Targeted capture and massively parallel sequencing in pediatric cardiomyopathy: development of novel diagnostics

Muhammad Tariq,[1] Thanh-Tam Le,[1] Patrick Putnam,[2] Steven Kindel,[1] Mehdi Keddache,[2] Stephanie M. Ware[1,2]

[1]Division of Molecular Cardiovascular Biology, [2]Division of Human Genetics, Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine, Cincinnati, OH, USA

Abstract

Pediatric cardiomyopathy is a genetically heterogeneous disease associated with significant morbidity. Although identification of underlying etiology is important for management, therapy, and screening of at risk family members, molecular diagnosis is difficult due to the large number of causative genes, the high rate of private mutations, and cost. In this study, we aimed to define the genetic basis of pediatric cardiomyopathy and test a novel diagnostic tool using a custom targeted microarray coupled to massively parallel sequencing. Three patients with cardiomyopathy were screened using a custom NimbleGen sequence capture array containing 110 genes and providing 99.9% coverage of the exons of interest. The sensitivity and specificity was over 99% as determined by comparison to long-range polymerase chain reaction (PCR)-based massively parallel sequencing, Sanger sequencing of missense variants, and single nucleotide polymorphisms genotyping using the Illumina Infinium Omni1 array. Overall, 99.73% of the targeted regions were captured and sequenced at over 10x coverage, allowing reliable mutation calling in all patients. Analysis identified a total of 165 non-synonymous coding single nucleotide polymorphisms (cSNPs) of which 89 were unique and 14 were novel. On average, each patient had 4 cSNPs predicted to be pathogenic. In conclusion, we report a cardiomyopathy sequencing array that allows simultaneous assessment of 110 genes. Comparison of targeted sequence capture *versus* PCR-based enrichment methods demonstrates that the former is more sensitive and efficient. Array-based sequence capture technology followed by massively parallel sequencing is promising as a robust and comprehensive tool for genetic screening of cardiomyopathy. These results provide important information about genetic architecture and indicate that improved annotation of variants and interpretation of clinical significance, particularly in cases with multiple rare variants, are important for clinical practice.

Introduction

Pediatric cardiomyopathies are clinically heterogeneous disorders of heart muscle that are responsible for significant morbidity and mortality. The primary cause of the majority of cases of pediatric cardiomyopathy is genetic. To date, more than 100 genes have been implicated in cardiomyopathy, but comprehensive genetic diagnosis has been problematic because of the large number of genes and private nature of mutations.

In the last two decades, polymerase chain reaction (PCR)-based enrichment of genomic DNA has been the dominant technique for targeted enrichment of specific genomic regions. Massively parallel (Next-Generation) sequencing has recently emerged as a powerful tool for sequencing utilizing PCR-based genomic enrichment. However, the high throughput benefits of massively parallel sequencing remain limited by the relatively labor intensive nature of PCR-based enrichment. As a result, novel non-PCR based enrichment procedures have emerged as promising alternatives.[1] Microarray-based sequence capture hybridization (SeqCap) uses a chip arrayed with specific probes and allows enrichment of multiple dispersed segments or long single genome segments.[2-4] Recently, this approach has been shown to be successful in identifying novel genes involved in inherited human monogenic (recessive or dominant) disorders[5-8] or polymorphic genetic variants associated with complex human diseases.[9,10] While these approaches have been proven to be useful in research-based testing, it is not clear whether these platforms are suitable for the higher stringency required for clinical molecular testing. In the present study, we applied a customized array-based SeqCap enrichment strategy coupled to massively parallel sequencing and successfully screened 110 genes in cardiomyopathy patients in a single sequencing run each and identified mutations associated with disease. We have evaluated the performance and quality of long-range PCR (LR-PCR) *versus* SeqCap enrichment methods using cardiomyopathy as a disease model. The results indicate that this is a promising platform for clinical diagnostics and provide important information about the complexity of the genetic architecture in this disease.

Materials and Methods

Patients and DNA samples

Three cardiomyopathy patients, CM0001, CM0002 (Pedigree A, Figure 1) and CM0006 (Pedigree B, Figure 1) were studied. The patients underwent clinical evaluation by a pediatric geneticist and pediatric cardiologist at Cincinnati Children's Hospital Medical Center (CCHMC). Written informed consent was obtained from the parents or study participants prior to initiation. This study design was approved by the CCHMC Institutional Review Board. Genomic DNA was extracted using a Qiagen purification kit (Puregene Blood Case kit C, QIAGEN Sciences, Maryland, USA). DNA quality and quantity were calculated with a NanoDrop spectrophotometer (Thermo-Scientific).

Selection of candidate genes

One hundred and ten nuclear genes were selected for analysis with microarray-based targeted SeqCap. Of these, a smaller subset of 31 genes was chosen for PCR-based enrichment (Table 1). Genes available on commercially available panels for cardiomyopathy at

Correspondence: Stephanie Ware, Cincinnati Children's Hospital Medical Center, 240 Albert Sabin Way, MLC 7020, Cincinnati, OH 45229, USA.
Tel. +513.636.9427 - Fax: +513.636-5958.
E-mail: stephanie.ware@cchmc.org

Key words: sequence capture, massively parallel sequencing, restrictive cardiomyopathy, desmin, missense mutation.

Acknowledgments: we thank the patients for their participation. We acknowledge the technical contribution of Anup Tilak. This study was supported by funding from the Children's Cardiomyopathy Foundation and Cincinnati Children's Hospital's Clinical and Translational Science Award NIH-ULlRR026314.

Contributions: MT and TL performed experiments; PP and MK performed sequence alignment of massively parallel sequencing data to RefSeq; PP and SMW performed statistical analysis; MT performed bioinformatic and mutational analysis, and Sanger validation; SK and SMW performed clinical diagnosis; MT and SMW wrote manuscript; SMW designed the project and received funding. All authors read and approved the final manuscript.

Conflict of interests: the authors have no conflict of interests.

Received for publication: 1 October 2011.
Revision received: 15 April 2012.
Accepted for publication: 17 April 2012.

Cardiogenetics 2012; 2:e7
doi:10.4081/cardiogenetics.2012.e7

Table 1. Cardiomyopathy custom array gene content.

Genes targeted by capture array enrichment	Total coding exons	Encoded protien (AA)	Targeted sequence (bp) on capture custom array	NCBI GenBank accession #	Chromosomal location
Sarcomere					
*MYH7**	38	1935	9419	NG_007884	14q11.2
*MYBPC3**	33	1274	7035	NG_007667	11p11.2
*TNNT2**	15	295	3517	NG_007556	1q32.1
*TPM1**	9	284	2285	NG_007557	15q22.2
*MYL3**	6	195	1093	NG_007555	3q21.31
*MYL2**	7	166	1512	NG_007554	12q24.11
*ACTC1**	6	377	1609	NG_007553	15q14
*TNNI3**	6	210	1785	NG_007866	19q13.4
*MYH6**	37	1939	8949	NC_000014	14q11.2
TNNC1	6	161	1455	NG_008963	3p21.1
Fatty acid oxidation					
ACADVL	19	633	1967	NG_007975	13p13.1
HADHA (α)	20	763	4680	NG_007121	2p23.3
HADHA (β)	15	474	3265	NG_007294	2p23.3
ACADM	12	421	3047	NG_007045	1p31.1
Carnitine transport					
*SLC22A5/OCTN2**	10	557	2761	NG_008982	5q31.1
CPT2	5	658	2331	NG_008035	1p32.3
CPT1A	18	773	4536	NG_011801	11q13.3
SLC25A20	9	301	2266	NG_008171	3p21.31
Storage disorder					
GAA	19	952	4613	NG_009822	17q25.3
GANC	24	914	4623	NM_198141.2	15q15.1
*LAMP-2**	9	411	2276	NG_007995	Xq24
*PRKAG2**	16	569	4080	NG_007486	7q36.1
GBE1	16	702	4022	NG_011810	3p12.2
GLA	7	429	1911	NG_007119	Xq22.1
PHKA2	33	1235	8124	NG_016622	Xp12.13
AGL	33	1532	7690	NG_012865	1p21.2
NAGA	9	411	2250	NG_009247	22q13.2
PYGM	20	842	4345	NG_013018	11q13.1
Metabolic					
HMGCL	9	325	2023	NG_013061	1p36.11
PMM2	8	246	2011	NG_009209	16p13.2
D2HGD	9	521	2361	NG_012012	2q37.3
MLYCD	5	493	1852	NG_009079	16q23.3
Desmosome					
JUP	9	563	3160	NG_009090	17q21.2
DSP	24	2871	11112	NG_008803	6p24.3
PKP2	14	881	3826	NG_009000	12p11.21
DSG2	15	1118	4617	NG_007072	18q12.1
DSC2	16	901	4068	NG_008208	18q12.1
Cytoskeleton, Z-disc etc					
ACTN2	21	894	5264	NG_009081	1q43
*DES**	9	470	2279	NG_008043	2q35
MYOZ2	5	264	1260	NM_016599.3	4q26
*LDB3**	13	732	3619	NG_008876	10q23.2
*CSRP3**	5	194	1265	NG_011932	11p15.1
*TCAP**	2	167	664	NG_008892	17q12
SGCB	6	318	1509	NG_008891	4q12
SGCD	8	290	2019	NG_008693	5q33.3
TTN	311	33423	123375	NG_011618	2q31.2
DMD	79	3385	9800	NG_012232.1	xp21
*MYPN**	19	1320	5933	NM_032578.2	10q21.3
*PLN**	1	52	250	NG_009082	6q22.31
VCL(meta)	22	1134	5524	NG_008868	10q22.2
*CRYAB**	3	175	758	NG_009824	11q23.1
CAV3	2	151	612	NG_008797	3p25.3
DTNA	21	743	5250	NG_009201	18q12.1
SNTA1	8	505	1989	NG_011622	20q11.21

To be continued on next page

Continued from previous page.

Genes targeted by capture array enrichment	Total coding exons	Encoded protien (AA)	Targeted sequence (bp) on capture custom array	NCBI GenBank accession #	Chromosomal location
Syndromic					
*TAZ**	11	292	1896	NG_009634	Xq28
ALMS1	23	4169	4566	NG_011690	2p13.1
PTPN11	15	593	3654	NG_007459	12q24.13
HRAS	4	189	919	NG_007666	11p15.5
KRAS	4	189	1001	NG_007524	12p12.1
SOS1	23	1333	6332	NG_007530	2p22.1
BRAF	18	766	4563	NG_007873	7q34
MEK1(MAP2K1)	11	393	2762	NG_008305	15q22.31
MEK2 or MAP2K2	11	400	2548	NG_007996	19p13.3
*RAF1**	16	648	3886	NG_007467	3p25.2
Mitochondrial					
*SCO2**	1	266	821	NG_016235	22q13.33
*SURF1**	9	300	1413	NG_008477	9q34.2
COX10	7	443	2026	NG_008034	17p12
*COX17**	2	63	502	NM_005694	3q13.33
COX15	9	410	2379	NG_008986	10q24.2
*TK2**	10	307	2508	NG_016862	16q21
*DGK (DGUOK)**	7	277	1764	NG_008044	2p13.1
FXN	5	210	1258	NG_008845	9q21.11
*ANT1 (SLC25A4)**	4	298	1262	NG_013001	4q35.1
SLC25A3	7	362	1638	NG_011702	12q23.1
*SOD2**	5	222	1277	NG_008729	6q25.3
*TSFM**	7	346	1929	NG_016971	12q14.1
NDUFV2	8	249	1750	NG_013355	18p11.22
*NDUFA11**	4	228	1000	NM_001193375	19p13.3
DNAJC19	6	116	1252	NG_022933	3q26.33
*NDUFS8**	6	210	992	NG_017040	11q13.2
NDUFS2	14	463	3131	NG_013352	1q23.3
NDUFS6	4	124	1000	NG_013354	5p15.33
C6ORF66	3	175	808	NG_013379	6q16.1
Miscellaneous					
EYA4	19	639	4545	NG_011596	6q23.2
*LMNA**	12	664	3218	NG_008692	1q22
EMD	6	254	1270	NG_008677	Xq28
TMPO	4	694	2190	NG_021393	12q23.1
PSEN1	10	467	2527	NG_007386	14q24.2
PSEN2	10	448	2507	NG_007381	1q42.13
RYR2	105	4967	28213	NG_008799	1q43
ABCC9	38	1549	9432	NG_012819.	12p12.1
SCN5A	27	2015	8377	NG_008934	3p22.2
TGFB3	7	412	1872	NG_011715	14q24.3
FKTN	9	461	2278	NG_008754	9q31.2
MYO6	34	1285	8320	NG_009934	6q14.1
TMEM43	12	400	3000	NG_008975	3p25.1
CMYA1/XIRP1	1	1843	3627	NM_194293	3p22.2
PTPLA	7	288	1535	NM_014241	10p12.33
FK506 bp12	4	108	755	NM_000801	20p13
SCRIB	30	1655	13492	NM_015356	8q24.3
GSK3B	12	433	3000	NG_012922	3q13.33
WNT5b	4	359	1310	NM_030775	12p13.33
WNT5a	5	380	1562	NM_003392	3p14.3
WNT11	5	354	1367	NM_004626	11q13.5
SFRP1	3	314	1157	NM_003012	8p11.21
SFRP3	6	325	1748	NG_017197	2q32.1
MAPK1 (ERK2)	8	360	2072	NC_000022	22q11.2
SERCA2	20	1042	5169	NG_007097	12q24.11
NCX1(SLC8A1)	7	937	4021	NC_000002	2p22.1
CABIN1	35	2141	10093	NC_000022	22q11.23
Total	1691	109802	503340		

*Genes also analyzed by long-range polymerase chain reaction followed by massively parallel sequencing.

the time of study initiation were included. The 110 genes and their encoded proteins are involved in sarcomere, cytoskeletal, or desmosome organization, fatty acid oxidation, carnitine transport, storage disorders, syndromic phenotypes, and mitochondrial organization. These genes have previously been reported as disease-causing in distinct human cardiomyopathy phenotypes.

Long-range PCR and DNA pooling

For PCR-based enrichment, genes of interest were divided into 84 long-range amplicons to provide coverage to all coding regions and splice-junctions. The custom design aimed to minimize the number of amplicons and generate a majority of amplicons ranging from 3 kb to 10 kb. As a result, the majority of introns less than 4 kb were covered in their entirety. Primer sets were designed using Primer3 program (sequences available upon request). PCR amplifications were performed in a 30 µL reaction mixture using 20 ng DNA, 20 pmoles of each primer, 400 µM dNTPs, and 3 units of LA Taq polymerase in 1x LA PCR buffer (Takara Inc., Japan). Thermocycling included an initial denaturation for 3 min at 94ºC followed by 34 cycles of denaturation at 94ºC for 30 s, annealing and extension at 68ºC for 3, 6 or 10 min depending on amplicon length, and a final extension at 72ºC for 8 min. PCR products were purified using QIAquick PCR purification kit (Cat # 28106, QIAGEN Sciences), quantified by Nanodrop and pooled in an equimolar fashion based on DNA concentration and amplicon size.

Cardiomyopathy capture array

Custom SeqCap 385K human arrays were designed to our specifications and manufactured by Roche NimbleGen (Madison, WI, USA) by printing 60-mer oligonucleotide probes complementary to the genomic sequences of the coding regions and splice sites of 110 candidate genes (NCBI build 36.1, hg18) onto glass slides. A total of 1505 target regions containing 384,000 probes were tiled onto arrays after exclusion of repetitive sequences. Supplementary Table 1 shows that 99.9% of targeted nucleotides are represented on the arrays.

Library preparation

Five genomic libraries (for SeqCap arrays) and 3 LR-PCR libraries (after PCR enrichment and pooling) were prepared for massively parallel sequencing in an identical fashion according to the manufacturer's recommendations (Illumina Inc. San Diego, CA, USA). Briefly, 5 µg of DNA (from genomic DNA or 84 pooled long-range PCR products) was fragmented by nebulization for 6 min using 38psi nitrogen gas. Fragments were concentrated and ligated to adaptors from the Illumina Paired-End Library construction kit. After separation by 2% agarose gel electrophoresis, the 300bp (±25bp) fraction was gel cut and purified using QIAquick gel purification kit (Cat # 28706, QIAGEN Sciences). Library construction was completed by enriching the size selected fragments through 18 cycles of PCR. At this stage, LR-PCR libraries are ready to sequence while SeqCap libraries will be enriched further after array hybridization and elution.

Enrichment of genomic libraries by SeqCap

Five massively parallel sequencing libraries from 3 patients (CM0001, CM0006 and CM0002, including 3 technical replicates) were generated from genomic DNA and hybridized to NimbleGen custom SeqCap microarrays for targeted enrichment. After elution from arrays, an additional enrichment PCR was performed for these libraries. The manufacturer's recommended protocol was modified as follows: i) 1 µg (~4 µmole) of library was used in the hybridization; and ii) supplemented with 300 µmole (75x excess) each of Illumina primers PE1.1 and PE2.1; iii) after NaOH based elution of captured library molecules, 18 cycles of PCR were performed.

Massively parallel sequencing

Massively parallel sequencing of the SeqCap array enriched genomic libraries or LR-PCR libraries (Table 2) was carried out on the Illumina Genome analyzer IIx according to the manufacturer's protocol as single end 35- and 50-bp reads or as paired end (PE) 72-bp reads. One lane of an Illumina flow cell was used for each SeqCap sample whereas 3 LR-PCR amplified samples were run in groups of 10 per lane (pooled samples with 10 different DNA barcodes including 7 other unrelated samples to this study). All sequence reads were mapped to the reference human genome (UCSC hg 18) using the Illumina Pipeline software version 1.5 featuring a gapped aligner (ELAND v2). For variant identification downstream of CASAVA, the bioinformatics and sequencing core at CCHMC uses a validated custom, post-alignment software for read-sequence visualization and analysis called SeqMate (P Putnam, unpublished data, 2011). The tool combines the aligned reads with the reference sequence and computes a distribution of call quality at each aligned base position taking into account strand bias. Variants are reported on the basis of a configurable formula. For the custom arrays, the following settings were used: 10x minimum coverage, quality score (Phred) greater than 25, heterozygote allelic ratios allowed from 50%/50% to 75%/25%, and a minimum of 3 distinct reads per location per base call. All variants were searched in dbSNP135 and 1000 genomes project (June 2011). Amino acid changes were identified by comparison with the UCSC RefSeq database track.

Sanger sequencing validation

Non-synonymous coding SNPs (cSNPs) and

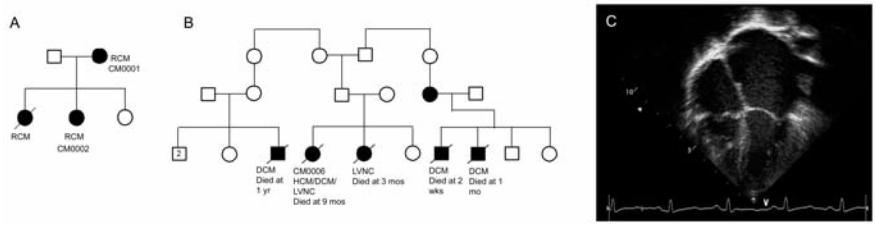

Figure 1. (A, B) Pedigrees of families A and B; (C) Echocardiographic findings of restrictive cardiomyopathy in CM0002.

Table 2. Summary of long-range polymerase chain reaction *versus* sequence capture based massively parallel sequencing.

Method	Total target size (bp)	Total exons (bp)	Avg total aligned reads	Avg depth of coverage	% exons avg>10x	% exon bp >10x
LR-PCR	379,285	41,486	751,512	71x	90.04	89.1
SeqCap	623,783	328,498	4,223,998	442x	99.65	99.73

LR-PCR, long-range polymerase chain reaction; SeqCap, sequence capture; Avg, average.

a splice-site mutation detected by SeqCap and PCR-enrichment were validated by Sanger sequencing. Primer sets for exons of interest were designed flanking the nucleotide change using ExonPrimer perl script available at Genome Browser UCSC homepage and Primer3 program. More than 9000 bases targeted by both platforms were sequenced by Sanger sequencing and results were analyzed via Bioedit sequence alignment editor, version 6.0.7. Primer sequences, amplicon size, and PCR conditions for individual cSNPs are available upon request.

Illumina Quad-Omni array

Samples were genotyped using the Illumina Infinium Omni1 SNP array using standard methods. 1045 markers intersect with the region captured by the custom arrays and 310 markers are located within the LR-PCR products (Table 3).

In silico functional assessment

The functional pathogenicity of all non-synonymous cSNPs was analyzed using the bioinformatic tools Polyphen, PANTHER, and SIFT (See Supplementary Methods S1).[11-14] Novel cSNPs predicted as pathogenic by one program were considered mutations. Rare variants present in dbSNP at a frequency of less than 0.05 and predicted as pathogenic by two bioinformatic programs were also considered mutations.

Results

Clinical findings

Two familial cases of cardiomyopathy were selected for genetic analysis. Pedigree A (Figure 1) shows 2 patients, CM0001 and CM0002, with restrictive cardiomyopathy (RCM) inherited as an autosomal dominant trait. CM0002, now 8 years of age, was diagnosed with RCM at the age of 4 years when she developed echocardiographic evidence of biatrial enlargement (left atrial z-score 4.13, Figure 1C) and diastolic dysfunction. Cardiac catheterization subsequently confirmed the diagnosis. Genetic testing for *MYBPC3, MYH7, TPM1, MYL2, MYL3, ACTC1, TNNT2*, and *TNNC1* genes was performed as part of the clinical diagnostic evaluation. Results showed no pathogenic mutation. CM0001 developed arrhythmia and diastolic dysfunction at the age of 16 years and was diagnosed with RCM after her first pregnancy. The oldest child died at the age of 8 years (status post transplantation for RCM).

Pedigree B (Figure 1) is consistent with an autosomal recessive inheritance pattern in a family of Amish descent. CM0006 was diagnosed with hypertrophic cardiomyopathy and probable left ventricular non-compaction (LVNC) at 5 months of age (LV diastolic septal thickness z-score 7.18; M-mode LV mass 8.17). The proband's sister died from LVNC with heart failure at 3 months. Her clinical course was complicated by tricuspid atresia. The proband showed an undulating cardiac phenotype with dilation and poor cardiac function. The patient met modified Walker criteria, a strict clinical scoring system, for mitochondrial disease. The patient had lactic acidosis. A skin biopsy sent for electron transport chain analysis showed markedly elevated citrate synthase activity, consistent with mitochondrial proliferation. Deficiencies in complexes I, III and IV were detected. The degree of deficiency was marked, especially given that the sample was from skin fibroblasts rather than muscle. CM0006 had clinical testing for one gene known to cause autosomal recessive mitochondrial disorders, SCO2. No mutation was identified. The family declined additional clinical or research testing. The patient died at the age of 9 months.

SeqCap and molecular genetic analysis

A SeqCap based hybridization approach was used to investigate 110 genes causing cardiomyopathy (Table 1). While mutations in sarcomeric and cytoskeletal genes have been well documented, mutations in metabolic genes are known to cause disease but prevalence is unknown. A custom 385K NimbleGen array provided 99.9% coverage of the exons of interest and hybridization was followed by massively parallel sequencing. On average, 4,223,998 mappable, aligned reads were generated for the custom array (PE, 72 bp) (Supplementary Table 1). The average exon depth of coverage was 442x. Alignment and coverage depth was visualized using a novel graphical user interface (Supplementary Figure 1). Read depth is a very important parameter for making accurate SNP calls. We set a 10x depth of coverage, a minimum for reliable calls. On average, 99.73% of exon bases were greater than or equal to 10x coverage with only 897 bases of a total 328,498 protein coding bases lacking sufficient depth for reliable calls (Supplementary Table 1).

The quality of the hybridization based approach and sequencing was assessed using four approaches to determine reproducibility, false positive and false negative rates. Three technical replicates using sample CM0002 showed high reproducibility with base pair coverage less than 10x occurring in 0.27%, 0.55%, and 0.36% (Supplementary Table 2). Somewhat surprisingly, increasing the overall depth of coverage did not correlate with increased comprehensive coverage at the base pair level, as the replicate with the highest overall depth of coverage (CM0002-2, 1105x exon coverage) had the lowest rate of coverage greater than 10x across all exon base pairs (99.45%). Second, we compared our results using the hybridization-based platform to a PCR enrichment approach (LR-PCR) for 31 of

Table 3. Comparison of coding single-nucleotide polymorphism identified by Illumina whole genome genotyping *versus* massively parallel sequencing (long-range polymerase chain reaction and sequence capture).

Method	Total shared cSNPs*	% Concordance	Discordant cSNPs		
LR-PCR			Het loss°	Het gain#	Allele switch§
Covg >5x	532	99.62	1	1	0
Covg >10x	504	100	0	0	0
Covg >20x	463	100	0	0	0
SeqCap					
Covg >5x	2966	99.43	10	6	1
Covg >10x	2944	99.49	8	6	1
Covg >20x	2888	99.51	7	6	1

LR-PCR, long-range polymerase chain reaction; cSNPs, coding single-nucleotide polymorphisms; SeqCap, sequence capture; Het, heterozygous; Covg, coverage. *Number of nucleotide positions genotoyped by Illumina whole genome SNP array that are also present in the indicated LR-PCR or SeqCap sequencing platforms. °Het loss is defined as a heterozygous position (*e.g.* A/C) identified by Illumina genotyping that is identified as homozygous by massively parallel sequencing; #Het gain is defined as a homozygous position (*e.g.* A/A) identified by Illumina genotyping that is identified as heterozygous by massively parallel sequencing; §Allele switch is defined as discrepant SNP calls between Illumina genotyping and massively parallel sequencing (*e.g* A/C vs G/T).

110 genes for all samples (Tables 1 and 2) and to Sanger sequencing for 9 genes for 2 samples. Comparison of the LR-PCR approach (Supplementary Table 3) with the SeqCap approach indicates that the areas with low depth of coverage (<10x) are distinct (data not shown), suggesting a platform specific basis. Third, we performed whole genome genotyping of samples CM0001, CM0002, and CM0006 using the Illumina Infinium Omni1 array and compared base calls at SNPs shared with those identified via resequencing. Concordance rate provides an overall assessment of specificity. Since the majority of pathogenic mutations are rare cSNPs, the heterozygous loss rate (SNPs identified by Illumina genotyping but not by massively parallel sequencing approaches) provides an estimate of false negatives whereas the heterozygous gain provides an estimate of false positives. Concordance rates between our variant calling algorithm and the SNP microarray position calls are indicated in Table 3. Taken together with the overall concordance, the results are consistent with over 99% sensitivity and specificity. Fourth, we performed Sanger sequencing of non-synonymous cSNPs identified by SeqCap and LR-PCR approaches and validated 98.64%, again consistent with a very low false positive rate.

Analysis of mutations

To distinguish potentially pathogenic DNA mutations from synonymous and other variants, we focused on non-synonymous cSNPs. A total of 3813 SNPs were detected by SeqCap-mediated massively parallel sequencing, including 479 coding changes in all 3 patients (Figure 2). Of these, 75 were present in dbSNP (in total 148 in 3 patients; Supplementary Table 4) and 14 were novel (in total 17 in 3 patients; Supplementary Table 5). In CM0001 and CM0002 from pedigree A, a previously described heterozygous mutation, p.A213V, was identified in *DES* gene using both SeqCap and LR-PCR platforms. In CM0001, 7 additional novel cSNPs were identified including 3 predicted to be potentially damaging (p.K435N in the *XIRP1* gene; p.T1351P and p.G1833W in the *CABIN1* gene; Supplementary Table 5). In CM0002, two additional cSNPs were identified.

In CM0006, a homozygous splice site mutation in the *MYBPC3* gene (c.3330+2T>G) was identified on both SeqCap and LR-PCR platforms (Table 4). This mutation has previously been described in the Amish population.[15] Furthermore, two predicted pathogenic novel cSNPs, p.V134A and p.P31A, were detected in *NDUFV2* and *NDUFAF4*, respectively on the SeqCap platform. The *NDUFAF4* p.P31A mutation was not identified in more than 200 control chromosomes in a mitochondrial clinical diagnostic laboratory (Lee-Jun Wong, personal communication, 2011). The SeqCap platform also identified a novel cSNP, p.R1346W, in the *MYH6* gene in CM0006. Mutations in this gene have been reported in patients with atrial septal defects[16] and idiopathic or familial dilated cardiomyopathy.[17]

In silico functional prediction of missense mutation effect was performed using PolyPhen, PANTHER, and SIFT programs for all missense mutations (See Supplementary Methods S1).[11-14] Fifty-four variants were pre-

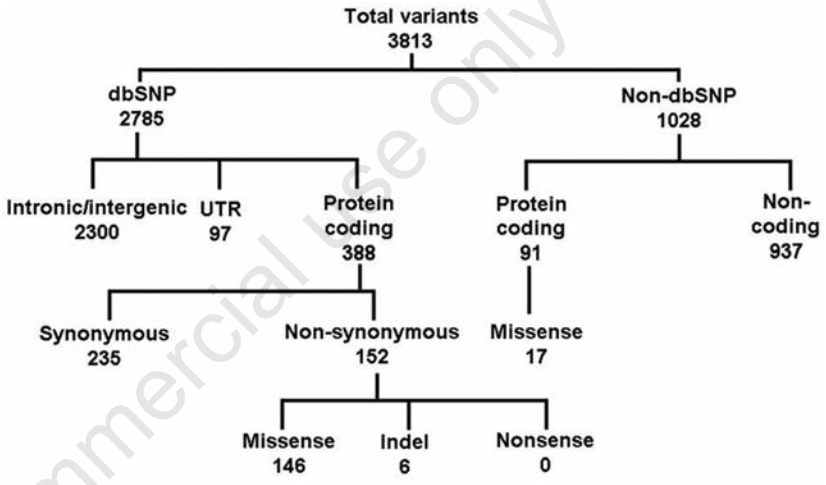

Figure 2. Summary of the total variants detected by sequence capture based massively parallel sequencing in all 3 patients.

Table 4. Mutations identified in each patient.

Patient ID	Gene	Mutation	MAF	Predicted Pathogenicity*	Previously reported with a disease
CM0001	*DES*	p.A213V	N/A	+++	Yes
	GBE1	p.V334I	0.0 17	+++	No
	XIRP1	p.K435N	N/A	+++	No
	CABIN1	p.T1351P	N/A	+++	No
	CABIN1	p.G1833W	N/A	+	No
CM0002	*DES*	p.A213V	N/A	+++	Yes
	GBE1	p.V334I	0.0 17	+++	No
	PYGM	p.A365V	N/A	++	Yes
CM0006	*NDUFV2*	p.V134A	N/A	++	No
	GBE1	p.V334I	0.017	+++	No
	XIRP1	p.Q754P	N/A	++	No
	NDUFAF4	p.P31A	N/A	+++	No
	MYH6	p.R1346W	N/A	++	No
	MYBPC3	c.3330+2T>G	N/A	Altered splicing	Yes

MAF, Minor allele frequency (HapMap/CEU); N/A, Not identified in 226 ethnically matched control alleles. *PolyPhen, SIFT, and PANTHER were used to assess pathogenicity. Each program that predicted a deleterious change is designated with +. See Supplementary Tables 4 and 5 for full details.

dicted as pathogenic by at least one bioinformatic tool (Figure 3) and 27 of these were predicted as pathogenic by more than one program, increasing the likelihood of functional significance. On average, each patient has 4 cSNPs predicted to be deleterious (Table 4). Figure 4 illustrates that amino acid residues altered in missense mutations in *NDUFAF4, XIRP1, PYGM, CABIN1* and *MYH6* are highly conserved across species, arguing for functional significance.

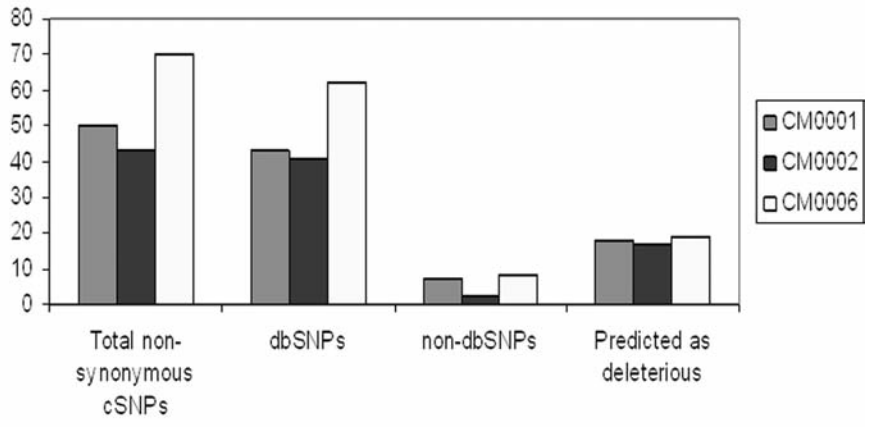

Figure 3. Non-synonymous coding single-nucleotide polymorphisms.

Discussion

Towards development of a novel cardiomyopathy SeqCap platform

Advances in genomic technologies have markedly accelerated the search for genetic causes of human disease and answered previously intractable questions regarding disease mechanisms. In this context, selected genomic regions, genes or exons are targeted by different enrichments methods (*e.g.* SeqCap, RainDance, Fluidigm) followed by massively parallel and cost-effective resequencing.[18-20] Although traditional Sanger sequencing remains the gold standard for mutation detection, it is not readily scalable for diseases with significant genetic heterogeneity such as cardiomyopathy. Our results indicate the utility of a novel custom cardiomyopathy chip for identification of rare variants and disease causing mutations. This custom array hybridization approach was efficient and less time consuming and costly compared to PCR amplification and Sanger sequencing of the same genes.

Analytical validity and clinical validity are two important metrics for genetic testing within a diagnostic laboratory. The SeqCap platform for the analysis of 110 genes in a single reaction showed excellent analytical validity, with 99.73% of base pairs called and over 99% sensitivity and specificity. There were few regions with less than adequate coverage. In clinical testing, increases above 99.73% call rates could potentially be achieved with double or triple probe coverage in regions known to be problematic for capture. In addition, it is likely that improvements in bioinformatic alignment programs and technical improvements in the SeqCap technology and massively parallel sequencing output will lead to overall improved technical performance in the future. While whole-exome and whole-genome sequencing are becoming increasingly affordable, one advantage of a more restricted panel of genes for clinical testing is the less problematic interpretation and improved clinical validity. Finally, the LR-PCR based approach, with 89.1% base pair coverage, does not meet analytical requirements for a clinical test.

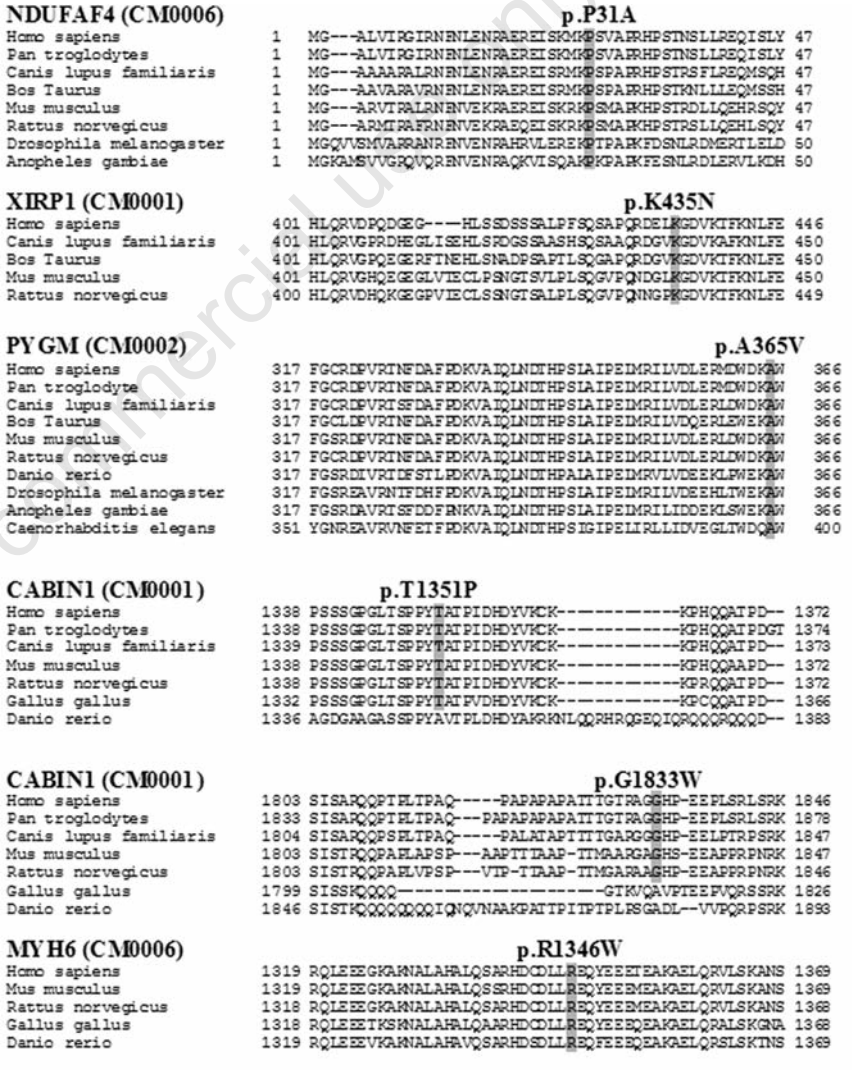

Figure 4. Novel missense mutations identified occur in highly conserved amino acids. Partial amino acid sequence comparisons of human NDUFAF4, XIRP1, PYGM, CABIN1 and MYH6 with other orthologs. The shaded residues indicate high conservation across species. Variations at protein level are shown above each shaded amino acid.

Nevertheless, we note that this degree of coverage is within the realm of most research based applications and given the high specificity and sensitivity, this approach remains a viable option for research based testing. Overall, the LR-PCR approach was much more labor intensive than SeqCap.

Identification of mutations in cardiomyopathy patients

The custom array was used to investigate two types of cardiomyopathy in which the molecular basis is not well understood: RCM and mitochondrial cardiomyopathy. RCM is a distinct cardiomyopathy characterized by diastolic dysfunction but intact systolic function until later stages of the disease. RCM accounts for less than 5% of all cardiomyopathies in the United States and Europe[21] and the prognosis is particularly poor in children. To date, dominant mutations causing RCM have been reported in *DES, ACTC1, TNNI3, TNNT2*, and *MYH7*, but the majority of cases are considered idiopathic.[22,23] In the present study, we identified the heterozygous mutation p.A213V in *DES* in the mother and daughter CM0001 and CM0002 from pedigree A. This mutation has been described previously as conditionally pathogenic causing late-onset dilated cardiomyopathy, familial RCM and desminopathy.[24,25] Desmin (*DES*) encodes a muscle-specific cytoskeletal protein found in smooth, cardiac, and skeletal muscles. Disease causing mutations have been reported in *DES* in dilated cardiomyopathy or desmin-related myopathy.[24,25]

The sample from CM0006 provided the opportunity to investigate the genetic basis of disease in a patient with a complex pedigree and phenotype consistent with mitochondrial disease. Mitochondrial cardiomyopathy is poorly understood at the molecular level. In children, the majority of mitochondrial cardiomyopathies are autosomal recessive and caused by mutations in the nuclear genome rather than the mitochondrial genome. Because many nuclear genes important for mitochondrial function have not yet been identified, molecular diagnosis is challenging. In CM0006, two rare cSNP predicted to be disease-causing were identified in genes encoding proteins important for complex I of the mitochondrial electron transport chain, p.V134A in *NDUFV2* and p.P31A in *NDUFAF4*. *NDUFAF4* encodes an assembly factor of mitochondrial complex I and has been previously described in complex I deficiency causing infantile mitochondrial encephalomyopathy and cardiomyopathy.[26] The functional significance of mutations for two complex I mitochondrial proteins requires further assessment.

Surprisingly, the patient was also found to be homozygous for a previously reported splice mutation in *MYBPC3* (c.3330+2T>G). This mutation has been associated previously with a severe neonatal hypertrophic cardiomyopathy in the Amish community when homozygous and hypertrophic cardiomyopathy with incomplete penetrance when heterozygous.[15,27,28] In addition, a novel cSNP in *MYH6* gene was identified that may play a role in the familial cardiomyopathy or history of congenital heart defect based on its published functions.[16,17] The finding of multiple putative disease-causing alleles in this family may explain the complex inheritance pattern noted in the pedigree. Unfortunately, family members declined further research testing, precluding the possibility of defining segregation with phenotype. Nevertheless, these results indicate the importance of screening multiple genes on a single platform, allowing examination of the potential interplay between genetic alterations in the structural apparatus of the cardiomyocyte and metabolic pathway-specific genes. Further development of these technologies will offer a unique opportunity to interrogate complex patterns of inheritance due to the involvement of more than one gene.

Novel sequencing technology identifies multiple rare variants

Phenotypic variability is frequently observed in closely related family members carrying the same cardiomyopathy causing mutations.[29] This phenotypic variability has been attributed to environmental influences (*e.g.* blood pressure, diet, activity) and the individual's specific genetic background. However, recognition of modifying genetic influences has significantly increased as the identification of disease-causing genes has expanded.

Recent studies on hypertrophic cardiomyopathy have shown that individuals may have 2 distinct pathogenic mutations.[30] The identification of more than one mutation in some children with cardiomyopathy has led to the suggestion that early onset is due to a *two hit* digenic mechanism with case reports indicating that these individuals are younger and have more aggressive disease.[31-33] However, this view is controversial with some stating that there is insufficient evidence of the functional effects of multiple mutations.[34] As demonstrated by this study, one of the challenges for clinical laboratories will be the interpretation of the multiple rare or novel cSNPs identified using massively parallel sequencing. In this small study, an average of 4 potential disease-causing cSNPs was identified in each patient. In the past, a *final common pathway* hypothesis had been proposed to explain similar phenotypes caused by different genes (genetic heterogeneity). This hypothesis suggests that mutated proteins of similar functions or signaling pathways[35-37] can lead to cardiomyopathy phenotypes that are clinically indistinguishable. This study documents both the technical ability to identify multiple rare variants or mutations in a single patient as well as the challenges of interpretation. In the future, it will be possible to analyze whether multiple rare mutations in the same or different genes act in concert to modulate the disease phenotype.

Whole exome sequencing has recently been successfully applied to identify the genetic basis of rare disorders with very specific clinical features while its application in finding genetic mechanisms for complex human diseases is under development.[38-41] Results with emerging genomic technologies indicate that approximately 400 novel protein coding variants are identified with full exome sequencing. In addition, bioinformatic programs predict that approximately 20% of common human non-synonymous SNPs damage the proteins.[12] Because of this, recognition of rare variants that significantly impact protein function in genetically and phenotypically heterogeneous diseases, such as cardiomyopathy, is challenging. Mutations identified in multiple genes in the present study provide insight into the large number of altered protein-protein interactions, potentially resulting in a combined effect on function. Although bioinformatics programs continue to improve, they do have limitations. For example, the variant p.R1745H in dystrophin is present in the population at a frequency of 43% but is nevertheless predicted to be deleterious. Use of multiple prediction programs, as in this study, can improve overall prediction but nevertheless cannot prove functional significance. High throughput sequencing technologies provide the technical basis to screen candidate genes or the complete exome in large cohorts of patients, However, the interpretation of the results of these studies will require further development of bioinformatics, mutation prediction, and functional screens. By addressing these issues with interpretation of massively sequencing data we will be able to provide genome-guided diagnosis and treatment.

Study limitations

Detection of deletions and insertions with the stringency required for clinical testing remains problematic with massively parallel sequencing. Incorporation of probes specific to previously described deletions and insertions associated with cardiomyopathy may be used to overcome this problem. In addition, further development of efficient and robust bioinformatic approaches addressing novel indel variations will aid mutation identification. The 13 cSNPs identified in this study are predicted to be pathogenic (*in silico*) as missense mutations (Table 4) but additional studies are required to more fully investigate their effect and combined functions.

Conclusions

In summary, we have investigated a novel genetic diagnostic platform for pediatric cardiomyopathy using array-based sequence technology followed by massively parallel sequencing. Furthermore, we compared the performance of targeted array-based and long-range PCR based genomic DNA enrichments for massively parallel sequencing and found array-based enrichment more cost and time efficient and more sensitive. Through this array-based sequencing, 89 different non-synonymous cSNPs and one splice site mutation were identified including multiple predicted disease-causing variants in each patient. This study highlights the difficulty of interpreting high-throughput sequencing data with a number of predicted pathogenic variants even in a small sequence capture experiment.

References

1. Mamanova L, Coffey AJ, Scott CE, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods 2010;7:111-8.
2. Okou DT, Locke AE, Steinberg KM, et al. Combining microarray-based genomic selection (MGS) with the Illumina Genome Analyzer platform to sequence diploid target regions. Ann Hum Genet 2009;73:502-13.
3. Albert TJ, Molla MN, Muzny DM, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007;4:903-5.
4. Porreca GJ, Zhang K, Li JB, et al. Multiplex amplification of large sets of human exons. Nat Methods 2007;4:931-6.
5. Ng SB, Bigham AW, Buckingham KJ, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 2010;42:790-3.
6. Nikopoulos K, Gilissen C, Hoischen A, et al. Next-generation sequencing of a 40 Mb linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreoretinopathy. Am J Hum Genet 2010; 86:240-7.
7. Volpi L, Roversi G, Colombo EA, et al. Targeted next-generation sequencing appoints c16orf57 as clericuzio-type poikiloderma with neutropenia gene. Am J Hum Genet 2010;86:72-6.
8. Rehman AU, Morell RJ, Belyantseva IA, et al. Targeted capture and next-generation sequencing identifies C9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness DFNB79. Am J Hum Genet 2010;86:378-88.
9. Raca G, Jackson C, Warman B, et al. Next generation sequencing in research and diagnostics of ocular birth defects. Mol Genet Metab 2010;100:184-92.
10. Summerer D, Schracke N, Wu H, et al. Targeted high throughput sequencing of a cancer-related exome subset by specific sequence capture with a fully automated microarray platform. Genomics 2010;95: 241-6.
11. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073-81.
12. Sunyaev S, Ramensky V, Koch I, et al. Prediction of deleterious human alleles. Hum Mol Genet 2001;10:591-7.
13. Thomas PD, Campbell MJ, Kejariwal A, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 2003;13:2129-41.
14. Thomas PD, Kejariwal A, Campbell MJ, et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res 2003;31:334-41.
15. Xin B, Puffenberger E, Tumbush J, et al. Homozygosity for a novel splice site mutation in the cardiac myosin-binding protein C gene causes severe neonatal hypertrophic cardiomyopathy. Am J Med Genet A 2007;143A:2662-7.
16. Ching YH, Ghosh TK, Cross SJ, et al. Mutation in myosin heavy chain 6 causes atrial septal defect. Nat Genet 2005;37: 423-8.
17. Carniel E, Taylor MR, Sinagra G, et al. Alpha-myosin heavy chain: a sarcomeric gene associated with dilated and hypertrophic phenotypes of cardiomyopathy. Circulation 2005;112:54-9.
18. Mondal K, Shetty AC, Patel V, et al. Targeted sequencing of the human X chromosome exome. Genomics 2011;98: 260-5.
19. Najmabadi H, Hu H, Garshasbi M, et al. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature 2011;478:57-63.
20. Jones MA, Bhide S, Chin E, et al. Targeted polymerase chain reaction-based enrichment and next generation sequencing for diagnostic testing of congenital disorders of glycosylation. Genet Med 2011;13:921-32.
21. Felker GM, Thompson RE, Hare JM, et al. Underlying causes and long-term survival in patients with initially unexplained cardiomyopathy. N Engl J Med 2000;342:1077-84.
22. Kaski JP, Syrris P, Burch M, et al. Idiopathic restrictive cardiomyopathy in children is caused by mutations in cardiac sarcomere protein genes. Heart 2008;94:1478-84.
23. Ware SM, Quinn ME, Ballard ET, et al. Pediatric restrictive cardiomyopathy associated with a mutation in beta-myosin heavy chain. Clin Genet 2008;73: 165-70.
24. Goudeau B, Rodrigues-Lima F, Fischer D, et al. Variable pathogenic potentials of mutations located in the desmin alpha-helical domain. Hum Mutat 2006;27:906-13.
25. Kostareva A, Gudkova A, Sjoberg G, et al. Desmin mutations in a St. Petersburg cohort of cardiomyopathies. Acta Myol 2006;25:109-15.
26. Saada A, Edvardson S, Rapoport M, et al. C6ORF66 is an assembly factor of mitochondrial complex I. Am J Hum Genet 2008;82:32-8.
27. Zahka K, Kalidas K, Simpson MA, et al. Homozygous mutation of MYBPC3 associated with severe infantile hypertrophic cardiomyopathy at high frequency among the Amish. Heart 2008;94:1326-30.
28. Wang H, Xin B. Hypertrophic cardiomyopathy in the Amish community - What we may learn from it. Prog Pediatr Cardiol 2011;31:129-34.
29. Hershberger RE, Cowan J, Morales A, Siegfried JD. Progress with genetic cardiomyopathies: screening, counseling, and testing in dilated, hypertrophic, and arrhythmogenic right ventricular dysplasia/cardiomyopathy. Circ Heart Fail 2009;2:253-61.
30. Kelly M, Semsarian C. Multiple mutations in genetic cardiovascular disease: a marker of disease severity? Circ Cardiovasc Genet 2009;2:182-90.
31. Ingles J, Doolan A, Chiu C, et al. Compound and double mutations in patients with hypertrophic cardiomyopathy: implications for genetic testing and counselling. J Med Genet 2005;42:e59.
32. Lekanne Deprez RH, Muurling-Vlietman JJ, Hruda J, et al. Two cases of severe neonatal hypertrophic cardiomyopathy caused by compound heterozygous mutations in the MYBPC3 gene. J Med Genet 2006;43:829-32.
33. Girolami F, Ho CY, Semsarian C, et al. Clinical features and outcome of hypertrophic cardiomyopathy associated with triple sarcomere protein gene mutations. J Am Coll Cardiol 2010;55:1444-53.
34. Ho CY, MacRae CA. Defining the pathogenicity of DNA sequence variation. Circ Cardiovasc Genet 2009;2:95-7.
35. Towbin JA. The role of cytoskeletal proteins in cardiomyopathies. Curr Opin Cell Biol 1998;10:131-9.
36. Towbin JA, Bowles KR, Bowles NE.

Etiologies of cardiomyopathy and heart failure. Nat Med 1999;5:266-7.
37. Bowles NE, Bowles KR, Towbin JA. The "final common pathway" hypothesis and inherited cardiovascular disease. The role of cytoskeletal proteins in dilated cardiomyopathy. Herz 2000;25:168-75.
38. Feng BJ, Tavtigian SV, Southey MC, Goldgar DE. Design considerations for massively parallel sequencing studies of complex human disease. PLoS One 2011;6:e23221.
39. Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A 2009;106:19096-101.
40. Hedges DJ, Burges D, Powell E, et al. Exome sequencing of a multigenerational human pedigree. PLoS One 2009;4:e8232.
41. Johnston JJ, Teer JK, Cherukuri PF, et al. Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. Am J Hum Genet 2010;86:743-8.