



## Article

# A Whole-Genome Sequencing-Based SNP Protocol for Accurate Plant Variety Identification: Application in Blueberry

Agnelo Furtado <sup>1</sup>, Tom Gunther <sup>2</sup> and Robert J. Henry <sup>1,\*</sup><sup>1</sup> Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, QLD 4072, Australia; a.furtado@uq.edu.au<sup>2</sup> Mountain Blue Orchards, Lindendale, NSW 2480, Australia; tgunther@mountainblue.com.au

\* Correspondence: robert.henry@uq.edu.au

## Abstract

Plant variety/genotype identification has many applications in establishing the identity of plants, including the protection of intellectual property rights and the management of ex situ conservation of genetic resources. The variety may be important for operational reasons based on field performance or post-harvest processing. Blueberry (*Vaccinium corymbosum* L.), an economically important crop, is propagated by cuttings and commercially important accessions require an accurate variety traceability regime for the maintenance of purity, protection and policing ownership. Genome sequencing methods have improved and are feasible for use, making examination of the whole genome for all possible information on the genotype the ultimate way to distinguish plant varieties. We identified 5.3–5.5 million high-confidence homozygous SNPs with over 99% accuracy, enabling the distinction of 41 blueberry varieties. We developed a novel data-noise identification and filtering framework, which correctly determined the identity of ten unknown samples to be the Masena variety with 100% accuracy. The approach of using a data-noise filtration step minimized the impact of sequencing errors and coincident sequencing of only one allele of any heterozygous base. This SNP-based protocol with the establishment of sequence databases for all varieties of important plant species can potentially be adopted in providing reliable variety identification in critical industrial or legal applications.

**Keywords:** intellectual-property-protection; variety-identification; policing; data-noise-filtering; high-confidence-SNPs



Academic Editor: Adriana Basile

Received: 28 August 2025

Revised: 5 November 2025

Accepted: 6 November 2025

Published: 12 November 2025

**Citation:** Furtado, A.; Gunther, T.; Henry, R.J. A Whole-Genome Sequencing-Based SNP Protocol for Accurate Plant Variety Identification: Application in Blueberry. *Appl. Biosci.* **2025**, *4*, 51. <https://doi.org/10.3390/applbiosci4040051>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Plant variety identification is central to the protection/preservation of plant varieties and the policing of the intellectual property rights of plant breeders [1]. Plant variety identification may also be important in the marketing of agricultural and food products, the testing of the purity of foods [2–17] and in managing ex situ plant conservation [18,19]. Biosecurity and food safety issues may also require the determination of the varietal identity of plants. The identity of clonally propagated material is key to commercialization, conservation and preservation [20–22]. Blueberry is a high-value clonally propagated species and requires an identification and traceability regime for accurate variety identification in germplasm maintenance and in the establishment and policing of plant breeders' intellectual property rights [9,23,24].

Plant genotyping has been conducted using a range of molecular tools that measure specific genetic loci [25–30]. Plant species have been widely distinguished by the

analysis of ribosomal genes [31–34]. Early genotype/variety-specific makers were based upon arbitrary primer PCR methods, microsatellite or short sequence repeat (SSR) analysis [35–40]. All these methods had issues with reproducibility, technical complexity or labor-intensive procedures, and shared a common limitation—they failed to capture all possible polymorphic positions in the genome, leading to low resolution/discrimination of samples [41].

More recently, single nucleotide polymorphism (SNP) markers, requiring sequence information for their discovery, have been widely used. SNP represents a nucleotide change at a specific position in the genome sequence [42,43]. Advances in DNA sequencing methods have resulted in the widespread use of SNP analysis in plant identification [44,45]. Many techniques for SNP discovery have been used and these can be divided into three broad classes, Polymerase Chain Reaction (PCR)-based; microarray-based; and next generation sequencing (NGS)-based, also referred to as genotyping by sequencing (GBS) [25,27,46–50]. These types of genotyping methods are suitable for interrogating a small number of target SNPs over a large number of samples.

NGS-based methods are better than PCR-based or microarray-based methods, as they are high throughput, allowing for a more comprehensive interrogation of the genome without the need for prior knowledge of the sequence, and providing much better resolution in identifying rare variants [51]. Early NGS-based methods involved the use of restriction enzymes DNA libraries designed as a reduced representation method to reduce genome complexity and capture a fraction of the genome to be sequenced. Restriction-site DNA sequencing (RAD-seq) marker genotyping [52] was later adapted for multiplexing [53]. Subsequently, a less complicated but highly multiplexed reduced representation method, genotyping by sequencing (GBS), was proposed for SNP discovery and SNP-based genotyping [25,54,55]. Missing data in GBS can be imputed using whole-genome resequencing data [56]. All the above NGS SNP-based methods have been employed in genotyping samples by interrogating a fraction of the SNPs present in a genome. A number of software algorithms [55,57,58] have been developed to address missing data from low coverage sequence data which can lead to wrong interpretations [59]. In the resolution of IP disputes, methods require highly discriminating and reliable analysis of the whole genome to definitely conclude if two samples are the same genotype. This would necessitate sequencing at a high depth to increase confidence in the SNPs generated and reduce missing data [55,60,61].

The goal should be to capture/identify all the SNPs in the entire genome and in as many genotypes as possible in a cost-effective and high-throughput manner. Identifying all the SNPs in the genome, instead of a subset, provides the best chance of capturing the unique identifying SNPs to distinguish the reference samples and identify any unknown varieties compared to the reference samples. Improvements in DNA sequencing technology have resulted in a convergence of DNA sequencing and genotyping approaches [17,49,55]. SNP analysis of the entire plant genome sequence enables the distinction of genotypes that vary by as little as a single nucleotide. In principle, a single homozygous SNP identified at high coverage and unique in candidate-A compared to other candidates of the reference set can be used for variety identification. It is better to have as many unique-candidate SNPs for a sample as possible, especially should the need arise to increase the set of reference samples when a new variety is introduced. Next generation sequencing allows for the detection of point mutations in plants that may be associated with useful phenotypes. Continuing advances in DNA sequencing have made the sequencing of the whole genome increasingly more cost-effective [27,30,49,62].

Several points require consideration to obtain the required resolution and confidence in applying NGS for variety identification using SNP markers. These include high-quality

NGS data at sufficient depth, a good-quality assembled genome, preferably of the same species, a read mapping software with options for the optimization of mapping, and a variant calling software capable of using a probabilistic approach (error model estimation and Bayesian model with a Maximum Likelihood approach, e.g., CLC Genomics). The variant caller should also have relevant settings to reduce/eliminate erroneous calls. However, at a polymorphic allelic position, the variant caller nonetheless makes the calls based on the proportion of mapped sequence read. Low read depth can lead to incorrect interpretation, where a true heterozygous SNP is erroneously called a homozygous SNP. The probability of sequence reads with both alleles, at a heterozygous position, increases with mapped read depth, with near 100% accurate homozygous calls at mapped coverage of above 20 [55,60]. A mapped sequence depth of 20 is likely to be cost-effective for variety identification, which involves tens of samples. Nonetheless, the variant caller makes the SNP variant call based on the alleles present in the mapped reads. Hence, reduced read depth can influence the accuracy of the call, or lead to no call being made, leading to missing data. When using SNP data for variety identification, careful filtering of polymorphic SNP calls is essential to analyze for missing data and to reject the data arising from factors such as reduced read depth.

Here, using short read sequence data from 41 *Vaccinium corymbosum* L. (blueberry) varieties, we present a protocol for routine high-confidence SNP variant-based blueberry variety identification which can be adopted for use in the distinction of plant varieties. This method applies filtering steps to eliminate data-noise leading to accurate plant variety identification.

## 2. Materials and Methods

### 2.1. Leaf Sample Collection

A total of fifty-seven sample collections, S01 to S57, (Table 1) of leaf tissue, each consisting of 20 to 25 leaves, were made from field-grown plants of 41 blueberry varieties (S01 to S40 and S45) at Mountain Blue Orchards, Tabulam NSW 2469, Australia. The 41 blueberry varieties included are commercially important varieties, with S40 and S45 representing the varieties Masena and Eureka, respectively. Leaf samples, collections S01 to S39, S45 and S50, were used as reference samples for variety identification analysis. Additionally, to determine variation in “variant analysis” and factors affecting it, leaf sample collections S40 to S57, from the two varieties Masena and Eureka, were used in replicate analysis comprising technical and biological replicates. Technical replicate analysis was based upon one technical replicate library (Tec-rep) preparation from each of five replicate leaf collections from the same plant (P1L1 to P1L5) and labeled “/P1/L1 to P1/L5”, both for the Masena variety with library #s B40 to B44 and for the variety Eureka with library #s B45 to B49. Biological replicate analysis was conducted for three biological replicate library (Bio1, Bio2 and Bio3) preparations, where each of the library preparations was prepared from replicate leaf collections, one each from five plants (P1L1 to P5L1), and labeled as “/P1/L1/” to “/P5/L1/”, both for the Masena variety with the Bio1 library consisting of B40, B50 to B53, Bio2 consisting of B58 to B62 and Bio3 consisting of B68 to B72 and for Eureka with the Bio1 library consisting of B45, B54 to B57, Bio2 consisting of B63 to B67 and Bio3 consisting of B73 to B77 (Table 1).

Additional details are provided in the Supplementary Materials.

Table 1. Details of blueberry varieties/cultivars.

| Variety/<br>Cultivar | Genotype # | Sample<br>Collection # | Sequencing File<br>ID # | Blueberry<br>Library Number<br># | Plant and Leaf # | Sample Role/Replicate #<br>(Unknown/Reference: Variety<br>Identification)<br>(Replicate Technical, Replicate<br>Biological Library 1/Library<br>2/Library 3: Variability in SNP<br>Variant Analysis) | Overall Sample Code             |
|----------------------|------------|------------------------|-------------------------|----------------------------------|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------|
| G46                  | G46        | S01                    | Blueberry-01            | B01                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G46/S01/B01/P1/L1/Ref-rep<br>#1 |
| G08                  | G08        | S02                    | Blueberry-02            | B02                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G08/S02/B02/P1/L1/Ref-rep<br>#1 |
| G29                  | G29        | S03                    | Blueberry-03            | B03                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G29/S03/B03/P1/L1/Ref-rep<br>#1 |
| G14                  | G14        | S04                    | Blueberry-04            | B04                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G14/S04/B04/P1/L1/Ref-rep<br>#1 |
| G45                  | G45        | S05                    | Blueberry-05            | B05                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G45/S05/B05/P1/L1/Ref-rep<br>#1 |
| G32                  | G32        | S06                    | Blueberry-06            | B06                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G32/S06/B06/P1/L1/Ref-rep<br>#1 |
| G19                  | G19        | S07                    | Blueberry-07            | B07                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G19/S07/B07/P1/L1/Ref-rep<br>#1 |
| G18                  | G18        | S08                    | Blueberry-08            | B08                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G18/S08/B08/P1/L1/Ref-rep<br>#1 |
| G06                  | G06        | S09                    | Blueberry-09            | B09                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G06/S09/B09/P1/L1/Ref-rep<br>#1 |
| G03                  | G03        | S10                    | Blueberry-10            | B10                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G03/S10/B10/P1/L1/Ref-rep<br>#1 |
| G04                  | G04        | S11                    | Blueberry-11            | B11                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G04/S11/B11/P1/L1/Ref-rep<br>#1 |
| G09                  | G09        | S12                    | Blueberry-12            | B12                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G09/S12/B12/P1/L1/Ref-rep<br>#1 |
| G39                  | G39        | S13                    | Blueberry-13            | B13                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G39/S13/B13/P1/L1/Ref-rep<br>#1 |

Table 1. Cont.

| Variety/<br>Cultivar | Genotype # | Sample<br>Collection # | Sequencing File<br>ID # | Blueberry<br>Library Number<br># | Plant and Leaf # | Sample Role/Replicate #<br>(Unknown/Reference: Variety<br>Identification)<br>(Replicate Technical, Replicate<br>Biological Library 1/Library<br>2/Library 3: Variability in SNP<br>Variant Analysis) | Overall Sample Code             |
|----------------------|------------|------------------------|-------------------------|----------------------------------|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------|
| G31                  | G31        | S14                    | Blueberry-14            | B14                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G31/S14/B14/P1/L1/Ref-rep<br>#1 |
| G37                  | G37        | S15                    | Blueberry-15            | B15                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G37/S15/B15/P1/L1/Ref-rep<br>#1 |
| G48                  | G48        | S16                    | Blueberry-16            | B16                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G48/S16/B16/P1/L1/Ref-rep<br>#1 |
| G36                  | G36        | S17                    | Blueberry-17            | B17                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G36/S17/B17/P1/L1/Ref-rep<br>#1 |
| G33                  | G33        | S18                    | Blueberry-18            | B18                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G33/S18/B18/P1/L1/Ref-rep<br>#1 |
| G26                  | G26        | S19                    | Blueberry-19            | B19                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G26/S19/B19/P1/L1/Ref-rep<br>#1 |
| G43                  | G43        | S20                    | Blueberry-20            | B20                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G43/S20/B20/P1/L1/Ref-rep<br>#1 |
| G34                  | G34        | S21                    | Blueberry-21            | B21                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G34/S21/B21/P1/L1/Ref-rep<br>#1 |
| G40                  | G40        | S22                    | Blueberry-22            | B22                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G40/S22/B22/P1/L1/Ref-rep<br>#1 |
| G12                  | G12        | S23                    | Blueberry-23            | B23                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G12/S23/B23/P1/L1/Ref-rep<br>#1 |
| G17                  | G17        | S24                    | Blueberry-24            | B24                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G17/S24/B24/P1/L1/Ref-rep<br>#1 |
| G28                  | G28        | S25                    | Blueberry-25            | B25                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G28/S25/B25/P1/L1/Ref-rep<br>#1 |
| G42                  | G42        | S26                    | Blueberry-26            | B26                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G42/S26/B26/P1/L1/Ref-rep<br>#1 |

Table 1. Cont.

| Variety/<br>Cultivar | Genotype # | Sample<br>Collection # | Sequencing File<br>ID # | Blueberry<br>Library Number<br># | Plant and Leaf # | Sample Role/Replicate #<br>(Unknown/Reference: Variety<br>Identification)<br>(Replicate Technical, Replicate<br>Biological Library 1/Library<br>2/Library 3: Variability in SNP<br>Variant Analysis) | Overall Sample Code             |
|----------------------|------------|------------------------|-------------------------|----------------------------------|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------|
| G44                  | G44        | S27                    | Blueberry-27            | B27                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G44/S27/B27/P1/L1/Ref-rep<br>#1 |
| G05                  | G05        | S28                    | Blueberry-28            | B28                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G05/S28/B28/P1/L1/Ref-rep<br>#1 |
| G16                  | G16        | S29                    | Blueberry-29            | B29                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G16/S29/B29/P1/L1/Ref-rep<br>#1 |
| G27                  | G27        | S30                    | Blueberry-30            | B30                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G27/S30/B30/P1/L1/Ref-rep<br>#1 |
| G38                  | G38        | S31                    | Blueberry-31            | B31                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G38/S31/B31/P1/L1/Ref-rep<br>#1 |
| G21                  | G21        | S32                    | Blueberry-32            | B32                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G21/S32/B32/P1/L1/Ref-rep<br>#1 |
| G35                  | G35        | S33                    | Blueberry-33            | B33                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G35/S33/B33/P1/L1/Ref-rep<br>#1 |
| G30                  | G30        | S34                    | Blueberry-34            | B34                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G30/S34/B34/P1/L1/Ref-rep<br>#1 |
| G02                  | G02        | S35                    | Blueberry-35            | B35                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G02/S35/B35/P1/L1/Ref-rep<br>#1 |
| G41                  | G41        | S36                    | Blueberry-36            | B36                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G41/S36/B36/P1/L1/Ref-rep<br>#1 |
| G49                  | G49        | S37                    | Blueberry-37            | B37                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G49/S37/B37/P1/L1/Ref-rep<br>#1 |
| G11                  | G11        | S38                    | Blueberry-38            | B38                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G11/S38/B38/P1/L1/Ref-rep<br>#1 |
| G47                  | G47        | S39                    | Blueberry-39            | B39                              | P1/L1            | Reference/Ref-rep #1                                                                                                                                                                                 | G47/S39/B39/P1/L1/Ref-rep<br>#1 |

Table 1. Cont.

| Variety/<br>Cultivar | Genotype # | Sample<br>Collection # | Sequencing File<br>ID # | Blueberry<br>Library Number<br># | Plant and Leaf # | Sample Role/Replicate #<br>(Unknown/Reference: Variety<br>Identification)<br>(Replicate Technical, Replicate<br>Biological Library 1/Library<br>2/Library 3: Variability in SNP<br>Variant Analysis) | Overall Sample Code                                    |
|----------------------|------------|------------------------|-------------------------|----------------------------------|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------|
| Masena               | G10        | S40                    | Blueberry-40            | B40                              | P1/L1            | Unknown<br>Replicate Technical/Tec-rep #1<br>Replicate Biological<br>Library1/BioL1-rep #1                                                                                                           | G10/S40/B40/P1/L1/Unk/Tec-<br>rep #1/Bio1-rep #1       |
| Masena               | G10        | S41                    | Blueberry-41            | B41                              | P1/L2            | Replicate Technical/Tec-rep #2                                                                                                                                                                       | G10/S41/B41/P1/L2/Tec-rep<br>#2                        |
| Masena               | G10        | S42                    | Blueberry-42            | B42                              | P1/L3            | Replicate Technical/Tec-rep #3                                                                                                                                                                       | G10/S42/B42/P1/L3/Tec-rep<br>#3                        |
| Masena               | G10        | S43                    | Blueberry-43            | B43                              | P1/L4            | Replicate Technical/Tec-rep #4                                                                                                                                                                       | G10/S43/B43/P1/L4/Tec-rep<br>#4                        |
| Masena               | G10        | S44                    | Blueberry-44            | B44                              | P1/L5            | Replicate Technical/Tec-rep #5                                                                                                                                                                       | G10/S44/B44/P1/L5/Tec-rep<br>#5                        |
| Eureka               | G15        | S45                    | Blueberry-45            | B45                              | P1/L1            | Reference/Ref-rep #1<br>Replicate Technical/Tec-rep #1<br>Replicate Biological<br>Library1/BioL1-rep #1                                                                                              | G15/S45/B45/P1/L1/Ref-rep<br>#1/Tec-rep #1/Bio1-rep #1 |
| Eureka               | G15        | S46                    | Blueberry-46            | B46                              | P1/L2            | Replicate Technical/Tec-rep #2                                                                                                                                                                       | G15/S46/B46/P1/L2/Tec-rep<br>#2                        |
| Eureka               | G15        | S47                    | Blueberry-47            | B47                              | P1/L3            | Replicate Technical/Tec-rep #3                                                                                                                                                                       | G15/S47/B47/P1/L3/Tec-rep<br>#3                        |
| Eureka               | G15        | S48                    | Blueberry-48            | B48                              | P1/L4            | Replicate Technical/Tec-rep #4                                                                                                                                                                       | G15/S48/B48/P1/L4/Tec-rep<br>#4                        |
| Eureka               | G15        | S49                    | Blueberry-49            | B49                              | P1/L5            | Replicate Technical/Tec-rep #5                                                                                                                                                                       | G15/S49/B49/P1/L5/Tec-rep<br>#5                        |
| Masena               | G10        | S50                    | Blueberry-50            | B50                              | P2/L1            | Reference<br>Replicate Biological<br>Library1/Bio1-rep #2                                                                                                                                            | G10/S50/B50/P2/L1/Ref/Bio1-<br>rep #2                  |

Table 1. Cont.

| Variety/<br>Cultivar | Genotype # | Sample<br>Collection # | Sequencing File<br>ID # | Blueberry<br>Library Number<br># | Plant and Leaf # | Sample Role/Replicate #<br>(Unknown/Reference: Variety<br>Identification)<br>(Replicate Technical, Replicate<br>Biological Library 1/Library<br>2/Library 3: Variability in SNP<br>Variant Analysis) | Overall Sample Code                   |
|----------------------|------------|------------------------|-------------------------|----------------------------------|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------|
| Masena               | G10        | S51                    | Blueberry-51            | B51                              | P3/L1            | Unknown/Replicate Biological<br>Library1/Bio1-rep #3                                                                                                                                                 | G10/S51/B51/P3/L1/Unk/Bio1-<br>rep #3 |
| Masena               | G10        | S52                    | Blueberry-52            | B52                              | P4/L1            | Unknown/Replicate Biological<br>Library1/Bio1-rep #4                                                                                                                                                 | G10/S52/B52/P4/L1/Bio1-rep<br>#4      |
| Masena               | G10        | S53                    | Blueberry-53            | B53                              | P5/L1            | Unknown/Replicate Biological<br>Library1/Bio1-rep #5                                                                                                                                                 | G10/S53/B53/P5/L1/Bio1-rep<br>#5      |
| Eureka               | G15        | S54                    | Blueberry-54            | B54                              | P2/L1            | Replicate Biological<br>Library1/Bio1-rep #6                                                                                                                                                         | G15/S54/B54/P2/L1/Bio1-rep<br>#2      |
| Eureka               | G15        | S55                    | Blueberry-55            | B55                              | P3/L1            | Replicate Biological<br>Library1/Bio1-rep #7                                                                                                                                                         | G15/S55/B55/P3/L1/Bio1-rep<br>#3      |
| Eureka               | G15        | S56                    | Blueberry-56            | B56                              | P4/L1            | Replicate Biological<br>Library1/Bio1-rep #8                                                                                                                                                         | G15/S56/B56/P4/L1/Bio1-rep<br>#4      |
| Eureka               | G15        | S57                    | Blueberry-57            | B57                              | P5/L1            | Replicate Biological<br>Library1/Bio1-rep#9                                                                                                                                                          | G15/S57/B57/P5/L1/Bio1-rep<br>#5      |
| Masena               | G10        | S40                    | Blueberry-58            | B58                              | P1/L1            | Unknown/Replicate Biological<br>Library2/Bio2-rep #1                                                                                                                                                 | G10/S40/B58/P1/L1/Unk/Bio2-<br>rep #1 |
| Masena               | G10        | S50                    | Blueberry-59            | B59                              | P2/L1            | Unknown/Replicate Biological<br>Library2/Bio2-rep #2                                                                                                                                                 | G10/S50/B59/P2/L1/Unk/Bio2-<br>rep #2 |
| Masena               | G10        | S51                    | Blueberry-60            | B60                              | P3/L1            | Unknown/Replicate Biological<br>Library2/Bio2-rep #3                                                                                                                                                 | G10/S51/B60/P3/L1/Unk/Bio2-<br>rep #3 |
| Masena               | G10        | S52                    | Blueberry-61            | B61                              | P4/L1            | Unknown/Replicate Biological<br>Library2/Bio2-rep #4                                                                                                                                                 | G10/S52/B61/P4/L1/Unk/Bio2-<br>rep #4 |
| Masena               | G10        | S53                    | Blueberry-62            | B62                              | P5/L1            | Unknown/Replicate Biological<br>Library2/Bio2-rep #5                                                                                                                                                 | G10/S53/B62/P5/L1/Unk/Bio2-<br>rep #5 |
| Eureka               | G15        | S45                    | Blueberry-63            | B63                              | P1/L1            | Replicate Biological<br>Library2/Bio2-rep #6                                                                                                                                                         | G15/S45/B63/P1/L1/Bio2-rep<br>#1      |

Table 1. Cont.

| Variety/<br>Cultivar | Genotype # | Sample<br>Collection # | Sequencing File<br>ID # | Blueberry<br>Library Number<br># | Plant and Leaf # | Sample Role/Replicate #<br>(Unknown/Reference: Variety<br>Identification)<br>(Replicate Technical, Replicate<br>Biological Library 1/Library<br>2/Library 3: Variability in SNP<br>Variant Analysis) | Overall Sample Code                   |
|----------------------|------------|------------------------|-------------------------|----------------------------------|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------|
| Eureka               | G15        | S54                    | Blueberry-64            | B64                              | P2/L1            | Replicate Biological<br>Library2/Bio2-rep #7                                                                                                                                                         | G15/S54/B64/P2/L1/Bio2-rep<br>#2      |
| Eureka               | G15        | S55                    | Blueberry-65            | B65                              | P3/L1            | Replicate Biological<br>Library2/Bio2-rep #8                                                                                                                                                         | G15/S55/B65/P3/L1/Bio2-rep<br>#3      |
| Eureka               | G15        | S56                    | Blueberry-66            | B66                              | P4/L1            | Replicate Biological<br>Library2/Bio2-rep#9                                                                                                                                                          | G15/S56/B66/P4/L1/Bio2-rep<br>#4      |
| Eureka               | G15        | S57                    | Blueberry-67            | B67                              | P5/L1            | Replicate Biological<br>Library2/Bio2-rep #10                                                                                                                                                        | G15/S57/B67/P5/L1/Bio2-rep<br>#5      |
| Masena               | G10        | S40                    | BBDNA-01                | B68                              | P1/L1            | Unknown/Replicate Biological<br>Library3/Bio3-rep #1                                                                                                                                                 | G10/S40/B68/P1/L1/Unk/Bio3-<br>rep #1 |
| Masena               | G10        | S50                    | BBDNA-02                | B69                              | P2/L1            | Replicate Biological<br>Library3/Bio3-rep #2                                                                                                                                                         | G10/S50/B69/P2/L1/Bio3-rep<br>#2      |
| Masena               | G10        | S51                    | BBDNA-03                | B70                              | P3/L1            | Replicate Biological<br>Library3/Bio3-rep #3                                                                                                                                                         | G10/S51/B70/P3/L1/Bio3-rep<br>#3      |
| Masena               | G10        | S52                    | BBDNA-04                | B71                              | P4/L1            | Replicate Biological<br>Library3/Bio3-rep #4                                                                                                                                                         | G10/S52/B71/P4/L1/Bio3-rep<br>#4      |
| Masena               | G10        | S53                    | BBDNA-05                | B72                              | P5/L1            | Replicate Biological<br>Library3/Bio3-rep #5                                                                                                                                                         | G10/S53/B72/P5/L1/Bio3-rep<br>#5      |
| Eureka               | G15        | S45                    | BBDNA-06                | B73                              | P1/L1            | Replicate Biological<br>Library3/Bio3-rep #6                                                                                                                                                         | G15/S45/B73/P1/L1/Bio3-rep<br>#1      |
| Eureka               | G15        | S54                    | BBDNA-07                | B74                              | P2/L1            | Replicate Biological<br>Library3/Bio3-rep #7                                                                                                                                                         | G15/S54/B74/P2/L1/Bio3-rep<br>#2      |

Table 1. Cont.

| Variety/<br>Cultivar | Genotype # | Sample<br>Collection # | Sequencing File<br>ID # | Blueberry<br>Library Number<br># | Plant and Leaf # | Sample Role/Replicate #<br>(Unknown/Reference: Variety<br>Identification)<br>(Replicate Technical, Replicate<br>Biological Library 1/Library<br>2/Library 3: Variability in SNP<br>Variant Analysis) | Overall Sample Code              |
|----------------------|------------|------------------------|-------------------------|----------------------------------|------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------|
| Eureka               | G15        | S55                    | BBDNA-08                | B75                              | P3/L1            | Replicate Biological<br>Library3/Bio3-rep #8                                                                                                                                                         | G15/S55/B75/P3/L1/Bio3-rep<br>#3 |
| Eureka               | G15        | S56                    | BBDNA-09                | B76                              | P4/L1            | Replicate Biological<br>Library3/Bio3-rep#9                                                                                                                                                          | G15/S56/B76/P4/L1/Bio3-rep<br>#4 |
| Eureka               | G15        | S57                    | BBDNA-10                | B77                              | P5/L1            | Replicate Biological<br>Library3/Bio3-rep #10                                                                                                                                                        | G15/S57/B77/P5/L1/Bio3-rep<br>#5 |

#, number. Unknown: Masena S40 was used as the unknown sample for variety identification; Replicate Technical: replicate analysis using technical replicate samples from the same plant; Replicate Biological1 or Biological2 or Biological3, replicate analysis using biological replicate samples from different plants and DNA processed to generate sequencing Library1 or Library2 or Library3. All sample libraries, excluding Library3 samples, were sequenced (150 bp PE reads) across six lanes of a 10B flow cell of a NovaSeq (Illumina, San Diego, CA, USA). Library3 samples were sequenced in a separate lane of a10B flow cell.

## 2.2. Structuring of Samples for Varietal, Technical and Biological Library Preparation

For the adequate traceability of sample collection, each sample collection was labeled with extensive information as nested codes (Table 1). These nested codes included a range of sample codes, involving leaf collections (S#s) and corresponding identification #s for genotype (G), blueberry library (B), plant (P) and leaf (L) (Table 1). For ease of understanding, sample collection codes (S#s) or blueberry library codes (B#s) were used throughout the manuscript.

Additional details are provided in the Supplementary Materials.

## 2.3. DNA Extraction

The finely ground leaf samples were subjected to DNA extraction using an established method [63]. Essentially, between 2 and 3 g of leaf samples, ground under cryogenic conditions, was added to a warm (65 °C) mixture of 8 ml of 2% CTAB and 2 mL of 5% Sarkosyl contained in a 50 mL tube and mixed by gently inverting the tube 20 times. After incubating at 65 °C for 30 min, the mixture was subjected to two rounds of chloroform washes and the aqueous phase collected for DNA precipitation by iso-propanol. Isolated DNA was dissolved in 100 µL of 10 mM Tris-HCL buffer pH 8.0. DNA quality was assessed by measuring the absorbance ratios (A260/A280 optimal is 1.9 to 2.0 and A260/A230 optimal is over 2.0) and absorbance profile using the Nanodrop Spectrophotometer (ThermoFisher Scientific, Waltham, MA, USA). The isolated DNA was resolved on a gel to assess the level of DNA breakage (shearing), with the DNA concentration and total amount assessed by comparing the brightness intensity of the resolved DNA to that of a known amount of resolved lambda DNA.

## 2.4. DNA Sequencing and Sequence Data Output

The extracted genomic DNA of all samples, used for variety analysis and variability in SNP variant analysis, were subjected to the “Illumina DNA PCR-free Prep” indexed library preparation and then subjected to short read Illumina sequencing at 150 bp paired end (PE) reads on the Illumina NovaSeq X sequencing machine (AGRF Ltd., Melbourne, Australia). Sixty-seven libraries, B01 to B67, corresponding to reference, reference/technical replicate/biological replicate set-1, technical replicate, biological replicate set-1 and biological replicate set-2 samples were sequenced across six 10B lanes. Ten libraries, B68 to B77, corresponding to biological replicate set-3 samples, were sequenced in one 10B lane. With a sequence data (150 bp PE reads) output range of 375 Gb/10 B lane, the expected raw sequence data yield per sample of above 33 Gb is equivalent to a coverage of above 55 times (×) of the blueberry monoploid genome size of 600 Mb.

## 2.5. Data Analysis

All genomic sequence data analysis was undertaken using the CLC Genomics Workbench software v. 24.0.1 (Qiagen, Germantown, MD, USA).

## 2.6. Raw Sequence Data Processing and Data Output

Raw sequence data (150 bp PE reads) was trimmed for quality scores at a quality limit of 0.01, resulting in over 99% of the trimmed data having an average PHRED score of over 34. Other parameters for trimming were as follows: trim ambiguous nucleotides is “Yes”; read-through adaptor trimming is “Yes”; discard reads with length is “<50 and >1000 bases”.

## 2.7. Stage-M1: Optimization of Read Mapping Parameters

In Stage-M1, the optimal combination of read length fraction (LF) and read similarity fraction (SF), both basic mapping parameters, was identified. The tool “Map Reads to

Reference” was used to map PE trimmed Illumina reads to the diploid blueberry genome sequence (*Vaccinium darrowii*, NCBI 2021, GCA\_020921065.1\_USDA Vadar 1.0, Primary assembly 582,668,408 bases) as a reference sequence.

Additional details are provided in the Supplementary Materials.

#### 2.8. Stage-M2: Use of Optimized Read Mapping Parameters and Correction of Mapped Reads

Stage-M2 involved using the optimized mapping length fraction and similarity fraction parameters (identified from Stage-M1) and mapping the PE trimmed reads of each of the blueberry samples to the reference genome sequence, followed by four sequential steps (Step-1 to Step-4), using as many analysis tools to correct and improve the mapping as used for the reads mapped at Stage-M1.

Additional details are provided in the Supplementary Materials.

#### 2.9. Stage-V1: Optimization of Variant Detection Parameters

In Stage-V1, the optimal combination of variant (allele) detection parameters, “Minimum Frequency” (MF) and “Minimum Count” (MC) of the variant allele was identified.

Additional details are provided in the Supplementary Materials.

#### 2.10. Stage-V2: Use of Optimized Variant Detection Parameters on All Samples

Stage-V2 involved applying the optimized FPV detection MF and MC parameters to all the blueberry samples’ trimmed sequence files to identify Hom-SNPs and Het-SNPs. Variant allele detection parameters that were not changed were the same as those used in Stage-V1. The variants identified were filtered to identify Hom-SNPs at a minimum read coverage of 20 and an allele frequency of 100%, as outlined in Stage-V1. The variants identified were also filtered to identify Het-SNPs.

#### 2.11. Stage-V3: Paired Variant Comparison Analysis with the Unknown and a Reference Sample

The blueberry sequence data file and the Hom-SNPs file of Masena B40, corresponding to leaf sample collection S40, sample library B40 and overall sample code of Masena/G10/S40/B40/P1/L1/Ref-rep#1/Tec-rep#1/Bio1-rep#1 (Table 1), was treated as the unknown (Unk) for the Paired Variant Comparison (PVC) analysis. The Hom-SNPs files of all other blueberry samples, corresponding to library #s B01 to B39 and B45 were treated as reference samples. The blueberry variety Masena used as a reference sample corresponded to leaf sample collection S-50, sample library B50 and overall sample code Masena/G10/S50/B50/P2/L1/Bio1-rep#2 (Table 1).

At this Stage, the Paired Variant Comparison analysis was undertaken pairwise using pairs of Hom-SNPs files: the Hom-SNPs file of the unknown blueberry (Masena B40) and the Hom-SNPs file of each of the blueberry reference samples.

Additional details are provided in the Supplementary Materials.

#### 2.12. Stage-V4: Subclassification of Not-Common (Putative Unique) Hom-SNPs Post PVC-Analysis

Actual mapping scenarios can be envisaged, leading to the possibility that the Not-common Hom-SNPs are not all unique (putative Unique), as they can contain both True-Unique and Data-noise Hom-SNPs (putative Unique = True-Unique&Data-noise Hom-SNPs). Two of the mapping scenarios lead to the identification of the True-Unique Hom-SNPs, while three of the mapping scenarios lead to the identification of the Data-noise Hom-SNPs. Hence, the Not-common Hom-SNPs in the putative Unique = True-Unique + Data-noise-Hom-SNPs-(Unk&Ref-B# PVC)-File and in the (Ref-B)-putative-Unique = True-Unique + Data-noise-Hom-SNPs-(Unk&Ref-B# PVC)-File were interrogated for the five possible mapping scenarios to explain why the Not-common Hom-SNPs are called in

one sample and absent in the other, and how many of these are True-Unique Hom-SNPs and Data-noise Hom-SNPs.

Additional details are provided in the Supplementary Materials.

#### *2.13. Stage-V5: Combining PVC Hom-SNPs Files of All Unknown and Reference Pairs and Filtering to Remove Data-Noise*

At this Stage, a combined-PVC Hom-SNPs file is generated without or after filtering out the data-noise.

Additional details are provided in the Supplementary Materials.

#### *2.14. Stage-V6: Variety Identification Using the Combined-PVC Hom-SNPs with/Without Data-Noise*

The variety identification process was applied to all three combined Hom-SNPs-Combined-PVC files generated in Stage-V5, “File-A (Combined Hom-SNPs)”, File-A (Combined Hom-SNPs)-minus-File-B(Data-noise:\*C, S1&S2) and File-A (Combined Hom-SNPs)-minus-File-B(Data-noise:\*C, S1&S2)-minus- File-C (Combined Hom-SNPs Persisting-Data-noise:F1&F2) (Figure 4). The Variety identification protocol, “Protocol 1”, outlines the method development for variety identification, including the concept around identifying the Data-noise and the Persisting-data-noise leading to variety identification. Variety identification “Alternate Protocol 2” and “Alternate Protocol 3”, may also be implemented to confirm the same outcome as obtained from variety identification “Protocol 1” towards variety identification (Figures 4 and S13).

The analysis relies on identifying Hom-SNPs uniquely shared between the unknown sample Masena B40 and any other reference sample. For example, uniquely shared Hom-SNPs between the unknown sample (Masena B40) and the Masena reference sample (B50) can be obtained by using the following filters: Sample count is “2”, Origin Tracks contains “–40” and Origin Tracks contains “–50”.

#### *2.15. Calculations of Call Quality Metrics Pairwise Between B50 and an Unknown Sample*

Application of the variety identification protocol correctly identified ten unknown samples to be the Masena variety and closest only to the B50 Masena reference and not any of the other 40 reference samples. We calculated “Sensitivity”, “Positive Predictive Value” and “F-score” as quality metrics [55,58,64] to qualify the Hom-SNPs of the ten unknown samples compared to the B50 treated as the Masena reference sample. Samples B40, B51-B53, B58–62 and B68 were the unknown samples. All analysis was undertaken using tools in CLC Genomics. Total positives (TPs), Hom-SNPs correctly detected in the unknown sample compared to the reference sample B50, were identified by filtering out Hom-SNP calls shared between the unknown and the reference sample. False negatives (FNs), Hom-SNPs present in the reference B50 sample but not called in the unknown sample, were identified by filtering Hom-SNPs in the reference B50 sample with no exact match in the unknown sample. False positive (FPs), Hom-SNPs present in the unknown but not in the reference sample, were identified from the Mutation test (Protocol 1 or Alternate 2), and represent the U2 (persisting Data-noise) SNPs in B50 from the comparison between B50 and the unknown sample. Sensitivity or Recall was calculated using the formula =  $TP/(TP + FN)$ . Precision or Positive Predictive Value (PPV) was calculated using the formula =  $TP/(TP + FP)$ . F-score, the harmonic mean of precision and recall, providing a balanced measure of accuracy, was calculated using the formula =  $2 \times (Precision \times Sensitivity)/(Precision + Sensitivity)$ . Accuracy was calculated =  $(TP + TN)/(TP + FP + TN + FN)$ . TN was deduced as either a) the total number of heterozygous SNPs in the reference sample B50 or b) the reference genome assembly length in bp minus the total number of SNPs detected in the

reference sample B50. Reference genome assembly used (*Vaccinium darrowii*, NCBI 2021, GCA\_020921065.1\_USDA Vadar 1.0, Primary assembly 582,668,408 bases).

3. Results

3.1. DNA Quality

The purity of the DNA, assessed by examining the absorbance ratios at A260/230 and A260/230 (Table S1), was found to be within the acceptable range for all but a few samples. DNA fragmentation assessed by resolving the DNA on an agarose gel indicated the DNA to be mainly of high molecular weight (Figure S1). These quality metrics were acceptable for all the DNA samples, as they were successfully sequenced on the Illumina sequencing platform to generate 150 bp paired end reads.

3.2. Consideration of Genome Coverage for Detecting High-Confidence Homozygous SNPs

A heterozygous SNP (Het-SNP) position represented by two different nucleotides at a position can be identified by a variant caller at a read depth as low as two and where both nucleotides are represented (assuming no errors in sequencing). However, the sequencing of DNA/library fragments with these alleles is random, and where a small number of DNA/library fragments are available to be sequenced, the same allele can by chance be sampled for sequencing each time, leading to a false assessment of homozygosity (Table 2). This means that a Het-SNP position can be erroneously called a homozygous SNP (Hom-SNP), where one of the two nucleotides is represented in all reads (assuming no errors in sequencing). A higher read depth results in a lower probability of repeatedly sequencing the same allele and leads to higher confidence variant calls. A high read mapping depth was used to increase the probability of correctly identifying a Het-SNP and consequently not erroneously identifying it to be a Hom-SNP.

**Table 2.** Numbers of false calls of homozygous differences due to chance sequencing of the same allele at a heterozygous position in a diploid.

| Genome Size            | Heterozygosity | Sequence Coverage |     |     |
|------------------------|----------------|-------------------|-----|-----|
|                        |                | 10-               | 20- | 25- |
| Small genome (500 Mbp) | 1%             | 4883              | 5   | 0   |
|                        | 3%             | 14,648            | 15  | 0   |
| Large genome (2 Gbp)   | 1%             | 19,531            | 19  | 1   |
|                        | 3%             | 58,593            | 57  | 2   |

For a given genome size (Mb) and % heterozygosity, the number of false homozygous SNP calls, a function of mapped read depth (sequence coverage), can be computed as follows. At a sequence coverage of 10, 20 or 25, there are 1024, 1,048,576 or 33,554,432 read mapping scenarios, respectively, per heterozygous SNP site, which equates to one erroneous SNP call out of 1024, 1048.576 or 33,554,432 heterozygous SNPs, respectively. If the genome size in Mbp (g) and % heterozygosity (h) is known, then at a given sequence depth of, for example, 10, the probable number of false homozygous calls can be calculated: (g) × (h)/number of erroneous homozygous calls at a sequence depth of 10.

Identification of high-confidence Hom-SNP calls relies on increasing the read mapping depth. At a read mapping depth of one, there is a 100% probability of erroneously identifying a Het-SNP as a Hom-SNP or a no-variant call made if the nucleotide allele is the same as in the reference genome sequence, which progressively reduces with increasing read mapping depth (Table S2). At a read mapping depth of two or three reads, the probability of erroneously identifying a Het-SNP as a Hom-SNP reduces to 33.33% and 14.29%, respectively (Table S2). Increasing the mapping read depth to 15 ensures a near-100% probability of correctly identifying a Het-SNP, with 31 Het-SNP incorrectly identified as Hom-SNPs/million Het-SNP sites (Table S3). Increasing the mapping read

depth to 21 ensures an effective 100% probability of correctly identifying a Het-SNP, with less than one Het-SNP incorrectly identified as Hom-SNPs/million Het-SNP sites (Table S3). Conversely, a read mapping depth of 21 also ensures a 100% probability of correctly identifying a Hom-SNP (Figure 1). Based on this data, we sequenced the blueberry samples at a minimum expected raw data depth of  $55\times$  of the monoploid genome size of 600 Mb, equating to 33 Gb, to achieve a minimum read mapping depth of 20.

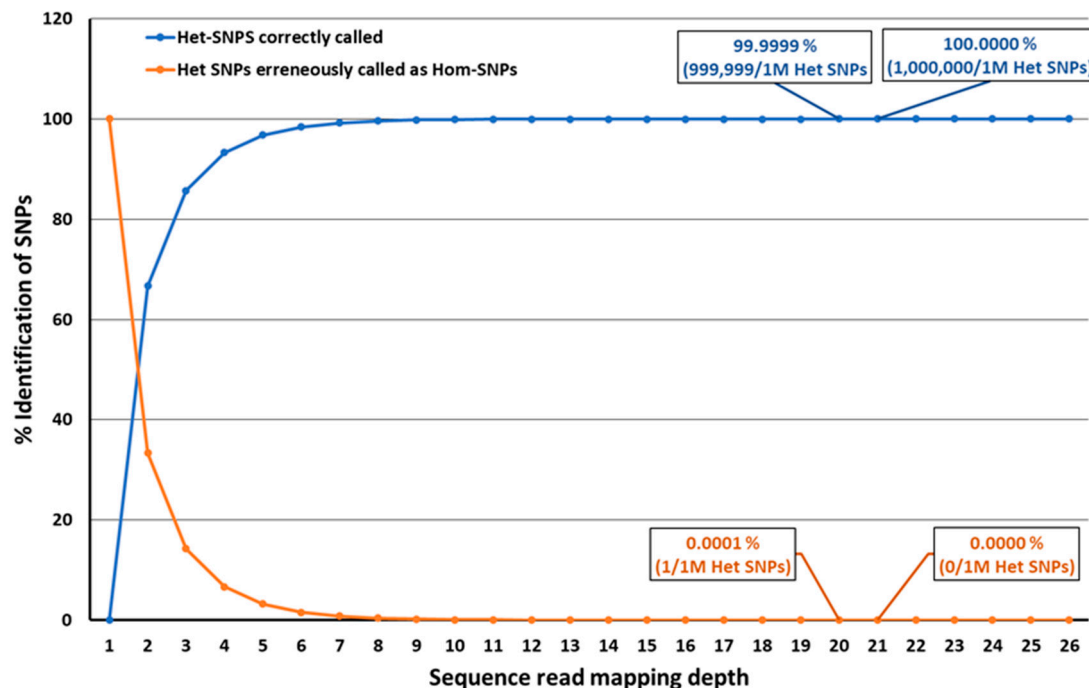


**Figure 1.** Probability (%) of heterozygous SNPs correctly called or erroneously identified as homozygous SNPs as a function of read mapping depth. At a read mapping depth of 21 and 1,048,576 read mapping scenarios, 100% of heterozygous SNPs are correct and 0% are erroneously called homozygous SNPs. At read mapping depth of 20, for every 1 million heterozygous SNPs, 1 is erroneously called a homozygous SNP.

### 3.3. Raw Sequence Data Quality and Size as Times Genome Size of 600 Mb

The quality of the raw sequence data was high, with over 90% of the sequence reads having an average PHRED quality score of above 30 (Table S4). The sequence data size of all samples, based on a genome size of 600 Mb ( $1\times$ ), obtained at a minimum of  $32.4\times$  (19.5 Gb) and a maximum of  $83.84\times$  (50.3 Gb) at an average of  $52.7\times$  (31.6 Gb), was above the expected data size of 33 Gb ( $55\times$ ) for most samples. Post trimming, more than 80% of the sequence data was recovered at an average PHRED score of above 37, and was at a minimum of  $26.8\times$  (16.1 Gb) and a maximum of  $70.8\times$  (42.5 Gb), which was above the expected data size of 33 Gb ( $55\times$ ) for most samples. Box plots of distribution of PHRED scores and read length of trimmed data, for all samples and sample groups indicated PHRED scores above 37, and 85% of the reads were above 120 bp (Table S4). The data size of the paired end component of the trimmed data, important for variant calling, was at a minimum of  $25.4\times$  (15.2 Gb), a maximum of  $67.62\times$  (40.6 Gb) and an average of  $42.1\times$  (25.3 Gb), which is above the minimum requirement of  $20\times$  (12 Gb) (Figure S2 and Table S4).

### 3.4. Optimal Read Mapping and Variant Calling Parameters

Titration of read mapping parameters, involving length and similarity fractions (LF and SF), including the implementation of additional mapping filters, indicated both the LF and SF at 0.9 to be the optimal settings (Figure S3A–E and Table S5).

Additional details are provided in the Supplementary Materials.

### 3.5. Mapping Coverage at Optimal Read Mapping Settings

Utilizing the optimal mapping setting of 0.9 each for LF and SF, the average read mapping coverage, excluding the regions with no mapped reads, was found to be above the desired 20 $\times$  for all the blueberry samples except three of the reference samples (B24, B34 and B37) and two samples (B59 and B69) used for replication analysis (Figure 2). The fraction of the reference genome sequence with mapped reads was above 0.86 for all samples tested (Figure S4), indicating high coverage of the reference genome sequence with mapped PE reads.

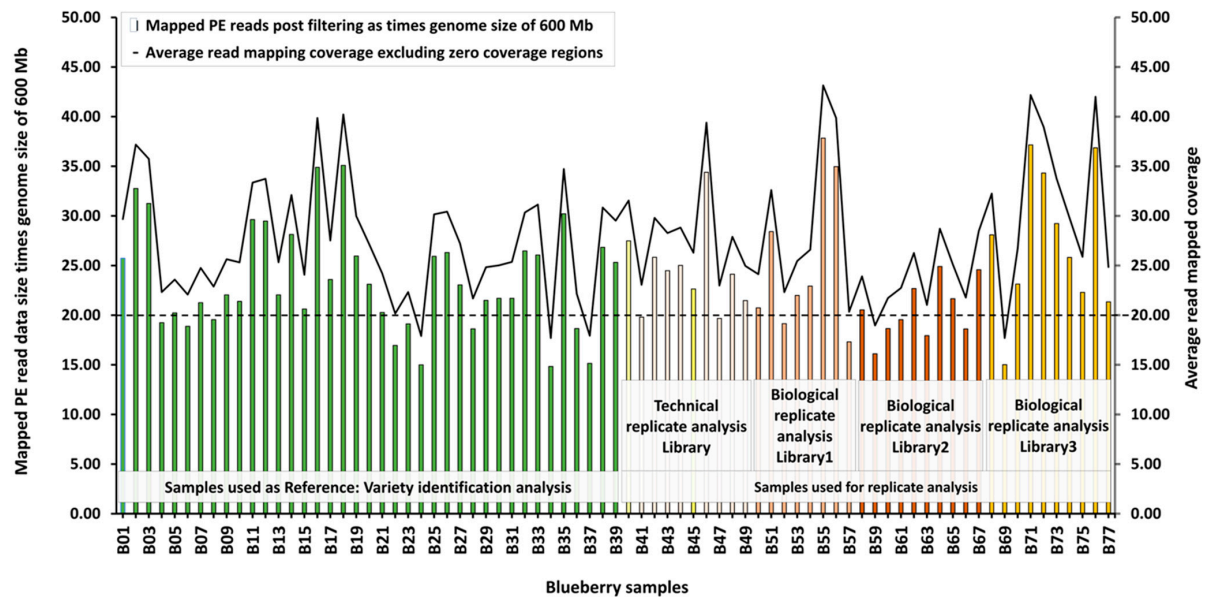


**Figure 2.** Mapped paired end (PE) trimmed read data size as times genome size of 600 Mb and as coverage depth of blueberry samples. The trimmed paired end (PE) reads were mapped to the genome sequence of *Vaccinium darrowii* (NCBI, GCA\_020921065.1) with optimized mapping settings and pipeline to address the mis-mapping of PE reads. The data size of the input PE reads for mapping shown as times ( $\times$ ) genome size of 600 Mb was lower than 20 $\times$  for some samples used as reference samples and used as technical and biological replicates. However, the average read mapping coverage, excluding the regions with zero coverage, was above 20 for all samples except three of the reference samples and the two samples Masena/G10/S50/B59/P2/L1/Bio2-rep#2 and Masena/G10/S50/B69/P2/L1/Bio3-rep#2 used for replicate analysis. Two samples within the technical replicate set, represented as yellow bars, were used for variety analysis, technical replicate analysis and biological replication analysis. Average read-mapped coverage from 20 is marked with a horizontal broken black line.

### 3.6. Identification of Variants

Het-SNPs were identified at a minimum read mapping coverage of 10, minor and major allele frequency range of 20% to 80%, respectively, and at a minimum variant allele read count of 3. The number of Het-SNPs ranged between 4.8 and 6.9 million (Figure S5A) and was strongly correlated with the sequence data size used for the variant analysis (Figure S5B). Hom-SNPs were filtered at a minimum read coverage of 20, at 100% allele frequency and at a minimum-variant read count of 5. The number of Hom-SNPs ranged between 0.7 M and 2.1 M (Figure S5A) and strongly correlated with the sequence data size used for the variant analysis (Figure S5C). Results also indicate that increasing the sequence data size above 45 $\times$  leads to diminishing returns in terms of sequencing cost versus aiming

to capture as many as possible or all of the Het-SNPs or Hom-SNPs present in the blueberry genome of a specific sample (Figure S5B,C).

### 3.7. Variability in Hom-SNPs Identified Within Technical and Biological Replicate Samples

Two blueberry samples, Masena and Eureka, were selected for technical and biological replication analysis. Technical replicates, P1L1 to P1L5, were represented as DNA extracted from five leaf samples from a single plant (Table 1). Biological replicates, P1L1 to P5L1, corresponded to DNA extracted from five leaf samples where each leaf sample was from a different plant (Table 1). The same DNA sample was subjected to three biological libraries, where library 1 and library 2 were sequenced in the same lane, while library 3 was sequenced in a separate lane. To assess variation in Hom-SNPs between samples of common library preparations, for each of the two samples for replication analysis, pairwise variant comparison (PVC)-analysis for Hom-SNPs was undertaken between 10 pairs of the five technical library samples and likewise between the five samples within each of the three biological libraries. Similarly, to assess variation in Hom-SNPs between samples of different library preparations, for each of the two samples for replication analysis, PVC-analysis for Hom-SNPs was undertaken between five pairs of the five samples between each of the three biological libraries.

The PVC-analysis between all the Masena samples and between all the Eureka samples identified both Common Hom-SNPs and Not-common Hom-SNPs (Table 3). The Common Hom-SNPs for Masena and Eureka technical library replicates samples were as high as 78.6% and 75.6%, respectively (Table 3, Figures S6A and S7A), with an average of  $73.6 \pm \text{SD } 4.9$  and  $70.2 \pm \text{SD } 4.1$ , respectively, indicating low variation between each of the ten paired comparisons for both the samples (Tables S5 and S7). In addition, the PVA analysis between the Masena technical replicate sample, Blueberry B41, and all other Masena technical replicate samples resulted in a consistently lower number of Common Hom-SNPs (Figure S6A) resulting possibly from mapped PE trimmed data size of under  $20 \times$  ( $1 \times$  is 600 Mb) and a read mapping coverage of under 25 (Figure 2).

The Common Hom-SNPs for Masena and Eureka Biological Library 1 replicate samples were as high as 79.2% and 82.6%, respectively (Table 3, Figures S6B and S7B), with an average of  $70.3 \pm \text{SD } 4.1$  and  $65.6 \pm \text{SD } 10.6$ , respectively, indicating low variation between each of the ten comparisons for the Masena but not the Eureka samples (Tables S6 and S7). The PVC-analysis between the Masena Biological library 1 sample Blueberry B52 and all other Masena library 1 replicate samples consistently resulted in low numbers of Common Hom-SNPs (Figure S6B), while for the Eureka sample it was Blueberry B57 (Figure S7B). Both the samples, Blueberry B52 and Blueberry B57, had a mapped PE trimmed data size of under  $20 \times$  ( $1 \times$  is 600 Mb) and a read mapping coverage of under and just above 20, respectively (Figure 2).

The Common Hom-SNPs for Masena and Eureka Biological library 2 replicates samples were as high as 72.1% and 76.5%, respectively (Table 3, Figures S6C and S7C), with an average of  $62.4 \pm \text{SD } 5.8$  and  $67.5 \pm \text{SD } 5.0$ , respectively, indicating low variation between each of the ten comparisons for the Masena and the Eureka samples (Tables S6 and S7). The PVC-analysis between the Masena Biological library 2 sample, Blueberry B59, and all other Masena library 2 replicate samples consistently resulted in low numbers of Common Hom-SNPs (Figure S6C), while for the Eureka sample set it was Blueberry B63 (Figure S7C). Both the samples, Blueberry B59 and Blueberry B63, had a mapped PE trimmed data size of under  $20 \times$  ( $1 \times$  is 600 Mb) and a read mapping coverage under 20 and 25, respectively (Figure 2).

**Table 3.** Percentage of Common and Not-common Hom-SNPs from PVC-analysis of the two blueberry control samples.

| Sample Library Used                          | Details of PVC-Analysis  |                           | Control Sample # | Number of Pairwise Comparisons | Masena vs. Masena Controls |         |       |      | Eureka vs. Eureka Controls |         |       |      |
|----------------------------------------------|--------------------------|---------------------------|------------------|--------------------------------|----------------------------|---------|-------|------|----------------------------|---------|-------|------|
|                                              | Samples Within A Library | Samples Between Libraries |                  |                                | Hom-SNPs                   |         |       |      | Hom-SNPs                   |         |       |      |
|                                              |                          |                           |                  |                                | Highest Value              | Average | ±SD   | ±SE  | Highest Value              | Average | ±SD   | ±SE  |
| Technical Library                            | yes                      | no                        | 1 to 5           | 10                             | 78.55                      | 73.59   | 4.85  | 1.54 | 75.61                      | 70.15   | 4.14  | 1.31 |
|                                              |                          |                           |                  |                                | 32.79                      | 26.41   | 4.85  | 1.54 | 38.99                      | 29.84   | 4.14  | 1.31 |
| Biological Library 1                         | yes                      | no                        | 1 to 5           | 10                             | 79.24                      | 70.28   | 4.05  | 1.28 | 82.57                      | 65.63   | 10.60 | 3.35 |
|                                              |                          |                           |                  |                                | 35.42                      | 29.72   | 4.05  | 1.28 | 51.67                      | 34.35   | 10.61 | 3.36 |
| Biological Library 2                         | yes                      | no                        | 1 to 5           | 10                             | 72.08                      | 62.40   | 5.83  | 1.84 | 76.51                      | 67.54   | 5.02  | 1.59 |
|                                              |                          |                           |                  |                                | 44.72                      | 37.60   | 5.83  | 1.84 | 38.37                      | 32.46   | 5.02  | 1.59 |
| Biological Library 3                         | yes                      | no                        | 1 to 5           | 10                             | 81.73                      | 61.04   | 18.86 | 5.96 | 79.42                      | 72.47   | 4.56  | 1.44 |
|                                              |                          |                           |                  |                                | 64.60                      | 38.96   | 18.86 | 5.96 | 35.45                      | 27.53   | 4.56  | 1.44 |
| Biological Library 1 and 2                   | no                       | yes                       | 1 to 5           | 5                              | 73.20                      | 64.98   | 7.42  | 3.32 | 75.48                      | 63.91   | 6.89  | 3.08 |
|                                              |                          |                           |                  |                                | 42.91                      | 35.02   | 7.42  | 3.32 | 41.67                      | 36.09   | 6.89  | 3.08 |
| Biological Library 1 and 3                   | no                       | yes                       | 1 to 5           | 5                              | 78.97                      | 65.38   | 12.44 | 5.56 | 82.25                      | 71.93   | 8.12  | 3.63 |
|                                              |                          |                           |                  |                                | 52.45                      | 34.62   | 12.44 | 5.56 | 37.92                      | 28.07   | 8.12  | 3.63 |
| Biological Library 2 and 3                   | no                       | yes                       | 1 to 5           | 5                              | 71.23                      | 62.27   | 9.13  | 4.08 | 76.61                      | 67.21   | 9.59  | 4.29 |
|                                              |                          |                           |                  |                                | 51.10                      | 37.73   | 9.13  | 4.08 | 43.23                      | 32.79   | 9.59  | 4.29 |
| Biological Library 1 and 2, 1 and 3, 2 and 3 | no                       | yes                       | 1                | 3                              | 78.97                      | 73.05   | 5.13  | 2.96 | 73.93                      | 64.72   | 8.60  | 4.96 |
|                                              |                          |                           |                  |                                | 30.10                      | 26.95   | 5.13  | 2.96 | 43.08                      | 35.28   | 8.60  | 4.96 |
|                                              | no                       | yes                       | 2                | 3                              | 57.09                      | 51.18   | 5.16  | 2.98 | 76.61                      | 75.98   | 0.58  | 0.33 |
|                                              |                          |                           |                  |                                | 52.45                      | 48.82   | 5.16  | 2.98 | 24.52                      | 24.02   | 0.58  | 0.33 |
|                                              | no                       | yes                       | 3                | 3                              | 73.16                      | 64.51   | 8.01  | 4.62 | 72.77                      | 67.24   | 4.91  | 2.84 |
|                                              |                          |                           |                  |                                | 42.65                      | 35.49   | 8.01  | 4.62 | 36.61                      | 32.76   | 4.91  | 2.84 |
|                                              | no                       | yes                       | 4                | 3                              | 66.98                      | 61.29   | 4.93  | 2.85 | 82.25                      | 65.78   | 14.28 | 8.25 |
|                                              |                          |                           |                  |                                | 41.72                      | 38.71   | 4.93  | 2.85 | 43.23                      | 34.22   | 14.28 | 8.25 |
|                                              | no                       | yes                       | 5                | 3                              | 73.20                      | 71.02   | 2.29  | 1.32 | 72.99                      | 64.70   | 7.34  | 4.24 |
|                                              |                          |                           |                  |                                | 31.38                      | 28.98   | 2.29  | 1.32 | 40.98                      | 35.30   | 7.34  | 4.24 |

#, number; PVC, Paiwise Variant Comparison analysis; Hom-SNPs, Homozygous SNPs. Common Hom-SNPs are in blue text and Not-common Hom-SNPs are in black text. Hom-SNPs with large standard deviation (SD) values are underlined.

The Common Hom-SNPs for Masena and Eureka Biological library 3 replicates samples were as high as 81.73% and 79.42%, respectively (Table 3, Figures S6D and S7D), with an average of  $61.04 \pm \text{SD } 18.86$  and  $72.47 \pm \text{SD } 4.56$ , respectively, indicating high variation between each of the ten PVC-analyses for the Masena samples but low for the Eureka samples (Tables S6 and S7). The PVC-analysis between the Masena Biological library 3 sample, Blueberry B69, and all other Masena Biological library 3 replicate samples consistently resulted in low numbers of Common Hom-SNPs (Figure S6D), and this sample had a mapped PE trimmed data size of under  $20\times$  ( $1\times$  is 600 Mb) and a read mapping coverage below 20 (Figure 2).

These results indicate that a mapped read data size below  $20\times$  ( $1\times$  is 600 Mb) and especially a read mapping depth below 20 can have an impact on maximizing the capture of SNP variants using NGS data. In a few of the samples used for replication analysis, the Hom-SNPs, at best ranging from 20% to 40%, and at worst at 60%, are captured as Not-common in the pairwise analysis of the same variety, indicating that not all of the Hom-SNPs are captured consistently between the replicates (Table 3).

### 3.8. Variability in PVC-Analysis Generated Hom-SNPs Between Biological Replicates

Within each of the three biological libraries, each of the five Masena samples were compared pairwise for Common Hom-SNPs, and this was likewise undertaken for the biological libraries of the Eureka samples (Figures S6E–G and S7E–G). For the Masena and Eureka samples, the Common Hom-SNPs from the PVC-analysis showed more variation for the samples within the biological replicate library 3 and biological replicate library 1, respectively, than from samples within the other two biological replicate libraries or from the technical replicate library (Tables S6 and S7). Comparison of samples 1 to 5 for Common Hom-SNPs between biological libraries (Library 1, 2 and 3) for the Masena and Eureka sample set showed more variation for Masena sample 3 and Eureka sample 4 (Tables S6 and S7).

### 3.9. Common and Not-Common Hom-SNPs Between Masena and Eureka Samples

PVC-analysis between two of the same blueberry variety, Masena samples B40 and B50 (biological replicates), identified 70.5% as Common Hom-SNPs and 29.5% as Not-common Hom-SNPs (Figure 3A). Similar results were obtained from the analysis between all the Masena samples (all 55 combinations of Masena sample B40 within technical replicates, within the biological replicates and between biological replicates) where 66.1% (average) were Common Hom-SNPs and 33.9% (average) were Not-common Hom-SNPs (Figure S8). Likewise, similar results were also obtained from PVC-analysis between all the Eureka samples, (all 55 combinations of Eureka sample B45 within technical replicates, within the biological replicates and between biological replicates), where 68.6% were Common Hom-SNPs and 31.4% were Not-common Hom-SNPs (Figure S8).

PVC-analysis between two different blueberry varieties, Masena sample B40 and Eureka sample B45, identified 38.5% as Common Hom-SNPs and 61.5% as Not-common Hom-SNPs (Figure 3A), and similar results were obtained between all the Masena and Eureka samples (all 100 combinations within technical replicates and within the biological replicates) where 35.9% (average) were Common Hom-SNPs and 64.1% (average) Not-common Hom-SNPs (Figure S8). Similar results were obtained from PVC-analysis between Masena B40 and each of the 39 different blueberry varieties, where 39.5% (average) were Common Hom-SNPs and 60.5% (average) Not-common Hom-SNPs (Figure S9).

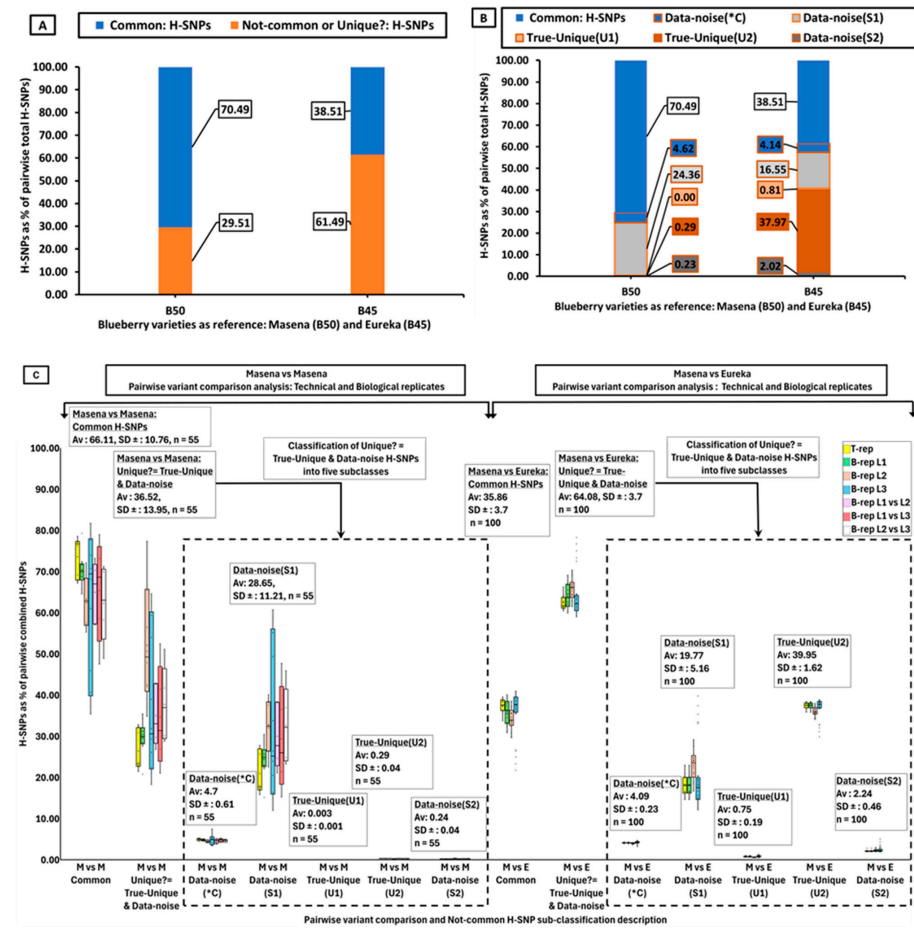


**Figure 3.** PVC-analysis between blueberry samples to identify Common, Not-common or putative Unique (Unique?) and sub-classes of Not-common Hom-SNPs. PVC, pairwise variant comparison analysis between a pair of samples; H-SNPs/Hom-SNPs, homozygous single nucleotide polymorphism; Common Hom-SNPs, Hom-SNPs present in both samples; M, Masena sample; E, Eureka variety; Not-common or putative Unique (Unique?); Hom-SNPs present in one of the sample with no variant call in the other although implied to be unique, with the possibility that some but not all are unique (U?); Data-noise (\*C), probably Common Hom-SNP as mapped at read depth  $\geq 20$  but failure of other variant quality filters; Data-noise (S1), spurious type 1 due to failure of mapped reads depth or mapped read depth and other variant quality filters; True-Unique(U1), True unique type 1 as variant is homozygous same as reference allele; True-Unique(U2), True unique type 2 as variant is heterozygous SNP; Data-noise(S2) indicating spurious type 2 (S2) due to the SNP call linked with indels or no reads mapped. As the Not-common or putative Unique Hom-SNPs are not all unique and divided into two groups of five subclasses. The True-Unique group comprises two subclasses, the True-Unique(U1) and True-Unique(U2). The Data-noise group comprises three subclasses, the Data-noise(\*C), Data-noise(S1) and Data-noise(S1). Hom-SNPs numbers indicated are a percentage of the total combined Hom-SNPs between a pair of samples taken for the PVC-analysis. (A) PVC-analysis between the unknown sample (Masena B-40) and the Masena reference sample (B-50) and the Eureka reference sample (B45), showing the Common Hom-SNPs to be higher between the B-40 and B-50 PVC and the Not-common or putative Unique Hom-SNPs to be higher between the B-40 and B-45. (B) The major subclasses of putative Unique Hom-SNPs are Data-noise (\*C) and Data-noise (S1) for the unknown (Masena) B-40 and Masena B-50 PVC and are Data-noise (\*C) and Data-noise (S1) and True-Unique (U2) for the unknown (Masena) B-40 and Eureka B-45 PVC. (C) Similar results as in B were obtained from fifty-five PVC analyses between Masena and Masena (technical and biological replicates) and from one-hundred PVC analyses between Masena and Eureka, including the subclasses of the Not-common or putative Unique Hom-SNPs.

The PVC-analysis results indicate that the number of Common Hom-SNPs were more when the pair of samples were the same variety (Masena vs. Masena), while the Not-common Hom-SNPs were more when the pair of samples were different varieties (Masena vs. Eureka). In addition, the PVC-analysis successfully identified the Not-common or Unique Hom-SNPs between two different blueberry varieties, which can be informative SNPs for variety identification. However, the identification of Not-common or Unique Hom-SNPs between two samples that are the same variety indicates that these Hom-SNPs must all be erroneous, as they cannot be Unique but should all be Not-Unique between samples of the same variety. Hence, it is possible that the Not-common Hom-SNPs identified between two samples that are different blueberry varieties may not all be Unique (putative Unique) and may contain both “True-Unique” and “Not-Unique” Hom-SNPs”, the latter representing Data-noise. Identifying the two components of the Not-common or putative Unique Hom-SNPs, the “True-Unique” and “Data-noise” (Not-Unique), and using only the True-Unique Hom-SNPs of the Not-common Hom-SNPs would be crucial in increasing the accuracy of variety identification.

**3.10. Sub-Classification of Not-Common Hom-SNPs Between the Same and Dissimilar Varieties**

Identifying the “True-Unique” and “Data-noise” Hom-SNPs, the two components of the Not-common Hom-SNPs of the same variety, was undertaken by visual inspection of reads mapping at locations corresponding to the Not-common or putative Unique Hom-Figure SNPs. This process led to the identification of read mapping scenarios in one sample, which led the variant caller to identify the Hom-SNPs as Not-common in the other. Five read mapping scenarios were identified, leading to five subclasses of the Not-common or putative Unique Hom-SNPs (Table 4), where three are grouped as “True-Unique” and two as “Data-noise” Hom-SNPs.

**Table 4.** Read mapping scenarios in one sample leading to five subclasses of Not-common SNPs identification in the other sample.

| Sub-Classes of Putative Unique = True-Unique and Data-Noise | Call Made by the Variant Caller | Sample 1 Sequence Reads Mapped and the Variant Is as Follows: |                                   |                                    |          |               | No Reads Mapped in Sample 1 at Locations of H-SNPs in Sample 2 |
|-------------------------------------------------------------|---------------------------------|---------------------------------------------------------------|-----------------------------------|------------------------------------|----------|---------------|----------------------------------------------------------------|
|                                                             |                                 | The Same H-SNP as in Sample 2                                 |                                   | Not the Same H-SNP in Sample 2     |          |               |                                                                |
|                                                             |                                 | Satisfies Required Read Depth                                 | Satisfies Variant Quality Filters | H-SNP Same as the Reference Allele | Het-SNPs | MNV or InDels |                                                                |
| Data-noise (*C)                                             | no                              | yes                                                           | no                                |                                    |          |               |                                                                |
| Data-noise (S1)                                             | no                              | no                                                            | yes                               |                                    |          |               |                                                                |
|                                                             | no                              | no                                                            | no                                |                                    |          |               |                                                                |
| True-Unique (U1)                                            | no                              |                                                               |                                   | yes                                |          |               |                                                                |
| True-Unique (U2)                                            | yes                             |                                                               |                                   |                                    | yes      |               |                                                                |
| Data-noise (S2)                                             | yes                             |                                                               |                                   |                                    |          | yes           |                                                                |
|                                                             | no                              |                                                               |                                   |                                    |          |               | yes                                                            |

H-SNPs, homozygous SNPs; Het-SNPs, heterozygous SNPs; MNV, multi nucleotide variants; putative Unique, Not-Common H-SNPs which in principle should all be Unique (U), but not all are Unique, comprising True-Unique and Data-noise H-SNPs. Data-noise (\*C), probably common H-SNP, as mapped read depth  $\geq 20$  but failure of other variant quality filters; Data-noise (S1), spurious type 1, due to failure of mapped reads depth or mapped read depth and other variant quality filters; True-Unique (U1), True unique type 1, as variant is homozygous and same as the reference allele; True-Unique (U2), True unique type 2, as variant is heterozygous SNP; Data-noise (S2) indicating spurious type 2, due to the H-SNP call linked with indels or no call is made as no reads are mapped.

The True-Unique group, within the Not-common Hom-SNPs, comprises two subclasses, True-Unique (U1) and True-Unique (U2). The read mapping scenario for subclass True-Unique (U1) is defined as reads that are mapped and satisfy the defined read depth

and variant filter settings, but where the identified variant is a Hom-SNP and is the same nucleotide allele as in the reference genome sequence, so the variant calling software does not record it as a variant. The read mapping scenario for the subclass True-Unique (U2) is defined as reads that are mapped and satisfy the read depth and variant filter settings, but the identified variant called is heterozygous, and although recorded by the variant caller, is filtered away when selecting only the Hom-SNPs. The Data-noise group, within the Not-common Hom-SNPs, comprises three subclasses, Data-noise (\*C), Data-noise (S1) and Data-noise (S2). The read mapping scenario for subclass Data-noise (\*C), is defined as, the mapped reads have the same Hom-SNP as in the other sample, and satisfies the read depth but not the variant filters settings resulting in the variant caller not recording the Hom-SNP variant. The read mapping scenario for subclass Data-noise (S1) is defined as the Hom-SNP being the same as in the other sample, where reads are mapped but these either do not pass the variant filters or both the read depth and variant filters, resulting in no variant called by the variant caller. The mapping scenario for the subclass Data-noise (S2) can be defined as either no reads are mapped at the corresponding location of the Hom-SNP, resulting in the variant caller not recording a variant, or reads are mapped, with the variant caller recording a multi nucleotide variant (which may not be reliable) or a heterozygous call involving a deletion (unreliable due to mis-mapped reads at a homopolymer region), and although recorded by the variant caller, the variant filtered away when selecting only Hom-SNPs.

The Not-common or putative Unique Hom-SNPs (29.51%) identified from the PVC-analysis between Masena vs. Masena (Figure 3A), where the two varieties are the same, were classified into the five subclasses as outlined above (Figure 3B). All three subclasses of the Data-noise (\*C, S1 and S2) Not-common Hom-SNPs are present, with Data-noise (S1) (24.36%) and Data-noise (\*C) (4.62%) as the first two major components. Of the True-Unique (U1 and U2) Not-common Hom-SNPs, no Hom-SNPs corresponded to subclass True-Unique (U1) but Hom-SNPs at 0.23% corresponded to the subclass True-Unique (U2), equating to 2300 True-Unique (U2) Hom-SNPs/million Hom-SNPs. Similar results were also obtained when the Not-common or putative Unique Hom-SNPs (average 36.52%) identified from PVC-analysis of all the 55 combinations of Masena vs. Masena (technical and biological libraries) were subclassified into its two components (Figure 3C). All three subclasses of the Data-noise (\*C, S1 and S2) Not-common Hom-SNPs are present, with Data-noise (S1) (average 28.65%) and Data-noise (\*C) (average 4.7%) as the first two major components. Of the two subclasses comprising the True-Unique (U1 and U2) Not-common Hom-SNPs, both subclasses True-Unique (U1) (average 0.003%) and True-Unique (U2) (average 0.29%) were present, equating to 30 True-Unique (U1) Hom-SNPs and 2900 True-Unique (U2) Hom-SNPs, respectively, per million Hom-SNPs (Figure 3C).

The Not-common or putative Unique Hom-SNPs (61.49%) identified from the PVC-analysis between Masena and Eureka (Figure 3A), where the two varieties are different, were classified into the five subclasses as outlined above (Figure 3B). All three subclasses of Data-noise (\*C, S1 and S2) Not-common Hom-SNPs are present, with Data-noise (S1) (16.55%) and Data-noise (\*C) (4.14%) as the first two major components. Of the two subclasses comprising the True-Unique (U1 and U2) Not-common Hom-SNPs, both subclasses True-Unique (U1) (0.81%) and True-Unique (U2) (37.97%) were present, with the latter the major component. Similar results were also obtained when the Not-common or putative Unique Hom-SNPs (average 64.08%) identified from PVC-analysis from all the 100 combinations of Masena vs. Eureka (technical and biological libraries) were subclassified into its two components (Figure 3C). All three subclasses of the Data-noise (\*C, S1 and S2) Not-common Hom-SNPs are present, with Data-noise (S1) (average 19.77%) and Data-noise (\*C) (average 4.09%) Hom-SNPs the first two major components. Of the two subclasses comprising the True-Unique (U1 and U2) Not-common Hom-SNPs, both subclasses True-

Unique (U1) (average 0.75%) and True-Unique (U2) (average 39.95%) Hom-SNPs were present (Figure 3C).

As with the subclasses obtained for the Not-common or putative Unique Hom-SNPs for Masena vs. Eureka comparison, where the two varieties were different (Figure 3B,C), similar results were obtained when the Masena sample was compared pairwise to all the other thirty-nine blueberry varieties (Figure S10). All three subclasses of the Data-noise (\*C, S1 and S2) Not-common Hom-SNPs are present, with Data-noise (S1) and Data-noise (\*C) the first two major components. Of the two subclasses comprising the True-Unique (U1 and U2) Not-common Hom-SNPs, both subclasses of True-Unique (U1) (average 0.75%) and True-Unique (U2) (average 39.95%) Hom-SNPs were present, with the latter the major component (Figure 3C).

The Data-noise Not-common Hom-SNPs, comprising the subclasses Data-noise (\*C), Data-noise (S1) and Data-noise (S2), are the non-informative Hom-SNPs and were filtered out from the Not-common or putative Unique Hom-SNPs before using them for varietal identification.

### 3.11. Variety Identification After Removal of Data-Noise Hom-SNPs

Identification of an unknown variety compared to a specific cohort of selected reference samples is based on identifying the Hom-SNPs that are uniquely shared pairwise between an unknown sample and a specific reference sample and not with any of the other reference samples within the cohort of reference samples used. The detailed process of variety identification, as explained below, involves several steps collectively referred to as Variety Identification: Protocol 1 (Figure 4), with alternate protocols referred to as Variety Identification: Alternate Protocol 2 (Figure 4) and Variety Identification: Alternate Protocol 3 (Figure S13) that can be implemented to confirm results.

We used a suite of forty-one blueberry reference samples including a Masena reference sample B50. PVC-analysis-generated Hom-SNPs files, between the unknown sample and each of the forty-one reference samples, were combined to obtain the “File-A (Combined Hom-SNPs)”. The Hom-SNPs corresponding to the Data-noise (\*C, S1 and S2) present within the Not-common or putative Unique Hom-SNPs between the unknown sample and each of the forty-one reference samples, were also combined to obtain “File-B (Combined Hom-SNPs Data-noise)”. Filtering out the Hom-SNPs in File-B (Combined Hom-SNPs Data-noise) from File-A (Combined Hom-SNPs) generated File-A (Combined Hom-SNPs)-minus-File-B(Data-noise: \*C, S1 and S2).

Using File-A (Combined Hom-SNPs) for identifying the unknown sample variety, the Hom-SNPs shared uniquely between the unknown (Masena B40) and each of the forty-one reference samples indicated the highest number (18,673) to be between the unknown (Masena B40) and the Masena reference B50 (Figure S11). The next highest number of uniquely shared Hom-SNPs is 2593, between the unknown sample and the blueberry sample B18 (not Masena), and for most of the other reference samples, the uniquely shared Hom-SNPs are under 500. Although this result clearly identifies the unknown sample to be the Masena variety, the presence of uniquely shared Hom-SNPs between the unknown (Masena B40) and the reference samples other than the Masena reference B50 is expected to be due to data-noise, as identified between the unknown and all the forty one reference samples used (Figures 3 and S10) and represented in File-B (Combined Hom-SNPs Data-noise), which was not filtered out.

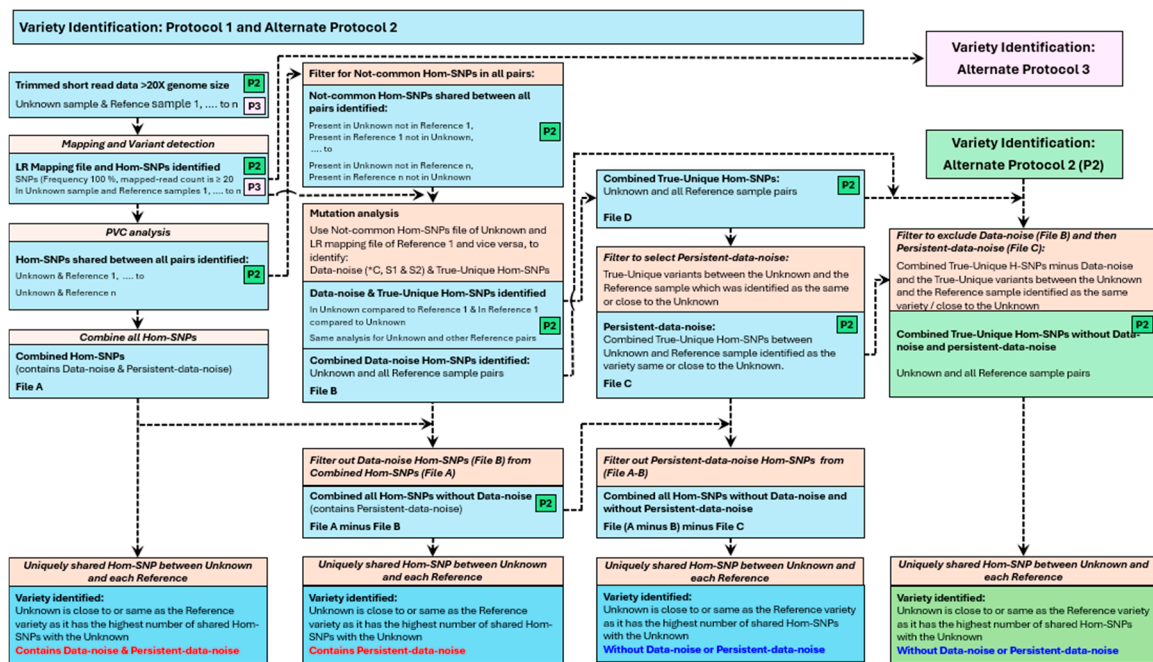


**Figure 4.** Schematic of Variety Identification Protocol 1 and Alternate Protocol 2. Hom-SNPs, Homozygous SNPs; X genome size, time genome size of 600 Mb; LR mapping, outcome of optimized read mapping to a reference genome sequence; PVC-analysis, Paired Variant Analysis between a pair of samples. Schematics for two protocols schemas are presented. Combining Hom-SNPs between the unknown and a reference sample generates Hom-SNPs common to both samples and unique Hom-SNPs present in one sample but not the other. Combining the unique Hom-SNPs from all sample pairs (File-A) can be processed to identify shared Hom-SNPs between the unknown sample and each of the reference samples. If the unknown sample variety matches one of the reference varieties, then this unknown–reference sample pair should be the only pair to share a set of Hom-SNPs unique to this pair of unknown–reference samples. However, using File-A, shared variants between the unknown and each of the reference samples are not unique, as some are also found to be shared Hom-SNPs between the unknown sample and another reference sample. This indicates an anomaly due to data-noise, which we have identified as Data-noise and Persisting-data-noise. When the Data-noise is identified (File-B) and deducted from File-A, data-noise is removed for most of the unknown/reference sample pair comparisons. When the Persisting-data-noise is identified and removed, the data-noise is completely removed, leaving one unknown/reference pair with uniquely shared Hom-SNPs, implying the reference in question to be the same as/close to the unknown sample. An “Alternate Protocol 2” with its steps labeled as P2 in green highlight (P2) is also presented here, leading to the same outcome. An “Alternate Protocol 3”, with some steps indicated here as P3 in pink highlight (P3) with the same outcome can be used but is not explained here (see Figure S13).

When File-A (Combined Hom-SNPs) is first subjected to removal of the Data-noise represented in File-B (Combined Hom-SNPs Data-noise), and then used for identifying the unknown sample variety, the Hom-SNPs shared uniquely between the unknown (Masena B40) and each of the forty-one reference samples indicated the highest number (5806) to be between the Unknown (Masena B40) and the Masena reference B50 (Figure 5A). This result also clearly identified the unknown sample to be the Masena variety. All but five of the other forty reference samples had no uniquely shared Hom-SNPs with the unknown (Masena B40) sample; B08, B24, B33 and B34 samples had one and B27 sample had two uniquely shared Hom-SNPs. However, the presence of uniquely shared variants, although only as high as two Hom-SNPs, indicates the presence of Persisting-data-noise other than Data-noise Hom-SNPs captured in File-B (Combined Hom-SNPs Data-noise), which were filtered out. The Persisting-data-noise was also observed when the analysis was undertaken with only six reference samples including the Masena reference B50 and

the Eureka reference B45, although the unknown was clearly identified to be the Masena variety, with 77,281 uniquely shared Hom-SNPs (Figure 5 insert).

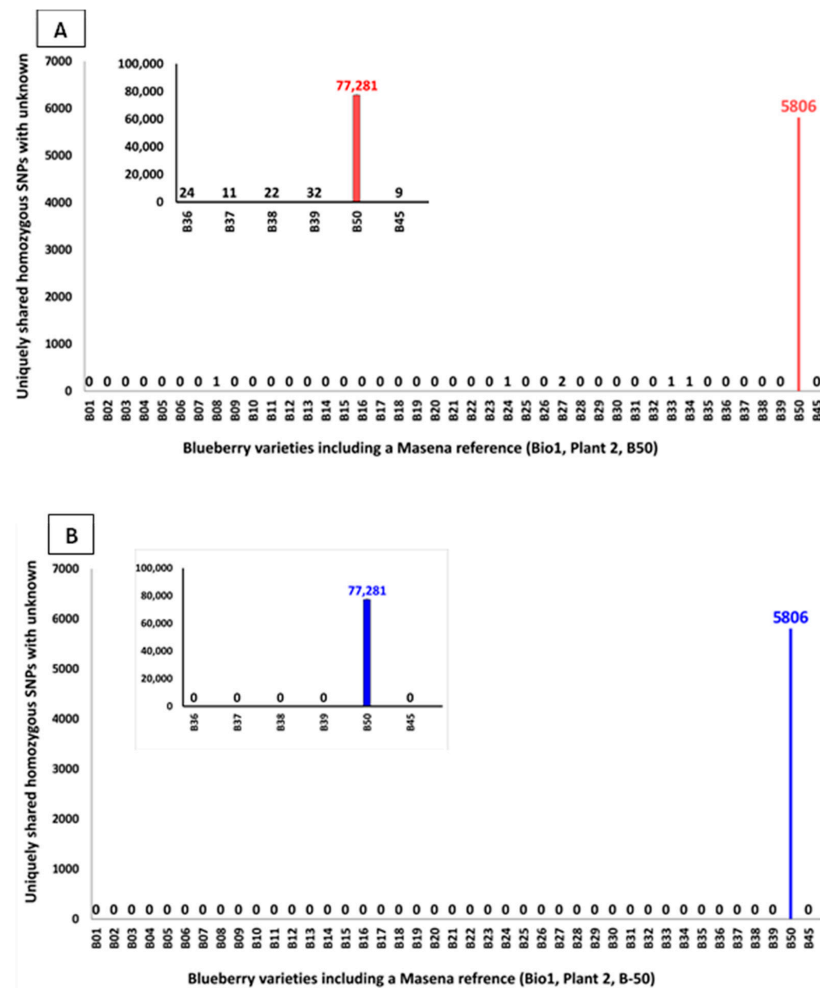


**Figure 5.** Identification of an unknown blueberry sample using uniquely shared Hom-SNPs after filtering Data-noise and Persisting-data-noise of subclass False-Unique (F1 and F2). H-SNPs/Hom-SNPs, homozygous SNPs; Bio1, Biological replicate library 1. Hom-SNPs of the unknown sample (Masena/G10/S40/B40/P1/L1/Unk/Tec-rep#1/Bio1-rep#1) was compared pairwise, using Paired Variant Comparison (PVC) analysis, with the Hom-SNPs from each of the forty-one blueberry reference samples (B01 to B39, B50 and B45) used to obtain a Hom-SNPs-(Unk&Ref-B# PVC)-file, one for each of the forty-one unknown and reference sample pairs. A shared Hom-SNPs file was generated by combining the Hom-SNPs-(Unk&Ref-B# PVC)-files of all the forty-one unknown and reference pairs. Shared Hom-SNPs identified between the unknown and each of the reference forty-one samples were then subjected to variety identification analysis, where (A) post-removal of Data-noise (\*C, S1 and S2), 5806 was the highest number of uniquely shared Hom-SNPs between the unknown sample (Masena, Bio1, plant1, B40) and the Masena reference sample (Plant2, Bio1, B50). In addition, between one and two Hom-SNPs were uniquely shared with five other reference samples. The presence of the uniquely shared Hom-SNPs between the unknown sample and with five other reference samples indicates the presence of another class of data-noise we refer to as Persisting-data-noise. (B) The Persisting-data-noise referred to as subclasses False-Unique (F1) and False-Unique (F2) was originally identified as True-Unique (U1 and U2) from the unknown and Masena reference (B-50) PVA comparison. Post-removal of Data-noise (\*C, S1 and S2) and the False-Unique (F1 and F2), the number of uniquely shared Hom-SNPs was found to be 5806 between the unknown sample (Masena, Bio1, plant1, B40) and the Masena reference sample (Plant2, Bio1, B50). Similar results were obtained when the analysis was undertaken with six reference blueberry varieties (figure-insert).

Identifying the nature of the Persisting-data-noise was important to its removal for variety identification with no data-noise. On examining the read mapping and variant track files corresponding to the reference samples B08, B24, B27, B33 and B34 with the Persisting-data-noise (Figure 5A), we identified the cause of the Persisting-data-noise to be due to the presence of SNPs erroneously called Hom-SNPs of the classes True-Unique (U1) and True-Unique (U2) in the unknown (Masena B40) but correctly called Het-SNPs in the Masena reference B50 (Figure S12). The PVC-analysis between the unknown (Masena B40) and the Masena reference B50 (Figure 3B) sample would have erroneously identified a number of SNPs to be Hom-SNPs in the unknown sample, leading to their inclusion as the Not-Common or putative Unique Hom-SNPs of the sub-classes True-Unique (U1) and True-Unique (U2), thus leading to the Persisting-data-noise. As expected, further filtering of the Persisting-data-noise Hom-SNPs captured in File-C (Combined Hom-SNPs Persisting-data-noise was undertaken from the file obtained by deducting the Data-noise captured in File-B (Combined Hom-SNPs Data-noise from File-A (Combined Hom-SNPs). Post-filtering of the Persisting-data-noise indicated the highest number of uniquely shared Hom-SNPs (5806) to be between the unknown (Masena B40) and the Masena reference sample B50 (Figure 5B), with zero uniquely shared Hom-SNPs between the unknown and any of the other reference samples. Similar results were obtained using the limited set of six reference samples including the Masena reference B50 and the Eureka reference B45 (Figure 5B insert).

### 3.12. Quality Metrics of Homozygous SNP Calls in Unknown Samples and Impact on Variety Identification

The variety identification protocol correctly identified the Unknown sample B40 as the Masena variety (Figure 5), in addition to nine other biological replicate samples of the Masena variety (B51-B53, B58-B62 and B68) when treated as unknown samples (Table S8). Read mapping coverage of the unknown samples was variable (Table S9), with samples B52, B59, B60 and B61 having an overall read mapping coverage under 20, while sample B59 had an average read mapping coverage excluding zero coverage of just under 20. The quality of the homozygous calls in the unknown samples compared to the B50 reference sample was determined using Quality metrics. The Sensitivity score was above 0.7 for all samples except 0.61 for B59, and with a low number of false positives in all samples; the PPV, F-score was high and the % accuracy was nearly 100% (Table S9). These results indicate the quality and depth of the sequence data, and the mapping and variant caller software programs lead to high-confidence homozygous SNPs calls being identified.

## 4. Discussion

Variety identification is important in plant breeding, as it provides plant breeders with a method to protect the variety generated, thus providing benefits to growers and producers/funders of the breeding programs. Variety identification in Blueberry, mainly propagated by cuttings, is vital for the protection, maintenance and policing of elite lines, important for plant breeders' rights and the protection of commercial rights. A number of technologies based on variation present in DNA have been used in plant identification [65] making use of single nucleotide polymorphisms (SNPs) [43,49,66]. Here, we present a protocol for accurate plant variety identification (Figure 4 and Figure S13), using SNPs identified from whole genome sequence data and applied on commercial blueberry (*Vaccinium*, spp.) samples.

The accurate identification of a plant sample using SNPs from NGS data involves best practice procedures, starting from sample collection to sequence data analysis [67]. Ensuring the identity of the varieties selected as reference samples and good quality DNA may require

testing several procedures [68–81]. This is critical for good quality NGS data. Sequence data at a mapping depth of over 20 (Figure 1), and a suitable variant caller, ensured the identification of high-confidence homozygous SNPs. The application of Data-noise filtering steps applied to the homozygous SNPs of an unknown sample and a set of reference samples ensured high accuracy of variety identification. Reliability of SNP calls at a data depth of at least 20 was demonstrated in other studies [55,60] ensuring that a Het-SNP is not erroneously called a Hom-SNP, or conversely, a Hom-SNP when called is a correct call, which can be due to the disproportionate or non-coincident representation of the Hom-SNP bases in the mapped sequence reads. Post read mapping, a variant titration (Figure S2) was used to define the minimum allele frequency required for Het-Variant calling, increasing the probability of equal representation of both the nucleotides of a Het-SNP location. Although high data depth ensured Hom-SNP calls with high accuracy (Table S9), the presence of data-noise indicating unique SNP markers between the unknown and a number of reference samples but not the matching reference sample (Figure S11) may be interpreted with doubt and may not provide the certainty required of variety identity in a legal sense for variety protection [12,24]. Data-noise is embedded within the putative Unique SNPs of dissimilar varieties (C\*, S1 and S2) and between similar varieties (U1 and U2). Data-noise identified between samples of the same variety, as in Masena samples at an average of 36.5% (Figure 3C), is mainly due to areas of coverage under 20, which can be reduced by using sequence data with mapping coverage over 20 in these regions. Missing data at 25% was reported for genotyping by a sequencing (GBS) approach in maize [82], which could be reduced to half or even lower using imputations approaches [82,83] but at the expense of an increased genotyping error rate, which was found to be 0.7% [84]. We identified missing data (U2) of 0.29% as False-True-Unique Hom-SNPs identified from the PVC-analysis between 55 Masena sample combinations (Figure 3C). Kmer-based heterozygosity has been reported to be 1.27% [85]. Based on the heterozygosity between 0.85% and 1%, equating to 5.1 to 6.0 M Het-SNP sites in the blueberry genome (600 Mb), we estimate 17,400 of the Het-SNPs to be erroneously called Hom-SNPs, corresponding to a 0.29% error rate (Data-noise U2). The application of the developed data-noise filtering framework (Figure 4), which was tested across nine other unknown samples (Table S8), eliminated any ambiguity in variety identity. The “Variety identification Protocol 1” (Figure 4), outlines the method development process for variety identification including the concept around identification and the removal of the Data-noise and the Persisting-data-noise. Application of the Data-noise framework eliminates all data-noise, resulting in Unique Hom-SNPs only between two samples, the unknown and a reference sample, but not with any of the other reference samples (Figure 4).

Prior to NGS, prevailing molecular marker technologies had limited use in understanding the genetic diversity in the blueberry group [86–88]. Microsatellite markers were used in a limited manner for genetic analysis of the various groups of blueberry [89]. The use of GBS aided in identifying SNPs for genetic diversity and population structure analysis and phenological traits [90]. Microsatellite markers have also been used for genetic diversity analysis varietal identification, although they require cumbersome non NGS approaches [91,92]. However, unlike NGS, these methods, in general, had limitations (Table 5). The protocol developed here (Figures 4 and S13), is economical, as the generated sequence data of the reference samples can be reused and is also forward compatible, as sequence data of additional reference samples can be added to expand the sequence data of the reference sample cohort. The cost of SNP genotyping per data point using NGS has been reduced [45,93], and thus this approach is suited to variety identification. High-throughput approaches using NGS and reduced representation approaches such as GBS [50,52,54] or SNP array-based approaches such as Axiom, Infinium and MassARRAY iPLEX [30] are ap-

proprate for screening large samples for genetic diversity and other genetic studies. Other methods for variety identification are not suitable as they lack reliability and reproducibility and require the re-analysis of all samples when new varieties are introduced.

**Table 5.** Comparison of methods for DNA-based plant variety distinction.

| Method                | Sensitivity to Distinguish a Closely Related Variety | Need for Prior Information                      |
|-----------------------|------------------------------------------------------|-------------------------------------------------|
| SSR                   | very low <sup>1</sup>                                | Requires prior identification of SSR loci       |
| Specific SNP          | very low <sup>1</sup>                                | Requires prior selection of distinguishing loci |
| Whole genome sequence | High <sup>2</sup>                                    | No prior analysis required same for all species |

<sup>1</sup> Marker set may not distinguish unknown variety. Match does not establish identity and failure to match would suggest not a variety in reference set. <sup>2</sup> Can distinguish any difference in genome regardless of presence in reference set. Match confirms identity and failure to match confirms not in reference set.

Accurately identifying a plant sample may also impact plant identification for other applications, such as the ex situ maintenance of plant collections, essential for the conservation of important species including elite crop plant lines [18,19]. Well-established methods, such as seed banking, tissue culture, cryopreservation and living plant collections [21,22,94–97], could also benefit. Longterm genetic testing may be required of ex situ, founder accessions and native individuals for maintenance and reintroduction for biodiversity or commercial applications [18,98]. The importance of accurate plant variety identification for maintenance, orchard management and multiplication and the commercialization/sale of elite lines cannot be underestimated in horticultural crops. The inadvertent sale of erroneous lines of clonally propagated fruit trees can adversely impact commercialization, especially if the life cycle continues for several years before the error is realized. The data-noise reduction method can have applications in SNPs for phenotypic trait association using individual or bulk samples association analysis, using whole genome sequence data studies [79,99] or using expression sequence data sets [100].

In addition, genetic testing for distinctiveness can be part of the DUS test (distinctness, uniformity and stability) and a means of establishing plant breeders’ rights [15]. Use of this method could support applications for plant variety protection and be used to police intellectual property rights. These analyses provide data to establish the identity of protected varieties in disputes over ownership. Other approaches not involving WGS may not provide definitive evidence. Claims of identity of varieties based upon incomplete genome analysis may be open to challenges in IP disputes.

5. Conclusions

The SNPs data for the 41 blueberry samples should allow any blueberry variety within this cohort of blueberry samples to be identified. All varieties were easily associated with one of the genotypes in this group of varieties. Additional varieties can be added to expand the reference sample cohort for the application of the technique to unknown samples that need to be considered. Using this protocol to identify varieties in specific regions would need to include all known varieties from that region. The SNP data can also be used to identify variety-specific SNPs which are not present in any of the other blueberry varieties within this cohort of forty-one reference samples. The variety-specific SNPs could be used to develop a cost-competitive assay test to identify blueberry varieties within a given cohort. This would be useful in the screening of larger numbers of samples in applications, such as when policing the sale of protected varieties on sale in a retail environment. The concept of identifying and the removal of data-noise is not restricted to the use of the CLC Genomics Workbench and can be developed and applied using any

software programs, including command-line approaches. This approach can potentially be applied in other plant species for which variety identification is required, providing a robust genetic approach to variety identification in support of the protection of biosecurity, food safety and intellectual property. Given the trend of the reduced cost of sequencing, it is possible to aim for over 20× sequencing for species with higher genome size. Alternately, it is possible to apply this method using a lower sequence data size, provided that enough unique candidate SNPs at a mapping depth of 20 and above are identified and informative in discriminating the required varieties. The method may also be applicable to validate varieties from closely related species.

**Supplementary Materials:** The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/applbiosci4040051/s1>, Figure S1: QC data of the extracted blueberry DNA; Figure S2: Short read paired end trimmed sequence read data of 77 Blueberry DNA samples; Figure S3: Titration analysis of mapping and SNP variant detection parameters using blueberry PE reads mapped to a reference genome; Figure S4: Fraction of the genome covered and average mapped coverage by paired end reads of blueberry samples; Figure S5: Heterozygous SNPs and homozygous SNPs across several blueberry samples with variable sequence data size; Figure S6: High confidence common homozygous SNPs between Masena samples within technical and biological libraries and between different biological libraries; Figure S7: High confidence common homozygous SNPs between Eureka samples within technical and biological libraries and between different biological libraries; Figure S8: PVC analysis between Masena and between Eureka blueberry samples to identify Common and Not-common Hom-SNPs; Figure S9: PVC analysis between the unknown (Masena, B40) and forty blueberry varieties including Masena (B50) to identify Common and Not-Common Hom-SNPs; Figure S10: PVC analysis and classification of the Not-common-Hom-SNPs between the unknown (Masena, B40) and forty-one blueberry varieties (B01 to B39, Masena B50 and Eureka B45); Figure S11: Identification of an Unknown blueberry sample using uniquely shared Hom-SNPs without filtering data-noise; Figure S12: Nature of the Persisting-data-noise in the variety identification analysis; Figure S13: Schematic of Variety Identification: Alternate Protocol 3; Table S1: QC metrics of all 77 blueberry DNA samples; Table S2: Percentage probability of heterozygous SNPs erroneously called as homozygous due to read mapping coverage and scenarios; Table S3: Optimal read mapping coverage to reduce the percentage probability of heterozygous SNPs erroneously called as homozygous; Table S4: QC of the sequence data of all blueberry samples; Table S5: Mapping titration of short read (Illumina @ 150bp paired end reads) of five blueberry varieties using *V. darrowii* genome sequence (NCBI 2021) as the reference; Table S6: Statistical analysis of the percentage common Hom-SNPs from pairwise analysis for Masena technical and biological replicates; Table S7: Statistical analysis of common Hom-SNPs from pairwise analysis for Eureka technical and biological replicates; Table S8: Polymorphic Hom-SNPs uniquely shared only between the unknown sample and a reference sample post filtering Data-noise and Persisting Data-noise; Table S9: Homozygous genotype-call quality metrics of Unknown (Masena) samples compared to the Masena reference B50.

**Author Contributions:** Conceptualization, A.F., T.G. and R.J.H.; data curation, A.F.; lab, A.F.; methodology, A.F. and R.J.H.; project administration, A.F. and R.J.H.; resources, R.J.H. and T.G.; supervision, R.J.H.; validation, A.F.; visualization, A.F.; writing—original draft, A.F. and R.J.H.; writing, review and editing, A.F., T.G. and R.J.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** Samples of blueberry varieties were provided by Mountain Blue Orchards. Mountain Blue Orchards and The University of Queensland provided funding for the project.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data is available at NCBI under BioProject #1259146 and BioSample Sub #s SAMN48354722 to SAMN48354798.

**Acknowledgments:** We acknowledge the computing infrastructure support provided by the Research Computing Centre at The University of Queensland, Australia.

**Conflicts of Interest:** Author Tom Gunther was employed by the company Mountain Blue Orchards. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The authors declare that this study received funding from Mountain Blue Orchards. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

## Abbreviations

The following abbreviations are used in this manuscript:

|         |                                  |
|---------|----------------------------------|
| Tec-rep | Technical replicate              |
| Bio-rep | Biological replicate             |
| PVC     | Paired variant comparison        |
| SNP     | Single nucleotide polymorphism   |
| LF      | Length fraction                  |
| SF      | Similarity fraction              |
| MF      | Minimum allele frequency         |
| MC      | Minimum read count               |
| Hom-SNP | Homozygous SNP                   |
| Het-SNP | Heterozygous SNP                 |
| MFAA    | Most frequent alternative allele |

## References

1. Lawson, C. Propagating materials and harvested materials: Clarifying the scope of plant variety or breeder's rights. *J. Intellect. Prop. Law Pract.* **2023**, *18*, 655–672. [CrossRef]
2. Letsiou, S.; Madesis, P.; Vasdekis, E.; Montemurro, C.; Grigoriou, M.E.; Skavdis, G.; Moussis, V.; Koutelidakis, A.E.; Tzakos, A.G. DNA Barcoding as a Plant Identification Method. *Appl. Sci.* **2024**, *14*, 1415. [CrossRef]
3. de Boer, H.J.; Ichim, M.C.; Newmaster, S.G. DNA Barcoding and Pharmacovigilance of Herbal Medicines. *Drug Saf.* **2015**, *38*, 611–620. [CrossRef] [PubMed]
4. Galimberti, A.; Casiraghi, M.; Bruni, I.; Guzzetti, L.; Cortis, P.; Berterame, N.M.; Labra, M. From DNA barcoding to personalized nutrition: The evolution of food traceability. *Curr. Opin. Food Sci.* **2019**, *28*, 41–48. [CrossRef]
5. Fanelli, V.; Mascio, I.; Miazzi, M.M.; Savoia, M.A.; De Giovanni, C.; Montemurro, C. Molecular Approaches to Agri-Food Traceability and Authentication: An Updated Review. *Foods* **2021**, *10*, 1644. [CrossRef]
6. Uncu, A.O.; Uncu, A.T. A barcode-DNA analysis method for the identification of plant oil adulteration in milk and dairy products. *Food Chem.* **2020**, *326*, 126986. [CrossRef]
7. Mendler-Drienyovszki, N.; Magyar-Tábori, K.; Mancinelli, R.; Black, L.; Brown, H.; Allam, M.; Udupa, S.M.; Atait, M.; Novarina, E.; Bardelli, T.; et al. Assessing the Opportunities and Risks of DUS and VCU Variety Testing for Sustainable Production through SWOT Analysis Results. *Agriculture* **2024**, *14*, 1817. [CrossRef]
8. Macdonald, H. Ownership and Objectivity: Judging varieties for plant breeder's rights in Australia. *Univ. N. S. W. Law J.* **2024**, *47*, 247–281. [CrossRef]
9. USDA. Plant Variety Protection. Available online: <https://www.ams.usda.gov/services/plant-variety-protection> (accessed on 9 July 2025).
10. Government, A. How to Label My Protected Plant Variety. Available online: <https://www.ipaustralia.gov.au/manage-my-ip/how-to-prevent-infringement-of-my-ip/how-to-label-my-protected-plant-variety> (accessed on 9 July 2025).
11. Macpherson, H.; Rennie, A. Plant Variety Key to Agri Future. 2021. Available online: <https://www.anz.com.au/bluenotes/2021/01/baker-mckenzie-agriculture-diversity-plant-matter-food-security/> (accessed on 9 July 2025).
12. UPOV. *International Convention for the Protection of New Varieties of Plants of December 2, 1961, as Revised at Geneva on November 10, 1972, on October 23, 1978, and on March 19, 1991*. Available online: [www.upov.int/edocs/pubdocs/en/upov\\_pub\\_221.pdf](http://www.upov.int/edocs/pubdocs/en/upov_pub_221.pdf) (accessed on 9 July 2025).
13. Mariani, S. Law-Driven Innovation in Cereal Varieties: The Role of Plant Variety Protection and Seed Marketing Legislation in the European Union. *Sustainability* **2021**, *13*, 8049. [CrossRef]

14. He, X.; Wu, L.; Hu, K.; Wan, L. An Evaluation Model of Plant Variety Rights Capitalization Operation Value Based on Big Data Analysis. *Sci. Program.* **2022**, *2022*, 3816306. [\[CrossRef\]](#)
15. Yu, J.K.; Chung, Y.S. Plant Variety Protection: Current Practices and Insights. *Genes* **2021**, *12*, 1127. [\[CrossRef\]](#)
16. Lanubile, A.; Stagnati, L.; Marocco, A.; Busconi, M. DNA-based techniques to check quality and authenticity of food, feed and medicinal products of plant origin: A review. *Trends Food Sci. Technol.* **2024**, *149*, 104568. [\[CrossRef\]](#)
17. Henry, R.J. (Ed.) *Plant Genotyping II: SNP Technology*; CAB International: Oxfordshire, UK, 2008; pp. 1–285.
18. Engels, J.M.M.; Ebert, A.W. A Critical Review of the Current Global Ex Situ Conservation System for Plant Agrobiodiversity. II. Strengths and Weaknesses of the Current System and Recommendations for its Improvement. *Plants* **2021**, *10*, 1904. [\[CrossRef\]](#)
19. Yenson, A.J.M.; Sommerville, K.D.; Guja, L.K.; Merritt, D.J.; Dalziel, E.L.; Auld, T.D.; Broadhurst, L.; Coates, D.J.; Commander, L.; Crawford, A.D.; et al. Ex situ germplasm collections of exceptional species are a vital part of the conservation of Australia's national plant treasures. *Plants People Planet* **2024**, *6*, 44–66. [\[CrossRef\]](#)
20. Vendrame, W.; Faria, R.; Sorace, M.; Sahyun, S. Orchid cryopreservation. *Ciência Agrotecnologia* **2014**, *38*, 213–229. [\[CrossRef\]](#)
21. Rai, M.K. Chapter 9—Plant tissue culture targeting germplasm conservation. In *Advances in Plant Tissue Culture*; Rai, A.C., Kumar, A., Modi, A., Singh, M., Eds.; Academic Press: Cambridge, MA, USA, 2022; pp. 205–221.
22. Benelli, C.; Tarraf, W.; Izgu, T.; De Carlo, A. In Vitro Conservation through Slow Growth Storage Technique of Fruit Species: An Overview of the Last 10 Years. *Plants* **2022**, *11*, 3188. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Getachew, S.E.; Bille, N.H.; Bell, J.M.; Gebreselassie, W. Genotyping by Sequencing for Plant Breeding—A Review. *Adv. Biotechnol. Microbiol.* **2019**, *14*, 555891. [\[CrossRef\]](#)
24. IP Australia Australian Government. Plant Breeder's Rights. 2025. Available online: <https://www.ipaustralia.gov.au/plant-breeders-rights/how-to-search-existing-plant-breeders-rights> (accessed on 30 July 2025).
25. Thomas, W.J.W.; Zhang, Y.; Amas, J.C.; Cantila, A.Y.; Zandberg, J.D.; Harvie, S.L.; Batley, J. Innovative Advances in Plant Genotyping. In *Plant Genotyping: Methods and Protocols*; Shavrukov, Y., Ed.; Springer: New York, NY, USA, 2023; pp. 451–465.
26. Henry, R.J. (Ed.) *Plant Genotyping: The DNA Fingerprinting of Plants*; CABI Publishing: New York, NY, USA, 2001; p. 321.
27. Poets, A.; Silverstein, K.; Pardey, P.; Hearne, S.; Stevenson, J. *DNA Fingerprinting for Crop Varietal Identification: Fit-For-Purpose Protocols, Their Costs and Analytical Implications*; CGIAR Advisory Services (CAS) Secretariat: Rome, Italy, 2020.
28. Varshney, R.K.; Nayak, S.N.; May, G.D.; Jackson, S.A. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* **2009**, *27*, 522–530. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Deschamps, S.; Llaca, V.; May, G.D. Genotyping-by-Sequencing in Plants. *Biology* **2012**, *1*, 460–483. [\[CrossRef\]](#)
30. Kockum, I.; Huang, J.; Stridh, P. Overview of Genotyping Technologies and Methods. *Curr. Protoc.* **2023**, *3*, e727. [\[CrossRef\]](#)
31. Stepanenko, A.; Chen, G.; Hoang, P.T.N.; Fuchs, J.; Schubert, I.; Borisjuk, N. The Ribosomal DNA Loci of the Ancient Monocot *Pistia stratiotes* L. (Araceae) Contain Different Variants of the 35S and 5S Ribosomal RNA Gene Units. *Front. Plant Sci.* **2022**, *13*, 819750. [\[CrossRef\]](#)
32. Morton, E.A.; Hall, A.N.; Kwan, E.; Mok, C.; Queitsch, K.; Nandakumar, V.; Stamatoyannopoulos, J.; Brewer, B.J.; Waterston, R.; Queitsch, C. Challenges and Approaches to Genotyping Repetitive DNA. *G3 Genes | Genomes | Genet.* **2020**, *10*, 417–430. [\[CrossRef\]](#)
33. Hemleben, V.; Grierson, D.; Borisjuk, N.; Volkov, R.A.; Kovarik, A. Personal Perspectives on Plant Ribosomal RNA Genes Research: From Precursor-rRNA to Molecular Evolution. *Front. Plant Sci.* **2021**, *12*, 797348. [\[CrossRef\]](#)
34. Ko, H.L.; Henry, R.J. Specific 5S ribosomal RNA primers for plant species identification in admixtures 3. *Plant Mol. Biol. Rep.* **1996**, *14*, 33–43. [\[CrossRef\]](#)
35. Vieira, M.L.; Santini, L.; Diniz, A.L.; Munhoz Cde, F. Microsatellite markers: What they mean and why they are so useful. *Genet. Mol. Biol.* **2016**, *39*, 312–328. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Mason, A.S. SSR Genotyping. In *Plant Genotyping: Methods and Protocols*; Batley, J., Ed.; Springer: New York, NY, USA, 2015; pp. 77–89.
37. Srivastava, D.S.; Mukherjee, S.; Datta, S.; Burman, S.; Khan, A.; Rana, M.; Kumar, R.; Katyayani, K.K.S.; Karnwal, A.; Kumar, S.; et al. Microsatellite markers for crop improvement: A review. *J. Appl. Nat. Sci.* **2023**, *15*, 1018–1035. [\[CrossRef\]](#)
38. Hodel, R.G.J.; Segovia-Salcedo, M.C.; Landis, J.B.; Crawl, A.A.; Sun, M.; Liu, X.; Gitzendanner, M.A.; Douglas, N.A.; Germain-Aubrey, C.C.; Chen, S.; et al. The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century. *Appl. Plant Sci.* **2016**, *4*, 1600025. [\[CrossRef\]](#)
39. Varshney, R.K.; Graner, A.; Sorrells, M.E. Genic microsatellite markers in plants: Features and applications. *Trends Biotechnol.* **2005**, *23*, 48–55. [\[CrossRef\]](#)
40. Henry, R.J.; Chopra, V.L.; Malik, V.S.; Bhat, S.R. Molecular techniques for the identification of plants. In *Applied Plant Biotechnology*; Science Publishers: Enfield, UK, 1999; pp. 269–283.
41. Collard, B.C.Y.; Jahufer, M.Z.Z.; Brouwer, J.B.; Pang, E.C.K. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* **2005**, *142*, 169–196. [\[CrossRef\]](#)
42. Vignal, A.; Milan, D.; SanCristobal, M.; Eggen, A. A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.* **2002**, *34*, 275. [\[CrossRef\]](#) [\[PubMed\]](#)

43. Brookes, A.J. The essence of SNPs. *Gene* **1999**, *234*, 177–186. [[CrossRef](#)]
44. Edward, K.J.; Poole, R.L.; Barker, G.L. SNP discovery in plants. In *Plant Genotyping II: Snp Technology*; Henry, R.J., Ed.; Cab International: Oxfordshire, UK, 2008; pp. 1–29.
45. Mammadov, J.; Aggarwal, R.; Buyyarapu, R.; Kumpatla, S. SNP markers and their impact on plant breeding. *Int. J. Plant Genom.* **2012**, *2012*, 728398. [[CrossRef](#)]
46. Harshitha, N.S.; Sandal, S.S. DNA fingerprinting and its applications in crop improvement: A review. *Pharma Innov. J.* **2022**, *11*, 792–797.
47. Shen, Y.; Wang, J.; Shaw, R.; Yu, H.; Sheng, X.; Zhao, Z.; Li, S.; Gu, H. Development of GBTS and KASP Panels for Genetic Diversity, Population Structure, and Fingerprinting of a Large Collection of Broccoli (*Brassica oleracea* L. var. *italica*) in China. *Front. Plant Sci.* **2021**, *12*, 655254. [[CrossRef](#)] [[PubMed](#)]
48. Zhang, J.; Yang, J.; Zhang, L.; Luo, J.; Zhao, H.; Zhang, J.; Wen, C. A new SNP genotyping technology Target SNP-seq and its application in genetic analysis of cucumber varieties. *Sci. Rep.* **2020**, *10*, 5623. [[CrossRef](#)] [[PubMed](#)]
49. Lawrie, R.D.; Massey, S.E. Agrigenomic Diversity Unleashed: Current Single Nucleotide Polymorphism Genotyping Methods for the Agricultural Sciences. *Appl. Biosci.* **2023**, *2*, 565–585. [[CrossRef](#)]
50. Pootakham, W. Genotyping by Sequencing (GBS) for Genome-Wide SNP Identification in Plants. In *Plant Genotyping. Methods in Molecular Biology*; Humana Press: New York, NY, USA, 2023; pp. 1–8. [[CrossRef](#)]
51. Satam, H.; Joshi, K.; Mangrolia, U.; Waghoo, S.; Zaidi, G.; Rawool, S.; Thakare, R.P.; Banday, S.; Mishra, A.K.; Das, G.; et al. Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology* **2023**, *12*, 997. [[CrossRef](#)]
52. Miller, M.R.; Dunham, J.P.; Amores, A.; Cresko, W.A.; Johnson, E.A. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* **2007**, *17*, 240–248. [[CrossRef](#)] [[PubMed](#)]
53. Baird, N.A.; Etter, P.D.; Atwood, T.S.; Currey, M.C.; Shiver, A.L.; Lewis, Z.A.; Selker, E.U.; Cresko, W.A.; Johnson, E.A. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* **2008**, *3*, e3376. [[CrossRef](#)]
54. Elshire, R.J.; Glaubitz, J.C.; Sun, Q.; Poland, J.A.; Kawamoto, K.; Buckler, E.S.; Mitchell, S.E. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* **2011**, *6*, e19379. [[CrossRef](#)]
55. Cheng, A.Y.; Teo, Y.-Y.; Ong, R.T.-H. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics* **2014**, *30*, 1707–1713. [[CrossRef](#)] [[PubMed](#)]
56. Torkamaneh, D.; Belzile, F. Scanning and Filling: Ultra-Dense SNP Genotyping Combining Genotyping-By-Sequencing, SNP Array and Whole-Genome Resequencing Data. *PLoS ONE* **2015**, *10*, e0131533. [[CrossRef](#)]
57. Nielsen, R.; Paul, J.S.; Albrechtsen, A.; Song, Y.S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **2011**, *12*, 443–451. [[CrossRef](#)]
58. Kobayashi, M.; Ohyanagi, H.; Takanashi, H.; Asano, S.; Kudo, T.; Kajiya-Kanegae, H.; Nagano, A.J.; Tainaka, H.; Tokunaga, T.; Sazuka, T.; et al. Heap: A highly sensitive and accurate SNP detection tool for low-coverage high-throughput sequencing data. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* **2017**, *24*, 397–405. [[CrossRef](#)]
59. Josia, C.; Mashingaidze, K.; Assefa, A.; Kondwakwenda, A.; Musvosvi, C.; Sibiya, J. SNP-based assessment of genetic purity and diversity in maize hybrid breeding. *PLoS ONE* **2021**, *16*, e0249505. [[CrossRef](#)]
60. Bentley, D.R.; Balasubramanian, S.; Swerdlow, H.P.; Smith, G.P.; Milton, J.; Brown, C.G.; Hall, K.P.; Evers, D.J.; Barnes, C.L.; Bignell, H.R.; et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **2008**, *456*, 53–59. [[CrossRef](#)]
61. Li, R.; Li, Y.; Fang, X.; Yang, H.; Wang, J.; Kristiansen, K.; Wang, J. SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **2009**, *19*, 1124–1132. [[CrossRef](#)] [[PubMed](#)]
62. Henry, R.J. Progress in Plant Genome Sequencing. *Appl. Biosci.* **2022**, *1*, 113–128. [[CrossRef](#)]
63. Furtado, A. DNA Extraction from Vegetative Tissue for Next-Generation Sequencing. In *Cereal Genomics*; Henry, R.J., Furtado, A., Eds.; Humana Press: New York, NY, USA, 2014; pp. 1–5.
64. Tahir, M.; Sardaraz, M. A Fast and Scalable Workflow for SNPs Detection in Genome Sequences Using Hadoop Map-Reduce. *Genes* **2020**, *11*, 166. [[CrossRef](#)]
65. Kumar, R.; Das, S.P.; Choudhury, B.U.; Kumar, A.; Prakash, N.R.; Verma, R.; Chakraborti, M.; Devi, A.G.; Bhattacharjee, B.; Das, R.; et al. Advances in genomic tools for plant breeding: Harnessing DNA molecular markers, genomic selection, and genome editing. *Biol. Res.* **2024**, *57*, 80. [[CrossRef](#)]
66. Varala, K.; Swaminathan, K.; Li, Y.; Hudson, M.E. Rapid Genotyping of Soybean Cultivars Using High Throughput Sequencing. *PLoS ONE* **2011**, *6*, e24811. [[CrossRef](#)]
67. de Boer, H.; Rydmark, M.O.; Verstraete, B.; Gravendeel, B. Molecular identification of plants: From sequence to species. *Adv. Books* **2022**, *1*, 396.
68. Rogers, S.O.; Bendich, A.J. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Mol. Biol.* **1985**, *5*, 69–76. [[CrossRef](#)] [[PubMed](#)]
69. Arseneau, J.-R.; Steeves, R.; Laflamme, M. Modified low-salt CTAB extraction of high-quality DNA from contaminant-rich tissues. *Mol. Ecol. Resour.* **2017**, *17*, 686–693. [[CrossRef](#)] [[PubMed](#)]

70. Inglis, P.W.; Pappas, M.d.C.R.; Resende, L.V.; Grattapaglia, D. Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PLoS ONE* **2018**, *13*, e0206085. [CrossRef]
71. Rezadoost, M.H.; Kordrostami, M.; Kumleh, H.H. An efficient protocol for isolation of inhibitor-free nucleic acids even from recalcitrant plants. *3 Biotech* **2016**, *6*, 61. [CrossRef]
72. Souza, H.A.V.; Muller, L.A.C.; Brandão, R.L.; Lovato, M.B. Isolation of high quality and polysaccharide-free DNA from leaves of *Dimorphandra mollis* (Leguminosae), a tree from the Brazilian Cerrado. *Genet. Mol. Res.* **2012**, *11*, 756–764. [CrossRef] [PubMed]
73. Sharma, P.; Purohit, S.D. An improved method of DNA isolation from polysaccharide rich leaves of *Boswellia serrata* Roxb. *Indian J. Biotechnol.* **2012**, *11*, 67–71.
74. Berensmeier, S. Magnetic particles for the separation and purification of nucleic acids. *Appl. Microbiol. Biotechnol.* **2006**, *73*, 495–504. [CrossRef]
75. Porebski, S.; Bailey, L.G.; Baum, B.R. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Rep.* **1997**, *15*, 8–15. [CrossRef]
76. Wijesundara, U.K.; Masouleh, A.K.; Furtado, A.; Dillon, N.L.; Henry, R.J. A chromosome-level genome of mango exclusively from long-read sequence data. *Plant Genome* **2024**, *17*, e20441. [CrossRef] [PubMed]
77. Al-Dossary, O.; Alsubaie, B.; Kharabian-Masouleh, A.; Al-Mssallem, I.; Furtado, A.; Henry, R.J. The jojoba genome reveals wide divergence of the sex chromosomes in a dioecious plant. *Plant J.* **2021**, *108*, 1283–1294. [CrossRef]
78. Healey, A.; Furtado, A.; Cooper, T.; Henry, R.J. Protocol: A simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods* **2014**, *10*, 21. [CrossRef]
79. Tran, H.T.M.; Furtado, A.; Vargas, C.A.C.; Smyth, H.; Slade Lee, L.; Henry, R. SNP in the *Coffea arabica* genome associated with coffee quality. *Tree Genet. Genom.* **2018**, *14*, 72. [CrossRef]
80. Kilby, N.J.; Furner, I.J. Another CTAB Protocol: Isolation of High Molecular Weight DNA from Small Quantities of Arabidopsis Tissue 2022 (Protocol). Available online: [https://www.arabidopsis.org/download\\_files/Protocols/compleat\\_guide/4\\_alt\\_CTAB\\_DNA\\_extr.pdf](https://www.arabidopsis.org/download_files/Protocols/compleat_guide/4_alt_CTAB_DNA_extr.pdf) (accessed on 25 October 2023).
81. Sop204-01 2023; High Molecular Weight (HMW) Total DNA Extraction from Non-Model Species. (gih.uq.edu.au). Standard Operating Procedure. The University Of Queensland: Brisbane, Australia, 2023.
82. Swarts, K.; Li, H.; Navarro, J.A.R.; An, D.; Romay, M.C.; Hearne, S.; Acharya, C.; Glaubitz, J.C.; Mitchell, S.; Elshire, R.J.; et al. Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. *Plant Genome* **2014**, *7*, 23. [CrossRef]
83. Bradbury, P.J.; Casstevens, T.; Jensen, S.E.; Johnson, L.C.; Miller, Z.R.; Monier, B.; Romay, M.C.; Song, B.; Buckler, E.S. The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. *Bioinformatics* **2022**, *38*, 3698–3702. [CrossRef] [PubMed]
84. Wang, N.; Yuan, Y.; Wang, H.; Yu, D.; Liu, Y.; Zhang, A.; Gowda, M.; Nair, S.K.; Hao, Z.; Lu, Y.; et al. The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. *Sci. Rep.* **2020**, *10*, 16308. [CrossRef]
85. Cui, F.; Ye, X.; Li, X.; Yang, Y.; Hu, Z.; Overmyer, K.; Brosché, M.; Yu, H.; Salojärvi, J. Chromosome-level genome assembly of the diploid blueberry *Vaccinium darrowii* provides insights into its subtropical adaptation and cuticle synthesis. *Plant Comm.* **2022**, *3*, 100307. [CrossRef] [PubMed]
86. McCallum, S.; Graham, J.; Jorgensen, L.; Rowland, L.J.; Bassil, N.V.; Hancock, J.F.; Wheeler, E.J.; Vining, K.; Poland, J.A.; Olmstead, J.W.; et al. Construction of a SNP and SSR linkage map in autotetraploid blueberry using genotyping by sequencing. *Mol. Breed.* **2016**, *36*, 41. [CrossRef]
87. Rowland, L.J.; Ogden, E.L.; Ehlenfeldt, M.K. EST-PCR markers developed for highbush blueberry are also useful for genetic fingerprinting and relationship studies in rabbiteye blueberry. *Sci. Hortic.* **2010**, *125*, 779–784. [CrossRef]
88. Boches, P.; Bassil, N.V.; Rowland, L. Genetic diversity in the highbush blueberry evaluated with microsatellite markers. *J. Am. Soc. Hort Sci.* **2006**, *131*, 674–686. [CrossRef]
89. Bassil, N.V.; Bunch, T.; Nyberg, A.; Hummer, K.; Zee, F.T. *Microsatellite Markers Distinguish Hawaiian Ohelo from Other Vaccinium L. Section Myrtillos Species*; International Society For Horticultural Science (ISHS): Leuven, Belgium, 2010.
90. Campa, A.; Ferreira, J.J. Genetic diversity assessed by genotyping by sequencing (GBS) and for phenological traits in blueberry cultivars. *PLoS ONE* **2018**, *13*, e0206361. [CrossRef] [PubMed]
91. Carvalho, M.; Matos, M.; Crespi, A.; Lopes, V.R.; Carnide, V. Genetic Diversity and Identification of *Vaccinium* Species Through Microsatellite Analysis. *Plants* **2024**, *13*, 3488. [CrossRef] [PubMed]
92. Miteca, H.; Castro, M.H.; Meneses, M.; Prat, L.; Muñoz, C.; Hinrichsen, P. Validation of a minimal panel of microsatellite markers for blueberry cultivar identification and frequency of spontaneous mutations. *Chil. J. Agric. Res.* **2024**, *84*, 28–42. [CrossRef]
93. Ganai, M.W.; Altmann, T.; Röder, M.S. SNP identification in crop plants. *Curr. Opin. Plant Biol.* **2009**, *12*, 211–217. [CrossRef]

94. Malhotra, N.; Panatu, S.; Singh, B.; Negi, N.; Singh, D.; Singh, M.; Chandora, R. Chapter 3—Genetic Resources: Collection, Conservation, Characterization And Maintenance. In *Lentils*; Singh, M., Ed.; Academic Press: Cambridge, MA, USA, 2019; pp. 21–41.
95. Liu, Y.; Huang, T.; Lin, S.; Jiang, Y. Programmed Cooling and Vitrification Method Applied to the Germplasm Preservation of *Eriobotrya* Plants. *HortScience* **2023**, *58*, 164–169. [[CrossRef](#)]
96. Kulak, V.; Longboat, S.; Brunet, N.D.; Shukla, M.; Saxena, P. In Vitro Technology in Plant Conservation: Relevance to Biocultural Diversity. *Plants* **2022**, *11*, 503. [[CrossRef](#)]
97. Fernández, A.; León-Lobos, P.; Contreras, S.; Ovalle, J.F.; Sershen; van der Walt, K.; Ballesteros, D. The potential impacts of climate change on ex situ conservation options for recalcitrant-seeded species. *Front. For. Glob. Change* **2023**, *6*, 1110431. [[CrossRef](#)]
98. Chen, C.; Chang, H.; Pang, X.; Liu, Q.; Xue, L.; Yin, C. Genetic diversity analysis and conservation strategy recommendations for ex situ conservation of *Cupressus chengiana*. *BMC Plant Biol.* **2025**, *25*, 552. [[CrossRef](#)]
99. Badro, H.; Ndjiondjop, M.-N.; Furtado, A.; Henry, R. SNPs Linked to Key Traits in Hybrids between African and Asian Rice. *Proceedings* **2019**, *36*, 25. [[CrossRef](#)]
100. Hodgson-Kratky, K.; Perlo, V.; Furtado, A.; Choudhary, H.; Gladden, J.M.; Simmons, B.A.; Botha, F.; Henry, R.J. Association of gene expression with syringyl to guaiacyl ratio in sugarcane lignin. *Plant Mol Biol.* **2021**, *106*, 173–192. [[CrossRef](#)] [[PubMed](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.