

RESEARCH ARTICLE

Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay

Michael J. Roach¹, Daniel L. Johnson¹, Joerg Bohlmann^{2,3}, Hennie J. J. van Vuuren^{2,3}, Steven J. M. Jones⁴, Isak S. Pretorius⁵, Simon A. Schmidt¹*, Anthony R. Borneman^{1,6}*

1 The Australian Wine Research Institute, Glen Osmond, South Australia, Australia, **2** Michael Smith Laboratories, The University of British Columbia, Vancouver, British Columbia, Canada, **3** Wine Research Centre, Faculty of Land and Food Systems, University of British Columbia, Vancouver, British Columbia, Canada, **4** Michael Smith Genome Sciences Centre, British Columbia Cancer Research Centre, Vancouver, British Columbia, Canada, **5** Chancellery, Macquarie University, Sydney, New South Wales, Australia, **6** Department of Genetics and Evolution, University of Adelaide, South Australia, Australia

* These authors contributed equally to this work.

* simon.schmidt@awri.com.au (SAS); anthony.borneman@awri.com.au (ARB)



OPEN ACCESS

Citation: Roach MJ, Johnson DL, Bohlmann J, van Vuuren HJJ, Jones SJM, Pretorius IS, et al. (2018) Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. *PLoS Genet* 14(11): e1007807. <https://doi.org/10.1371/journal.pgen.1007807>

Editor: Paul M. Richardson, MicroTrek Incorporated, UNITED STATES

Received: August 9, 2018

Accepted: November 2, 2018

Published: November 20, 2018

Copyright: © 2018 Roach et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Raw sequencing reads, and the Chardonnay assembly and annotations are available under the BioProject accession PRJNA399599 at the National Center for Biotechnology Information. Additional Chardonnay genome annotations and files are available under DOI: [10.5281/zenodo.1480037](https://doi.org/10.5281/zenodo.1480037). All custom code and analysis workflows required to reproduce the results presented are included in [S1 Archive](#).

Funding: This project is supported by Australia's grapegrowers and winemakers through their

Abstract

Chardonnay is the basis of some of the world's most iconic wines and its success is underpinned by a historic program of clonal selection. There are numerous clones of Chardonnay available that exhibit differences in key viticultural and oenological traits that have arisen from the accumulation of somatic mutations during centuries of asexual propagation. However, the genetic variation that underlies these differences remains largely unknown. To address this knowledge gap, a high-quality, diploid-phased Chardonnay genome assembly was produced from single-molecule real time sequencing, and combined with re-sequencing data from 15 different Chardonnay clones. There were 1620 markers identified that distinguish the 15 clones. These markers were reliably used for clonal identification of independently sourced genomic material, as well as in identifying a potential genetic basis for some clonal phenotypic differences. The predicted parentage of the Chardonnay haplotypes was elucidated by mapping sequence data from the predicted parents of Chardonnay (Gouais blanc and Pinot noir) against the Chardonnay reference genome. This enabled the detection of instances of heterosis, with differentially-expanded gene families being inherited from the parents of Chardonnay. Most surprisingly however, the patterns of nucleotide variation present in the Chardonnay genome indicate that Pinot noir and Gouais blanc share an extremely high degree of kinship that has resulted in the Chardonnay genome displaying characteristics that are indicative of inbreeding.

Author summary

Phenotypic variation within a grapevine cultivar arises from an accumulation of mutations from serial vegetative propagation. Old cultivars such as Chardonnay have been

investment body Wine Australia, with matching funds from the Australian Government. Grapevine genomics work at the AWRI is also supported by Bioplatforms Australia as part of the National Collaborative Research Infrastructure Strategy, an initiative of the Australian Government. Funding was also obtained from Genome British Columbia and the Wine Research Centre at The University of British Columbia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

propagated for centuries resulting in hundreds of available ‘clones’ containing unique genetic mutations and a range of various phenotypic peculiarities. The genetic mutations can be leveraged as genetic markers and are useful in identifying specific clones for authenticity testing, or as breeding markers for new clonal selections where particular mutations are known to confer a phenotypic trait. We produced a high-quality genome assembly for Chardonnay, and using re-sequencing data for 15 popular clones, were able to identify a large selection of markers that are unique to at least one clone. We identified mutations that may confer phenotypic effects, and were able to identify clones from material independently sourced from nurseries and vineyards. The marker detection framework we describe for authenticity testing would be applicable to other grapevine cultivars or even other agriculturally important woody-plant crops that are vegetatively propagated such as fruit orchards. Finally, we show that the Chardonnay genome contains extensive evidence for parental inbreeding, such that its parents, Gouais blanc and Pinot noir, may even represent first-degree relatives.

Introduction

Chardonnay is known for the production of some of the world’s most iconic wines and is predicted to be the result of a cross between the *Vitis vinifera* cultivars Pinot noir and Gouais blanc [1, 2]. Since first appearing in European vineyards, Chardonnay has spread throughout the world and has become one of the most widely cultivated wine-grape varieties [3]. For much of the 20th century, grapevine cultivars were generally propagated by mass selection. High genetic variability therefore existed between individual plants within a single vineyard and this heterogeneity often lead to inconsistent fruit quality, production levels, and in some wine-producing regions, poor vine health [4]. Clonal selection arose as a technique to combat these shortcomings, preserving the genetic profile of superior plants, while amplifying favourable characteristics and purging viral contamination, leading to improved yields [4, 5].

Chardonnay’s global expansion throughout commercial vineyards, which started to accelerate rapidly in the mid-1980s, coincided with the maturation of several clonal selection programmes based in France, the USA and Australia. As a result, there are now many defined clones of Chardonnay available that exhibit differences in key viticultural and oenological traits [3, 6–11]. For example, clone I10V1—also known as FPS06 [12]—showed early promise as a high-yielding clone with moderate cluster weight and vigorous canopy [13]. The availability of virus-free clonal material of I10V1 helped cement productivity gains in the viticultural sector and I10V1 quickly dominated the majority of the Australian Chardonnay plantings [4, 5].

Since the concurrent publication of two draft Pinot noir genomes in 2007 [14, 15] grapevine genomics has increasingly contributed to the understanding of this woody plant species. However, the haploid Pinot noir reference genome does not represent the typical complexity of commercial wine-grape cultivars and the heterozygous Pinot noir sequence remains highly fragmented [16]. In recent years, the maturation of single molecule long-read sequencing technologies such as those developed by PacBio [17] and Oxford Nanopore [18], and the development of diploid-aware assemblers such as FALCON [19] and CANU [20] have given rise to many highly-contiguous genome assemblies, including a draft genome assembly for the grapevine variety Cabernet sauvignon [19, 21–24]. Furthermore, whole genome phasing at the assembly level is possible with assemblers such as FALCON Unzip [19], allowing both haplotypes of a diploid organism to be characterised. For heterozygous diploid organisms, such as

Chardonnay, this is especially important for resolving haplotype-specific features that might otherwise be lost in a traditional genome assembly.

The aim of this work was to explore the diversity extant within Chardonnay clones. A reference genome for Chardonnay was assembled *de novo* from PacBio long-read sequence data against which short-read clonal sequence data was mapped. This led to the identification of clone diagnostic single-nucleotide polymorphisms (SNP) and Insertions/Deletions (InDel) that show little shared clonal heritage. Furthermore, comparison of the Chardonnay reference with Pinot noir revealed some unexpected complexities in haplotype features with implications for the pedigree of this important grapevine variety.

Results

Assembly and annotation of a high quality, heterozygous phased Chardonnay genome

Of the many Chardonnay clones available, clone I10V1 was chosen as the basis for the reference genome due to its prominent use in the Australian wine industry. The initial I10V1 genome was assembled, phased and polished using subreads generated from 54 PacBio RS II SMRT cells and the FALCON Unzip, Quiver pipeline [19, 25]. While this assembly method should produce an assembly in which the primary contigs represent the haploid genome content of the organism in question, the size of the initial assembly (580 Mb) significantly exceeded that expected for *V. vinifera* (450–500 Mb). Both analysis with BUSCO [26] and short-read mapping indicated that this increased size was primarily due to both copies of many genomic regions (rather than only a single haplotype) being represented in the primary contigs (S1 Table and S1 Fig). This is a common problem when performing *de novo* assembly of highly heterozygous genomes [19, 27–29]. The primary contigs and haplotigs were processed with Purge Haplotigs [29]. This pipeline identifies contigs that are syntenic and reassigns contigs to get a complete haploid representation with minimal duplication. This approach reassigned 694 primary contigs as haplotigs (100 Mb) and 36 haplotigs as primary contigs (11 Mb), while also purging 18 artefactual contigs (1.3 Mb). Manual curation, based upon alignments to the PN40024 assembly [14] and subread mapping were used to address several mis-assemblies.

The final curated Chardonnay assembly consists of 854 primary contigs (N_{50} of 935 kb) and 1883 haplotigs, totalling 490 Mb and 378 Mb, respectively (Table 1). There were approximately 95% complete, and only 1.6% fragmented BUSCO-predicted genes (S1 Table). BUSCO duplication is also predicted to be low for both the primary contigs and the associated haplotigs (4% and 2% respectively). A custom repeat library was constructed for Chardonnay and used to

Table 1. Quast-based assembly statistics for the Chardonnay clone I10V1 genome.

	Primary contigs	Haplotigs
Contigs	854	1883
Contigs (≥ 50 kb)	838	1614
Assembly Size (Mb)	490.0	378.0
Largest Contig (Mb)	6.35	1.91
GC (%)	34.41	34.45
N_{50} (kb)	935.8	318.4
N_{75} (kb)	502.7	165.3
L50	145	335
L75	321	749

<https://doi.org/10.1371/journal.pgen.1007807.t001>

annotate 336 Mb (38.7%) of the diploid genome as repetitive. RNAseq data were used to annotate potential coding regions of the primary contigs using Maker [30], which predicted 29 675 gene models (exclusive of repetitive regions) and 66 548 transcripts in total.

Phasing coverage, and identification of homozygous and hemizygous regions

While long-range connectivity information is needed to produce a more contiguous scaffold or chromosomal assembly for Chardonnay, it was possible to assign contigs to chromosomes using the PN40024 Pinot Noir assembly for the purposes of downstream analysis and visualisation. A total of 614 primary contigs (397 Mb) and 1502 haplotigs (305 Mb) were confidently placed in chromosomal order using the PN40024 scaffold as a reference. To analyse the degree and distribution of heterozygosity across the genome, read depth (from mapped RS II subreads to the diploid assembly), heterozygous variant density (from mapped Illumina short-reads) and phasing coverage (regions of primary contigs that are homologous to haplotigs) were calculated for the primary contigs (Fig 1A).

Chromosomes 2, 3, 7, 15, 17, and 18 contain runs of homozygosity greater than 500 kb (intersect of lack of phasing coverage, double read-depth, low heterozygous variant density), and are indicative of inbreeding or gene conversion events. There are a further 22.8 Mb that lacked phasing coverage, had low heterozygous variant density, but had a *median* read-depth. These regions presumably result from either hemizyosity of these genomic regions, or undetected allelic duplicates remaining in the primary contigs, as opposed to homozygosity resulting from inbreeding or gene conversions.

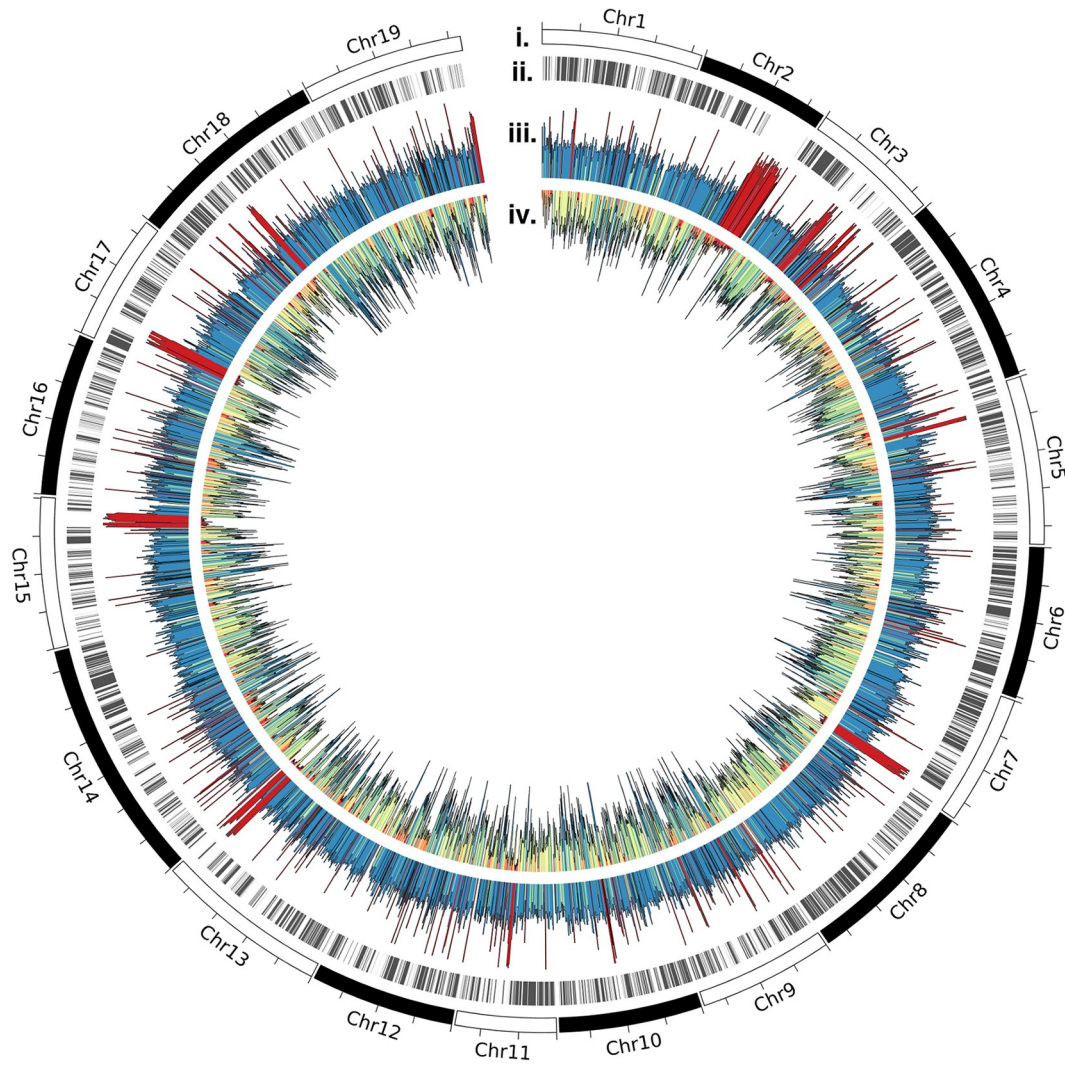
The largest run of homozygosity identified resides on Chromosome 2 (from 12.5 Mbp to the end of Chr2) and aligns closely to the Pinot noir assembly at over 99.8% identity (Fig 1B). There is also a region of synteny remaining in the primary contigs (from 5.6 to 6.5 Mbp on Chr2), evidenced by the ends of neighbouring primary contigs aligning to the same region in Pinot noir. In addition, there were two regions of low heterozygous variant density, poor phasing coverage, and median read-depth (9.9 to 10.3 Mbp and 11.5 to 12.6 Mbp on Chr2). BLAST searches of these regions to the remaining primary contigs and haplotigs were performed and the alignments were manually inspected; there were no significant alignments to other contigs or haplotigs and as such these regions appear to be hemizygous.

Defining parental contributions to the Chardonnay genome

To further refine the relationship between Chardonnay and the two varieties previously reported to be its parents [1, 2, 31], an attempt was made to identify the parental origin of each allele in the diploid Chardonnay assembly. Phase blocks were assigned across the genome by stringently aligning and trimming both the primary contigs and haplotigs into pairs of closely aligning syntenic sequence blocks (P and H alleles). This produced 1153 phase-blocks covering 270 Mb of the genome (71% of the haplotigs). Each pair of phase blocks should have one allele inherited from each parent. To assign likely genomic parentage within each phase block, short-reads from Gouais blanc, and a merged dataset comprising sequencing reads from several different genetic clones of Pinot noir (Pinot blanc, gris, and meunier, hereafter referred to as Pinot) [32] were mapped to the phase block sequences. The proportion of inherited nucleotide variation (using heterozygous variant loci) was then used to attribute the likely parentage of each block.

It was possible to confidently assign parentage to 197 Mb of the 244 Mb of chromosome-ordered phase-blocks (Fig 2A). Interestingly, rather than a 1:1 ratio of Gouais blanc to Pinot matches, Pinot was shown to match a higher proportion of the phase blocks (49% versus 34%

A



B

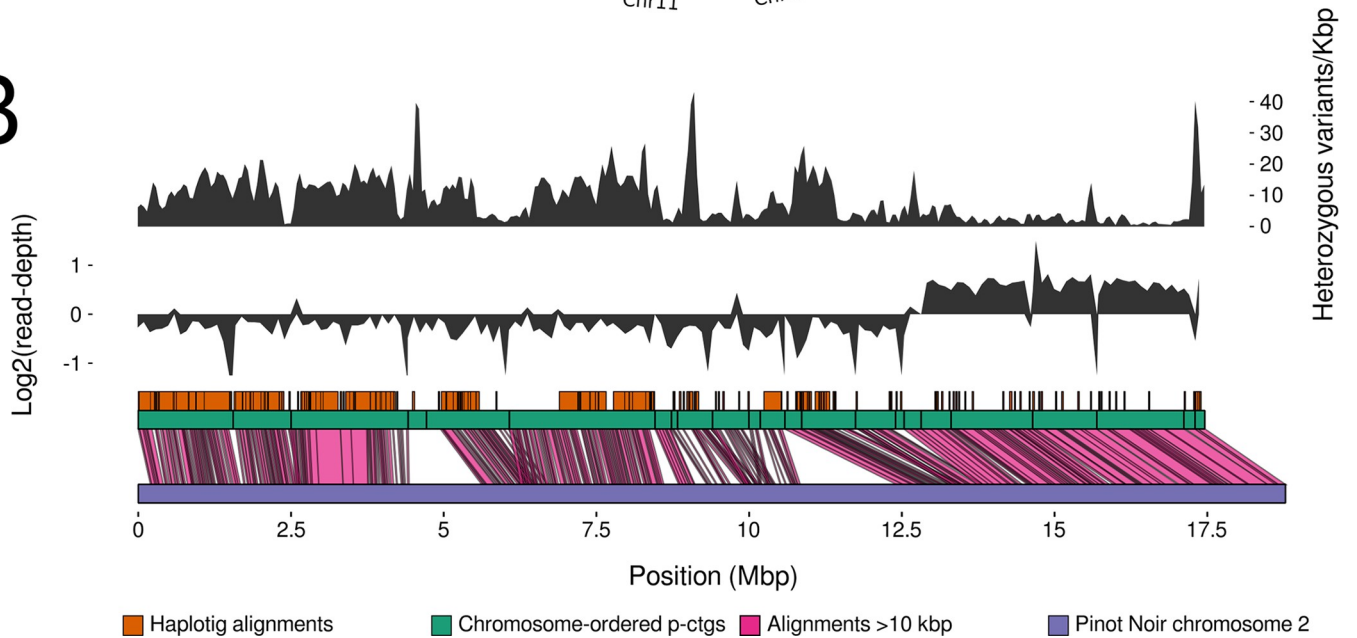


Fig 1. The *Vitis vinifera* cultivar Chardonnay reference genome. (A) A circos plot showing chromosome-ordered primary contigs (i), haplotig alignments (ii), read-depth of RS II subreads mapped to diploid assembly (read-depth colour scale: yellow, low; blue, high; red, double) (iii), and heterozygous variant density (SNP density colour scale: red, low; blue, high) (iv). (B) An expanded view of Chardonnay Chromosome 2 showing heterozygous variant density (top track), log₂ read-depth (middle track), and alignment with Pinot noir Chromosome 2 (bottom track).

<https://doi.org/10.1371/journal.pgen.1007807.g001>

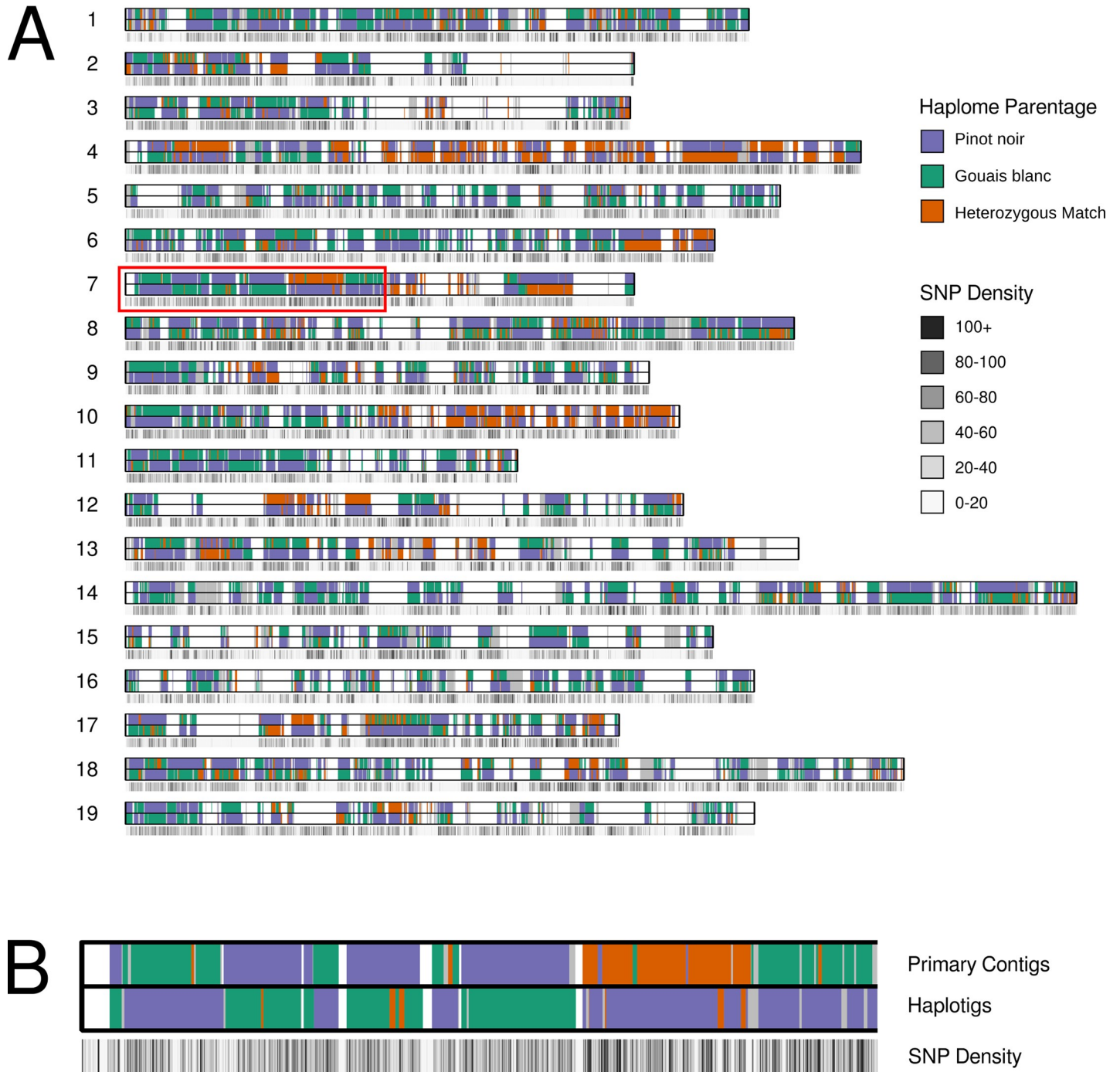


Fig 2. Parental architecture of the Chardonnay genome. (A) An ideogram of the Chardonnay reference assembly with the positions of both primary contig and haplotig phase-blocks indicated and juxtaposed with a SNP density track (for the primary contigs). Gaps in phase-blocks are indicated in white. (B) An enlargement of a region of *Vitis vinifera* Chromosome 7 (red box) in Fig 2A.

<https://doi.org/10.1371/journal.pgen.1007807.g002>

Gouais blanc), suggesting that the Pinot noir genome has contributed a higher proportion of genetic material to Chardonnay than Gouais blanc. However, further complicating this imbalance was the observation that in the remaining 17% of assigned regions, the pattern of nucleotide variation across the two heterozygous Chardonnay haplotypes matched both haplotypes of Pinot, with one of these haplotypes also matching one of the Gouais haplotypes. These ‘double Pinot haplotype’ regions are in some cases many megabases in size and are indicative of a common ancestry between Pinot and Gouais blanc.

While reciprocity (one Gouais blanc haplotype, one Pinot haplotype) was observed between allelic phase-blocks for over 95% of the parentage-assigned sequence, frequent haplotype switching (a known characteristic of FALCON-based assemblies) was observed between the haplomes, producing a haplotype mosaic which is observable as a ‘checkerboard’ pattern that alternates between the primary contigs and haplotigs for each chromosome (Fig 2B). Finally, an orthogonal kmer-based approach was used to call IBS over the primary contigs and haplotigs to ensure that the data were not biased by analysing only the phase blocks or by placing to the Pinot noir reference chromosomes. There was generally excellent concordance between the SNP- and kmer-based IBS methodologies (S2 Fig).

Parental-specific genomic variation

With parental contributions delineated in the Chardonnay assembly, it was possible to determine the parental origins of variation between orthologous chromosomes, including parent-specific gene family expansions. The Pinot noir reference genome was leveraged to identify the contribution of Gouais blanc; tandem pairs of orthologous proteins were defined in Chardonnay and filtered to identify tandem orthologs that were both expanded in Chardonnay compared to the Pinot noir reference assembly, and which resided in regions of the Gouais blanc haplome that did not match the Pinot haplotype. Chromosome alignments containing gene expansion candidates were inspected for features consistent with tandem gene duplication. Using this analysis, an expansion of Fatty Acyl-CoA Reductase 2-like (*FAR2-like*) genes on Chromosome 5 was identified, with the arrangement of *FAR2-like* open reading frames (ORFs) consistent with a tandem duplication event (S3 Fig).

A protein-based phylogeny was produced that encompassed the four *FAR2-like* ORFs present in the Chardonnay assembly, in addition to the homologous proteins from the Pinot noir PN40024 assembly (Fig 3A). Using these data, the Chardonnay haplotig sequence was identified as being derived from Pinot (nucleotide sequences of *FAR2-like* genes from Pinot noir and the Chardonnay haplotig were identical). However, rather than having an orthologous set of protein-coding regions, the genomic sequence derived from Gouais blanc (present in the primary contig) is predicted to encode two additional copies of *FAR2-like* homologues and an extra *FAR2-like* pseudogene (Fig 3B). While the ORF that was orthologous to the Pinot *FAR2-like* gene was closely related to the Pinot noir *FAR2-like* gene (98% identity), the two additional ORFs from Gouais blanc were more distantly related (93–94% identity).

Clonal nucleotide variation within a grapevine cultivar

As for many commercial grapevine varieties, there are currently many clones of Chardonnay, with each exhibiting a unique range of phenotypic traits. However, unlike varietal development, all of these genetic clones were established through the repeated asexual propagation of cuttings that presumably trace back to an original Chardonnay plant. It is therefore an accumulation of somatic mutations, that has contributed to phenotypic differences that uniquely define each clone and which provide an avenue for the confirmation of a clone’s identity. While clonal variation has so far been ill-defined in grapevine, the availability of the

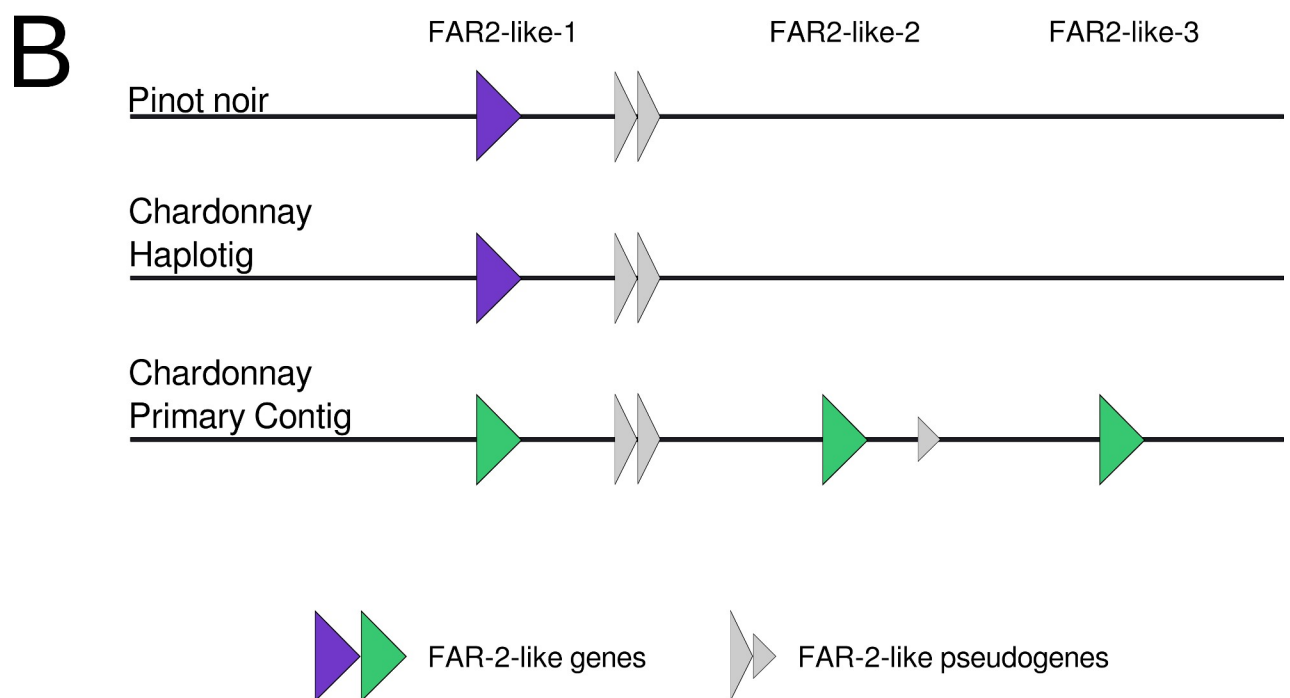
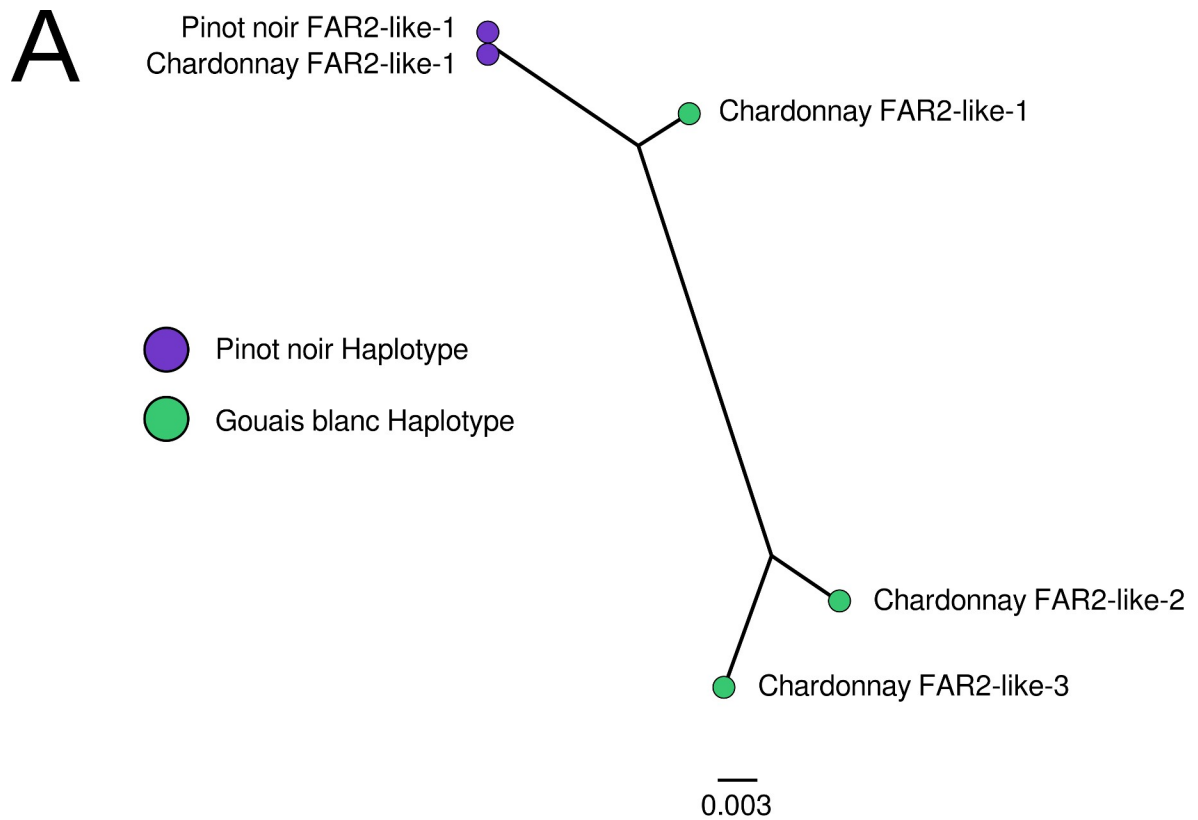


Fig 3. A FAR2-like expanded gene family in Chardonnay. (A) An unrooted tree of *FAR2-like* genes. (B) A schematic of the predicted genomic arrangement of the *FAR2-like* genes in Chardonnay. Both Pinot noir-derived (purple) and Gouais blanc-derived genes (green) are shown.

<https://doi.org/10.1371/journal.pgen.1007807.g003>

Chardonnay reference genome provides an opportunity to investigate the SNP spectrum that has arisen during the long history of Chardonnay propagation.

To begin to catalogue the diversity that exists across the clonal landscape of Chardonnay, short-read re-sequencing was used to define single nucleotide variation across 15 different Chardonnay clones; the known origins, yields and prominent traits of these clones are summarised in [S2 Table](#). The analysis of these highly related genomes (separated by a low number of true SNPs) was facilitated through the use of a marker discovery pipeline developed to call variants while applying a stringent kmer-based filter to remove false positives (including those calls due to sequencing batch or individual library size distribution at the expense of some false negative calls). Similar kmer approaches have been reported with excellent fidelity [33]. After filtering, 1620 high confidence marker variants evenly distributed across the Chardonnay genome were identified ([Table 2](#), [S4 Fig](#), and Sheet 1 in [S1 Dataset](#)). Variant calls were concatenated and used to generate a Chardonnay clone phylogeny ([Fig 4](#)).

‘CR Red’ and ‘Waite Star’, suspected phenotypic mutants of I10V1 that have red-skinned and seedless berries respectively [34, 35], formed a tight clade with only 40 variants (36 SNPs, 4 InDels) separating the three samples ([Fig 4](#)). The tight grouping of these clones confirms that

Table 2. A summary of Chardonnay clonal marker variants.

Sample Group	Number of SNPs and InDels
Mendoza	221
809	187
95	150
G9V7	143
277	137
76	137
548	133
352	121
1066	120
96	61
CR Red, Waite Star, I10V1	60
118	27
Waite Star	26
118, 352, 124	24
CR Red, Waite Star, I10V1, 277	18
CR Red	14
352, G9V7	11
76, 548	10
CR Red, Waite Star, I10V1, Mendoza, G9V7, 95, 277, 352, 96, 1066, 124, 118, 809	8
118, 124, 96	4
CR Red, Waite Star, I10V1, Mendoza, 809, 95, 277	2
CR Red, I10V1	2
118, 1066, 124, 96	2
124	2*

*alternate base calls were very low for these two variants.

<https://doi.org/10.1371/journal.pgen.1007807.t002>

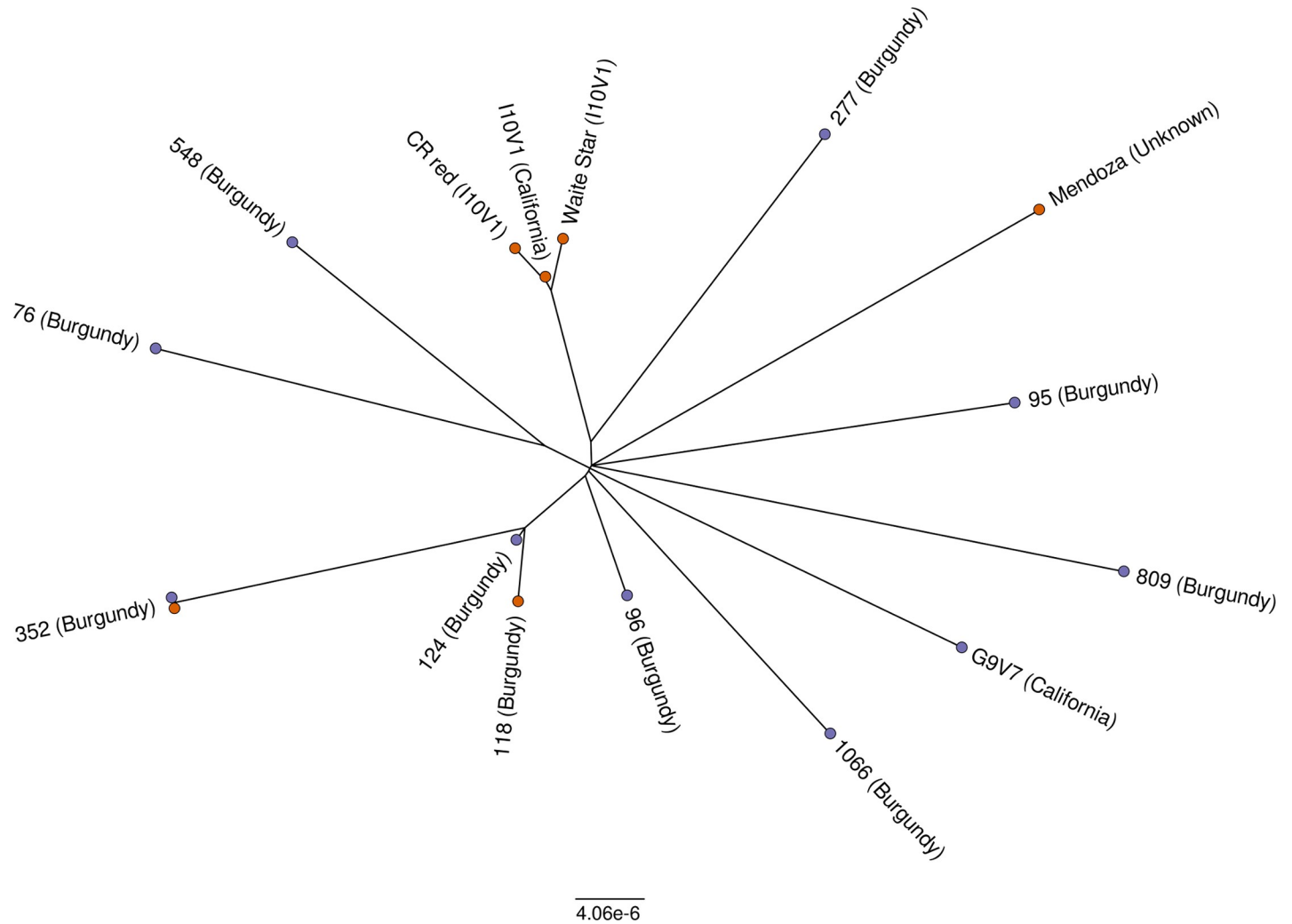


Fig 4. Genetic diversity in Chardonnay clones. An unrooted tree of Chardonnay clones based upon bi-allelic SNPs. Sequencing batches are designated by coloured terminal nodes (orange, sequencing batch #1; purple, sequencing batch #2).

<https://doi.org/10.1371/journal.pgen.1007807.g004>

the variant discovery pipeline can reliably detect recent clonal relationships from independent tissue samples. There were no further *a priori* relationships known for the remaining clones. However, the variant analysis would suggest that clones 124 and 118 also share some common ancestry as they are separated by only 23 SNPs.

The accumulation of SNPs can also lead to phenotypic differentiation that underlies the clonal selection process. For example, ‘Muscat character’, one of the most influential clone-specific phenotypic variants of Chardonnay, results from one of several single nucleotide substitutions that produce non-synonymous amino acid changes in 1-deoxy-D-xylulose-5-phosphate synthase 1 (*DXS1*) gene and are associated with the production of higher levels of monoterpenoids [36]. A combination of Annovar [37] and Provean (Choi *et al.*, 2012) were therefore used to annotate and predict the potential protein-coding consequences of each of the marker variant mutations identified. This pipeline correctly identified a previously characterised Muscat mutation (S272P) in *DXS1* in clone 809, the only Chardonnay clone in this study that is known to display the Muscat character. Provean scored this mutation at -3.37 , where values less than -2.5 generally signify an increased likelihood that the mutation impacts

the function of the enzyme. In addition to this known Muscat mutation, an additional 55 marker mutations were identified that displayed a high chance of impacted protein function (Sheet 2 in [S1 Dataset](#)). However, further work is required to investigate the links between known inter-clonal phenotypic variation and these specific mutations.

The application of SNP and InDel-based markers for clone-specific genotyping

While various phenotypic characteristics (known as ampelography) and microsatellite based genetic tests can be used to positively identify grapevines, the accurate identification of specific clones is extremely difficult and to date, microsatellite-based marker systems have proven unreliable for the identification of clonal material [38, 39]. Uncertainties can therefore exist as to the exact clone that has been planted in many vineyards. To enable a rapid clonal re-identification methodology, a kmer-approach was developed (similar to the method described in Shajii, Yorukoglu [40]) for screening raw short-read sequence data from unknown Chardonnay samples against the pre-identified clone-specific variants. This method queries known marker variants against a kmer count database generated from the unknown sample. The matching markers and sample groups are returned allowing the potential identification of the unknown Chardonnay sample.

The marker detection pipeline was tested using data from a variety of different samples and sequencing methods (Sheet 3 in [S1 Dataset](#)). Chardonnay clones 76, 124, 548, 809, 1066, and G9V7 were independently sequenced at a second location from the same genomic DNA that was used for the original identification of nucleotide variants; however, both a different library preparation and sequencing platform were used. This enabled an evaluation of marker suitability and false discovery rates in a best-case scenario (i.e. when the source DNA was the same). The pipeline reports the number of markers that were found, and reports the number of markers that were missing as either 'missing', or as 'insufficient coverage' where the mean coverage at the marker locus was less than 12-fold. When screened with the pipeline ([Fig 5A](#)), between 30% (clone 548) and 77% (clone 1066) of the markers were detected for each of the samples. Total coverage for this dataset varied between 11.7-fold (clone 548) and 148-fold (clone 1066). However, uneven read-coverage resulted in low coverage at some marker loci regardless of total coverage. As such, nearly all the missing markers coincided with poor coverage at marker loci.

To validate suitability of the markers for clonal identification, Chardonnay clones 76, 95, 96, 277, 352, 548 and G9V7 were independently sourced and sequenced. High-coverage (142- to 244-fold) sequencing was performed for three of these independently-sourced clones ([Fig 5B](#)). Kmer analysis identified between 55% (clone 96) and 83% (clone 76) of the expected markers for each sample. However, despite being the same clones, there were a significant proportion (14% to 44%) of the expected markers in each of these three samples that were not found in the independent material and which could not be attributed to insufficient marker loci coverage. This indicates that there may be intra-clonal genetic variation that has accumulated during the independent passaging of clonal material. The markers shared between the discovery and validation datasets form the validated clonal identification markers; however, the unshared markers may prove useful in assigning lineage or regionality in supplement of clonal identification.

As the level of sequencing coverage ultimately impacts the economics of clonal testing, the impact of sequencing depth on marker identification was further assessed. Data consisting of the pooled results of two sequencing batches for independently-sourced clone 95 was subsampled to a range of coverages and then screened for clonal identification effectiveness ([Fig 5C](#)).

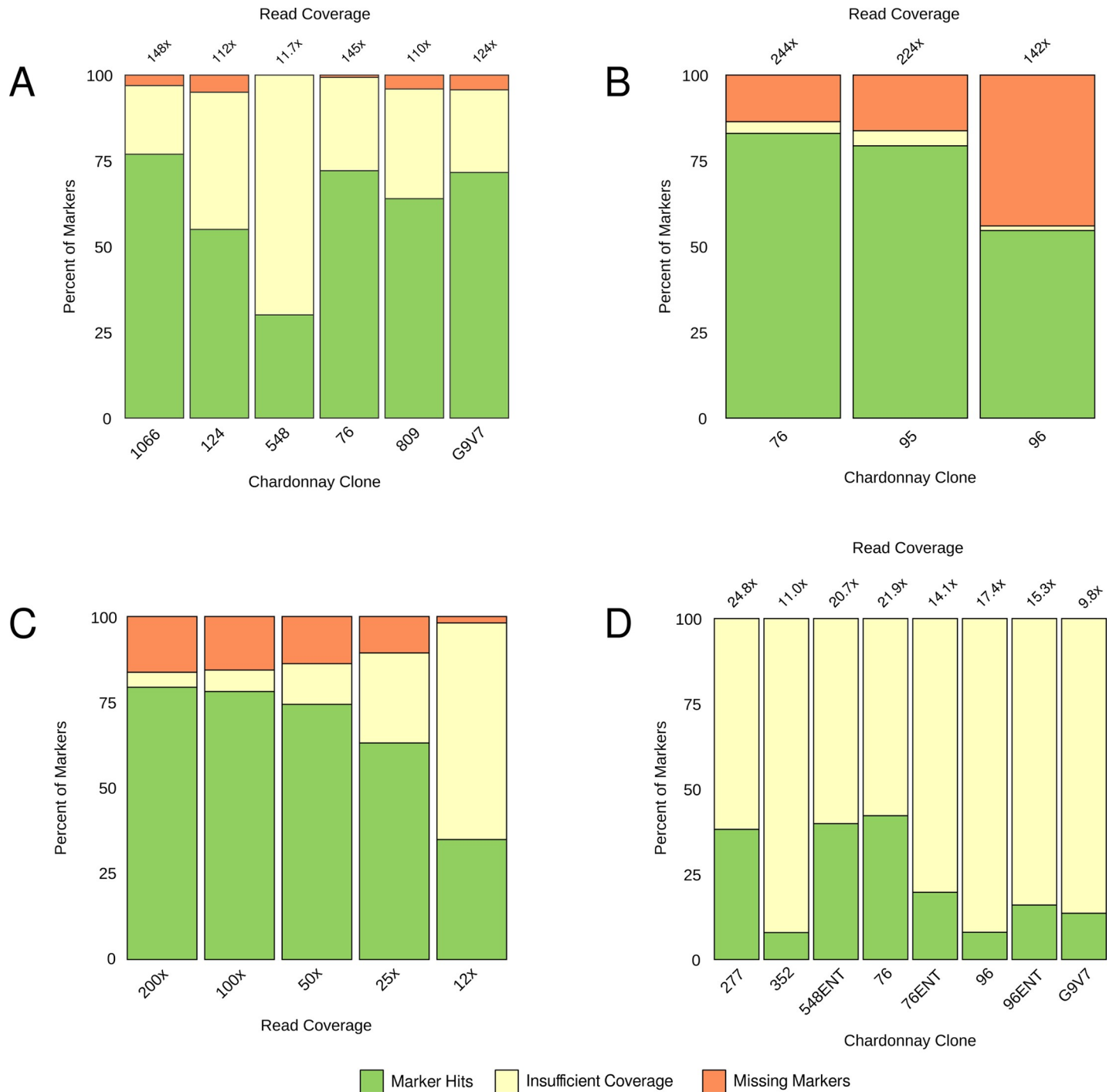


Fig 5. Validation of clonal markers for Chardonnay. Results of marker detection pipeline run on four WGS sequencing datasets. The percentage breakdown of clonal markers is shown as either markers identified in the datasets (marker hits), missing from the datasets (missing markers), or missing with low coverage at marker loci (insufficient coverage). (A) Moderate coverage sequencing replicates of marker discovery clonal material. (B) High-coverage sequencing of independently-sourced clonal material. (C) Coverage subsampling of independently-sourced clonal material (clone 95 from Fig 5B). (D) Low-coverage sequencing of DNA derived from several independently-sourced clones ('ENT' denotes ENTAV-INRA source material).

<https://doi.org/10.1371/journal.pgen.1007807.g005>

There was little difference in the number of discoverable markers from 200-fold down to 25-fold coverage (79% and 63% respectively), and only a 6% decrease in markers confidently-

flagged as missing. At 12-fold coverage it was still possible to detect 35% of the markers for this clone.

Given the successful results of the coverage titration, low coverage (9.8- to 24.8-fold) datasets were obtained from independent material of six clones, with clones 76 and 96 each sourced from proprietary and generic selections (Fig 5D). Despite the combination of independent material and low coverage it was still possible to detect between 7.9% (clone 352) and 42% (clone 76) of the expected markers for each sample.

Discussion

The genomic complexity of grapevine, combined with its clonal mode of propagation (absence of outcrossed populations), has so far limited classical genetic approaches to understanding inherited traits in this valuable crop. The availability of a reference genome for *V. vinifera* [14, 15] has facilitated genetic studies through the provision of additional microsatellite markers for parentage and other studies [31, 41–43] and more recently has driven the development of dense SNP arrays that are being used for analysis of population structure and genome wide association studies [44–46]. While not subject to the same technical limitations of microsatellite analysis [47], using predefined sets of SNPs also has its limitations, particularly with regard to discovery of novel genomic features. Recent advances in sequencing technology, and specifically read length, have provided a way forward, enabling repeat-rich genomes, such as grapevine, to be considered in their native state, without having to strip its inherent genomic variability in order to achieve a genome model with moderate contiguity.

A reference genome for Chardonnay was produced using long-read single-molecule sequence data in order to more precisely and accurately define the differences between the almost identical derivatives (clones) of a single cultivar. The Chardonnay assembly reported here exhibits a high level of contiguity and predicted completeness and provides a fundamental platform for the in-depth investigation of Chardonnay's genome function and, more generally, of grapevine evolution and breeding.

Heterosis (the improved biological function of a hybrid offspring) has been reported to have played a large role in the prominence of Gouais blanc and Pinot noir crosses in wine grapevines [5]. Deleterious mutations in inbred lines can lead to increased susceptibility to pests and diseases, reduced stress tolerance, and poorer biomass production [5]. This can be offset with the introduction of novel genes and gene families by crossing with a genetically dissimilar sample. The inheritance of an expanded family of *FAR2-like* genes from Gouais blanc represents one example of where this may have occurred in Chardonnay. The sequence divergence in *FAR2-like* copies and haplotypes suggests that the gene expansion event was not a recent occurrence. The increased gene copy number and sequence diversity potentially enriches the Chardonnay genome for both redundancy and functionality of this gene.

Fatty Acyl-CoA Reductase (FAR) enzymes catalyze the reaction: long-chain acyl-CoA + 2 NADPH → CoA + a long-chain alcohol + 2 NADP⁺ [48, 49]. There are numerous copies of FARs in plants and each tends to be specific for long-chain acyl-CoA molecules of a certain length [50]. FARs form the first step in wax biosynthesis and are associated with many plant surfaces, most notably epicuticular wax. Epicuticular waxes are important for protecting plants against physical damage, pathogens, and water loss [51–55]. It was reported in Konlechner and Sauer [56] that Chardonnay has a very high production and unique pattern of epicuticular wax; this might be attributed to novel FARs. The fatty alcohols produced by *FAR2* are associated with production of sporopollenin, which forms part of the protective barrier for pollen [57]. More work is needed to determine if the expanded family of *FAR2-like* genes identified here influences fertility or epicuticular wax levels in Chardonnay.

The Chardonnay genome enables thorough characterization of inter-clonal genetic variation. Attempts have been made in the past to use whole genome shotgun sequencing (WGS) to characterize inter-clonal diversity in other grapevine cultivars. These were ultimately limited by either available sequencing technology [58] or a lack of a reference genome for the particular grapevine variety under investigation [58, 59], although both studies were able to identify a small number of inter-clonal nucleotide variants. By taking advantage of both a reference genome for Chardonnay and increased read coverage, this study was able to identify 1620 high quality inter-clone nucleotide variants. There were limited shared somatic mutations among the Chardonnay clones, especially outside of the highly-related I10V1 group (I10V1, CR-Red and Waite Star). Clones 118 and 124, varieties from Burgundy used predominantly for sparkling wine production, were the exceptions to this, with 56% of their mutations being common between the two clones. Otherwise, the Chardonnay clones do not share a significant number of common mutations. This is likely the result of the centuries-long history of mass selection propagation. The clonal varieties of today likely represent a very small fraction of the genetic diversity that existed for Chardonnay after generations of serial propagation. The end result of this is that clear genetic distinctions can be made for the many clones of Chardonnay that originated as isolates from mass-selected vineyards.

Furthermore, as the marker discovery pipeline developed in this study was limited in scope to detecting nucleotide polymorphisms within non-repetitive areas of the genome, there are likely to be structural variants, such as transposon insertions, that also impact on clone-specific phenotypes. Nevertheless, marker mutations were identified for most of the clones that are predicted to impact gene function and could account for some of the clone specific phenotypes in Chardonnay.

Inter-clonal genetic variation provides an avenue for testing clone authenticity. The clone detection pipeline provides a fast and simple method to detect defined markers from a range of WGS library chemistries and platforms. Markers were reliably detected at coverages as low as 9.8-fold. Validation using independently-sourced clonal material indicated that most of the genetic variants were likely suitable for use in the identification of clones. Furthermore, there were a significant portion of markers that appeared to be variable across independently-sourced clonal material. This suggests that there might be region-specific genetic variation between clonal populations and this could potentially be exploited to further pinpoint the source of Chardonnay clones to specific regions, or to split clones into divergent subsets. The marker discovery and marker detection pipelines together form a solid framework for the future use of SNP- and InDel-based markers for the identification of unknown vegetatively propagated plant clones.

While the diploid Chardonnay reference genome enabled a much deeper understanding of the variation that has occurred since the initial establishment of this variety, it has also provided the means to unravel the detailed genetic ancestry of this variety and its parents, Pinot noir and Gouais blanc. Chardonnay matches both haplotypes of Pinot noir across approximately one fifth of its genome and these areas include large tracks of both homozygous and heterozygous variation. While the presence of the homozygous 'double-Pinot noir' regions could be the result of a high number of large-scale gene conversion events early in Chardonnay's history, the numerous heterozygous double-Pinot noir regions are only possible if the haplotype inherited from Gouais blanc was almost identical to the non-inherited allele of Pinot noir. Gouais blanc sequencing indeed confirms that within these 'double-Pinot noir' regions, one of the two Pinot noir haplotypes is a match for an allele of Gouais blanc.

The data reported in this work therefore supports a more complicated pedigree for Chardonnay than simply a sexual cross between two distantly related parents (Fig 6). The two parents of Chardonnay are predicted to share a large proportion of their genomes; this is

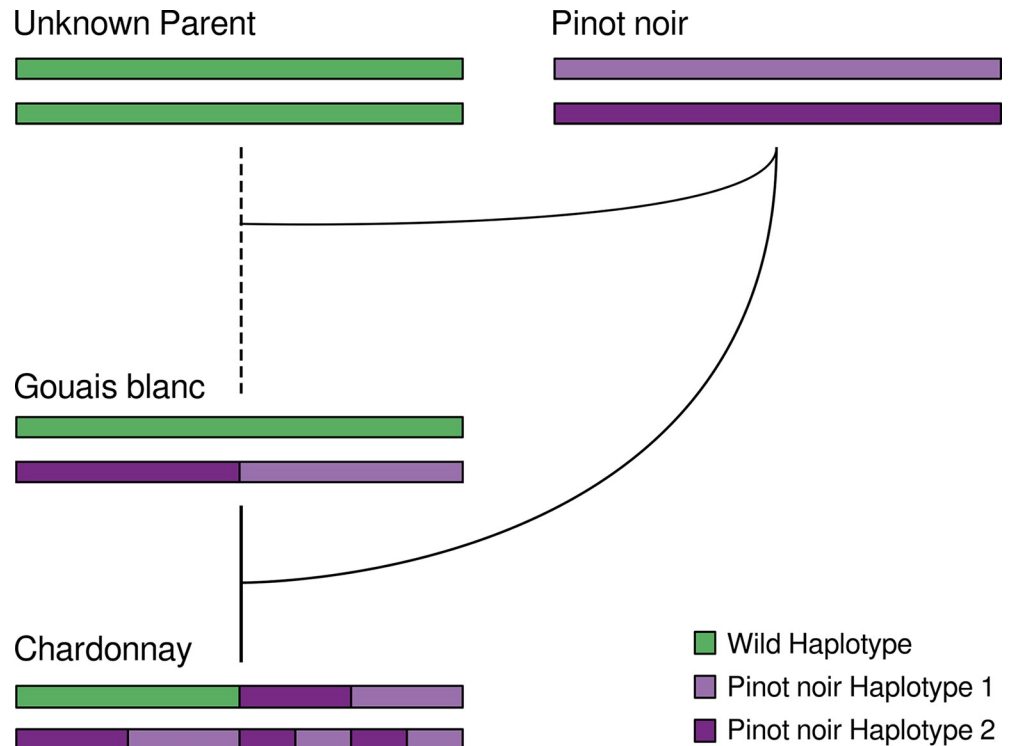


Fig 6. A schematic model for the complex pedigree of Chardonnay, Gouais blanc and Pinot noir. Two crossing events (akin to a standard genetic backcross) with Pinot noir would result in the homozygous and heterozygous Pinot noir regions present in Chardonnay.

<https://doi.org/10.1371/journal.pgen.1007807.g006>

suggestive of a previous cross between Pinot noir and a very recent ancestor of Gouais blanc (Pinot noir might even be a direct parent of Gouais blanc). Surprisingly, data supporting this complicated relationship between Gouais blanc and Pinot noir have appeared in previous low-resolution DNA marker analyses, with the two varieties sharing marker alleles at over 60% of marker loci in two separate studies [1, 31]. However, the potential kinship between the two ancient varieties could not have been discovered without the insights provided by this diploid-phased Chardonnay genome.

A high-quality, diploid-phased Chardonnay assembly provided the means to assess several interesting facets of grapevine biology. It was possible to detect instances of heterosis, with differentially-expanded gene families being inherited from the parents of Chardonnay and to define the nucleotide variation that has accumulated during asexual propagation of this woody-plant species. However, most surprisingly, the completed genome indicates that the parents of Chardonnay shared a high degree of kinship, suggesting that the pedigree of this important wine-grape variety might be more complicated than originally thought.

Methods

All custom scripts used for analysis, along with detailed workflows are available in [S1 Archive](#).

DNA preparation and sequencing

Nuclear DNA was isolated from early season, disease free, field grown Chardonnay leaves taken from plants at a nursery vineyard (Oxford Landing, Waikerie, South Australia). DNA was extracted by Bio S&T (Quebec, Canada) from nuclear-enriched material using a CTAB/

Chloroform method. DNA from clone I10V1 was enriched using a 1:0.45 Ampure cleanup prior to being used to build 15–50 kb SMRT Bell libraries with Blue Pippin size selection following library preparation (Ramaciotti Centre for Genomics, UNSW, Sydney, Australia). These libraries were sequenced on a PacBio RS II using 54 SMRT cells to give a total sequencing yield of 51,921 Mb (115-fold coverage) with an N_{50} length of 14.4 kb. Short-read sequencing of clones for marker discovery was performed on Illumina HiSeq 2000 and HiSeq X-Ten platforms from TruSeq libraries (100 and 150 bp paired end read chemistries). Short-read sequencing of clones for marker validation was performed on Illumina HiSeq 2500 and MiSeq platforms from Nextera libraries made from material sourced from both Foundation Plant Services (University of California, Davis) and Mission Hill Family Estate, Quail's Gate and Burrowing Owl wineries in British Columbia, Canada.

Assembly

The FASTA subreads were used to assemble the genome using FALCON (commit: 103ca89). Length cut-offs of 18 000 bp and 9 000 bp were used for the subread error correction and error-corrected reads respectively. FALCON Unzip (commit: bfa5e6e) was used with default parameters to phase the assembly from the FASTA subreads and Quiver-polish from the raw sequencing data.

The Purge Haplotigs pipeline (commit: f63c180) [29] was developed to automate the identification and reassignment of syntenic contigs from highly heterozygous long-read based assemblies. The PacBio RS II subreads were mapped to the diploid assembly (primary contig and haplotigs) using BLASR (packaged with SMRT-Link v3.1.0.180439) [60] and sorted with SAMtools v1.3.1. As required by Purge Haplotigs, read-depth thresholds were chosen to capture both peaks (diploid and haploid coverage levels) from the bimodal read-depth histogram and a contig-by-contig breakdown of average read-depth was calculated. Purge Haplotigs takes the read-depth summary and uses sequence alignments to reassign contigs. Curated primary contigs were assigned to *V. vinifera* chromosomes by using the PN40024 Pinot noir reference genome for scaffolding and for the identification of possible mis-assemblies. Several mis-assemblies were identified and manually corrected. The haploid and diploid curated assemblies were evaluated with BUSCO v3.0.1 using the embryophyta ODB v9 database.

Annotation

A custom repeat library was produced for Chardonnay for use with RepeatMasker, similar to the method described in Fallon, Lower [61]. Miniature inverted-repeat transposable element (MITE) sequences for *V. vinifera* were obtained from the P-MITE database [62]. Repeats were predicted using RepeatModeler open-1.0.10 [63], and the RepeatModeler predictions and MITE sequences were concatenated to produce the custom Chardonnay repeat library. Repeats were annotated using RepeatMasker open-4.0.7 using this custom library.

RNA-seq was performed on total RNA extracted from I10V1 leaf tissue, extracted using a Spectrum Plant Total RNA Kit (Sigma), and sequenced using Illumina paired-end 75 bp chemistry on the HiSeq 2500 platform (Michael Smith Genome Sciences Centre, British Columbia Cancer Research Centre, British Columbia). Additional RNA-seq data from Chardonnay berry skins were obtained from the Sequence Read Archive (BioProject: PRJNA260535). All RNA-seq reads were mapped to the Chardonnay genome using STAR v2.5.2b [64], with transcripts predicted using Cufflinks v2.2.1 [65]. Initial transcript predictions and repeat annotations were then used in the Maker gene prediction pipeline (v2.31.9) using Augustus v3.2.3 [66]. The predicted proteins were assigned OrthoMCL [67] and KEGG annotations [68] for orthology and pathway prediction. Draft names for the predicted proteins

were obtained from protein BLAST v2.2.31+ [69] hits against the Uniprot knowledgebase [70, 71] using an evalue cutoff of $1e-10$.

Parental mapping

Using BLAST and MUMmer v4.0.0beta [72] alignments, the primary contigs were aligned to the PN40024 reference, and the haplotigs were aligned to the primary contigs. The primary contig alignments were used for placing primary contigs in chromosome-order. The haplotig alignments (to the primary contigs) were used to trim and extract the closely aligning phase-blocks between the primary contigs and haplotigs. Coordinate tables were produced that could be used for mapping the phase-blocks to the primary contigs and then to the chromosome-ordered scaffolds.

To identify the most likely parent for each phase-block pair, publicly-available short-read sequencing data were obtained for three clonally-derived variants of Pinot noir; Pinot blanc, Pinot gris, and Pinot meunier (BioProject: PRJNA321480); data for Pinot noir were not available at the time of analysis. To avoid potential issues with data from any single Pinot variety, pooled reads from all three were used for mapping. The sequencing data for Pinot, Chardonnay, and Gouais blanc were mapped to the primary contig and haplotig phase-block sequences using BWA-MEM v0.7.12 [73]. PCR duplicates and discordantly-mapped reads were removed, and poorly mapping regions were masked using a window coverage approach. Heterozygous SNPs were called using VarScanv2.3 [74] (p -value $< 1e-6$, coverage > 10 , alt reads $> 30\%$) and Identity By State (IBS) was assessed over 10 kb windows (5 kb steps) at every position where a heterozygous Chardonnay SNP was found.

Where the parent (Pinot or Gouais blanc) was homozygous and matched the reference base, an IBS of 2 was called. Where the parent was homozygous and did *not* match the reference base, an IBS of 0 was called. Finally, where the parent and Chardonnay had identical heterozygous genotype, an IBS of 1 was called. The spread of IBS calls was used to assign windows as 'Pinot', 'Gouais blanc', or 'double-match'. The window coordinates were transformed to chromosome-ordered scaffold coordinates and neighbouring identically called windows were chained together. Complementary Pinot/Gouais blanc calls from the parent datasets were merged and clashing calls removed. For ease of visualisation, the 'double-match' calls from the Pinot dataset were merged with the Gouais blanc calls (and vice versa). A SNP density track for the Chardonnay primary contigs was created over 5-kb windows from previously-mapped Illumina reads. The chromosome ideograms with SNP densities and IBS assignments were produced in Rstudio using ggplot2.

An orthologous kmer method for assigning parentage was developed to assign parentage over the entire genome. All 27-bp-long kmers (27mers) were counted using JELLYFISH v2.2.6 [75] (canonical representation, singletons ignored) directly from Pinot, Gouais blanc, and Chardonnay I10V1 paired-end reads to create 27mer count databases. Non-overlapping 1-kb windows were generated for the primary contigs and the haplotigs. For each window all 27mers were extracted from the contig sequences, queried against the kmer count databases using JELLYFISH and the number of kmers not appearing in each were returned. The Pinot/Gouais blanc missing kmer counts were normalised against the Chardonnay counts and averaged over 10 kb windows with 5-kb steps. Windows with 150 or more missing kmers (approximately 0.56 SNPs/Kb) were classified as mismatch (missing kmer density was visualized over the genome to determine an appropriate cut-off) and neighbouring complementary windows were merged.

Gene expansion

Protein-based BLAST alignments of Chardonnay proteins were performed against the Pinot noir (PN40024) reference proteome. Chardonnay Maker GFF annotations and PN40024 GFF

annotations were converted to BED format. Chardonnay annotations were then transformed to scaffold coordinates. The 'blast_to_raw.py' script (from github.com/tanghaibao/quota-alignment) was used to flag tandem repeat homologues for both the primary contig and haplotig Chardonnay proteins against the Pinot noir reference. Illumina paired-end reads for Pinot were mapped to the Chardonnay primary and haplotig assemblies, BED annotations were created for regions with poor mapping, and these annotations were transformed to the scaffold coordinates for use with filtering. Predicted expanded gene families in Chardonnay that resided in Gouais blanc regions (identified using the kmer parental mapping) that had multiple gene models with poor read-coverage of Pinot mapped reads were returned as a filtered list. Dotplots were produced with MUMmer and the chromosomes were visually assessed for evidence of tandem sequence duplication at the filtered gene expansion candidate loci. The genomic sequences of *FAR2-like* ORFs from the Pinot noir assembly and from the Chardonnay primary contigs and haplotigs were aligned using MUSCLE v3.8.31 [76] within AliView v1.20 [77]. Phylogenies were calculated within Rstudio using Ape [78] and Phangorn [79].

Marker variant discovery

Paired-end reads for each clone were manually quality trimmed using Trimmomatic v0.36 [80] (SLIDINGWINDOW:5:20, TRAILING:20, CROP:100), mapped to the Chardonnay reference genome using BWA-MEM and filtered for concordant and non-duplicated reads. Variant calls were made using VarScan (p-value < 1e-3, alt reads > 15%) with variants across each clone pooled into a combined set. The combined variant set was then compared against leniently-scored variant calls for each clone (number of alt reads > 5 and alt reads > 5% of all reads), with differences in genotype between clones resulting in that variant being flagged as a potential clonal marker.

Kmers were used to filter false positives from the pool of potential clonal markers. Kmer count databases (27mers) were created for each clone from the sequencing reads using JELLYFISH. For each potential marker, all possible kmers at the marker loci, from all samples, were extracted from the sequencing reads in the BAM alignment files. The kmer counts were queried from the kmer databases for each sample. Where a set of unique kmers were present for the matching samples, that variant was confirmed as a marker. This unique set of kmers that contain the marker mutation forms the 'marker kmers', whereas the kmers that were common to all of the clones form the 'reference kmers'. The marker variants, marker kmers, and reference kmers were output in a table for use with querying unknown Chardonnay clones. The heterozygous variant calls from the marker discovery pipeline (excluding the filtered variants) were concatenated and Phylogenies were calculated within Rstudio using Ape [78] and Phangorn [79].

Marker detection pipeline

Markers are detected directly from short-read sequencing data using kmers. A kmer count database (27mers) is calculated from raw sequencing reads as previously-described. The marker kmers and reference kmers that were identified in the marker discovery pipeline are then queried from the kmer database. A marker is flagged as a 'hit' if the marker kmers are present at a mean depth of 3 or greater, and if the ratio of the mean reference kmer depth to mean marker kmer depth was less than 10. Markers are flagged as 'insufficient read coverage' if the marker both fails to satisfy the conditions for a 'hit' and the reference kmers are present at a mean read-depth of less than 12.

Supporting information

S1 Archive. Scripts and workflows for data analysis. Extract with tar for linux or Mac, or with 7zip (7zip.org) for Windows. Contents: **bin/**, All custom scripts used for analysis; **lib/**,

Custom Perl library for scripts; **src/**, Source code for window coverage masking program; **workflows/**, Commands used with comments for all data analysis; **Makefile**, The GNU Make pipeline for marker discovery.

(GZ)

S1 Dataset. Chardonnay clonal-specific markers. Chardonnay clone-specific markers with read-counts (Sheet 1), Markers in gene models with Annovar and Provean predictions (Sheet 2), Summaries of clonal marker detection screening against validation datasets (Sheet 3).

(XLSX)

S1 Fig. Redundant primary contig reduction. (A) Circular representations of FALCON Unzip Chardonnay assembly and (B) the same assembly after curation. Tracks are: length-ordered contigs (i), read depth of mapped PacBio RS II subreads, coloured by read-depth (blue, high; yellow, median; red, low) (ii) and heterozygous SNP density, coloured by SNP density (blue, low; yellow, median; red, high) (iii).

(TIFF)

S2 Fig. Parental architecture of the Chardonnay genome: Comparison of SNP- and kmer-based IBS methods. (A) an ideogram of the Chardonnay reference assembly with the positions of primary contigs (but not haplotigs) for both the kmer- and SNP-based IBS methods, juxtaposed with homozygous annotations (fewer than 10 SNPs per 5 kb window). Gaps in phase-blocks are indicated in white. (B) An enlargement of a region of *Vitis vinifera* Chromosome 7 (red box in A).

(TIFF)

S3 Fig. Gene expansion of Chardonnay Chromosome 5 (primary contigs) region containing FAR2-like genes. Alignments are indicated as black lines (dotplot), the ORFs for FAR2-like genes and pseudogenes are indicated for both Pinot noir and Chardonnay.

(TIFF)

S4 Fig. Distribution of clonal markers over the Chardonnay assembly. Circos plot derived from Fig 1A with the inclusion of the clonal marker locations. The tracks shown are chromosome-ordered primary contigs (i), haplotig alignments (ii), read-depth of RS II subreads mapped to diploid assembly (read-depth colour scale: yellow, low; blue, high; red, double) (iii), locations of clonal markers (iv), and heterozygous variant density (SNP density colour scale: red, low; blue, high) (v).

(TIFF)

S1 Table. BUSCO analysis of the Chardonnay FALCON Unzip assembly before and after curation.

(PDF)

S2 Table. Information relating to Chardonnay clones used in this study.

(PDF)

Acknowledgments

We thank Nick Dry (Yalumba Nursery) and Michael McCarthy (South Australian Research and Development Institute) for provision of plant material; Andrew Lonie and Torsten Seemann (University of Melbourne), Cihan Altinay and Derek Benson (University of Queensland), Stephen Crawley (QCIF), QRIScloud, and Melbourne Bioinformatics for assistance with computing resources; Jason Chin, Gregory Concepcion and Emily Hatas (Pacific Biosciences) for early access and guidance on FALCON Unzip; Sean Myles who acted as an advisor to

Genome BC; Samantha Turner (University of British Columbia) for outstanding administrative duties; and Andrew Gilbert (Bioplatforms Australia) and Bioplatforms Australia for supporting this work.

Author Contributions

Conceptualization: Daniel L. Johnson, Joerg Bohlmann, Hennie J. J. van Vuuren, Steven J. M. Jones, Isak S. Pretorius, Simon A. Schmidt, Anthony R. Borneman.

Data curation: Michael J. Roach, Steven J. M. Jones, Simon A. Schmidt, Anthony R. Borneman.

Formal analysis: Simon A. Schmidt, Anthony R. Borneman.

Funding acquisition: Daniel L. Johnson, Joerg Bohlmann, Hennie J. J. van Vuuren, Steven J. M. Jones, Isak S. Pretorius.

Investigation: Michael J. Roach.

Methodology: Anthony R. Borneman.

Project administration: Simon A. Schmidt.

Resources: Michael J. Roach, Steven J. M. Jones, Simon A. Schmidt.

Software: Michael J. Roach.

Supervision: Simon A. Schmidt, Anthony R. Borneman.

Validation: Michael J. Roach, Simon A. Schmidt.

Visualization: Michael J. Roach.

Writing – original draft: Michael J. Roach, Simon A. Schmidt, Anthony R. Borneman.

Writing – review & editing: Michael J. Roach, Daniel L. Johnson, Joerg Bohlmann, Hennie J. J. van Vuuren, Steven J. M. Jones, Isak S. Pretorius, Simon A. Schmidt, Anthony R. Borneman.

References

1. Bowers J, Boursiquot J-M, This P, Chu K, Johansson H, Meredith C. Historical Genetics: The Parentage of Chardonnay, Gamay, and Other Wine Grapes of Northeastern France. *Science*. 1999; 285(5433):1562–5. PMID: [10477519](https://pubmed.ncbi.nlm.nih.gov/10477519/)
2. Hunt HV, Lawes MC, Bower MA, Haeger JW, Howe CJ. A banned variety was the mother of several major wine grapes. *Biology Letters*. 2010; 6(3):367–9. <https://doi.org/10.1098/rsbl.2009.0810> PMID: [20015860](https://pubmed.ncbi.nlm.nih.gov/20015860/)
3. Anderson K, Aryal NR. Database of Regional, National and Global Winegrape Bearing Areas by Variety, 2000 and 2010. Wine Economics Research Centre, University of Adelaide (third revision July 2014). 2013.
4. Bernard R. Aspects of clonal selection in Burgundy. *Proceedings of the International Symposium on Clonal Selection*. 1995:17–9.
5. Olmo HP. Selecting and breeding new grape varieties. *California Agriculture*. 1980; 34(7):23–4.
6. Bettiga LJ. Comparison of Seven Chardonnay Clonal Selections in the Salinas Valley. *American Journal of Enology and Viticulture*. 2003; 54(3):203–6.
7. Reynolds AG, Cliff M, Wardle DA, King M. Evaluation of Winegrapes in British Columbia: ‘Chardonnay’ and ‘Pinot noir’ Clones. *HortTechnology*. 2004; 14(4):594–602.
8. Fidelibus MW, Christensen LP, Katayama DG, Verdenal P-T. Yield Components and Fruit Composition of Six Chardonnay Grapevine Clones in the Central San Joaquin Valley, California. *American Journal of Enology and Viticulture*. 2006; 57(4):503–7.

9. Vouillamoz JF, Grando MS. Genealogy of wine grape cultivars: 'Pinot' is related to 'Syrah'. *Heredity*. 2006; 97:102–10. <https://doi.org/10.1038/sj.hdy.6800842> PMID: 16721391
10. Anderson MM, Smith RJ, Williams MA, Wolpert JA. Viticultural Evaluation of French and California Chardonnay Clones Grown for Production of Sparkling Wine. *American Journal of Enology and Viticulture*. 2008; 59(1):73–7.
11. Duchêne E, Legras JL, Karst F, Merdinoglu D, Claudel P, Jaegli N, et al. Variation of linalool and geraniol content within two pairs of aromatic and non-aromatic grapevine clones. *Australian Journal of Grape and Wine Research*. 2009; 15(2):120–30.
12. Nicholas PP. National register of grapevine varieties and clones. 2006 ed ed: Australian Vine Improvement Association; 2006.
13. Wolpert JA, Kasimatis AN, Weber E. Field Performance of Six Chardonnay Clones in the Napa Valley. *American Journal of Enology and Viticulture*. 1994; 45(4):393–400.
14. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*. 2007; 449(7161):463–7. <https://doi.org/10.1038/nature06148> PMID: 17721507
15. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, et al. A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. *PLoS ONE*. 2007; 2(12): e1326. <https://doi.org/10.1371/journal.pone.0001326> PMID: 18094749
16. Minio A, Lin J, Gaut BS, Cantu D. How Single Molecule Real-Time Sequencing and Haplotype Phasing Have Enabled Reference-Grade Diploid Genome Assembly of Wine Grapes. *Frontiers in Plant Science*. 2017; 8(826). <https://doi.org/10.3389/fpls.2017.00826> PMID: 28567052
17. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009; 323(5910):133–8. <https://doi.org/10.1126/science.1162986> PMID: 19023044
18. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, proteomics & bioinformatics*. 2016; 14(5):265–79.
19. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. 2016; 13:1050–4. <https://doi.org/10.1038/nmeth.4035> PMID: 27749838
20. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*. 2017; 27(5):722–36. <https://doi.org/10.1101/gr.215087.116> PMID: 28298431
21. Fu X, Li J, Tian Y, Quan W, Zhang S, Liu Q, et al. Long-read sequence assembly of the firefly *Pyrocoelia pectoralis* genome. *GigaScience*. 2017; 6(12):1–7.
22. Khost DE, Eickbush DG, Larracuenta AM. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome Research*. 2017; 27(5):709–21. <https://doi.org/10.1101/gr.213512.116> PMID: 28373483
23. Yoshida Y, Koutsovoulos G, Laetsch DR, Stevens L, Kumar S, Horikawa DD, et al. Comparative genomics of the tardigrades *Hypsibius dujardini* and *Ramazzottius variegatus*. *PLoS Biology*. 2017; 15(7): e2002266. <https://doi.org/10.1371/journal.pbio.2002266> PMID: 28749982
24. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*. 2018; 36:338–45 <https://doi.org/10.1038/nbt.4060> PMID: 29431738
25. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*. 2013; 10:563–9. <https://doi.org/10.1038/nmeth.2474> PMID: 23644548
26. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015; 31(19):3210–2. <https://doi.org/10.1093/bioinformatics/btv351> PMID: 26059717
27. Vinson JP, Jaffe DB, O'Neill K, Karlsson EK, Stange-Thomann N, Anderson S, et al. Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. *Genome Research*. 2005; 15(8):1127–35. <https://doi.org/10.1101/gr.3722605> PMID: 16077012
28. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*. 2014; 24(8):1384–95. <https://doi.org/10.1101/gr.170720.113> PMID: 24755901
29. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: Synteny Reduction for Third-gen Diploid Genome Assemblies. *bioRxiv*. 2018:286252.

30. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*. 2008; 18(1):188–96. <https://doi.org/10.1101/gr.6743907> PMID: 18025269
31. Lacombe T, Boursiquot J-M, Laucou V, Di Vecchi-Staraz M, Péros J-P, This P. Large-scale parentage analysis in an extended set of grapevine cultivars (*Vitis vinifera* L.). *Theoretical and Applied Genetics*. 2013; 126(2):401–14. <https://doi.org/10.1007/s00122-012-1988-2> PMID: 23015217
32. Marroni F, Scaglione D, Pinosio S, Policriti A, Miculan M, Di Gaspero G, et al. Reduction of heterozygosity (ROH) as a method to detect mosaic structural variation. *Plant Biotechnology Journal*. 2017; 15(7):791–3. <https://doi.org/10.1111/pbi.12691> PMID: 28055146
33. Gómez-Romero L, Palacios-Flores K, Reyes J, García D, Boege M, Dávila G, et al. Precise detection of de novo single nucleotide variants in human genomes. *Proceedings of the National Academy of Sciences*. 2018; 115(21):5516–21.
34. Boss PK, Davies C, Robinson S P. Anthocyanin composition and anthocyanin pathway gene expression in grapevine sports differing in berry skin colour. *Australian Journal of Grape and Wine Research*. 1996; 2(3):163–70.
35. Longbottom ML, Dry PR, Sedgley M. Observations on the morphology and development of star flowers of *Vitis vinifera* L. cvs Chardonnay and Shiraz. *Australian Journal of Grape and Wine Research*. 2008; 14(3):203–10.
36. Emanuelli F, Battilana J, Costantini L, Le Cunff L, Boursiquot JM, This P, et al. A candidate gene association study on muscat flavor in grapevine (*Vitis vinifera* L.). *BMC plant biology*. 2010; 10:241. <https://doi.org/10.1186/1471-2229-10-241> PMID: 21062440
37. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. 2010; 38(16):e164–e. <https://doi.org/10.1093/nar/gkq603> PMID: 20601685
38. Cipriani G, Marrazzo MT, Di Gaspero G, Pfeiffer A, Morgante M, Testolin R. A set of microsatellite markers with long core repeat optimized for grape (*Vitis* spp.) genotyping. *BMC plant biology*. 2008; 8(1):127.
39. Laucou V, Lacombe T, Dechesne F, Siret R, Bruno JP, Dessup M, et al. High throughput analysis of grape genetic diversity as a tool for germplasm collection management. *Theoretical and Applied Genetics*. 2011; 122(6):1233–45. <https://doi.org/10.1007/s00122-010-1527-y> PMID: 21234742
40. Shajji A, Yorukoglu D, William Yu Y, Berger B. Fast genotyping of known SNPs through approximate k-mer matching. *Bioinformatics*. 2016; 32(17):i538–i44. <https://doi.org/10.1093/bioinformatics/btw460> PMID: 27587672
41. Cipriani G, Spadotto A, Jurman I, Di Gaspero G, Crespan M, Meneghetti S, et al. The SSR-based molecular profile of 1005 grapevine (*Vitis vinifera* L.) accessions uncovers new synonymy and parentages, and reveals a large admixture amongst varieties of different geographic origin. *Theoretical and Applied Genetics*. 2010; 121(8):1569–85. <https://doi.org/10.1007/s00122-010-1411-9> PMID: 20689905
42. Gd Lorenzis, Imazio S, Brancadoro L, Failla O, Scienza A. Evidence for a Sympatric Origin of Ribolla gialla, Gouais Blanc and Schiava cultivars (*V. vinifera* L.). *South African Journal of Enology and Viticulture*. 2014; 35(1):149–56.
43. Maul E, Eibach R, Zyprian E, Töpfer R. The prolific grape variety (*Vitis vinifera* L.) 'Heunisch Weiss' (= 'Gouais blanc'): bud mutants, "colored" homonyms and further offspring. *Journal of Grapevine Research*. 2015; 54(2):79–86.
44. Lijavetzky D, Cabezas JA, Ibanez A, Rodriguez V, Martinez-Zapater JM. High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics*. 2007; 8:424. <https://doi.org/10.1186/1471-2164-8-424> PMID: 18021442
45. Myles S, Chia JM, Hurwitz B, Simon C, Zhong GY, Buckler E, et al. Rapid genomic characterization of the genus *vitis*. *PLoS One*. 2010; 5(1):e8219. <https://doi.org/10.1371/journal.pone.0008219> PMID: 20084295
46. Laucou V, Launay A, Bacilieri R, Lacombe T, Adam-Blondon AF, Berard A, et al. Extended diversity analysis of cultivated grapevine *Vitis vinifera* with 10K genome-wide SNPs. *PLoS One*. 2018; 13(2): e0192540. <https://doi.org/10.1371/journal.pone.0192540> PMID: 29420602
47. This P, Jung A, Boccacci P, Borrego J, Botta R, Costantini L, et al. Development of a standard set of microsatellite reference alleles for identification of grape cultivars. *Theoretical and Applied Genetics*. 2004; 109(7):1448–58. <https://doi.org/10.1007/s00122-004-1760-3> PMID: 15565426
48. Kunst L, Samuels AL. Biosynthesis and secretion of plant cuticular wax. *Progress in Lipid Research*. 2003; 42(1):51–80. PMID: 12467640
49. Cheng JB, Russell DW. Mammalian wax biosynthesis. I. Identification of two fatty acyl-Coenzyme A reductases with different substrate specificities and tissue distributions. *Journal of biological Chemistry*. 2004; 279(36):37789–97. <https://doi.org/10.1074/jbc.M406225200> PMID: 15220348

50. Rowland O, Domergue F. Plant fatty acyl reductases: Enzymes generating fatty alcohols for protective layers with potential for industrial applications. *Plant Science*. 2012; 193–194:28–38. <https://doi.org/10.1016/j.plantsci.2012.05.002> PMID: 22794916
51. Krauss P, MarktÄDter C, Riederer M. Attenuation of UV radiation by plant cuticles from woody species. *Plant, Cell & Environment*. 2005; 20(8):1079–85.
52. Leide J, Hildebrandt U, Reussing K, Riederer M, Vogg G. The Developmental Pattern of Tomato Fruit Wax Accumulation and Its Impact on Cuticular Transpiration Barrier Properties: Effects of a Deficiency in a β -Ketoacyl-Coenzyme A Synthase (LeCER6). *Plant physiology*. 2007; 144(3):1667–79. <https://doi.org/10.1104/pp.107.099481> PMID: 17468214
53. Isaacson T, Kosma DK, Matas AJ, Buda GJ, He Y, Yu B, et al. Cutin deficiency in the tomato fruit cuticle consistently affects resistance to microbial infection and biomechanical properties, but not transpirational water loss. *the Plant Journal*. 2009; 60(2):363–77. <https://doi.org/10.1111/j.1365-313X.2009.03969.x> PMID: 19594708
54. Yeats TH, Rose JKC. The Formation and Function of Plant Cuticles. *Plant physiology*. 2013; 163(1):5–20. <https://doi.org/10.1104/pp.113.222737> PMID: 23893170
55. Xue D, Zhang X, Lu X, Chen G, Chen Z-H. Molecular and Evolutionary Mechanisms of Cuticular Wax for Plant Drought Tolerance. *Frontiers in Plant Science*. 2017; 8:621. <https://doi.org/10.3389/fpls.2017.00621> PMID: 28503179
56. Konlechner C, Sauer U. Ultrastructural leaf features of grapevine cultivars (*Vitis vinifera* L. ssp. *vinifera*). *OENO One*. 2016; 50(4).
57. Dobritsa AA, Shrestha J, Morant M, Pinot F, Matsuno M, Swanson R, et al. CYP704B1 is a long-chain fatty acid omega-hydroxylase essential for sporopollenin synthesis in pollen of Arabidopsis. *Plant physiology*. 2009; 151(2):574–89. <https://doi.org/10.1104/pp.109.144469> PMID: 19700560
58. Carrier G, Le Cunff L, Dereeper A, Legrand D, Sabot F, Bouchez O, et al. Transposable Elements Are a Major Cause of Somatic Polymorphism in *Vitis vinifera* L. *PLoS ONE*. 2012; 7(3):e32973. <https://doi.org/10.1371/journal.pone.0032973> PMID: 22427919
59. Gambino G, Dal Molin A, Boccacci P, Minio A, Chitarra W, Avanzato CG, et al. Whole-genome sequencing and SNV genotyping of ‘Nebbiolo’ (*Vitis vinifera* L.) clones. *Scientific Reports*. 2017; 7(1):17294. <https://doi.org/10.1038/s41598-017-17405-y> PMID: 29229917
60. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*. 2012; 13(1):238.
61. Fallon TR, Lower SE, Chang C-H, Bessho-Uehara M, Martin GJ, Bewick AJ, et al. Firefly genomes illuminate parallel origins of bioluminescence in beetles. *eLife*. 2018; 7:e36495. <https://doi.org/10.7554/eLife.36495> PMID: 30324905
62. Chen J, Hu Q, Zhang Y, Lu C, Kuang H. P-MITE: a database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Research*. 2014; 42(Database issue):D1176–81. <https://doi.org/10.1093/nar/gkt1000> PMID: 24174541
63. Smit A, Hubley R. RepeatModeler Open-1.0. 2008–2015 [Available from: <http://www.repeatmasker.org>.
64. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013; 29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635> PMID: 23104886
65. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*. 2010; 28(5):511–5. <https://doi.org/10.1038/nbt.1621> PMID: 20436464
66. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*. 2011; 27(6):757–63. <https://doi.org/10.1093/bioinformatics/btr010> PMID: 21216780
67. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*. 2003; 13(9):2178–89. <https://doi.org/10.1101/gr.1224503> PMID: 12952885
68. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. 2000; 28(1):27–30. PMID: 10592173
69. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009; 10:421. <https://doi.org/10.1186/1471-2105-10-421> PMID: 20003500
70. Pundir S, Martin MJ, O'Donovan C. UniProt Protein Knowledgebase. In: Wu CH, Arighi CN, Ross KE, editors. *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*. New York, NY: Springer New York; 2017. p. 41–55.
71. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*. 2017; 45(D1):D158–D69. <https://doi.org/10.1093/nar/gkw1099> PMID: 27899622

72. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome biology*. 2004; 5(2):R12. <https://doi.org/10.1186/gb-2004-5-2-r12> PMID: 14759262
73. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997v1. 2013.
74. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*. 2012; 22(3):568–76. <https://doi.org/10.1101/gr.129684.111> PMID: 22300766
75. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011; 27(6):764–70. <https://doi.org/10.1093/bioinformatics/btr011> PMID: 21217122
76. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 2004; 5:113. <https://doi.org/10.1186/1471-2105-5-113> PMID: 15318951
77. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*. 2014; 30(22):3276–8. <https://doi.org/10.1093/bioinformatics/btu531> PMID: 25095880
78. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. 2004; 20(2):289–90. PMID: 14734327
79. Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics*. 2011; 27(4):592–3. <https://doi.org/10.1093/bioinformatics/btq706> PMID: 21169378
80. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014; 30(15):2114–20. <https://doi.org/10.1093/bioinformatics/btu170> PMID: 24695404