) ^{1}H-NMR spectroscopy, as well as gas chromatography-mass spectrometry (GC-MS) [85–87]. In general, metabolites associated with the TCA cycle were found to be lower in malignant tissues; and intermediates of the urea cycle, purines, pyrimidines, amino acids, and choline containing compounds were more abundant, consistent with the higher metabolic requirements of rapidly dividing cells.

Fecal water extracts have also been submitted to metabolomic analysis. Monleon *et al.* reported that patients with CRC had low fecal concentrations of short chain fatty acids such as acetate and butyrate and higher levels of proline and cysteine [88]. While there may be some diagnostic utility in this observation, further study is required to understand the origin of these alterations. CRC-associated changes in fecal metabolites may be related to differences in gut mucosa, malabsorption of certain nutrients, or alterations in the gut microflora; any of these differences may represent factors that predispose to CRC or that occur secondary to CRC. In an attempt to delineate the contribution of intestinal bacteria, Weir and coworkers simultaneously assessed stool metabolome (by GC-MS) and gut microbiome [89]. Indeed, in patients with CRC, butyrate-producing species were under-represented and a mucin-degrading species (*Akkermansia muciniphila*) was present at a higher level.

A number of reports have appeared describing the serum metabolome associated with CRC [90–96]. Subtle differences in disease-associated metabolomic changes may reflect population-based differences based on dietary, environmental and genetic factors, although the contribution of these factors to results are currently difficult to identify because of the diversity of analytical platforms used in each of the studies. Qiu *et al.* compared 64 Chinese patients with CRC to healthy controls; metabolomic profiles were determined by GC-MS and liquid chromatography-mass spectrometry (LC-MS) [90]. A distinct metabolomic signature for CRC was identified. In a follow-up replication study, similar findings were observed, which demonstrated alterations in the TCA cycle, urea cycle, glutamine metabolism, and gut flora metabolism [96]. Similar efforts have emerged from Japan, using GC-MS [94]. CRC is associated with changes in the composition of serum fatty acid profile (evaluated by GC-MS) [91]. The serum amino acid profile as identified by electrospray tandem mass spectrometry differs from normal controls [93]. The biological meaning of alterations in fatty acid and amino acid profile will require further interrogation. Stage of CRC as well as site of metastasis may affect metabolomic profile [95]. In addition, there is evidence that the serum metabolomic profile may be related to survival, although

it is unclear whether it is of prognostic or predictive significance [97]. In all, these studies demonstrate the feasibility of using metabolomics biomarkers to diagnose CRC, and also suggest that it may be possible to identify subgroups based on metabolomic profile.

Efforts at describing the metabolomic changes that occur in various tissues and biofluids in CRC have largely been descriptive in nature. Much more work will be required to understand the biological implications of any of the metabolomic alterations that typify CRC. Diet, environment and genetic background can represent confounding factors. In metabolomic studies using blood, urine, and stool, it is difficult to determine whether metabolic perturbations are derived from tumor, host factors (including response to tumor), or gut microbiome. To complicate the analysis further, it is possible that some of the metabolic perturbations are not actually a result of CRC, but rather reflect the metabolic state of an individual who is predisposed to CRC. Therefore, to understand the biological basis of metabolomic studies, more detailed experiments and bioinformatic analyses will be required.

5. Linking Genotypic Subsets with Functional Subsets of CRC

As we improve our understanding of the metabolic changes associated with CRC, it will be important to dissect the genetic, epigenetic and transcriptional events that accompany these changes. Using the multiplexed information derived from metabolomic studies, bioinformatic approaches have been described to identify pathways that are putatively involved in the metabolic derangements of CRC [98–100]. This is facilitated by the fact that many metabolites behave in a collinear fashion due to their relationships in metabolic processes. Such an approach generates hypotheses, facilitating a more focused interrogation based on experiments. Indeed, even the limited information derived from a proteomic screen can be used to seed a bioinformatic search for functional networks at the protein and transcriptional levels [101].

Alternatively, multiplexed genomic, transcriptomic, proteomic, and metabolomic data sets can be generated in parallel to catalogue relationships in genotype and phenotype. Integrating “omics” information has been attempted by several groups [100,102–105]. The Cancer Genome Atlas (TCGA) Project represents a large-scale effort at synchronously cataloguing genome sequence, DNA copy number, promoter methylome, and transcriptome for a number of tumor types, including CRC [9]. The tools to analyze these parallel “omics” data sets are evolving quickly. For example, the cBio Cancer Genomics Portal [106] is an online open-access application that enables public access to the multidimensional raw data derived from TCGA, as well as evaluation for possible molecular relationships [107]. The University of California, Santa Cruz (UCSC) Interaction Browser [108] facilitates the simultaneous visualization and analysis of multiple “omics” data sets, enabling integration of biological networks at multiple levels [105]. So far, metabolomic data have not been a large part of such large-scale efforts, but the means to perform such research is now available. The Subpathway-GM method for identifying important metabolic subpathways using genomic and metabolomic data sets represents one example of an effort to integrate metabolomics with upstream molecular events [100].

While integrative analysis of multiple “omics” data sets is attractive, there are some limitations. Even alterations in a single metabolic enzyme or any other molecular event may affect multiple

pathways, and such effects can be modified by consequential molecular events. Therefore, such an approach will prove challenging from a bioinformatic perspective. Importantly, opposing molecular events may appear at different biological levels (for example, at the genomic and transcriptional levels), and the net effect of those events cannot be predicted without the final phenotypical and functional information. Finally, it is important to keep in mind that bioinformatic analyses are largely hypothesis-generating, and specific experiments remain an essential means to accentuate our biological knowledge.

To more efficiently derive an understanding of the linkages between genome, epigenome, transcriptome and metabolome, targeted analyses of specific events in large data sets will be required. Moreover, to derive mechanistic insight, discreet experimental systems must be utilized. This represents an important direction in the field, and publications taking this approach are beginning to appear. One such recent effort involved the analysis of 376 surgical samples of CRC and adjacent normal colon from US and Chinese patients, using GC-MS [109]. Fifteen metabolites were found to be differentially abundant in tumor and normal colonic epithelium. Investigators observed several metabolic variations to support proliferation; findings consistent with the Warburg effect (glycolysis and aerobic fermentation); and activation of the pentose phosphate pathway (providing substrates for nucleic acid and fatty acid synthesis). A subsequent targeted transcriptomic analysis was performed based on the pattern of metabolites seen to be differentially abundant in CRC tissue. Fatty acid synthase (FASN) and stearoyl-CoA desaturase-1 (SCD1) were among the most highly upregulated transcripts. Thus, using a combination of bioinformatic interrogation and experimental work, transcriptional and metabolic linkages could be identified.

Manna and coworkers analyzed the urine metabolome of *Apc*$^{Min/+}$ mice as well as mice with azoxymethane (AOM) induced tumors by LC-MS [110]. In mice with colon-specific disruption of *APC*, urinary excretion of amino acid metabolites (e.g., glutamine, proline, Nα-acetyl lysine) and nucleic acid metabolites (e.g., xanthosine, inosine, xanthine, cytidine, deoxyuridine, thymidine) increased progressively during tumorigenesis. Similar changes occurred in mice with AOM-induced colorectal tumors, although there were some differences in individual metabolites in this model. In *Apc*$^{Min/+}$ mice, these metabolomic changes were associated with expression of key genes involved in related pathways. For example, there was overexpression of a number of genes involved in amino acid metabolism, urea cycle and polyamine metabolism. The interconnectivity of these events suggested that the pathogenesis of CRC involved a coordinated reprogramming of metabolic pathways during tumorigenesis.

Finally, Tessem *et al.* utilized HR-MAS ^{1}H-NMR spectroscopy to determine differences in CRC tissue between MSI-H tumors and MSS tumors [87]. The metabolomic profiles were easily distinguished. MSI-H tumors were characterized by higher levels of lactate, glycine, taurine, creatine, and choline; myo-inositol and glucose were decreased. Interestingly, there were also differences seen in adjacent normal colon between MSI-H tumors and MSS tumors. The biological significance of these findings might become apparent with a more comprehensive analysis of the metabolic perturbations seen in each of these CRC subtypes.

Further studies will clearly be required to connect genomic, epigenomic, transcriptomic, and proteomic events to alterations in metabolism in CRC. A combined approach including

bioinformatics and targeted experimental analysis appears to be quite constructive. It is possible, since metabolism is a terminal event preceding function, that new phenotypical subtypes can be identified, which may aid in individualizing systemic therapy for CRC.

6. Metabolomics as a Means to Discover Novel Therapeutic Targets

Genotype is known to affect sensitivity to treatment. For example, tumors containing KRAS or NRAS mutations are resistant to EGFR inhibitors [111,112]. Tumors that are MSI-H are resistant to fluoropyrimidines [26]. While genotype has some clinical utility as a predictive biomarker, it does not predict with any certainty whether an individual tumor will respond to any antineoplastic agent, for coincidental mutations and molecular alterations may additionally influence chemosensitivity.

Information derived from metabolomic studies may identify improved ways of targeting the disordered metabolism seen in CRC, which is particularly critical for those tumors that are resistant to other agents. Defining the metabolic phenotype may also enhance therapeutic efforts in individuals and in subgroups. For example, one CRC variant may have dysregulation of specific metabolic pathways that provide it with a growth advantage; other variants may have different metabolic disorders that could be targeted. Indeed, in CRC several metabolic variations have been described [87,109].

The recognition that disordered metabolism is a hallmark of cancer has spurred some interest in therapies targeting metabolism. Cytotoxic drugs, such as fluoropyrimidines, target metabolism and they have been used in practice for years. Because tumors frequently have disordered mitochondrial function, drugs have been developed that affect mitochondrial function [113]. Interventions influencing the disordered carbohydrate metabolism that characterizes most cancers are also attracting interest. For example, oral hypoglycemics used to treat diabetes are being investigated, and retrospective studies have demonstrated reductions in cancer-related mortality in diabetics taking metformin [114,115]. Metformin inhibits the mTOR pathway. Interestingly, metformin is toxic to cancer stem cells [116]. In breast cancer patients, metformin is associated with higher response rates to cytotoxic chemotherapy [117]. A clinical trial is in progress assessing the role of metformin in colorectal adenoma formation [118]. Other mTOR inhibitors are also being tested in CRC [119–121], as are other drugs targeting the insulin-like growth factor pathway [122,123]. These are only some of the examples of pharmacologic agents that target specific metabolic processes that are being evaluated for cancer.

Genotype may aid in identifying the metabolic disorders that are likely to be contained in any particular tumor. For example, tumors with *KRAS* mutations are known to have typical metabolic derangements. *KRAS* transformed fibroblasts lose their proliferative ability with glutamine deprivation [124]. In preclinical models, targeting metabolic enzymes to disrupt glucose metabolism is effective in the treatment of tumors driven by *KRAS* [64,125]. Other common genotypes may similarly be treatable by a specific metabolism-targeted therapy.

The main challenge in designing agents that target metabolism will be to avoid toxicity related to targeting metabolic pathways in normal proliferating cells. Therefore, it will be vital to identify pathways that are redundant in normal cells but absent in cancer cells. Identification of such a

therapeutic window may be facilitated by comprehensive analysis of the metabolome in cancer cells and normal cells.

7. Conclusions

It has become apparent that cataloguing the static structural and sequence alterations in the CRC genome merely represents a start to understanding the biology of CRC. It is essential to develop a greater understanding of the dynamic functional perturbations that accompany the genomic changes that characterize CRC and its subtypes, including the multitude of changes in the transcriptome, the proteome, and the metabolome. Moreover, the interactions of each of these elements that comprise each of these downstream molecular events must be dissected. Understanding the functional (or phenotypic) implications of each genotype is imperative to the clinician for a number of reasons. The biological behavior of subsets of CRC can be defined for the purpose of prognostication; and therapies targeting specific biological events can be better engineered. Metabolomics allows a comprehensive analysis of some of the most fundamental biological processes that typify CRC and its subtypes, and is perhaps the closest molecular representation of phenotype currently available.

Author Contributions

Oliver F. Bathe and Farshad Farshidfar both conceived of the paper, researched relevant literature, and wrote the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Miranda, E.; Destro, A.; Malesci, A.; Balladore, E.; Bianchi, P.; Baryshnikova, E.; Franchi, G.; Morenghi, E.; Laghi, L.; Gennari, L.; *et al.* Genetic and epigenetic changes in primary metastatic and nonmetastatic colorectal cancer. *Br. J. Cancer* **2006**, *95*, 1101–1107.
2. Oliveira, C.; Velho, S.; Moutinho, C.; Ferreira, A.; Preto, A.; Domingo, E.; Capelinha, A.F.; Duval, A.; Hamelin, R.; Machado, J.C.; *et al.* KRAS and BRAF oncogenic mutations in MSS colorectal carcinoma progression. *Oncogene* **2007**, *26*, 158–163.
3. Gunther, K.; Leier, J.; Henning, G.; Dimmler, A.; Weissbach, R.; Hohenberger, W.; Forster, R. Prediction of lymph node metastasis in colorectal carcinoma by expression of chemokine receptor CCR7. *Int. J. Cancer* **2005**, *116*, 726–733.
4. Artinyan, A.; Essani, R.; Lake, J.; Kaiser, A.M.; Vukasin, P.; Danenberg, P.; Danenberg, K.; Haile, R.; Beart, R.W., Jr. Molecular predictors of lymph node metastasis in colon cancer: Increased risk with decreased thymidylate synthase expression. *J. Gastrointest. Surg.* **2005**, *9*, 1216–1221.

5. Lin, Y.M.; Furukawa, Y.; Tsunoda, T.; Yue, C.T.; Yang, K.C.; Nakamura, Y. Molecular diagnosis of colorectal tumors by expression profiles of 50 genes expressed differentially in adenomas and carcinomas. *Oncogene* **2002**, *21*, 4120–4128.
6. Arango, D.; Wilson, A.J.; Shi, Q.; Corner, G.A.; Aranes, M.J.; Nicholas, C.; Lesser, M.; Mariadason, J.M.; Augenlicht, L.H. Molecular mechanisms of action and prediction of response to oxaliplatin in colorectal cancer cells. *Br. J. Cancer* **2004**, *91*, 1931–1946.
7. Mariadason, J.M.; Arango, D.; Shi, Q.; Wilson, A.J.; Corner, G.A.; Nicholas, C.; Aranes, M.J.; Lesser, M.; Schwartz, E.L.; Augenlicht, L.H. Gene expression profiling-based prediction of response of colon carcinoma cells to 5-fluorouracil and camptothecin. *Cancer Res.* **2003**, *63*, 8791–8812.
8. Li, M.; Lin, Y.M.; Hasegawa, S.; Shimokawa, T.; Murata, K.; Kameyama, M.; Ishikawa, O.; Katagiri, T.; Tsunoda, T.; Nakamura, Y.; *et al.* Genes associated with liver metastasis of colon cancer, identified by genome-wide cDNA microarray. *Int. J. Oncol.* **2004**, *24*, 305–312.
9. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **2012**, *487*, 330–337.
10. Ogino, S.; Goel, A. Molecular classification and correlates in colorectal cancer. *J. Mol. Diagn.* **2008**, *10*, 13–27.
11. Jass, J.R. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology* **2007**, *50*, 113–130.
12. Ogino, S.; Cantor, M.; Kawasaki, T.; Brahmandam, M.; Kirkner, G.J.; Weisenberger, D.J.; Campan, M.; Laird, P.W.; Loda, M.; Fuchs, C.S. Cpg island methylator phenotype (CIMP) of colorectal cancer is best characterised by quantitative DNA methylation analysis and prospective cohort studies. *Gut* **2006**, *55*, 1000–1006.
13. Samowitz, W.S.; Albertsen, H.; Herrick, J.; Levin, T.R.; Sweeney, C.; Murtaugh, M.A.; Wolff, R.K.; Slattery, M.L. Evaluation of a large, population-based sample supports a cpg island methylator phenotype in colon cancer. *Gastroenterology* **2005**, *129*, 837–845.
14. Weisenberger, D.J.; Siegmund, K.D.; Campan, M.; Young, J.; Long, T.I.; Faasse, M.A.; Kang, G.H.; Widschwendter, M.; Weener, D.; Buchanan, D.; *et al.* Cpg island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with braf mutation in colorectal cancer. *Nat. Genet.* **2006**, *38*, 787–793.
15. Ogino, S.; Odze, R.D.; Kawasaki, T.; Brahmandam, M.; Kirkner, G.J.; Laird, P.W.; Loda, M.; Fuchs, C.S. Correlation of pathologic features with CpG island methylator phenotype (CIMP) by quantitative DNA methylation analysis in colorectal carcinoma. *Am. J. Surg. Pathol.* **2006**, *30*, 1175–1183.
16. Soreide, K.; Janssen, E.A.M.; Soiland, H.; Korner, H.; Baak, J.P.A. Microsatellite instability in colorectal cancer. *Br. J. Surg.* **2006**, *93*, 395–406.
17. Grady, W.M.; Carethers, J.M. Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology* **2008**, *135*, 1079–1099.
18. Walther, A.; Houlston, R.; Tomlinson, I. Association between chromosomal instability and prognosis in colorectal cancer: A meta-analysis. *Gut* **2008**, *57*, 941–950.

19. Popat, S.; Hubner, R.; Houlston, R.S. Systematic review of microsatellite instability and colorectal cancer prognosis. *J. Clin. Oncol.* **2005**, *23*, 609–618.
20. Guastadisegni, C.; Colafranceschi, M.; Ottini, L.; Dogliotti, E. Microsatellite instability as a marker of prognosis and response to therapy: A meta-analysis of colorectal cancer survival data. *Eur. J. Cancer* **2010**, *46*, 2788–2798.
21. Laghi, L.; Malesci, A. Microsatellite instability and therapeutic consequences in colorectal cancer. *Dig. Dis.* **2012**, *30*, 304–309.
22. Ward, R.L.; Cheong, K.; Ku, S.L.; Meagher, A.; O'Connor, T.; Hawkins, N.J. Adverse prognostic effect of methylation in colorectal cancer is reversed by microsatellite instability. *J. Clin. Oncol.* **2003**, *21*, 3729–3736.
23. Ogino, S.; Meyerhardt, J.A.; Kawasaki, T.; Clark, J.W.; Ryan, D.P.; Kulke, M.H.; Enzinger, P.C.; Wolpin, B.M.; Loda, M.; Fuchs, C.S. CpG island methylation, response to combination chemotherapy, and patient survival in advanced microsatellite stable colorectal carcinoma. *Virchows Arch.* **2007**, *450*, 529–537.
24. Watanabe, T.; Wu, T.T.; Catalano, P.J.; Ueki, T.; Satriano, R.; Haller, D.G.; Benson, A.B., 3rd; Hamilton, S.R. Molecular predictors of survival after adjuvant chemotherapy for colon cancer. *N. Engl. J. Med.* **2001**, *344*, 1196–1206.
25. Jung, B.; Smith, E.J.; Doctolero, R.T.; Gervaz, P.; Alonso, J.C.; Miyai, K.; Keku, T.; Sandler, R.S.; Carethers, J.M. Influence of target gene mutations on survival, stage and histology in sporadic microsatellite unstable colon cancers. *Int. J. Cancer* **2006**, *118*, 2509–2513.
26. Ribic, C.M.; Sargent, D.J.; Moore, M.J.; Thibodeau, S.N.; French, A.J.; Goldberg, R.M.; Hamilton, S.R.; Laurent-Puig, P.; Gryfe, R.; Shepherd, L.E.; *et al.* Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. *N. Engl. J. Med.* **2003**, *349*, 247–257.
27. Lao, V.V.; Grady, W.M. Epigenetics and colorectal cancer. *Nat. Rev. Gastroenterol. Hepatol.* **2011**, *8*, 686–700.
28. Matsui, M.; Chu, Y.; Zhang, H.; Gagnon, K.T.; Shaikh, S.; Kuchimanchi, S.; Manoharan, M.; Corey, D.R.; Janowski, B.A. Promoter RNA links transcriptional regulation of inflammatory pathway genes. *Nucleic Acids Res.* **2013**, *41*, 10086–10109.
29. Svoboda, M.; Slyskova, J.; Schneiderova, M.; Makovicky, P.; Bielik, L.; Levy, M.; Lipska, L.; Hemmelova, B.; Kala, Z.; Protivankova, M.; *et al.* HOTAIR long non-coding RNA is a negative prognostic factor not only in primary tumors, but also in the blood of colorectal cancer patients. *Carcinogenesis* **2014**, *35*, 1510–1515.
30. Qi, P.; Xu, M.D.; Ni, S.J.; Shen, X.H.; Wei, P.; Huang, D.; Tan, C.; Sheng, W.Q.; Zhou, X.Y.; Du, X. Down-regulation of ncRAN, a long non-coding RNA, contributes to colorectal cancer cell migration and invasion and predicts poor overall survival for colorectal cancer patients. *Mol. Carcinog.* **2014**, doi:10.1002/mc.22137.
31. Chen, T.; Yao, L.Q.; Shi, Q.; Ren, Z.; Ye, L.C.; Xu, J.M.; Zhou, P.H.; Zhong, Y.S. MicroRNA-31 contributes to colorectal cancer development by targeting factor inhibiting HIF-1alpha (FIH-1). *Cancer Biol. Ther.* **2014**, *15*, 516–523.

32. Pichler, M.; Ress, A.L.; Winter, E.; Stiegelbauer, V.; Karbiener, M.; Schwarzenbacher, D.; Scheideler, M.; Ivan, C.; Jahn, S.W.; Kiesslich, T.; *et al.* MIR-200a regulates epithelial to mesenchymal transition-related gene expression and determines prognosis in colorectal cancer patients. *Br. J. Cancer* **2014**, 110, 1614–1621.
33. Cappuzzo, F.; Sacconi, A.; Landi, L.; Ludovini, V.; Biagioni, F.; D'Incecco, A.; Capodanno, A.; Salvini, J.; Corgna, E.; Cupini, S.; *et al.* Microrna signature in metastatic colorectal cancer patients treated with anti-EGFR monoclonal antibodies. *Clin. Colorectal Cancer* **2014**, *13*, 37–45.
34. Nosho, K.; Igarashi, H.; Nojima, M.; Ito, M.; Maruyama, R.; Yoshii, S.; Naito, T.; Sukawa, Y.; Mikami, M.; Sumioka, W.; *et al.* Association of microRNA-31 with BRAF mutation, colorectal cancer survival and serrated pathway. *Carcinogenesis* **2014**, *35*, 776–783.
35. Pizzini, S.; Bisognin, A.; Mandruzzato, S.; Biasiolo, M.; Facciolli, A.; Perilli, L.; Rossi, E.; Esposito, G.; Rugge, M.; Pilati, P.; *et al.* Impact of micrornas on regulatory networks and pathways in human colorectal carcinogenesis and development of metastasis. *BMC Genomics* **2013**, *14*, doi:10.1186/1471-2164-14-589.
36. Poliseno, L.; Salmena, L.; Zhang, J.; Carver, B.; Haveman, W.J.; Pandolfi, P.P. A coding-independent function of gene and pseudogene mrnas regulates tumour biology. *Nature* **2010**, *465*, 1033–1038.
37. Tay, Y.; Kats, L.; Salmena, L.; Weiss, D.; Tan, S.M.; Ala, U.; Karreth, F.; Poliseno, L.; Provero, P.; di Cunto, F.; *et al.* Coding-independent regulation of the tumor suppressor pten by competing endogenous mrnas. *Cell* **2011**, *147*, 344–357.
38. Matassa, D.S.; Amoroso, M.R.; Agliarulo, I.; Maddalena, F.; Sisinni, L.; Paladino, S.; Romano, S.; Romano, M.F.; Sagar, V.; Loreni, F.; *et al.* Translational control in the stress adaptive response of cancer cells: A novel role for the heat shock protein trap1. *Cell Death Dis.* **2013**, *4*, e851.
39. Dixon, D.A. Dysregulated post-transcriptional control of COX-2 gene expression in cancer. *Curr. Pharm. Des.* **2004**, *10*, 635–646.
40. Pedersen, J.W.; Blixt, O.; Bennett, E.P.; Tarp, M.A.; Dar, I.; Mandel, U.; Poulsen, S.S.; Pedersen, A.E.; Rasmussen, S.; Jess, P.; *et al.* Seromic profiling of colorectal cancer patients with novel glycopeptide microarray. *Int. J. Cancer* **2011**, *128*, 1860–1871.
41. Nakagawa, H.; Chadwick, R.B.; Peltomaki, P.; Plass, C.; Nakamura, Y.; de la Chapelle, A. Loss of imprinting of the insulin-like growth factor II gene occurs by biallelic methylation in a core region of H19-associated CTCF-binding sites in colorectal cancer. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 591–596.
42. Levine, A.J.; Ihenacho, U.; Lee, W.; Figueiredo, J.C.; Vandenberg, D.J.; Edlund, C.K.; Davis, B.D.; Stern, M.C.; Haile, R.W. Genetic variation in insulin pathway genes and distal colorectal adenoma risk. *Int. J. Colorectal Dis.* **2012**, *27*, 1587–1595.
43. Guo, S.T.; Jiang, C.C.; Wang, G.P.; Li, Y.P.; Wang, C.Y.; Guo, X.Y.; Yang, R.H.; Feng, Y.; Wang, F.H.; Tseng, H.Y.; *et al.* Microrna-497 targets insulin-like growth factor 1 receptor and has a tumour suppressive role in human colorectal cancer. *Oncogene* **2013**, *32*, 1910–1920.

44. Janku, F. Tumor heterogeneity in the clinic: Is it a real problem? *Ther. Adv. Med. Oncol.* **2014**, *6*, 43–51.
45. Mittal, D.; Gubin, M.M.; Schreiber, R.D.; Smyth, M.J. New insights into cancer immunoediting and its three component phases-elimination, equilibrium and escape. *Curr. Opin. Immunol.* **2014**, *27C*, 16–25.
46. Bathe, O.; Dalyot-Herman, N.; Malek, T. Therapeutic limitations in tumor-specific $CD8^+$ memory T cell engraftment. *BMC Cancer* **2003**, *3*, doi:10.1186/1471-2407-3-21.
47. Schwitalla, S. Tumor cell plasticity: The challenge to catch a moving target. *J. Gastroenterol.* **2014**, *49*, 618–627.
48. Navin, N.E. Tumor evolution in response to chemotherapy: Phenotype *versus* genotype. *Cell Rep.* **2014**, *6*, 417–419.
49. Wu, D.; Wu, P.; Huang, Q.; Liu, Y.; Ye, J.; Huang, J. Interleukin-17: A promoter in colorectal cancer progression. *Clin. Dev. Immunol.* **2013**, doi:10.1155/2013/436307.
50. Grivennikov, S.I.; Wang, K.; Mucida, D.; Stewart, C.A.; Schnabl, B.; Jauch, D.; Taniguchi, K.; Yu, G.Y.; Osterreicher, C.H.; Hung, K.E.; *et al.* Adenoma-linked barrier defects and microbial products drive IL-23/IL-17-mediated tumour growth. *Nature* **2012**, *491*, 254–258.
51. Pozza, A.; Scarpa, M.; Ruffolo, C.; Polese, L.; Erroi, F.; Bridda, A.; Norberto, L.; Frego, M. Colonic carcinogenesis in ibd: Molecular events. *Ann. Ital. Chir.* **2011**, *82*, 19–28.
52. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674.
53. Warburg, O. On respiratory impairment in cancer cells. *Science* **1956**, *124*, 269–270.
54. Lee, H.S.; Kim, H.O.; Hong, Y.S.; Kim, T.W.; Kim, J.C.; Yu, C.S.; Kim, J.S. Prognostic value of metabolic parameters in patients with synchronous colorectal cancer liver metastasis following curative-intent colorectal and hepatic surgery. *J. Nucl. Med.* **2014**, *55*, 582–589.
55. Miles, K.A.; Ganeshan, B.; Rodriguez-Justo, M.; Goh, V.J.; Ziauddin, Z.; Engledow, A.; Meagher, M.; Endozo, R.; Taylor, S.A.; Halligan, S.; *et al.* Multifunctional imaging signature for V-Ki-RAS2 Kirsten rat sarcoma viral oncogene homolog (KRAS) mutations in colorectal cancer. *J. Nucl. Med.* **2014**, *55*, 386–391.
56. Babbar, M.; Sheikh, M.S. Metabolic stress and disorders related to alterations in mitochondrial fission or fusion. *Mol. Cell. Pharmacol.* **2013**, *5*, 109–133.
57. Grills, C.; Jithesh, P.V.; Blayney, J.; Zhang, S.D.; Fennell, D.A. Gene expression meta-analysis identifies VDAC1 as a predictor of poor outcome in early stage non-small cell lung cancer. *PLoS One* **2011**, *6*, e14635.
58. Uyttenhove, C.; Pilotte, L.; Theate, I.; Stroobant, V.; Colau, D.; Parmentier, N.; Boon, T.; van den Eynde, B.J. Evidence for a tumoral immune resistance mechanism based on tryptophan degradation by indoleamine 2,3-dioxygenase. *Nat. Med.* **2003**, *9*, 1269–1274.
59. Ino, K.; Yamamoto, E.; Shibata, K.; Kajiyama, H.; Yoshida, N.; Terauchi, M.; Nawa, A.; Nagasaka, T.; Takikawa, O.; Kikkawa, F. Inverse correlation between tumoral indoleamine 2,3-dioxygenase expression and tumor-infiltrating lymphocytes in endometrial cancer: Its association with disease progression and survival. *Clin. Cancer Res.* **2008**, *14*, 2310–2317.

60. Sucher, R.; Kurz, K.; Weiss, G.; Margreiter, R.; Fuchs, D.; Brandacher, G. IDO-mediated tryptophan degradation in the pathogenesis of malignant tumor disease. *Int. J. Tryptophan Res.* **2010**, *3*, 113–120.
61. Cai, F.; Dupertuis, Y.M.; Pichard, C. Role of polyunsaturated fatty acids and lipid peroxidation on colorectal cancer risk and treatments. *Curr. Opin. Clin. Nutr. Metab. Care* **2012**, *15*, 99–106.
62. McMillan, D.C.; Crozier, J.E.; Canna, K.; Angerson, W.J.; McArdle, C.S. Evaluation of an inflammation-based prognostic score (GPS) in patients undergoing resection for colon and rectal cancer. *Int. J. Colorectal Dis.* **2007**, *22*, 881–886.
63. Crozier, J.E.; McKee, R.F.; McArdle, C.S.; Angerson, W.J.; Anderson, J.H.; Horgan, P.G.; McMillan, D.C. Preoperative but not postoperative systemic inflammatory response correlates with survival in colorectal cancer. *Br. J. Surg.* **2007**, *94*, 1028–1032.
64. Yun, J.; Rago, C.; Cheong, I.; Pagliarini, R.; Angenendt, P.; Rajagopalan, H.; Schmidt, K.; Willson, J.K.; Markowitz, S.; Zhou, S.; *et al.* Glucose deprivation contributes to the development of KRAS pathway mutations in tumor cells. *Science* **2009**, *325*, 1555–1559.
65. Weinberg, F.; Hamanaka, R.; Wheaton, W.W.; Weinberg, S.; Joseph, J.; Lopez, M.; Kalyanaraman, B.; Mutlu, G.M.; Budinger, G.R.; Chandel, N.S. Mitochondrial metabolism and ROS generation are essential for Kras-mediated tumorigenicity. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 8788–8793.
66. Chatr-aryamontri, A.; Breitkreutz, B.-J.; Heinicke, S.; Boucher, L.; Winter, A.; Stark, C.; Nixon, J.; Ramage, L.; Kolas, N.; O'Donnell, L.; *et al.* The biogrid interaction database: 2013 Update. *Nucl. Acids Res.* **2013**, *41*, D816–D823.
67. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucl. Acids Res.* **2000**, *28*, 235–242.
68. Magrane, M.; Consortium, U. Uniprot knowledgebase: A hub of integrated protein data. *Database* **2011**, *2011*, doi:10.1093/database/bar009.
69. Furlan, D.; Sahnane, N.; Carnevali, I.; Cerutti, R.; Uccella, S.; Bertolini, V.; Chiaravalli, A.M.; Capella, C. Up-regulation and stabilization of HIF-1alpha in colorectal carcinomas. *Surg. Oncol.* **2007**, *16*, S25–S27.
70. Semenza, G.L. Targeting HIF-1 for cancer therapy. *Nat. Rev. Cancer* **2003**, *3*, 721–732.
71. Levine, A.J.; Puzio-Kuter, A.M. The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. *Science* **2010**, *330*, 1340–1344.
72. Kim, J.W.; Tchernyshyov, I.; Semenza, G.L.; Dang, C.V. HIF-1-mediated expression of pyruvate dehydrogenase kinase: A metabolic switch required for cellular adaptation to hypoxia. *Cell Metab.* **2006**, *3*, 177–185.
73. Bluemlein, K.; Gruning, N.M.; Feichtinger, R.G.; Lehrach, H.; Kofler, B.; Ralser, M. No evidence for a shift in pyruvate kinase PKM1 to PKM2 expression during tumorigenesis. *Oncotarget* **2011**, *2*, 393–400.
74. Christofk, H.R.; Vander Heiden, M.G.; Harris, M.H.; Ramanathan, A.; Gerszten, R.E.; Wei, R.; Fleming, M.D.; Schreiber, S.L.; Cantley, L.C. The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* **2008**, *452*, 230–233.

75. Maddocks, O.D.; Vousden, K.H. Metabolic regulation by p53. *J. Mol. Med. (Berl.)* **2011**, *89*, 237–245.
76. Sinha, S.; Ghildiyal, R.; Mehta, V.S.; Sen, E. ATM-NFkappab axis-driven tigar regulates sensitivity of glioma cells to radiomimetics in the presence of TNFalpha. *Cell Death Dis.* **2013**, *4*, e615.
77. Molinari, F.; Frattini, M. Functions and regulation of the PTEN gene in colorectal cancer. *Front. Oncol.* **2013**, *3*, e326.
78. Chalhoub, N.; Baker, S.J. PTEN and the PI3-kinase pathway in cancer. *Annu. Rev. Pathol.* **2009**, *4*, 127–150.
79. Elstrom, R.L.; Bauer, D.E.; Buzzai, M.; Karnauskas, R.; Harris, M.H.; Plas, D.R.; Zhuang, H.; Cinalli, R.M.; Alavi, A.; Rudin, C.M.; *et al.* AKT stimulates aerobic glycolysis in cancer cells. *Cancer Res.* **2004**, *64*, 3892–3899.
80. Buzzai, M.; Bauer, D.E.; Jones, R.G.; Deberardinis, R.J.; Hatzivassiliou, G.; Elstrom, R.L.; Thompson, C.B. The glucose dependence of AKT-transformed cells can be reversed by pharmacologic activation of fatty acid beta-oxidation. *Oncogene* **2005**, *24*, 4165–4173.
81. Ogino, S.; Kawasaki, T.; Ogawa, A.; Kirkner, G.J.; Loda, M.; Fuchs, C.S. TGFBR2 mutation is correlated with CpG island methylator phenotype in microsatellite instability-high colorectal cancer. *Hum. Pathol.* **2007**, *38*, 614–620.
82. Fleming, N.I.; Jorissen, R.N.; Mouradov, D.; Christie, M.; Sakthianandeswaren, A.; Palmieri, M.; Day, F.; Li, S.; Tsui, C.; Lipton, L.; *et al.* SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. *Cancer Res.* **2013**, *73*, 725–735.
83. Bellam, N.; Pasche, B. TGF-beta signaling alterations and colon cancer. *Cancer Treat. Res.* **2010**, *155*, 85–103.
84. Kim, Y.S.; Yi, Y.; Choi, S.G.; Kim, S.J. Development of TGF-beta resistance during malignant progression. *Arch. Pharm. Res.* **1999**, *22*, 1–8.
85. Chan, E.C.; Koh, P.K.; Mal, M.; Cheah, P.Y.; Eu, K.W.; Backshall, A.; Cavill, R.; Nicholson, J.K.; Keun, H.C. Metabolic profiling of human colorectal cancer using high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy and gas chromatography mass spectrometry (GC/MS). *J. Proteome Res.* **2009**, *8*, 352–361.
86. Denkert, C.; Budczies, J.; Weichert, W.; Wohlgemuth, G.; Scholz, M.; Kind, T.; Niesporek, S.; Noske, A.; Buckendahl, A.; Dietel, M.; *et al.* Metabolite profiling of human colon carcinoma—Deregulation of TCA cycle and amino acid turnover. *Mol. Cancer* **2008**, *7*, doi:10.1186/1476-4598-7-72.
87. Tessem, M.B.; Selnaes, K.M.; Sjursen, W.; Trano, G.; Giskeodegard, G.F.; Bathen, T.F.; Gribbestad, I.S.; Hofsli, E. Discrimination of patients with microsatellite instability colon cancer using 1H HR MAS MR spectroscopy and chemometric analysis. *J. Proteome Res.* **2010**, *9*, 3664–3670.
88. Monleon, D.; Morales, J.M.; Barrasa, A.; Lopez, J.A.; Vazquez, C.; Celda, B. Metabolite profiling of fecal water extracts from human colorectal cancer. *NMR Biomed.* **2009**, *22*, 342–348.

89. Weir, T.L.; Manter, D.K.; Sheflin, A.M.; Barnett, B.A.; Heuberger, A.L.; Ryan, E.P. Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS One* **2013**, *8*, e70803.
90. Qiu, Y.; Cai, G.; Su, M.; Chen, T.; Zheng, X.; Xu, Y.; Ni, Y.; Zhao, A.; Xu, L.X.; Cai, S.; *et al.* Serum metabolite profiling of human colorectal cancer using GC-TOFMS and UPLC-QTOFMS.
J. Proteome Res. **2009**, *8*, 4844–4850.
91. Kondo, Y.; Nishiumi, S.; Shinohara, M.; Hatano, N.; Ikeda, A.; Yoshie, T.; Kobayashi, T.; Shiomi, Y.; Irino, Y.; Takenawa, T.; *et al.* Serum fatty acid profiling of colorectal cancer by gas chromatography/mass spectrometry. *Biomark. Med.* **2011**, *5*, 451–460.
92. Ludwig, C.; Ward, D.G.; Martin, A.; Viant, M.R.; Ismail, T.; Johnson, P.J.; Wakelam, M.J.; Gunther, U.L. Fast targeted multidimensional NMR metabolomics of colorectal cancer. *Magn. Reson. Chem.* **2009**, *47*, S68–S73.
93. Leichtle, A.B.; Nuoffer, J.M.; Ceglarek, U.; Kase, J.; Conrad, T.; Witzigmann, H.; Thiery, J.; Fiedler, G.M. Serum amino acid profiles and their alterations in colorectal cancer. *Metabolomics* **2012**, *8*, 643–653.
94. Nishiumi, S.; Kobayashi, T.; Ikeda, A.; Yoshie, T.; Kibi, M.; Izumi, Y.; Okuno, T.; Hayashi, N.; Kawano, S.; Takenawa, T.; *et al.* A novel serum metabolomics-based diagnostic approach for colorectal cancer. *PLoS One* **2012**, *7*, e40459.
95. Farshidfar, F.; Weljie, A.M.; Kopciuk, K.; Buie, W.D.; Maclean, A.; Dixon, E.; Sutherland, F.R.; Molckovsky, A.; Vogel, H.J.; Bathe, O.F. Serum metabolomic profile as a means to distinguish stage of colorectal cancer. *Genome Med.* **2012**, *4*, doi:10.1186/gm341.
96. Tan, B.; Qiu, Y.; Zou, X.; Chen, T.; Xie, G.; Cheng, Y.; Dong, T.; Zhao, L.; Feng, B.; Hu, X.; *et al.* Metabonomics identifies serum metabolite markers of colorectal cancer. *J. Proteome Res.* **2013**, *12*, 3000–3009.
97. Bertini, I.; Cacciatore, S.; Jensen, B.V.; Schou, J.V.; Johansen, J.S.; Kruhoffer, M.; Luchinat, C.; Nielsen, D.L.; Turano, P. Metabolomic NMR fingerprinting to identify and predict survival of patients with metastatic colorectal cancer. *Cancer Res.* **2012**, *72*, 356–364.
98. Xia, J.; Psychogios, N.; Young, N.; Wishart, D.S. Metaboanalyst: A web server for metabolomic data analysis and interpretation. *Nucl. Acids Res.* **2009**, *37*, W652–W660.
99. Ingenuity Systems Pathway Analysis. Availbale online: http://www.ingenuity.com/ (aceessed on 23 December 2014).
100. Li, C.; Han, J.; Yao, Q.; Zou, C.; Xu, Y.; Zhang, C.; Shang, D.; Zhou, L.; Zou, C.; Sun, Z.; *et al.* Subpathway-GM: Identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways. *Nucl. Acids Res.* **2013**, *41*, e101.
101. Nibbe, R.K.; Koyutürk, M.; Chance, M.R. An integrative-omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Comput. Biol.* **2010**, *6*, e1000639.
102. Kanani, H.; Dutta, B.; Klapa, M.I. Individual *vs.* Combinatorial effect of elevated CO_2 conditions and salinity stress on arabidopsis thaliana liquid cultures: Comparing the early molecular response using time-series transcriptomic and metabolomic analyses. *BMC Syst. Biol.* **2010**, *4*, doi:10.1186/1752-0509-4-177.

103. Grimplet, J.; Cramer, G.R.; Dickerson, J.A.; Mathiason, K.; van Hemert, J.; Fennell, A.Y. Vitisnet: "Omics" integration through grapevine molecular networks. *PLoS One* **2009**, *4*, e8365.
104. Oberbach, A.; Bluher, M.; Wirth, H.; Till, H.; Kovacs, P.; Kullnick, Y.; Schlichting, N.; Tomm, J.M.; Rolle-Kampczyk, U.; Murugaiyan, J.; *et al.* Combined proteomic and metabolomic profiling of serum reveals association of the complement system with obesity and identifies novel markers of body fat mass changes. *J. Proteome Res.* **2011**, *10*, 4769–4788.
105. Wong, C.K.; Vaske, C.J.; Ng, S.; Sanborn, J.Z.; Benz, S.C.; Haussler, D.; Stuart, J.M. The UCSC interaction browser: Multidimensional data views in pathway context. *Nucl. Acids Res.* **2013**, *41*, W218–W224.
106. Cbioportal for cancer genomics. Avaible online: http://www.cbioportal.org/public-portal/ (accessed on 14 January 2014).
107. Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; *et al.* The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2012**, *2*, 401–404.
108. Interaction browser. Avaible online: http://sysbio.soe.ucsc.edu/nets/ (accessed on 14 January 2014).
109. Qiu, Y.; Cai, G.; Zhou, B.D.O.M.; Li, D.; Zhao, A.; Xie, G.; Li, H.; Cai, S.; Xie, D.; Huang, C.; *et al.* A distinct metabolic signature of human colorectal cancer with prognostic potential. *Clin. Cancer Res.* **2014**, *20*, 2136–2146.
110. Manna, S.K.; Tanaka, N.; Krausz, K.W.; Haznadar, M.; Xue, X.; Matsubara, T.; Bowman, E.D.; Fearon, E.R.; Harris, C.C.; Shah, Y.M.; *et al.* Biomarkers of coordinate metabolic reprogramming in colorectal tumors in mice and humans. *Gastroenterology* **2014**, *146*, 1313–1324.
111. Douillard, J.Y.; Oliner, K.S.; Siena, S.; Tabernero, J.; Burkes, R.; Barugel, M.; Humblet, Y.; Bodoky, G.; Cunningham, D.; Jassem, J.; *et al.* Panitumumab-FOLFOX4 treatment and RAS mutations in colorectal cancer. *N. Engl. J. Med.* **2013**, *369*, 1023–1034.
112. Karapetis, C.S.; Khambata-Ford, S.; Jonker, D.J.; O'Callaghan, C.J.; Tu, D.; Tebbutt, N.C.; Simes, R.J.; Chalchal, H.; Shapiro, J.D.; Robitaille, S.; *et al.* K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N. Engl. J. Med.* **2008**, *359*, 1757–1765.
113. Leanza, L.; Zoratti, M.; Gulbins, E.; Szabo, I. Mitochondrial ion channels as oncological targets. *Oncogene* **2014**, doi:10.1038/onc.2013.578.
114. Evans, J.M.; Donnelly, L.A.; Emslie-Smith, A.M.; Alessi, D.R.; Morris, A.D. Metformin and reduced risk of cancer in diabetic patients. *Br. Med. J.* **2005**, *330*, 1304–1305.
115. Bowker, S.L.; Majumdar, S.R.; Veugelers, P.; Johnson, J.A. Increased cancer-related mortality for patients with type 2 diabetes who use sulfonylureas or insulin. *Diabetes Care* **2006**, *29*, 254–258.
116. Hirsch, H.A.; Iliopoulos, D.; Tsichlis, P.N.; Struhl, K. Metformin selectively targets cancer stem cells, and acts together with chemotherapy to block tumor growth and prolong remission. *Cancer Res.* **2009**, *69*, 7507–7511.

117. Jiralerspong, S.; Palla, S.L.; Giordano, S.H.; Meric-Bernstam, F.; Liedtke, C.; Barnett, C.M.; Hsu, L.; Hung, M.C.; Hortobagyi, G.N.; Gonzalez-Angulo, A.M. Metformin and pathologic complete responses to neoadjuvant chemotherapy in diabetic patients with breast cancer. *J. Clin. Oncol.* **2009**, *27*, 3297–3302.
118. Higurashi, T.; Takahashi, H.; Endo, H.; Hosono, K.; Yamada, E.; Ohkubo, H.; Sakai, E.; Uchiyama, T.; Hata, Y.; Fujisawa, N.; *et al.* Metformin efficacy and safety for colorectal polyps: A double-blind randomized controlled trial. *BMC Cancer* **2012**, *12*, doi:10.1186/1471-2407-12-118.
119. Jardim, D.L.; Wheler, J.J.; Hess, K.; Tsimberidou, A.M.; Zinner, R.; Janku, F.; Subbiah, V.; Naing, A.; Piha-Paul, S.A.; Westin, S.N.; *et al.* FBXW7 mutations in patients with advanced cancers: Clinical and molecular characteristics and outcomes with mtor inhibitors. *PLoS One* **2014**, *9*, e89388.
120. Francipane, M.G.; Lagasse, E. mTOR Pathway in colorectal cancer: An update. *Oncotarget* **2014**, *5*, 49–66.
121. Francipane, M.G.; Lagasse, E. Selective targeting of human colon cancer stem-like cells by the mTOR inhibitor Torin-1. *Oncotarget* **2013**, *4*, 1948–1962.
122. Ewing, G.P.; Goff, L.W. The insulin-like growth factor signaling pathway as a target for treatment of colorectal carcinoma. *Clin. Colorectal Cancer* **2010**, *9*, 219–223.
123. Golan, T.; Javle, M. Targeting the insulin growth factor pathway in gastrointestinal cancers. *Oncology (Williston Park)* **2011**, *25*, 518–526, 529.
124. Gaglio, D.; Soldati, C.; Vanoni, M.; Alberghina, L.; Chiaradonna, F. Glutamine deprivation induces abortive S-phase rescued by deoxyribonucleotides in K-ras transformed fibroblasts. *PLoS One* **2009**, *4*, e4715.
125. Clem, B.; Telang, S.; Clem, A.; Yalcin, A.; Meier, J.; Simmons, A.; Rasku, M.A.; Arumugam, S.; Dean, W.L.; Eaton, J.; *et al.* Small-molecule inhibition of 6-phosphofructo-2-kinase activity suppresses glycolytic flux and tumor growth. *Mol. Cancer Ther.* **2008**, *7*, 110–120.

An Efficient Estimator of the Mutation Parameter and Analysis of Polymorphism from the 1000 Genomes Project

Yunxin Fu

Abstract: The mutation parameter θ is fundamental and ubiquitous in the analysis of population samples of DNA sequences. This paper presents a new highly efficient estimator of θ by utilizing the phylogenetic information among distinct alleles in a sample of DNA sequences. The new estimator, called Allelic BLUE, is derived from a generalized linear model about the mutations in the allelic genealogy. This estimator is not only highly accurate, but also computational efficient, which makes it particularly useful for estimating θ for large samples, as well as for a large number of cases, such as the situation of analyzing sequence data from a large genome project, such as the 1000 Genomes Project. Simulation shows that Allelic BLUE is nearly unbiased, with variance nearly as small as the minimum achievable variance, and in many situations, it can be hundreds- or thousands-fold more efficient than a previous method, which was already quite efficient compared to other approaches. One useful feature of the new estimator is its applicability to collections of distinct alleles without detailed frequencies. The utility of the new estimator is demonstrated by analyzing the pattern of θ in the data from the 1000 Genomes Project.

Reprinted from *Genes*. Cite as: Fu, Y. An Efficient Estimator of the Mutation Parameter and Analysis of Polymorphism from the 1000 Genomes Project. *Genes* **2014**, *5*, 561–575.

1. Introduction

The pattern of genetic polymorphism in a population can be influenced by a number of factors, among which the mutation parameter (commonly denoted by θ) plays a central role. θ is defined as $4Nu$ and $2Nu$ for diploid and haploid genomes, respectively, where N is the effective population size and u is the mutation rate per sequence per generation. Almost all existing summary statistics for polymorphism are related to θ. Well-known examples include the number of alleles in a sample [1] the number (K) of segregating sites (or polymorphic sites) [2], mean number (Π) of nucleotide differences between two sequences [3] and the number of mutations of various sizes [4].

Due to the fundamental nature of this parameter for understanding both population dynamics, as well as the mechanism of evolution, it is important that it can be estimated as accurately as possible. Classical estimators include Watterson's estimator [2], Tajima's estimator [3], Ewens' estimator based on the number of alleles [5] in the sample and the heterozygosity estimator [6]. Under the assumption of a single random mating population evolving according to the Wright–Fisher model with constant population size and neutral mutations, these estimators are all either unbiased or nearly unbiased. However, their variances, which are the primary measure of accuracy of an estimator, can differ considerably and, furthermore, are substantially larger than the minimum achievable variance [7]:

$$V_{min} = \theta \left[\sum_{i=1}^{n-1} \frac{1}{\theta + i} \right]^{-1} \tag{1}$$

where n is the sample size. Realizing the limitations of these classical estimators, several new approaches were developed in the 1990s, all utilizing the fine structural result of coalescent theory [3,8,9]. Representative are Griffiths and Tavare's Markov Chain Monte Carlo (MCMC) estimator [10,11] based on recurrent equations for the probability of the polymorphism configuration, Knuher and Felsenstein's MCMC method [12] based on Metropolitan-Hasting sampling and Fu's BLUE estimators [13,14] based on linear regression taking advantage of the linear relationship between mutations in the genealogy of a sample and the mutation parameter. These new groups of estimators can all achieve substantially smaller variances and may even reach the minimum variance [13]. One common feature of these estimators is that they are all computationally intensive and, as a result, are suitable for only relatively smaller samples. Such limitations are particularly serious for the MCMC-based approach.

The potential for genetic research based on population samples has been greatly enhanced by the steady reduction in the cost of sequencing. As a result, sample sizes in these studies are substantially larger than before, and the trend will continue with the arrival of next generation sequencers. Already, it is commonplace to see sequenced samples of many hundreds of individuals and even thousands (such as the sample in the 1000 Genomes Project [15]). The reduction of sequencing cost also leads to a larger region of the genome or even the entire genome being sequenced (e.g., 1000 Genomes Project). Consequently, new approaches that are both highly accurate and efficient in computation are desirable. This paper presents one such method and demonstrates its utility by analyzing polymorphism from the 1000 Genomes Project.

2. Theory and Method

2.1. The Theory

Assume that a sample of n DNA sequences at a locus without recombination is taken from a single population evolving according to the Wright–Fisher model and all mutations are selectively neutral. The sample genealogy thus consists of $2(n-1)$ branches, each spanning at least one coalescent time (Figure 1). The number of mutations that occurred in a branch is thus the sum of the numbers of mutations in the coalescent time it spans. Consider one branch, and without loss of generality, assume it spans the i-th coalescent time. Then, during the i-th coalescent time, the number of mutations occurred in the branch has expectation and variance equal to:

$$\frac{\theta}{i(i-1)} \qquad \text{and} \qquad \frac{\theta}{i(i-1)} + \frac{1}{i^2(i-1)^2}\theta^2 \tag{2}$$

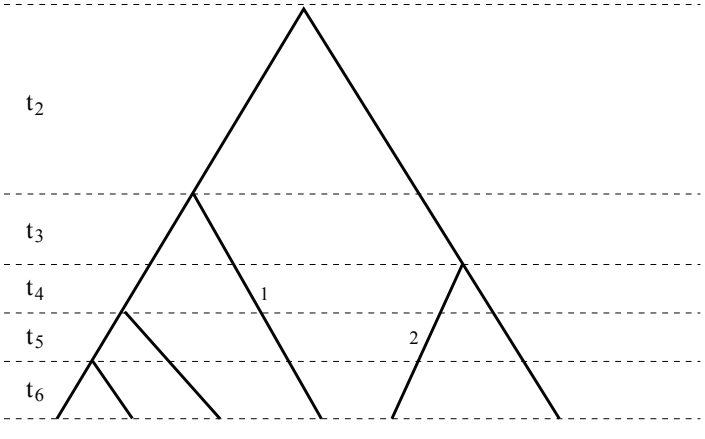
respectively. These are consequences of the coalescent time being exponentially distributed and the number of mutations in a given number of generations following a Poisson distribution. Consider

the number of mutations in another branch that spans the j-th coalescent time. Then, the covariance between the two numbers of mutations is equal to:

$$\frac{\theta^2}{i^2(i-1)^2} \tag{3}$$

if $i = j$, and zero otherwise.

Figure 1. A sample genealogy with different coalescent times separated by dashed lines. Branch 1 spans the third to the sixth coalescent times, $\chi_1(2) = 0,\ \chi_1(i) = 1,$ for $i = 3, ..., 6,$ while Branch 2 spans the fourth to the sixth coalescent times, $\chi_2(2) = \chi_2(3) = 0,\ \chi_2(4) = \chi_2(5) = \chi_2(6) = 1.$ Combining Branches 1 and 2 results in $\phi(2) = 0, \phi(3) = 1, \phi(4) = \phi(5) = \phi(6) = 2.$



For the branch $k(k = 1, ..., 2(n-1))$ in the genealogy, define an index $\chi_k(i)$, such that it takes value one if the branch spans the i-th coalescent time and zero otherwise. Then, m_k, the number of mutations, has its expectation and variance equal to:

$$E(m_k) = \theta \sum_{i=2}^{n} \frac{\chi_k(i)}{i(i-1)} \tag{4}$$

$$V(m_k) = E(m_k) + \theta^2 \sum_{i=2}^{n} \frac{\chi_k^2(i)}{i^2(i-1)^2} \tag{5}$$

respectively, and for two different branches a and b:

$$Cov(m_a, m_b) = \theta^2 \sum_{i=2}^{n} \frac{\chi_a(i)\chi_b(i)}{i^2(i-1)^2} \tag{6}$$

The previous results are readily generalized. Instead of considering the mutations in different branches separately, one can combine mutations in several branches. Suppose branches $(k_1, ..., k_t)$ are combined. Define for the combined branches a variable ϕ as:

$$\phi(i) = \sum_{j=1}^{t} \chi_{k_j}(i) \tag{7}$$

Consider a population dynamics model in which the effective population size can change only at the time a coalescent event occurs. Although such a model does not stem from biological reality, its laddered changes in population sizes allow a reasonable approximation of reality and makes the mathematics simpler. Let θ_i represent the θ during the i-th coalescent period. Suppose the combined branches is denoted by branch (group) k, then m_k, the number of mutations in branch k has expectation and variance equal to:

$$E(m_k) = \sum_{i=2}^{n} \frac{\phi_k(i)\theta_i}{i(i-1)} \tag{8}$$

$$V(m_k) = E(m_k) + \sum_{i=2}^{n} \frac{\phi_k^2(i)\theta_i^2}{i^2(i-1)^2} \tag{9}$$

respectively, and for two such branches a and b, we have:

$$Cov(m_a, m_b) = \sum_{i=2}^{n} \frac{\phi_a(i)\phi_b(i)\theta_i^2}{i^2(i-1)^2} \tag{10}$$

Suppose that the $2(n-1)$ branches of the sample genealogy are divided into M $(\leq 2(n-1))$ disjoint groups (*i.e.*, each branch belongs to one and only one group). Let m_k represent the number of mutations in branch group k and $\boldsymbol{m} = (m_1, ..., m_k)^T$. Then, similar to the previous result [13], the relationship between $\boldsymbol{\theta} = (\theta_2, ..., \theta_n)^T$ and $\boldsymbol{m}$ can be expressed by a generalized linear model:

$$\boldsymbol{m} = \boldsymbol{\alpha}\boldsymbol{\theta} + \boldsymbol{e} \tag{11}$$

where $\boldsymbol{\alpha}$ is a matrix of dimension $M \times n$ with: $\alpha_{ij} = \frac{\phi_i(j)}{j(j-1)}$ and $\boldsymbol{e}$ a vector of length M representing error terms. Let $\Gamma(\boldsymbol{\theta}) = Var(\boldsymbol{m})$. Then:

$$\Gamma(\boldsymbol{\theta}) = \boldsymbol{\gamma}(\boldsymbol{\theta}) + \boldsymbol{\beta}(\boldsymbol{\theta}) \tag{12}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are both $M \times M$ matrices defined as:

$$\boldsymbol{\gamma}(\boldsymbol{\theta}) = Diag\{\boldsymbol{\alpha}_1\boldsymbol{\theta}, ..., \boldsymbol{\alpha}_M\boldsymbol{\theta}\} \tag{13}$$

$$\boldsymbol{\beta}(\boldsymbol{\theta}) = \left\{ \sum_{k=2}^{n} \frac{\phi_i(k)\phi_j(k)\theta_k^2}{k^2(k-1)^2} \right\} \tag{14}$$

where $\boldsymbol{\alpha}_k$ represents the k-th row vector of $\boldsymbol{\alpha}$. Following the previous approach [13,14], a best linear unbiased estimator of $\boldsymbol{\theta}$ can be obtained as the limit of the series:

$$\boldsymbol{\theta}^{(k+1)} = [\boldsymbol{\alpha}^T\Gamma(\boldsymbol{\theta}^{(k)})^{-1}\boldsymbol{\alpha}]^{-1}\boldsymbol{\alpha}^T\Gamma(\boldsymbol{\theta}^{(k)})^{-1}\boldsymbol{m} \tag{15}$$

with $\boldsymbol{\theta}^{(0)}$ being the initial estimate of $\boldsymbol{\theta}$ (for example, setting all θ_i equal to Watterson's estimate of θ).

The above formulation allows maximal $n-1$ different values of θ corresponding to the $n-1$ coalescent periods. Although very flexible, such an extreme model may lead to reduced accuracy of estimation for individual parameters, so some compromise is likely to be useful. When two or more consecutive θ values are assumed to be the same, the length of the $\boldsymbol{\theta}$ vector will be reduced. At the extreme, if all of the θs are the same, $\boldsymbol{\theta}$ is reduced to a single quantity, and when $M = 2(n-1)$, it further defaults to BLUE [13]. Since efficient estimators for a single value of θ will continue to be useful in the analysis of the whole genome polymorphism of large samples, we will focus on developing one such scheme in this paper.

2.2. Allelic BLUE estimator

In order to take advantage of the BLUE estimator, sample genealogy needs to be inferred. Furthermore, the key to developing a highly efficient BLUE estimator is to define the M groups of branches, which not only retains the detail mutational information, but also satisfies the relationship $M << 2(n-1)$. Fu's UPBLUE [13] corresponds to the extreme in which $M = 2(n-1)$, *i.e.*, each branch belongs to its own group. While this may retain the maximal mutational information, it leads to computational inefficiency. Fu's [14] approach is more or less equivalent to $M = n-1$, with groups defined by the size of mutations. This achieves computational efficiency with the expense of reduced accuracy due to over condensation of mutational information. Thus, the goal here is to strive for a balance.

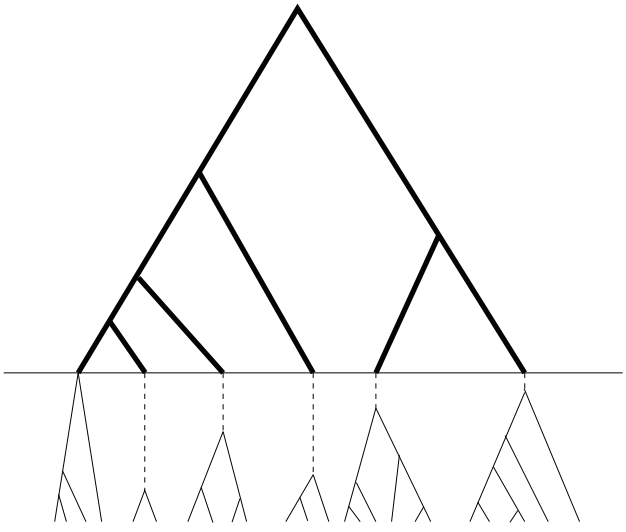
We recognize that much of the phylogenetic information in a sample resides in the pattern of differences between distinct alleles. The phylogenetic method, UPGMA (e.g., [16]), which was found to be appropriate in Fu [13], will continue to be used in our new method. Since UPGMA is a sequential method, which at each step joins two sequences (or two groups of sequences) that differ the least. As a result, copies of sequences of the same allele will be joined together before any pair of sequences of distinct alleles. The resulting sample genealogy can be roughly divided into two portions (see Figure 2), with the bottom portion corresponding to the coalescent within allelic class and the upper portion the coalescent among allelic classes. Combine all the branches (or segments of branches) underneath the dashed line into one group, which will be referred to as the within allele branch. Suppose there are L distinct alleles in the sample; then we have for the within allele branch:

$$\phi(i) = \begin{cases} 0, & i \le L \\ i, & i > L \end{cases} \tag{16}$$

Then, the number of mutations in the within allele branch have expectation and variance equal respectively to:

$$(a_n - a_L)\theta \qquad \text{and} \qquad (a_n - a_L)\theta + (b_n - b_L)\theta^2 \,. \tag{17}$$

Figure 2. Schematic relationship between sample genealogy and allelic genealogy. The dark portion is the genealogy of distinct alleles, while the light portion (which is below the horizontal line) is the coalescent within alleles and contains no mutation or only a few in the dashed segments of the branches.



Furthermore, we assume that there is no mutation in the within allele branch (which should be a good approximation, although technically, the assertion may not be true). Since the within allele branch does not span any coalescent time that overlaps with those of branches in the allelic genealogy, we have (assumed that the last branch group represents the within allele subtree) that:

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}^* \\ a_n - a_L \end{pmatrix} \tag{18}$$

where $\boldsymbol{\alpha}^*$ a vector of length $2(L-1)$ with the k-th element equal to $\sum_{i=2}^{L} \frac{\phi_k(i)}{i(i-1)}$. The inverse of the covariance matrix of $\boldsymbol{m}$ is:

$$\begin{pmatrix} \boldsymbol{\Gamma}^*(\theta)^{-1} & \mathbf{0} \\ 0 & [(a_n - a_L)\theta + (b_n - b_L)\theta^2]^{-1} \end{pmatrix} \tag{19}$$

where $\boldsymbol{\Gamma}^*$ is defined for branch groups of the allelic genealogy. Let $\boldsymbol{m}^*$ be the vector of mutations in branches of the allelic genealogy (the dark portion in the genealogy of Figure 2). Then, Equation (15) becomes:

$$\theta^{(k+1)} = \frac{(\alpha^*)^T \boldsymbol{\Gamma}^*(\theta^{(k)})^{-1}}{(a_n - a_L)^2[(a_n - a_L) + (b_n - b_L)\theta^{(k)}]^{-1} + (\alpha^*)^T \boldsymbol{\Gamma}^*(\theta^{(k)})^{-1}\boldsymbol{\alpha}^*} \boldsymbol{m}^* \tag{20}$$

This limit will be referred to as the Allelic BLUE estimator denoted by θ_{ab}. Since, for large samples, the number of distinct alleles is typically much smaller than the sample size; thus, the new estimator is expected to be highly efficient computationally.

To determine if it is indeed true that merging those branches representing within allele coalescent does not lead to significant loss of information and, thus, would not reduce the accuracy of estimation, we compared Allelic BLUE with the original BLUE using simulated samples for a number of combinations of θ and n. The correlation between the two estimates is around 0.99. Therefore, Allelic BLUE is expected to be as accurate as BLUE without merging branches.

2.3. Bias-Corrected Allelic BLUE Estimator

Since UPGMA systematically introduces bias in the inferred sample genealogy, the resulting Allelic BLUE estimate is expected to be biased similar to the BLUE estimator [13]. Therefore, it is necessary to correct the bias. Similar to Fu [13], we used simulated samples to derive understanding of the pattern of biases. A total of 550 combinations of θ and n were examined with 25 different θ values: $0.5, 0.75, 1, 1.5, 2(1)5, 6(2)12, 15(5), 50, 60(10)100$ and 150, and 25 different sample sizes n: $10(5)25, 30(10)60, 80, 100(25)200, 250, 300, 400, 500, 750, 1000(1000)5000$. For each combination of the parameters, 1000 samples were simulated, and for each simulated sample, θ_{ab} was obtained and their mean value computed over all simulated samples. Similar to those in Fu [13], the estimates in almost all situations are underestimates of the true θ. In general, the underestimate is the result of systematic bias of the UPGMA algorithm used to construct the genealogy, because UPGMA leads to early coalescent for more similar sequences and, thus, has a tendency to place more mutations in branches that are deeper into the tree. In the current situation, it is further compounded by our simplification that, up to the $i + 1$ coalescent, there are no mutations.

Using regression analysis, Fu [13] showed that the relationship:

$$\sqrt{\theta_u} = -0.0336\sqrt{n-2} + 1.002\sqrt{\theta} \tag{21}$$

summarizes well the BLUE estimate (with $M = 2(n-1)$), θ and sample size n, which is not larger than 100. When larger sample sizes were examined, the above equation is not adequate, and log transformation, rather than square-root transformation, can lead to a better regression [17]. Therefore, log-transformation was chosen in our regression analysis. Table 1 showed that $ln(\theta_{ab})$ can be summarized very well by the following equation:

$$\begin{aligned} ln(\theta_{ab}) = &- 0.1981 + 0.0080ln(n) + 1.0043ln(\theta) \\ &- 0.0019ln(n)ln(\theta) + 0.0108ln^2(\theta) - 0.0012ln(n)ln^2(\theta) \end{aligned} \tag{22}$$

which leads to an estimate of θ from the solution for the above quadratic equation with regard to $ln(\theta_a)$:

$$\hat{\theta}_a = \exp\left[\frac{-b + \sqrt{b^2 - 4ac}}{2a}\right] \tag{23}$$

where:

$$a = 0.0108 - 0.0012ln(n) \tag{24}$$

$$b = 1.0043 - 0.0019ln(n) \tag{25}$$

$$c = -ln(\theta_{ab}) - 0.1981 + 0.008ln(n) \tag{26}$$

Although this estimator in most situations is excellent, we found that regression equations for a narrower range of sample sizes provides estimates that are more robust in some situations (particularly when θ is large). As a result, we derive our final estimator $\hat{\theta}_a$ of θ using Equation (23) with values of a, b and c, as provided in Table 2.

Table 1. Summary of regression analysis between θ_{ab}, θ and n.

Source	Sum of Squares	Degrees of freedom	Mean Square	
Model	1,715.227	5	343.0454	
Residual	0.074	644	0.0001	
Total	1,715.301	649	2.6430	
Term	Coefficient	Standard Error	t test	$P > \|t\|$
$ln(n)$	0.0080	0.0005	16.73	0.000
$ln(\theta)$	1.0043	0.0025	398.16	0.000
$ln(n)ln(\theta)$	-0.0019	0.0004	-4.19	0.000
$ln^2(\theta)$	0.0108	0.0005	19.80	0.000
$ln(n)ln^2(\theta)$	-0.0012	0.0001	-12.74	0.000
constant	-0.1981	0.0027	-72.99	0.000

Note: Number of points for regression = 650, $R^2 = 1.000$ and MSE = 0.0107

Table 2. Coefficients for estimating θ using Equation (23).

	Coefficients ($n' = ln(n)$)		
Sample Size	**a**	**b**	**c**
$n < 50$	$0.0112 - 0.0012n'$	$1.0076 - 0.0026n'$	$-ln(\theta_a) - 0.2101 + 0.0107n'$
$50 \leq n < 500$	$0.0131 - 0.0017n'$	$1.0009 - 0.0016n'$	$-ln(\theta_a) - 0.1980 + 0.0088n'$
$n \geq 500$	$0.0069 - 0.0007n'$	$0.9850 - 0.0008n'$	$-ln(\theta_a) - 0.1581 + 0.0025n'$

Figure 3 plots the relationship between θ, sample size (n) and the allelic BLUE estimates (θ_{ab}) for a subset of these parameter combinations. It is easy to see that the match between prediction and the mean value of θ_a is excellent.

Figure 4 shows the distributions of the estimate $\hat{\theta}_a$ from simulated samples in the case of $n = 500$ with $\theta = 5$, and $n = 2000$ with $\theta = 50$, respectively. It appears in both cases that the normalities are sufficiently accurate approximations, which is indeed expected from the theory of least squares estimators.

The ultimate measure of the quality of an estimator is its bias and standard deviation for samples independent of those used to derive the estimator. Therefore, we simulated another set of samples for a number of combinations of θ and n and applied $\hat{\theta}_a$, as well as UPBLUE to these samples. Table 3 presents the summary of these simulations, particularly the efficiency of the new approach.

Figure 3. The relationship between Allelic Blue estimate θ_a, θ and sample size. Solid lines represents the prediction of θ_a based on Equation (23) with the coefficients given in Table 2.

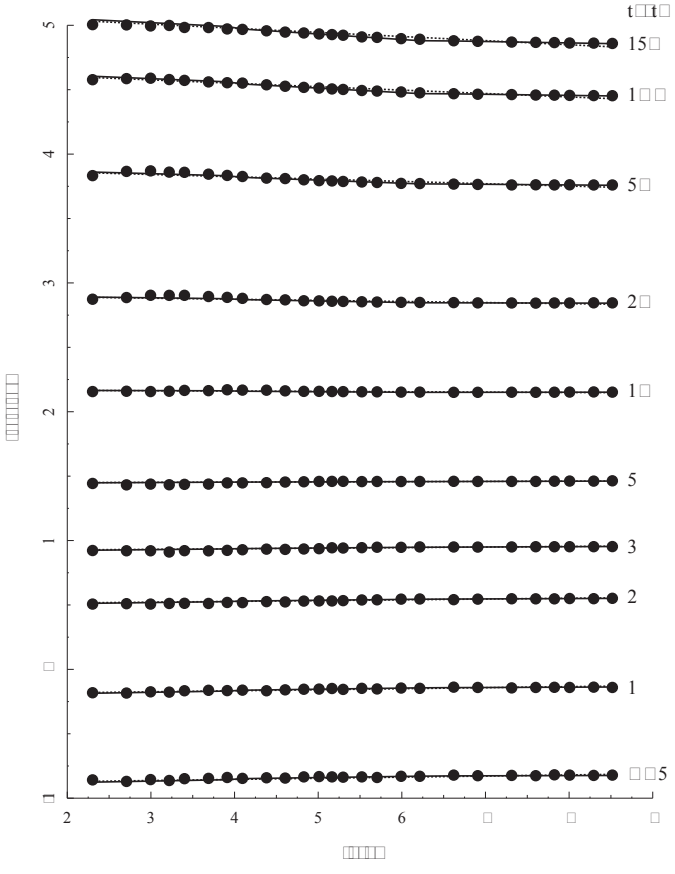


Figure 4. Distribution of $\hat{\theta}_a$ based on 1000 simulated samples. (**Left**) $n = 500$ and $\theta = 5$; (**right**) $n = 2000$ and $\theta = 50$.

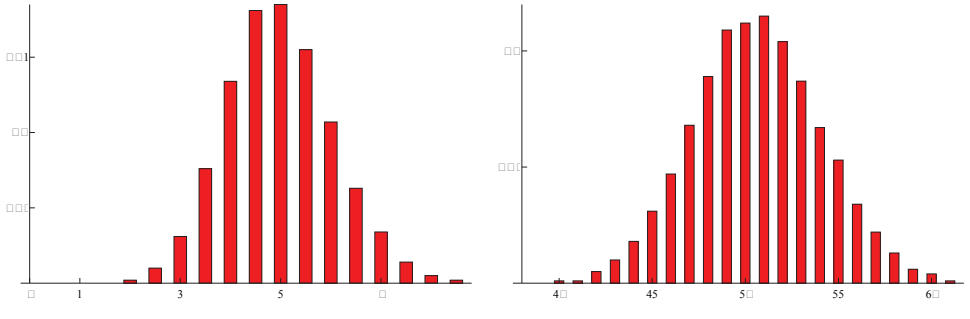


Table 3. Performance of $\hat{\theta}_a$.

θ	n	mean $\hat{\theta}_a$	SE	SD_{min}	Speed	Ratio
2	20	1.97	1.00	0.97	0.00	10
	50	1.98	0.84	0.81	0.00	59
	100	1.98	0.75	0.74	0.00	280
	500	2.00	0.63	0.61	0.02	1,569
	1,000	2.00	0.58	0.58	0.27	1,615
	2,500	2.00	0.54	0.54	2.59	3,104
5	20	4.93	1.90	1.83	0.00	4.5
	50	4.96	1.52	1.48	0.00	23
	100	4.99	1.34	1.30	0.00	78
	500	5.01	1.08	1.05	0.04	970
	1,000	5.00	1.00	0.98	0.47	1,306
	2,500	4.91	0.87	0.90	4.90	1,565
20	20	20.33	5.84	5.52	0.00	2.0
	50	20.27	4.20	4.05	0.01	4.4
	100	20.06	3.47	3.37	0.06	8.7
	500	20.04	2.56	2.49	0.18	359
	1,000	19.99	2.32	2.26	0.91	727
	2,500	19.96	2.07	2.04	16	842
50	20	50.92	12.58	12.51	0.01	1.6
	50	50.45	8.68	8.59	0.02	2.5
	100	50.09	6.99	6.79	0.07	3.8
	500	50.06	4.70	4.58	0.97	72
	1,000	49.90	4.15	4.06	3.6	206
	2,500	49.80	3.94	3.57	42	356
100	20	100.25	23.47	24.03	0.01	1.5
	50	100.20	15.52	15.87	0.15	1.7
	100	99.97	12.24	12.08	0.20	2
	500	100.38	7.86	7.48	4	22
	1,000	99.89	6.59	6.47	16	49
	2,500	99.69	5.58	5.54	75	202

Note: Speed is the average CPU time (in seconds) for obtaining $\hat{\theta}_a$ for a simulated sample in a Linux machine with a 2.3-Ghz CPU. SD_{min} is the minimum standard deviation computed as the square root of Equation (28) in [18]. Ratio is the ratio of speed for UPBLUE [13] and speed of $\hat{\theta}_a$.

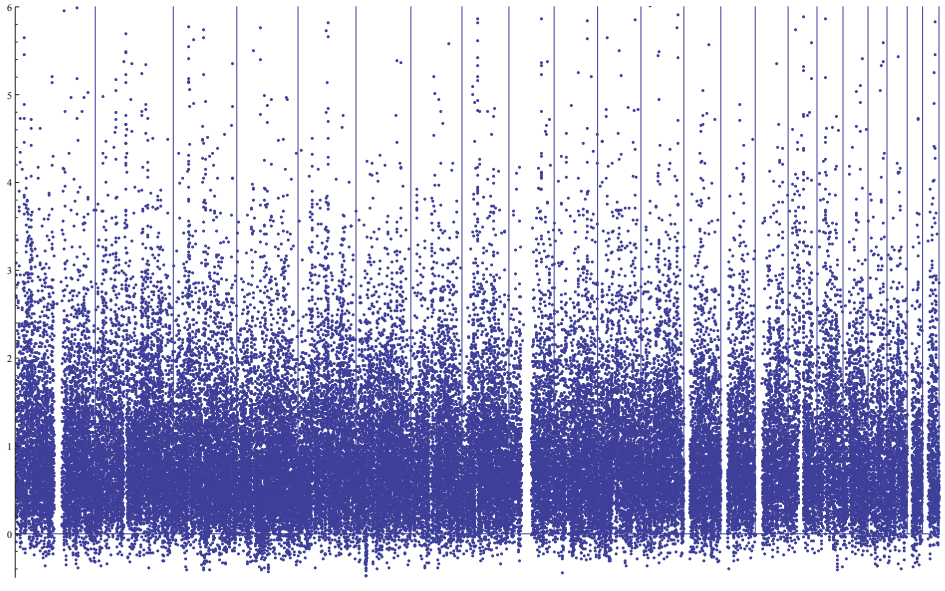
Table 3 shows that the speed of $\hat{\theta}_a$ increases with the sample size slowly, while it increases faster with θ. This is because $\hat{\theta}_a$'s speed is dependent on the number of alleles in the sample, which is more closely related to θ than sample size. In comparison, UPBLUE is considerably less efficient, particularly with increasing sample size. Take the case of $\theta = 100$ and $n = 5000$, it takes on

average about one minute for $\hat{\theta}_a$ to complete the estimation, while it takes about 6 h for UPBLUE to do the same.

3. Exploring θ in Data from the 1000 Genomes Project

The 1000 Genomes Project generated a very valuable set of genome-wide polymorphism data [15] for which the newly developed Allelic Blue estimator is applicable. Phase I (May, 2012, release) contains polymorphism, as well as inferred phases compiled from 1092 individuals from 14 different populations. The rich information captured by the genome-wide polymorphism deserves to be analyzed from various angles [20], and our main purpose here is to illustrate that our efficient estimator of θ provides additional insight into the pattern of polymorphism in addition to the conventional estimates. We chose to report the analysis for a subset of samples, which consists of the three African samples (YRI, LWKand ASW) with 246 individuals (thus, a sample size of $n = 492$).

Figure 5. Plot of $(\hat{\theta}_W - \hat{\theta}_T)/\hat{\theta}_T$ along 22 autosomes with windows of a size of 2000 bps (each dot represents a mean over 10 consecutive windows). Chromosomes 1 to 22 are presented from left to right separated by vertical lines. The overall mean value is 0.96.

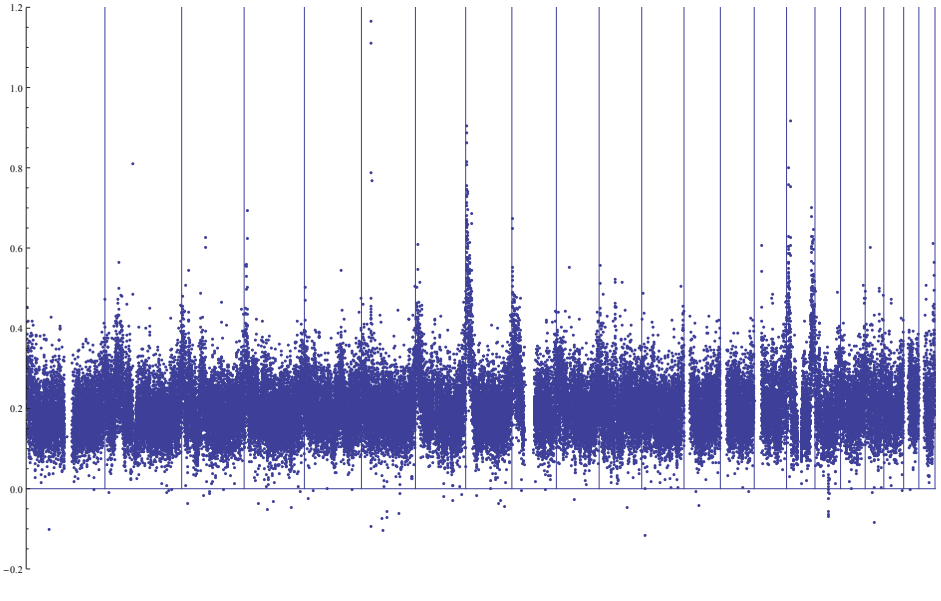


In our analysis, each of the 22 autosomal chromosomes was divided into non-overlapping consecutive windows of 2000 bps (within which the average impact of recombination should be negligible), and θ was estimated for each window. Since the phases of SNPs for each individual were the result of inference, there are some segments in which the quality of inference appears to be poor due to an unreasonably larger number of inferred alleles than the number of SNPs. We thus removed all of the segments in which there is evidence of either recombination or a poor quality of inference,

that is, when the number of alleles is larger than the number of SNPs plus one. It should be noted that the SNPs reported in the 1000 Genomes Project data set are those that passed various quality controls and filtering. In our analysis, no further filtering is applied, except for the aforementioned exclusion of segments that are suspected to be the result of poor phrase inference. A total of 648,903 windows were analyzed. This analysis required about 3 h to complete the estimates of θ in a desktop computer equipped with an Intel Xeon CPUs at 3.33 Ghz. In comparison, UBPLUE ran a couple days without being able to finish the same task.

Since both Watterson's estimator $\hat{\theta}_W$ and Tajima's estimator ($\hat{\theta}_T$) are widely used, we present our results in terms of the relative values with regard to $\hat{\theta}_W$ and contrast them with the relative values of $\hat{\theta}_W$ to $\hat{\theta}_T$. Since testing neutrality is not the purpose, we do not employ testing statistics, such as Tajima's test [19] or Fu and Li's tests [7]. Figure 5 plots the proportional difference between $\hat{\theta}_W = K/a_n$ (K is the number of segregating sites and a_n is a constant depending on the sample size) and $\hat{\theta}_T$. The overwhelming characteristic of the plot is that θ_W is, in general, larger than $\hat{\theta}_T$, with an average of 1.96-times the value of $\hat{\theta}_T$; similar patterns were observed previously (for example, [20]). In general, an estimated θ can be viewed as a weighted average of SNPs of various sizes. $\hat{\theta}_T$ gives on average more weight to SNPs that occurred long ago than those arisen recently, while $\hat{\theta}_W$ gives equal weight to every SNP. Therefore, elevated $\hat{\theta}_W$ values across the whole genome compared to $\hat{\theta}_T$ were considered as evidence of recent population growth. It should be noted that there is no obvious extended regions with smaller or larger values for θ_W or θ_T. In comparison, Figure 6 plots the proportional difference between $\hat{\theta}_a$ and $\hat{\theta}_W$.

Figure 6. Plot of $(\hat{\theta}_a - \hat{\theta}_W)/\hat{\theta}_W$ along 22 autosomes with windows of a size of 2000 bps (each dot represents a mean over 10 consecutive windows). Chromosomes 1 to 22 are presented from left to right separated by vertical lines. The overall mean value is 0.20.



The overwhelming pattern shown in Figure 6 is again that $\hat{\theta}_a$ in general is larger than $\hat{\theta}_W$, which means that the difference to $\hat{\theta}_T$ will be more pronounced than that of $\hat{\theta}_T$. This is the result that more weight is given to recent mutations than the old ones in $\hat{\theta}_a$. Beside some sporadic large values, there are regions at either the beginning or the end of some chromosomes that yield considerably elevated values of $\hat{\theta}_a$ (for example, for chromosomes 7, 8, 9, 16 and 22). We are not sure how to interpret these patterns, but suspect that they may partially suggest the decreased quality of phase inference at the beginning and end of chromosomes.

4. Discussion and Conclusions

The Allelic BLUE estimator of θ presented in this paper is a high quality estimator with little bias and its variance nearly as small as the minimum achievable variance. Furthermore, it is highly efficient computationally, because its speed depends on the number of distinct alleles in a sample rather than the sample size. This later characteristic makes it very useful for estimating θ for large samples and for situations in which a large region (or the whole genome) is sequenced, while θ needs to be estimated for successive windows of a genome, such as the case of 1000 Genomes Projects. Since θ_a and UPBLUE are both based on the same idea of utilizing phylogenetic information in a sample with generalized linear regression, their estimates are highly correlated, which are seen in both the simulation and in real data. However, since θ_a is computationally much more efficient, it is thus superior to UPBLUE [13] and, thus, can replace UPBLUE for general use. The analysis of the polymorphism from the 1000 Genomes Project shows that although UPBLUE is a relatively efficient estimator among sophisticated estimators; it has nearly reached its limit for exploratory data analysis for large genome projects. Therefore, the new Allelic BLUE estimator arrival is timely.

One additional utility of the new estimator $\hat{\theta}_a$ is for estimating θ from a collection of distinct alleles, which are collected without recording the multiplicity of each allele, as long as the number of alleles examined is roughly known. Such situations sometime arise when the collection of data is focused on identifying distinct alleles, such as in the survey of infectious pathogens or when data are collected over years and pooled from multiple sources. To illustrate this utility, we simulated samples of size 200 with $\theta = 10$. If only the distinct alleles are recorded (which implicitly assumes that the sample size is the same as the number of distinct alleles), then θ_a yields an average value of 22.8, which is more than twice as large as the true value. However, if the sample size used in the estimation is 20% smaller or larger than the actual value (thus, 160 and 240, respectively), the corresponding $\hat{\theta}_a$ are 10.6 and 9.5, respectively, both of which are quite close to the true θ value.

The Allelic BLUE estimator is developed under the assumption of one single random mating population evolving according to the Wright–Fisher model with a constant effective population size. The restriction to constant population size results in an estimator of average θ since the MRCAof a sample, which is comparable to some classical estimators and, thus, provides an informative contrast to other estimators (and may be used to construct hypothesis tests in the future). However, the restriction does make the method unsuitable for exploring historical changes in effective population sizes. On the other hand, the theoretical foundation for exploring changes in population sizes in

the linear regression framework has been established in this paper, and we will study its statistical properties, as well as its application elsewhere.

The Java programs for performing the Allelic Blue estimation can be downloaded from the author's web page [21].

Acknowledgments

This work was supported in part by grants from NIH (5U01HG005728 Fu, 2U54 HG003273 Gibbs), Betty Wheless Trotter Endowment Fund from The University of Texas Health Science Center at Houston (Fu) and the fund from Yunnan University, Kunming, China.

Conflicts of Interest

The author declares no conflicts of interest.

References

1. Ewens, W.J. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **1972**, *3*, 87–112.
2. Watterson, G.A. On the number of segregation sites. *Theor. Popul. Biol.* **1975**, *7*, 256–276.
3. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **1983**, *105*, 437–460.
4. Fu, Y.X. Statistical properties of segregating sites. *Theor. Popul. Biol.* **1995**, *48*, 172–197.
5. Ewens, W.J. *Mathematical Population Genetics*; Springer-Verlag: New York, NY, USA, 2004.
6. Xu, H.; Fu, Y.X. Estimating effective population size or mutation rate with microsatellites. *Genetics* **2004**, *166*, 555–563.
7. Fu, Y.X.; Li, W.H. Statistical tests of neutrality of mutations. *Genetics* **1993**, *133*, 693–709.
8. Kingman, J.F.C. The coalescent. *Stoch. Process. Their Appl.* **1982**, *13*, 235–248.
9. Hudson, R.R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **1983**, *23*, 183–201.
10. Griffiths, R.C.; Tavaré, S. Ancestral inference in population genetics. *Statist. Sci.* **1994**, *9*, 307–319.
11. Griffiths, R.C.; Tavaré, S. Monte Carlo inference methods in population genetics. *Math. Comput. Model.* **1996**, *23*, 141–158.
12. Kuhner, M.K.; Yamato, Y.; Felsenstein, J. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **1995**, *140*, 1421–1430.
13. Fu, Y.X. A phylogenetic estimator of effective population size or mutation rate. *Genetics* **1994**, *136*, 685–692.
14. Fu, Y.X. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. *Genetics* **1994**, *138*, 1375–1386.
15. 1000 Genomes Project Consortium. Website for the data generated by 1000 Genomes Project: http://www.1000genomes.org/ (accessed on 1 July 2014).

16. Nei, M. *Molecular Evolutionary Genetics*; Columbia University Press: New York, NY, USA, 1987.
17. Zhang, F. Statistical Methods for Estimating Mutation Rate and Effective Population Size from Samples of DNA Sequences. Ph.D. Dissertation, The University of Texas Health Science Center at Houston, Houston, TX, 2003.
18. Fu, Y.X.; Li, W.H. Maximum likelihood estimation of population parameters. *Genetics* **1993**, *134*, 1261–1270.
19. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **1989**, *123*, 585–595.
20. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1092 human genomes. *Nature* **2012**, *49*, 56–65.
21. Fu, Y.X. The Java programs for performing the Allelic Blue estimation. Availiable online: https://sph.uth.edu/yfu/ (accessed on 1 July 2014).

The Challenges of Genome Analysis in the Health Care Setting

Anneke Lucassen and Richard S. Houlston

Abstract: Genome sequencing is now a sufficiently mature and affordable technology for clinical use. Its application promises not only to transform clinicians' diagnostic and predictive ability, but also to improve preventative therapies, surveillance regimes, and tailor patient treatment to an individual's genetic make-up. However, as with any technological advance, there are associated fresh challenges. While some of the ethical, legal and social aspects resulting from the generation of data from genome sequencing are generic, several nuances are unique. Since the UK government recently announced plans to sequence the genomes of 100,000 Health Service patients, and similar initiatives are being considered elsewhere, a discussion of these nuances is timely and needs to go hand in hand with formulation of guidelines and public engagement activities around implementation of sequencing in clinical practice.

Reprinted from *Genes*. Cite as: Lucassen, A.; Houlston, R.S. The Challenges of Genome Analysis in the Health Care Setting. *Genes* **2014**, *5*, 576-585.

1. Introduction

The speed by which a person's genome can be analysed has increased phenomenally over recent years, while the attendant costs have plummeted. As a result, genetic testing is shifting from a targeted approach analysing specific genes based on particular symptoms or family histories to sequencing of an entire exome or genome (whole exome sequencing [WES], whole genome sequencing [WGS]). Targeted approaches characteristically have a high yield for penetrant monogenic conditions; whole genome approaches have the potential to unravel a much larger proportion of genetic disease burden. Whole genome analyses, therefore, are likely not only to transform a clinician's diagnostic and predictive ability, but also to improve preventative therapies, surveillance regimes, and tailor patient treatment to an individual's genetic make-up.

The improved diagnostic yields of genome sequencing are to be welcomed; however as with any technological advance there are associated fresh challenges. While the ethical, legal and social aspects resulting from the generation of data from genome sequencing are not unique, several nuances merit serious consideration. Since the UK government recently announced plans to sequence the genomes of 100,000 Health Service patients [1], and similar initiatives are being considered elsewhere, a discussion of these nuances is timely and needs to go hand in hand with formulation of guidelines and public engagement activities around the implementation of genome sequencing. Box 1 lists some of the ethical, legal, and social and practical issues that we consider merit consideration.

Box 1. Some of the overlapping ethical, legal, social and practical issues that need to be addressed as genome analysis enters clinical practice.

Complexity

- Genome analysis can provide many different predictions about diagnoses, or susceptibilities to conditions. However, it will do so with varying degrees of certainty or confidence intervals around the predictions. Such predictions are likely to change substantially over time as evidence about epistatic factors accumulates.
- Providing consent to genomic testing is therefore complex. Should consent be sought to any answer that genome analysis might provide? Or should there be cut-off for levels of certainty? Or should a genome analysis be used solely to answering a current clinical question? Should some results be staged? (e.g., risk of adult onset conditions diagnosed in children?)

Familial Aspects

- Although genomic information is on the one hand very personal, on the other, it may be relevant to relatives who have not sought medical advice but may be identified as being at risk from the results in another person. How can health services best record, store and communicate such familial information?

Re-Contacting/Follow Up Policies

- Who should be re-contacted and when, in the light of evolving knowledge? Who might be liable if a patient remains unaware of new evidence and therefore interpretation of previous test results?

Data Management

- What should be stored: the DNA sample, the DNA sequence, the interpretation of the sequence? Or combinations of these? What is to be stored in medical record systems, and how can these be compliant with relevant data protection—and other—legislation? How can/should these be linked with biobank or research databases, and how can the security issues around identifiable data best be managed?

Research/Clinical Divide

- The traditional route of research to clinic evolution is not necessarily applicable in rapidly evolving technologies.

Public Perceptions of Genetics

- Currently, this is often thought as a clear cut, or deterministic result than there is evidence for.
- Analytical validity not the same as clinical validity or utility; $1000 genome analysis is a reality soon, yet the cost of interpretation is much greater.

2. The Promise of Whole Genome Analysis

Genetic testing has traditionally been restricted to analysing small numbers of genes usually picked on the basis of a high prior probability of being mutated. However, this approach has several limitations. Firstly, many inherited diseases are genetically heterogeneous and sequential mutational analysis of individual genes is slow and expensive. Secondly, while subsets of some common diseases can be caused by mutations in a single gene, traditional methods of selecting whom to test on basis of disease characteristics or family history are crude and have a high false negative rate. Finally, analysis may not ultimately be diagnostic if the disease is a consequence of a hitherto

unknown disease-causing gene. Collectively these issues make whole genome approaches at competitive prices an attractive proposition.

Most next-generation sequencing (NGS) technologies are based on the fragmentation of genomic DNA with the oversampling of reads providing the necessary linking information for whole-genome assembly algorithms. For analysis of a gene to be of diagnostic quality using NGS there needs to be sufficient read depth for any mutation to be called with a high degree of confidence. While WGS or WES typically provide good overall coverage for most regions of the genome, for other regions it may be poor; sequencing some regions of the genome is problematic because of repetitive sequence and other features leading to systematic error [2]. Such limitations have, in part, been the motivation for developing targeted sequencing approaches focusing on panels of genes relevant to specific diseases states; for example, cancer gene and nervous system disease panels. Such technical shortcomings are likely to be addressed in the near future so that a "one-stop-shop" test will replace the sequential approaches to genetic diagnoses which were time and labour intensive.

3. Analytical Validity *versus* Interpretation of WGS Approaches

Whilst the analytical validity of WGS approaches is high, and improving at a rapid pace, the clinical validity of the output from WGS is much more complex than commonly perceived and the utility has often been evaluated only in very small groups. There is much genome variation that is either: uninterpretable; probably benign; or only pathogenic in certain circumstances, for example, in the presence of as yet unknown epistatic factors. This gap between technological advances and the interpretation of any NGS output, is neatly encapsulated by the phrase "$1000 genome; $1 million interpretation [3], yet, little recognised in the popular discourse around whole genome technologies.

In the clinical setting, certainly in the short term, diagnostic accuracy will therefore continue to depend on additional factors such as clinical history and, therefore, pre-test probability. Attempts to overcome these issues include use of gene panels or analyses of selected portions of the genetic code; an apparently anachronistic step in the evolution of whole genome approaches. However, if WGS approaches are to be used to answer clinical questions, some sort of filtering of sequence output will need to take place. Although targeted approaches are commonplace in health care, this has usually involved a targeting of the investigations. In WGS the targeting will have to be at the analysis stage—the results require targeted analysis—and this raises novel issues about what constitutes a result, what is disclosed to the patient, and what is recorded in a patient's medical records.

4. The Data Interpretation Problem

Much of the misperception about the diagnostic value of genome sequencing results from an oversimplification in which it is assumed there is "a gene" for the condition, when in fact any increase in risk conferred by a mutation may be subtle, or only manifest in the context of specific genetic background or environmental exposure. For many common diseases there are multiple risk

factors and while the identification of susceptibility genes has often provided novel insights in disease biology, their clinical utility in an individual may be very low because their predictive power in isolation is very poor.

There is, however, also a risk of over-interpretation even for mutations with seemingly large effects. For affected patients where there is a strong prior probability of the gene mutation being causal because of a positive family history and or specific clinical phenotype, interpretation can be straightforward. However, if mutations are not fully penetrant, there will be carriers in the population who are healthy. Much of our knowledge about the penetrance of mutations to date is based on family data and, hence, suffers from ascertainment bias [4]. Without unbiased knowledge of the effect of mutations, interpretation at the population level will be inherently problematic. Whilst policies to restrict genetic testing to high risk populations were initially driven by budget restraints, and the more widespread availability of testing thought to be an advantage of declining costs, another consequence is that the interpretation of the clinical significance of a mutation is much more difficult if found without the ascertainment bias noted above. That is to say, predicting the effects of a novel *BRCA2* mutation in the context of a strong family history of the mutation segregating with disease in the family, is far easier than when it is discovered in a population screen (see Box 2 for an illustrative example).

Box 2. Difficulty of clinical interpretation of genomic findings in absence of clear clinical phenotype or family history.

A two-year old boy was investigated for "absence spells". He had no loss of consciousness, was investigated in detail for epilepsy and no abnormalities were found. Paediatric cardiologists also found no abnormalities, his baseline ECG was defined as within normal limits and he had no family history (to 3rd degree relatives) of any cardiac problems. The cardiologist had been to a presentation about mainstreaming genetics and realised that long QT (LQT) interval gene carriers can be difficult to detect in childhood. He therefore requested genetic testing of LQT genes "to exclude LQT syndrome". A LQT1—associated mutation was identified, described on the laboratory report as "highly likely to be pathogenic". A reveal device was inserted but no abnormalities in his QT interval were recorded during subsequence "absence spells". Nevertheless, it was thought appropriate to treat him with beta blockers. Cascade testing of his family revealed his three-year-old sister, father, paternal aunt (and her two children, aged four and eight) and paternal grandfather all carried the same mutation. Cardiac investigations of their phenotype, at rest, with exercise, and pharmacological challenge were normal or equivocal. All carriers in the family were prescribed beta blockade and two members of the family were referred for possible implantable cardiac defibrillator insertion.

In Box 2 the assumption that this LQT1 mutation depicts a high future risk of clinical symptoms from LQT syndrome is based on the laboratory description of its likely pathogenicity and the previous finding in families with symptomatic LQT. The intensive therapy is in part because the first presentation of LQT can be sudden cardiac death. However, this family was not ascertained on the basis of any relevant clinical symptoms and clear clinical predictions for the seven asymptomatic

carriers are extremely difficult. However, if the mutation was found in a family with a segregating LQT phenotype, preventative therapy would be justifiable on clinical grounds. These cases serve to illustrate that the predictive powers of genetics require more than information about genotype, for the effects of any genotype are dependent on a range of other factors. Importantly, the penetrance of different mutations in the same gene can vary substantially and assigning a likelihood of a mutation being disease-causing will increasingly be based on the synthesis of multiple forms of evidence.

5. Determining Clinical Utility of Sequence Variants

The translation of genome sequence into medically actionable information is a key challenge. Without support from segregation in families, assigning pathogenicity can be problematic; notably large duplications, most synonymous and some missense mutations, intronic variants, and most variants in promoter and enhancers are particularly difficult to interpret. Predicting the functional consequences of variants which disrupt protein-coding sequence can also be challenging. A variant might affect a transcription factor binding site, a microRNA target site, affect RNA-splicing or stability or truncate a protein. Finally the issue of linkage disequilibrium (where benign variants lie close to a disease predisposing variant) can complicate interpretation of recurrent risk variants.

Irrespective of whether animal models can adequately mimic human disease such model systems are inherently unsuited to determining the consequences of specific mutations as a routine activity. While yeast and cell line systems can be used to assess the functionality of DNA repair gene mutations the general applicability of such model systems is limited. In view of these factors increasing reliance will be placed on the implementation of *in silico* tools to infer the functional consequences of mutations. Although such algorithms can help to predict the likely pathogenicity of variants, often different tools conclude in opposite directions and without an established relationship between gene dysfunction and disease phenotype, robust risk prediction is problematic.

6. The Need to Systematically Catalogue Sequence Variation with Phenotype

Several initiatives are cataloguing and assigning pathogenicity to variants/mutations in various specific genes. Examples of such databases include InSiGHT (International Society for Gastrointestinal Hereditary Tumours Incorporated) [5], LoVd (Leiden open variant database) [6] Decipher [7] and DMuDB [8] (the diagnostic mutation database), and the Locus Reference Genomic Collaboration [9]. These resources provide health care professionals with valuable information for decision making processes. While published reports are valuable sources for such databases their stewardship depends heavily on the submission of individual variants and associated clinic-pathological data by sequencing laboratories using some form of incentivization. Currently these databases are limited to curation of restricted number of genes. Even here translating genomic sequence into medically actionable information can be highly time consuming.

To meet the future needs, comprehensive resources with a far more overarching remit will need to be developed and maintained. This needs to be coupled with adoption of automated machine learning, support vector machines and other technologies to create systematic and efficient

mechanisms to assess the impact of variants found by genomic sequencing. All of this will require substantial investment before it becomes a reality and has not been factored into the $1000 genome analysis headlines.

7. Diagnostics *versus* Population Screening

Given the significant limitations to our current understanding of the impact of genetic variation, we believe that clinical genome sequencing should for now be focused on particular clinical presentations compatible with a genetic aetiology, rather than engaging in opportunistic population screening. For example, the identification of an *APC* mutation in a person with colonic polyposis is diagnostic and highly predictive for family members. In contrast the identification of variants, such as *LQT1* described in Box 1, in a population screen do not have sufficient certainty to infer as much, resulting in difficult clinical management issues. Such contextual differences may be difficult to grasp if genetics is portrayed as being clear cut, and clinical interventions may therefore be offered without sufficient evidence for their benefit.

Intelligent interrogation of genomic outputs in the clinic should initially therefore be restricted to specific genes or diseases for which there is a high prior likelihood of diagnosis. Any opportunistic screening should in the first instance be limited to known epistatic factors for particular conditions, e.g., low risk genes for breast cancer in the investigation of a family history of breast cancer, and formal evaluation of the benefits should not be leap frogged just because of the rapidly decreasing costs of the technologies involved.

8. The Need for Large Scale Genotype-Phenotype Linkages

Before more widespread population genome screening is to be contemplated, large-scale systematic and longitudinal investigation of variants in categorised populations would need to take place and their penetrance robustly determined. Depending on variant prevalence the ongoing international biobank sequencing projects are likely to provide a rich source of such data. Additionally, variants identified through clinical testing or research projects, could together with associated phenotypic information, be submitted to publicly accessible databases cataloguing genomic variation. Many of the current databases are however relatively ad hoc affairs and disease-specific. If the full potential of genomics is to be realised there is a need for the development of big data centres which have an overarching remit. However, the development and establishment of such initiatives brings with it the significant issue of data-storage and allied security requirements. These linkages will have to be undertaken within legislative frameworks relating to data protection within host countries and adapted to any changes to such legislation. For example, proposed changes by the European Commission to the data protection directive may have far reaching consequences for the gathering of such linkages [10].

9. The Need for Public-Professional Engagement

In parallel with the acquisition and curation of genetic data there needs to be an ongoing dialogue with health care professionals and the public around understandings and interpretations of

genomic data so that expectations of new WGS approaches are realistic and grounded in evidence. In the wake of public anxiety around large scale databases, e.g., care.data in UK [11], this dialogue urgently needs to incorporate the importance of data sharing to realise the clinical utility of whole genome approaches. It also needs to incorporate the issues around shifting the point of targeting, as outlined in Section 3. For example, international recommendations suggest that children should not be offered genetic testing for adult-onset conditions (unless a result would alter their medical management). However, once such a result is available many would opine it should be disclosed, even if they would not have tested for it in the first place [12,13].

10. Incidental Findings

Any broad, highly sensitive investigation has the ability to occasionally detect abnormalities that are incidental to the reason for the test. Whole genome approaches are much more likely to detect asymptomatic or silent abnormalities that have nothing to do with the current clinical reason for a test. Such findings have been variably termed "secondary", "non-pertinent", "unexpected" or "incidental" belying the fact that the appropriate adjective may vary according to the situation [14]. A genome test can, however, only have an incidental finding (IF) if it is used to answer, for example, a particular clinical question. If the question is "what are the abnormalities in this genome?" then there can be no IFs.

There has been much recent debate about the management of IFs in clinical applications of WG technologies [15–20]. The American College of Medical Genetics and Genomics (ACMG) produced guidelines that recommended the active search for particular IFs if using WGS/WES approaches [21]. The heated debate that ensued was largely focused on patient/parental choice regarding such IF searches with their purported "right not to know" being exercised by such guidance. A subsequent amendment now argues for decision about IF search to be made at the time of testing, but still recommends search for additional mutations not indicated by the clinical symptoms. The European Society of Human Genetics (ESHG) responded that WG approaches should be targeted to the clinical question, but there is still widespread debate about the management of IFs in practice and whether real up-front patient choice is feasible or preferable.

11. Familial Consequences of IFs

A family history of a particular disease usually means that unaffected relatives have some idea they too might be at risk. In contrast, if something is found incidentally there is unlikely to be awareness of the suspected condition. Furthermore, a new variant may only be found to be clinically significant once it has been studied alongside phenotypes in a family and the absence of a family history is likely to make the need for such cascade screening more difficult to comprehend. Furthermore, professionals may be uncertain what, if any, duties they have to alert relatives about risks that may only be clarified after cascade screening.

12. Return of Results from Genomic Testing

As the pace and scale of genetic testing increases, it is inevitable there will be less time to prepare individuals for potential test results. Since the implications of some variants, particularly IFs, may fall outside the expertise of the professional who requested the test, referral to another health care professional may be necessary. This process is likely to add to anxiety of families burdened with unexpected genetic information and means that consent and disclosure practices become dissociated. Training in genomic medicine should be expanded to all medical specialties so that the complexities of genomic information can be adequately communicated but we do not underestimate the size of this task in a rapidly changing environment. We suggest that clinical genetic professionals, although relatively few in number, will need to take on greater liaison activities to facilitate this training.

Opinions about disclosure of IFs vary, ranging from full disclosure to disclosing only those with established clinical significance, and/or which have an intervention can impact on disease. In reality clinically significant, because further investigations of the patient, and their relatives, may be required it can be extremely difficult to withhold details of IFs, even if a conclusion is that they are not to arrive at this conclusion. Even if the pathogenicity of an IF is established, disease onset may not be for many years. Hence robust mechanisms are required to identify, re-contact, and review family members when health care interventions become appropriate. Current health-care systems are, however ill-equipped to deal with the recording of familial information, future risks to health, or the monitoring of multiple family members. We consider that genome results need to be considered as a resource that can be accessed over time [22], rather than as one result that needs to be disclosed as one at the point of testing.

13. Consent for Genome Testing

Providing individuals with sufficient information in order to make decisions about investigations or interventions is a key element of good clinical practice. Achieving a balance between providing sufficient information but avoiding overload can be a challenge, especially for tests where multiple different outcomes are possible. Individuals need to understand what genome tests can reveal, but also that some degree of uncertainty is likely. The possible need to investigate relatives to assign pathogenicity of variants found is a difficult issue to incorporate into any consent process. Obtaining adequate consent to disclose an IF for which there is no prior suspicion on the basis of family history or symptoms is likewise problematic, especially if such an IF is unlikely to have clinical consequences for some time. All of this is set against a background of media coverage that generally portrays genetics as clear-cut and highly determinative.

14. Is Personalised Medicine a Helpful Term to Promote Genomics?

Although genomic analyses will help to stratify individuals into subpopulations with common characteristics so that particular variants might have greater predictive value, this is not the same as individualisation. A concern about describing genomics as leading to personalised medicine is that it may encourage views of genetic determinism or reductionism. There has been much professional

and public discussion regarding which parts of a whole genome sequence should be communicated, with emotive discussions about rights to personal information. On the one hand there is a public perception that some form of medical paternalism might be exhibited where useful information would be withheld, on the other there is acceptance that most of the three billion letter output of a genome sequence has no personal clinical relevance [23]. Some advocate that anyone sequenced should have the right to be appraised of "all of" the test results, even if the clinical relevance is indeterminate. Whilst full disclosure is thought to respect a person's autonomy it may do the opposite if it delivers outputs that are uninterpretable.

15. Conclusions

Whilst the technology of genome sequencing is now a sufficiently mature and affordable technology for it to be implemented clinically, significant challenges around interpretation and implementation remain. We believe that clinical genome analyses should be directed to delivering diagnoses for patients and that integration or linkage with biobanks and other research ventures will be crucial for better clinical translation in the future. We do not underestimate the practical challenges such a statement results in but hope that by delineating some of the complexities and aligning them next to common perceptions of genetics will lead to intelligent international debate about consent and disclosure practices, long-term follow-up arrangements, appropriate communication with relatives and linkages between clinical practice and research.

Author Contributions

This paper was jointly conceived and written by the authors.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Government. HMs. Strategy for UK Life Sciences: One Year on. Available online: http:// www.bisgovuk/assests/biscore/inovation/docs/s/12-1346-stragy-for-uk-life-sciences-one-year-onpdf/ (accessed on 4 July 2014).
2. Meacham, F.; Boffelli, D.; Dhahbi, J.; Martin, D.I.; Singer, M.; Pachter, L. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinform*. **2011**, *12*, 451.
3. Davies, K. The $1,000,000 genome interpretation. *Bio-IT World* **2010**, *8*, 50–54.
4. Zollner, S.; Pritchard, J.K. Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *Am. J. Hum. Genet*. **2007**, *80*, 605–615.
5. Ou, J.; Niessen, R.C.; Vonk, J.; Westers, H.; Hofstra, R.M.; Sijmons, R.H. A database to support the interpretation of human mismatch repair gene variants. *Hum. Mutat*. **2008**, *29*, 1337–1341.
6. Fokkema, I.F.; Taschner, P.E.; Schaafsma, G.C.; Celli, J.; Laros, J.F.; den Dunnen, J.T. LOVD v.2.0: The next generation in gene variant databases. *Hum. Mutat*. **2011**, *32*, 557–563.

7. Decipher Database. Available online: https://decipher.sanger.ac.uk/ (accessed on 4 July 2014).
8. Diagnostic Mutation Database. Available online: https://secure.dmudb.net/ngrl-rep/Home.do/ (accessed 4 July 2014).
9. Dalgleish, R.; Flicek, P.; Cunningham, F.; Astashyn, A.; Tully, R.E.; Proctor, G.; Chen, Y.; McLaren, W.M.; Larsson, P.; Vaughan, B.W.; *et al*. Locus Reference Genomic sequences: An improved basis for describing human DNA variants. *Genome Med*. **2010**, *2*, 24.
10. Act EDP. EU Data Protection Act Proposed Changes & Reforms. Updates to the Current EU Data Protection Acts Look Set to Radically Change the Way We Deal with Data. Available online: http://www.eudataprotectionlaw.com/ (accessed on 4 July 2014).
11. Handling of NHS patient data. Available online: http://www.parliament.uk/business/committees/committees-a-z/commons-select/health-committee/inquiries/parliament-2010/cdd-2014/ (accessed on 4 July 2014).
12. Clayton, E.W.; McCullough, L.B.; Biesecker, L.G.; Joffe, S.; Ross, L.F.; Wolf, S.M.; Clinical Sequencing Exploratory Research (CSER) Consortium Pediatrics Working Group. Addressing the ethical challenges in genetic testing and sequencing of children. *Am. J. Bioeth*. **2014**, *14*, 3–9.
13. Lucassen, A.; Widdershoven, G.; Metselaar, S.; Fenwick, A.; Parker, M. Genetic testing of children: The need for a family perspective. *Am. J. Bioeth*. **2014**, *14*, 26–28.
14. Crawford, G.; Fenwick, A.; Lucassen, A. A more fitting term in the incidental findings debate: One term does not fit all situations. *Eur. J. Hum. Genet*. **2013**, doi:10.1038/ejhg.2013.266.
15. Wolf, S.M. The past, present, and future of the debate over return of research results and incidental findings. *Genet. Med*. **2012**, *14*, 355–357.
16. Wolf, S.M.; Lawrenz, F.P.; Nelson, C.A.; Kahn, J.P.; Cho, M.K.; Clayton, E.W.; Fletcher, J.G.; Georgieff, M.K.; Hammerschmidt, D.; Hudson, K.; *et al*. Managing incidental findings in human subjects research: Analysis and recommendations. *J. Law Med. Ethics* **2008**, *36*, 219–248.
17. Bredenoord, A.L.; Kroes, H.Y.; Cuppen, E.; Parker, M.; van Delden, J.J. Disclosure of individual genetic data to research participants: The debate reconsidered. *Trends Genet*. **2011**, *27*, 41–47.
18. Gliwa, C.; Berkman, B.E. Do researchers have an obligation to actively look for genetic incidental findings? *Am. J. Bioeth*. **2013**, *13*, 32–42.
19. Holtzman, N.A. ACMG recommendations on incidental findings are flawed scientifically and ethically. *Genet. Med*. **2013**, *15*, 750–751.
20. Crawford, G.; Foulds, N.; Fenwick, A.; Hallowell, N.; Lucassen, A. Genetic medicine and incidental findings: It is more complicated than deciding whether to disclose or not. *Genet Med*. **2013**, *15*, 896–899.
21. Green, R.C.; Berg, J.S.; Grody, W.W.; Kalia, S.S.; Korf, B.R.; Martin, C.L.; McGuire, A.L.; Nussbaum, R.L.; O'Daniel, J.M.; Ormond, K.E.; *et al*. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med*. **2013**, *15*, 565–574.
22. Biesecker, L.G. Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: Lessons from the ClinSeq project. *Genet. Med*. **2012**, *14*, 393–398.
23. Genomics England. Genomics England Town Hall Event. Available online: http://www.genomicsengland.co.uk/town-hall-engagement-event/ (accessed on 4 July 2014).

MDPI AG
Klybeckstrasse 64
4057 Basel, Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
http://www.mdpi.com/

Genes Editorial Office
E-mail: genes@mdpi.com
http://www.mdpi.com/journal/genes

MDPI • Basel • Beijing • Wuhan
ISBN 978-3-03842-171-9
www.mdpi.com