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

Short Title: Chardonnay genome reveals clonal diversity and ancestral inbreeding

Michael J. Roach[1], Daniel L. Johnson[1], Joerg Bohlmann[2], Hennie J.J. van Vuuren[2,3], Steven J. M. Jones[4], Isak S. Pretorius[5], Simon A. Schmidt[1#*] and Anthony R. Borneman[1,6#*]

1. The Australian Wine Research Institute, PO Box 197, Glen Osmond, South Australia, 5046, 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, BC V6T 1Z4, Canada.
4. Michael Smith Genome Sciences Centre, British Columbia Cancer Research Centre, Vancouver, British Columbia, BC V6T 1Z4, Canada.
5. Chancellery, Macquarie University, Sydney, New South Wales, 2109. Australia
6. Department of Genetics and Evolution, University of Adelaide, South Australia, 5000, Australia

[#]Authors contributed equally to this work

*Address correspondence to:

Dr. Simon Schmidt; +61 8 8313 6600; simon.schmidt@awri.com.au

Dr. Anthony Borneman; +61 8 8313 6600; anthony.borneman@awri.com.au

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 commercial Chardonnay clones. There were 1620 markers identified that distinguish the 15 Chardonnay clones. These markers were reliably used for clonal identification of validation genomic material, as well as in identifying a potential genetic basis for some clonal phenotypic differences. The predicted parentage of the Chardonnay haplomes 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 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 fully 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 technology 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) has 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.

98 # Results

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

101 Of the many Chardonnay clones available, clone I10V1 was chosen as the basis for the
102 reference genome due to its prominent use in the Australian wine industry. The initial I10V1
103 genome was assembled, phased and polished using subreads generated from 54 PacBio
104 RS-II SMRT cells and the FALCON Unzip, Quiver pipeline (19, 25). While this assembly
105 method should produce an assembly in which the primary contigs represent the haploid
106 genome content of the organism in question, the size of the initial assembly (580 Mb)
107 significantly exceeded that expected for *V. vinifera* (450–500 Mb). Both analysis with
108 BUSCO (26) and short-read mapping indicated that this increased size was primarily due to
109 both copies of many genomic regions (rather than only a single haplotype) being
110 represented in the primary contigs (S1 Table and S2 Fig), a situation that is common in
111 heterozygous diploid genome assemblies (19, 27-29). To address these assembly artefacts,
112 the initial primary contig pool was aggressively de-duplicated, with small primary contigs that
113 were allelic to larger primary contigs being re-assigned to the haplotig pool. This approach
114 reassigned 694 primary contigs (100 Mb) and added 36 haplotigs (11 Mb), while also
115 purging 18 repeat-rich artefactual contigs (1.3 Mb). Manual curation, based upon alignments
116 to the PN40024 assembly (14) and subread mapping were used to address several
117 remaining mis-assemblies.

118 The final curated Chardonnay assembly consists of 854 primary contigs (N_{50} of 935 kb) and
119 1883 haplotigs, totalling 490 Mb and 378 Mb, respectively (Table 1). There were
120 approximately 95% complete, and only 1.6% fragmented BUSCO-predicted genes
121 (Supplementary Table S1). BUSCO duplication is also predicted to be low for both the
122 primary contigs and the associated haplotigs (4% and 2% respectively). A custom repeat

123 library was constructed for Chardonnay and used to annotate 336 Mb (38.7%) of the diploid
124 genome as repetitive. RNAseq data were used to annotate potential coding regions of the
125 primary contigs using Maker (30), which predicted 29 675 gene models (exclusive of
126 repetitive regions) and 66 548 transcripts in total.

127 **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
N50 (kb)	935.8	318.4
N75 (kb)	502.7	165.3
L50	145	335
L75	321	749

128

129 **Phasing coverage, and identification of homozygous and hemizygous regions**
130 A total of 614 primary contigs (397 Mb) and 1502 haplotigs (305 Mb) were confidently placed
131 in chromosomal order using the PN40024 scaffold as a reference. To analyse the degree
132 and distribution of heterozygosity across the genome, read depth (from mapped RS II
133 subreads), heterozygous variant density (from mapped Illumina short-reads) and phasing
134 coverage (from contig alignments) was calculated for the assembly (Fig 1A).

135 Chromosomes 2, 3, 7, 15, 17, and 18 contain runs of homozygosity greater than 500 kb
136 (intersect of lack of phasing coverage, double read-depth, low heterozygous variant density).
137 There are a further 22.8 Mb that lacked phasing coverage, had low heterozygous variant
138 density, and median read-depth. These regions presumably result from either hemizygosity
139 of these genomic regions, or undetected allelic duplicates remaining in the primary contigs.

The largest homozygous run identified resides on Chromosome 2 and aligns closely to the Pinot noir assembly at over 99.8% identity (Fig 1B). A region of synteny remaining in the primary contigs is present (CH.chr2:5570000– 6520000), 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 (CH.chr2:99000000–10300000 and CH.chr2:11450000–12600000). BLAST searches for these regions within the remaining primary contigs and haplotigs did not reveal any significant alignments. 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 aligning and trimming both the primary contigs and haplotigs into pairs of syntenic sequence blocks (P and H alleles). This produced 1153 phase-blocks covering 270 Mb of the genome (55%). 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 Pinot varieties (32) (hereafter referred to as Pinot) 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% 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

166 regions, the pattern of nucleotide variation across the two heterozygous Chardonnay
167 haplotypes matched both haplotypes of Pinot, with one of these haplotypes also matching
168 one of the Gouais haplotypes. These ‘double Pinot haplotype' regions are in some cases
169 many megabases in size and are indicative of a common ancestry between Pinot and
170 Gouais blanc.

171 While reciprocity (one Gouais blanc haplotype, one Pinot haplotype) was observed between
172 allelic phase-blocks for over 95% of the parentage-assigned sequence, frequent haplotype
173 switching (a known characteristic of FALCON-based assemblies) was observed between the
174 haplomes, producing a haplotype mosaic which is observable as a ‘checkerboard’ pattern
175 that alternates between the primary contigs and haplotigs for each chromosome (Fig 2B).

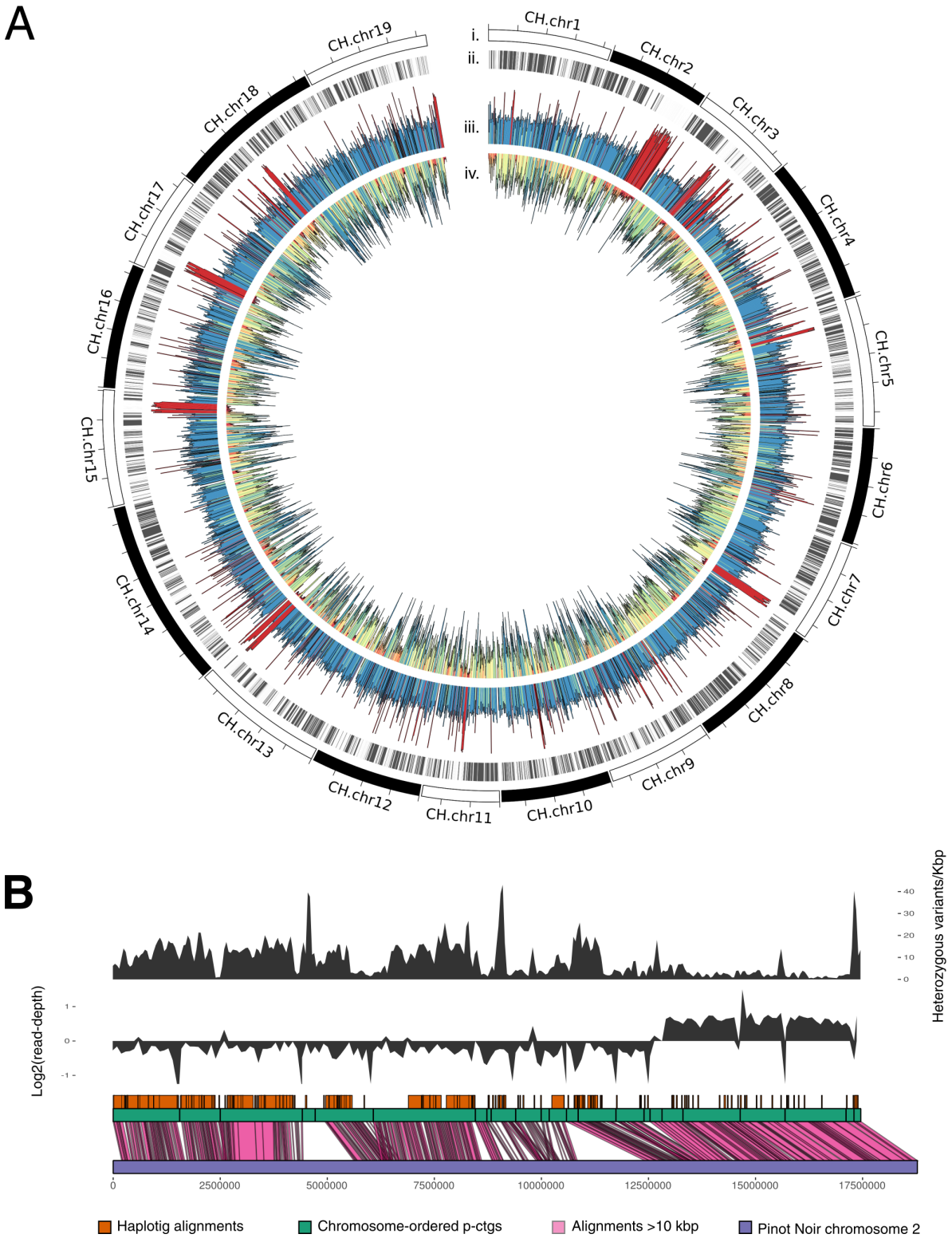

Fig 1. The *Vitis vinifera* cultivar Chardonnay reference genome. (*A*) A circos plot showing chromosome-ordered primary contigs (*i*), haplotig alignments (*ii*), read-depth of RS II subreads mapped to diploid assembly (read-depth colour scale: yellow, low; blue, high; red, double) (*iii*), and heterozygous 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).

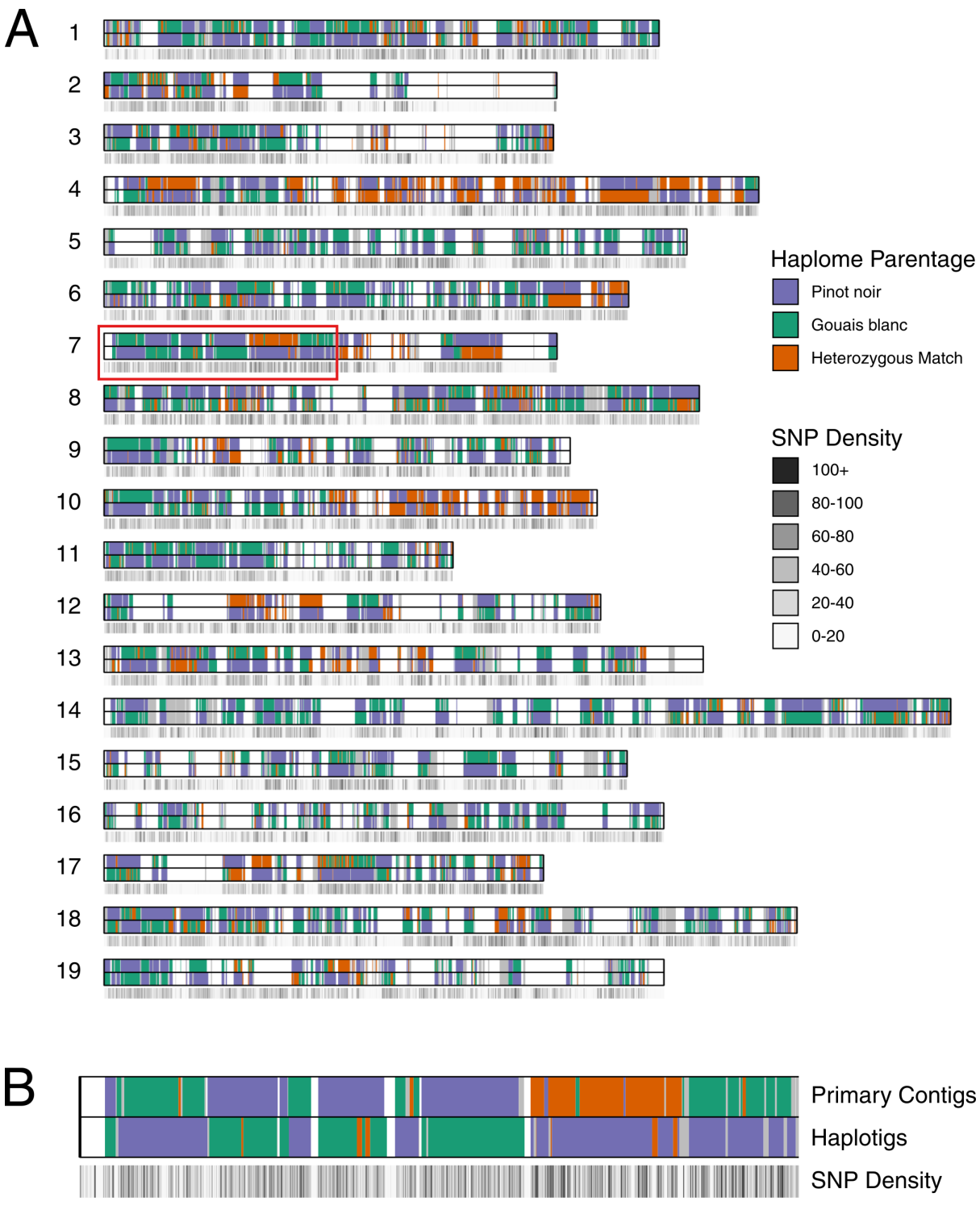

Fig 2. Parental architecture of the Chardonnay genome. (*A*) An ideogram of the Chardonnay reference assembly with the positions of both primary contig and haplotig phase-blocks indicated and juxtaposed with a SNP density track (for the primary contigs). Gaps in phase-blocks are indicated in white. (*B*)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

188 **Parental-specific genomic variation**

189 With parental contributions delineated in the Chardonnay assembly, it was possible to
190 determine the parental origins of structural variation between orthologous chromosomes,
191 including parent-specific gene family expansions. Tandem pairs of orthologous proteins were
192 defined in Chardonnay and filtered to identify tandem orthologs that were both expanded in
193 Chardonnay compared to the Pinot noir reference assembly, and which resided in the
194 Gouais blanc haplome. Chromosome alignments containing gene expansion candidates
195 were inspected for features consistent with tandem gene duplication. Using this analysis, an
196 expansion of Fatty Acyl-CoA Reductase 2-like (*FAR2-like*) genes on Chromosome 5 was
197 identified, with the arrangement of *FAR2-like* open reading frames (ORFs) consistent with a
198 tandem duplication event (S3 Fig).

199 A protein-based phylogeny was produced that encompassed the four *FAR2-like* ORFs
200 present in the Chardonnay assembly, in addition to the homologous proteins from the Pinot
201 noir PN40024 assembly (Fig 3A). Using these data, the Chardonnay haplotig sequence was
202 identified as being derived from Pinot (nucleotide sequences of *FAR2-like* genes from Pinot
203 noir and the Chardonnay haplotig were identical). However, rather than having an
204 orthologous set of protein-coding regions, the genomic sequence derived from Gouais blanc
205 (present in the primary contig) is predicted to encode two additional copies of *FAR2-like*
206 homologues and an extra *FAR2-like* pseudogene (Fig 3B). While the ORF that was
207 orthologous to the Pinot *FAR2-like* gene was closely related to the Pinot noir *FAR2-like* gene
208 (98% identity), the two additional ORFs from Gouais blanc were more distantly related (93–
209 94% identity), suggesting that this gene expansion was not a recent event.

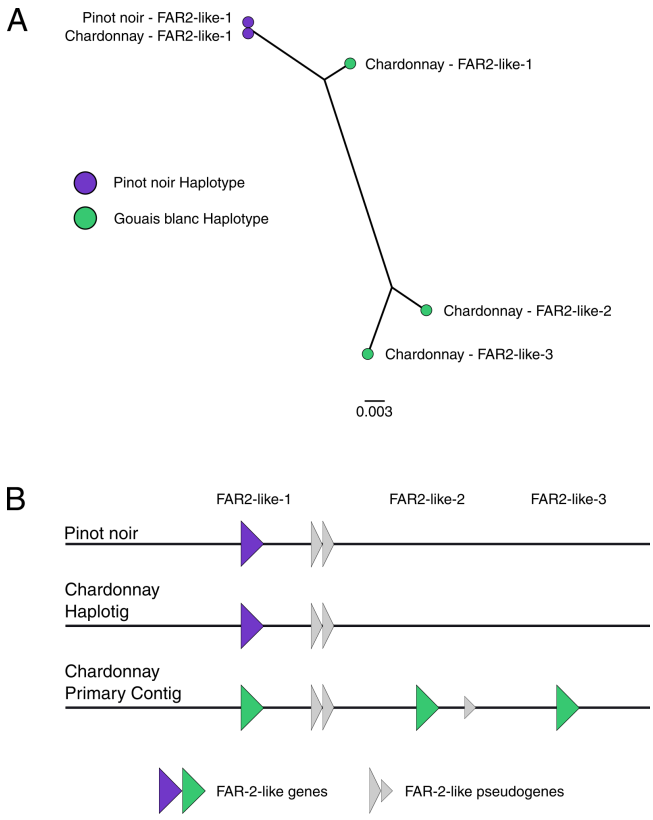

Fig 3. A FAR2-like expanded gene family in Chardonnay. (*A*) An unrooted tree of *FAR2-like* genes. (*B*) A schematic of the predicted genomic arrangement of the *FAR2-like* genes in Chardonnay. Both Pinot noir-derived (purple) and Gouais blanc-derived genes (green) are shown.

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 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 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 were identified and evenly distributed across the Chardonnay genome (Table 2, S4 Fig, and Sheet 1 in S5 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. The tight grouping of these clones confirms that 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.

244 **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*

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

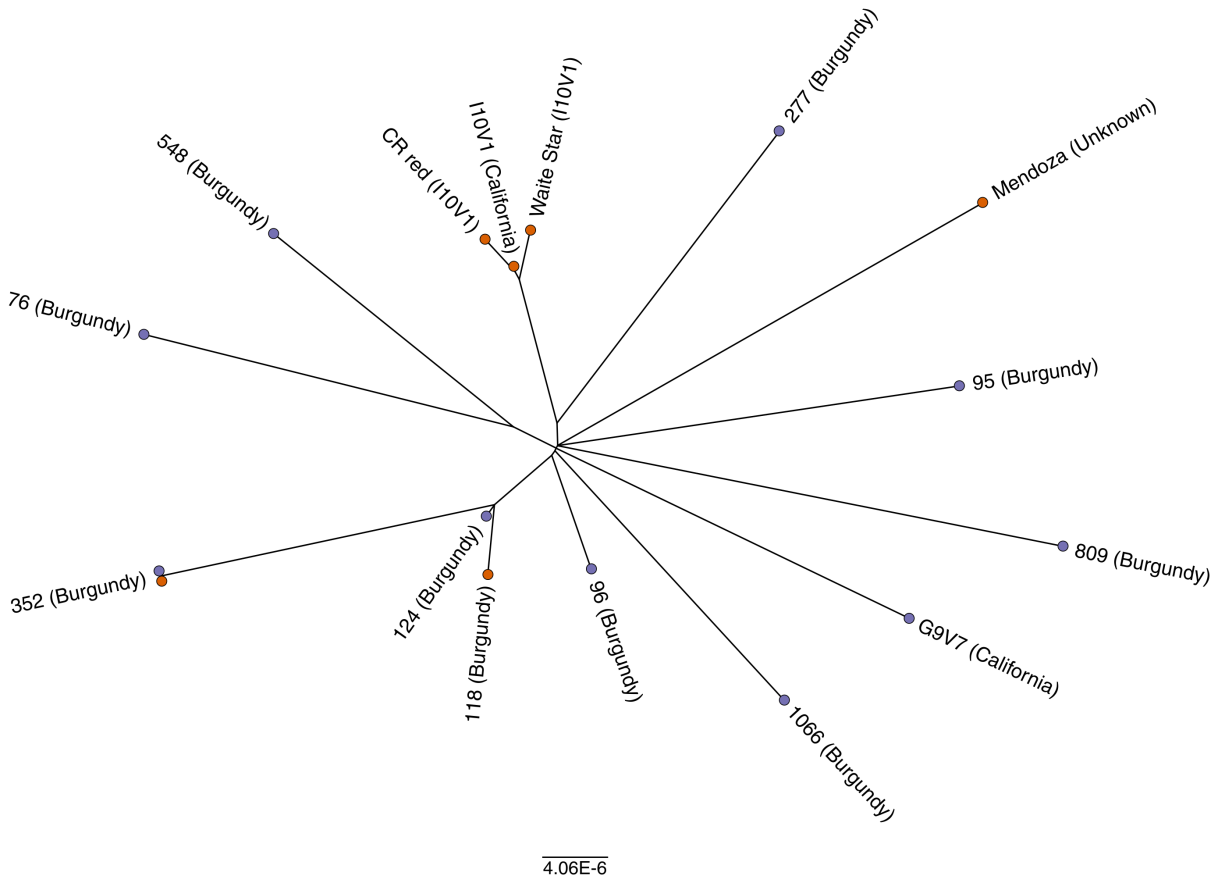

Fig 4. Genetic diversity in Chardonnay clones. An unrooted tree of Chardonnay clones based upon bi-allelic SNPs. Sequencing batches are designated by coloured terminal nodes (orange, sequencing batch #1; purple, sequencing batch #2).

The accumulation of SNPs can also lead to phenotypic differentiation that underlies the clonal selection process. For example, the major clonal-specific phenotypic variant of Chardonnay, “Muscat character”, 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 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 S5 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 clonal-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 S5 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). When screened with the pipeline (Fig 5A), between 29% and 76% of the markers were detected for each of the samples and 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 62% and 81% of the expected markers for each sample, with minimal (1.2% to 2.6%) off-clone variants detected. However, despite being the same clones, there were a significant proportion (12% to 36%) 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, which indicates that there may be intra-clonal genetic variation that has accumulated during the independent passaging of clonal material.

As the level of sequencing coverage ultimately impacts the economics of clonal testing, the impact of sequencing depth on marker identification was 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). At 200-fold coverage there were only 2 (low confidence) off-target hits, and none at lower coverages. There was little difference in the number of discoverable markers from 200-fold down to 25-fold coverage (79% and 73% respectively), and only a 4% decrease in markers confidently-flagged as missing. At 12-fold coverage it was still possible to detect 58% 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% and 38% of the expected markers for each sample, with no off-target hits.

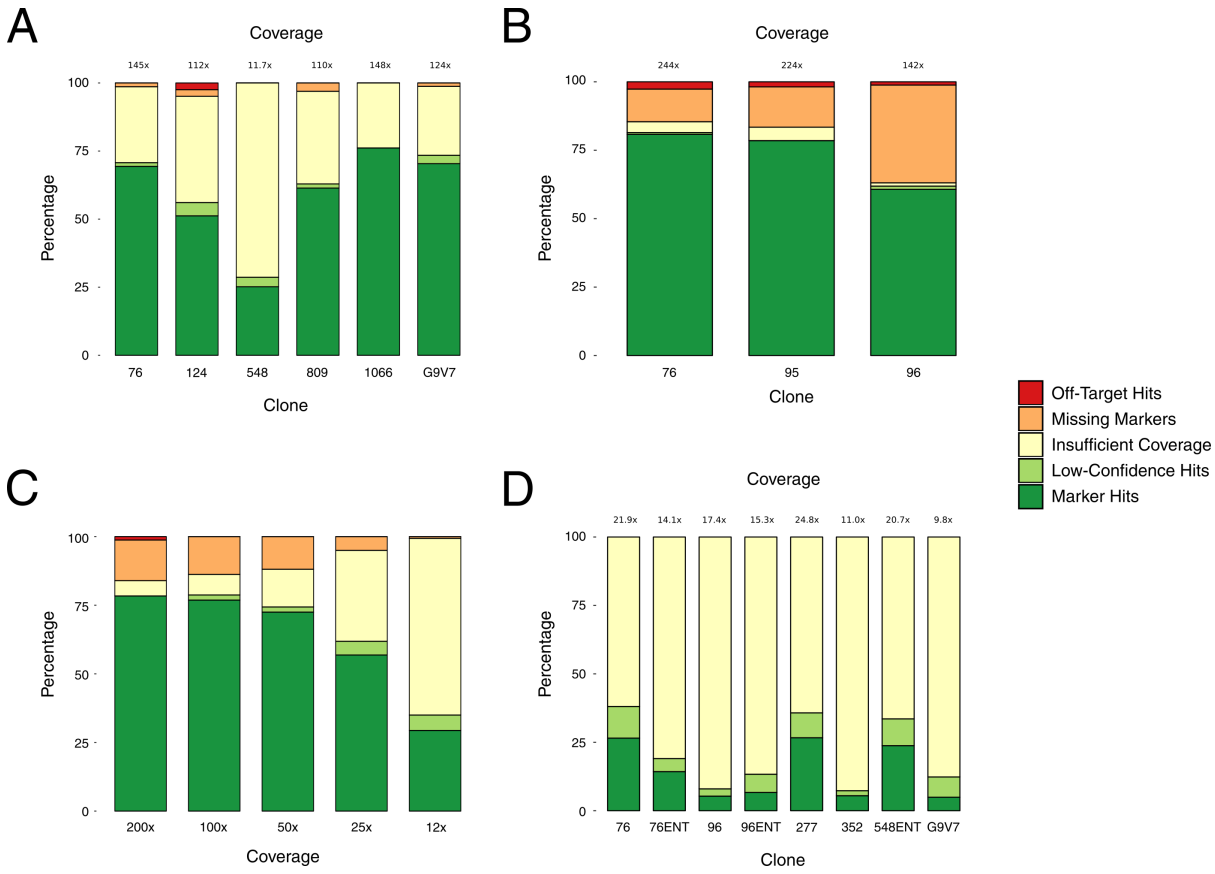

Fig 5: Marker screening using WGS data in Chardonnay. (*A*) Moderate coverage sequencing of clonal material used in marker discovery. (*B*) High-coverage sequencing of independently-sourced clonal material. (*C*) Subsampling of High-Coverage, independently-sourced clonal material (clone 95 from Figure 5B). (*D*) Low-coverage sequencing of independently-sourced clonal material ('ENT' denotes ENTAV-INRA® source material).

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 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 catalyse 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 the many clones that were isolated from mass-selected vineyards appear to be genetically quite distinct from one another.

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

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

392 While the diploid Chardonnay reference genome enabled a much deeper understanding of
393 the variation that has occurred since the initial establishment of this variety, it has also
394 provided the means to unravel the detailed genetic ancestry of this variety and its parents,
395 Pinot noir and Gouais blanc. Chardonnay matches both haplotypes of Pinot noir across
396 approximately one fifth of its genome and these areas include large tracks of both
397 homozygous and heterozygous variation. While the presence of the homozygous 'double-
398 Pinot noir' regions could be result of a high number of large-scale gene conversion events
399 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 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 Gouias 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.

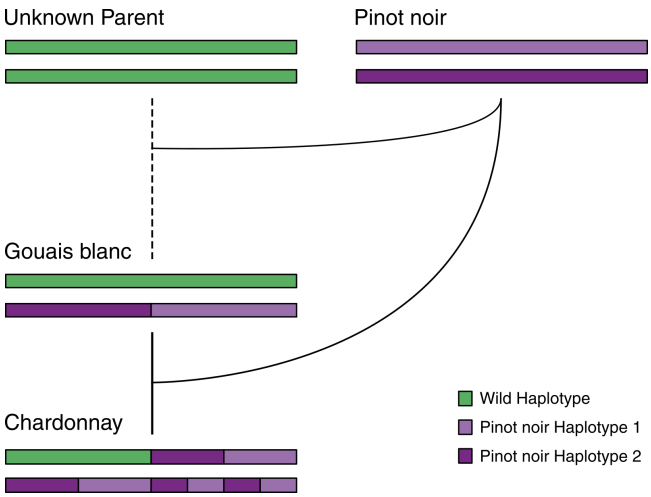

Fig 6: A schematic model for the complex pedigree of Chardonnay, Gouais blanc and Pinot noir. Two crossing events (akin to a standard genetic backcross) with Pinot noir would result in the homozygous and heterozygous Pinot noir regions present in Chardonnay.

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

425 ## Methods

426 All custom scripts used for analysis, along with detailed workflows are available in
427 S6 Archive. All sequencing data and the genome assembly have been lodged at the
428 National Center for Biotechnology Information under the BioProject accession:
429 PRJNA399599.

430 **DNA preparation and sequencing**

431 Nuclear DNA was isolated from early season, disease free, field grown Chardonnay leaves
432 taken from plants at a nursery vineyard (Oxford Landing, Waikerie, South Australia). DNA
433 was extracted by Bio S&T (Quebec, Canada) from nuclear-enriched material using a
434 CTAB/Chloroform method. DNA from clone I10V1 was enriched using a 1:0.45 Ampure
435 cleanup prior to being used to build 15-50 kb SMRT Bell libraries with Blue Pippin size
436 selection following library preparation (Ramaciotti Centre for Genomics, UNSW, Sydney,
437 Australia). These libraries were sequenced on a PacBio RSII using 54 SMRT cells to give a
438 total sequencing yield of 51,921 Mb (115-fold coverage) with an N50 length of 14.4 kb.
439 Short-read sequencing of clones for marker discovery was performed on Illumina HiSeq
440 2000 and HiSeq X-Ten platforms from TruSeq libraries (100 and 150 bp paired end read
441 chemistries). Short-read sequencing of clones for marker validation was performed on
442 Illumina HiSeq 2500 and MiSeq platforms from Nextera libraries made from material sourced

443 from both Foundation Plant Services (University of California, Davis) and Mission Hill Family
444 Estate, Quail's Gate and Burrowing Owl wineries in British Columbia, Canada.

445 **Assembly**

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

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

464 **Annotation**

465 A custom repeat library was produced for Chardonnay for use with RepeatMasker, similar to
466 the method described in Fallon, Lower (61). Miniature inverted-repeat transposable element
467 (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 alignment coordinates were used to trim and extract the closely aligning phase-blocks between the primary contigs and haplotigs. BED files were produced that could be used for mapping the phase-blocks to the primary contigs and 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

493 variety, pooled reads from all three were used for mapping. The sequencing data for Pinot,
494 Chardonnay, and Gouais blanc were mapped to the primary contig and haplotig phase-block
495 sequences using BWA-MEM v0.7.12 (73). PCR duplicates and discordantly-mapped reads
496 were removed, and poorly mapping regions were masked using a window coverage
497 approach. Heterozygous SNPs were called using VarScanv2.3 (p-value < 1e–6, coverage >
498 10, alt reads > 30%)(74) and Identity By State (IBS) was assessed over 10 kb windows (5 kb
499 steps) at every position where a heterozygous Chardonnay SNP was found.

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

512 An orthologous kmer method for assigning parentage was developed to assign parentage
513 over the entire genome. All 27-bp-long kmers (27mers) were counted using
514 JELLYFISH v2.2.6 (canonical representation, singletons ignored)(75) directly from Pinot,
515 Gouais blanc, and Chardonnay I10V1 PE reads to create 27mer count databases. Non-
516 overlapping 1-kb windows were generated for the primary contigs and the haplotigs. For
517 each window all 27mers were extracted from the contig sequences, queried against the
518 kmer count databases using JELLYFISH and the number of kmers not appearing in each

519 were returned. The Pinot/Gouais blanc missing kmer counts were normalised against the
520 Chardonnay counts and averaged over 10 kb windows with 5-kb steps. Windows with 150 or
521 more missing kmers were classified as mismatch (missing kmer density was visualized over
522 the genome to determine an appropriate cut-off) and neighbouring complementary windows
523 were merged.

524 **Gene expansion**

525 Protein-based BLAST alignments of Chardonnay proteins were performed against the Pinot
526 noir (PN40024) reference proteome. Chardonnay Maker GFF annotations and PN40024
527 GFF annotations were converted to BED format. Chardonnay annotations were then
528 transformed to scaffold coordinates. The 'blast_to_raw.py' script (from
529 github.com/tanghaibao/quota-alignment) was used to flag tandem repeat homologues for
530 both the primary contig and haplotig Chardonnay proteins against the Pinot noir reference.
531 Illumina paired-end reads for Pinot were mapped to the Chardonnay primary and haplotig
532 assemblies, BED annotations were created for regions with poor mapping, and these
533 annotations were transformed to the scaffold coordinates for use with filtering. Predicted
534 expanded gene families in Chardonnay that resided in Gouais blanc regions (identified using
535 the kmer parental mapping) that had multiple gene models with poor read-coverage of Pinot
536 mapped reads were returned as a filtered list. Dotplots were produced with MUMmer and the
537 chromosomes were visually assessed for evidence of tandem sequence duplication at the
538 filtered gene expansion candidate loci. The genomic sequences of *FAR2-like* ORFs from the
539 Pinot noir assembly and from the Chardonnay primary contigs and haplotigs were aligned
540 using MUSCLE v3.8.31 (76) within AliView v1.20 (77). Phylogenies were calculated within
541 Rstudio using Ape (78) and Phangorn (79).

542 **Marker variant discovery**

543 Paired-end reads for each clone were manually quality trimmed using Trimmomatic v0.36
544 (SLIDINGWINDOW:5:20, TRAILING:20, CROP:100)(80), 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 (alt reads > 5, alt reads > 5%), 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. The marker variants, marker kmers, and shared kmers were output in a table for use with querying unknown Chardonnay clones.

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 shared kmers that were identified in the marker discovery pipeline are then queried from the kmer database. A marker is flagged as a ‘hit’ if >80% of the marker kmers are present (at a depth of at least 3), and as ‘low-confidence hit’ if 30–80% of the marker kmers are present. Markers are flagged as ‘insufficient read coverage’ if fewer than 80% of the shared kmers are present at a depth of at least 12.

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

570 Information. All custom code and analysis workflows required to reproduce the results
571 presented are included in S6 Archive. Intermediate files can be made available on request.

572 Acknowledgements

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

581 This project is supported by Australia's grapegrowers and winemakers through their
582 investment body Wine Australia, with matching funds from the Australian Government.
583 Grapevine genomics work at the AWRI is also supported by Bioplatforms Australia as part of
584 the National Collaborative Research Infrastructure Strategy, an initiative of the Australian
585 Government. Funding was also obtained from Genome British Columbia and the Wine
586 Research Centre at The University of British Columbia.

587 Authors' contributions

588 SAS, ARB, DLJ, HJJVV, SJJ, JB and ISP conceived and outlined the original approaches for
589 the project. SAS sourced clonal material for reference sequencing and marker discovery.
590 HJJVV, SJJ and JB provided material and sequence data for clonal validation. MJR, SAS
591 and ARB designed and implemented the approaches used for genome assembly, marker
592 discovery, and genome analysis. MJR, SAS and ARB wrote the manuscript, which was
593 reviewed by all authors.

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.

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.

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.

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.

619 11. Duchêne E, Legras JL, Karst F, Merdinoglu D, Claudel P, Jaegli N, et al. Variation of
620 linalool and geraniol content within two pairs of aromatic and non-aromatic grapevine clones.
621 Australian Journal of Grape and Wine Research. 2009;15(2):120-30.
622 12. Nicholas PP. National register of grapevine varieties and clones. 2006 ed ed:
623 Australian Vine Improvement Association; 2006.
624 13. Wolpert JA, Kasimatis AN, Weber E. Field Performance of Six Chardonnay Clones in
625 the Napa Valley. American Journal of Enology and Viticulture. 1994;45(4):393-400.
626 14. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine
627 genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature.
628 2007;449(7161):463-7.
629 15. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, et al. A High
630 Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety.
631 PLoS ONE. 2007;2(12):e1326.
632 16. Minio A, Lin J, Gaut BS, Cantu D. How Single Molecule Real-Time Sequencing and
633 Haplotype Phasing Have Enabled Reference-Grade Diploid Genome Assembly of Wine
634 Grapes. Frontiers in Plant Science. 2017;8(826).
635 17. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from
636 single polymerase molecules. Science. 2009;323(5910):133-8.
637 18. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome
638 Assembly. Genomics, proteomics & bioinformatics. 2016;14(5):265-79.
639 19. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al.
640 Phased diploid genome assembly with single-molecule real-time sequencing. Nature
641 Methods. 2016;13:1050-4.
642 20. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable
643 and accurate long-read assembly via adaptive k-mer weighting and repeat separation.
644 Genome Research. 2017;27(5):722-36.

645 21. Fu X, Li J, Tian Y, Quan W, Zhang S, Liu Q, et al. Long-read sequence assembly of
646 the firefly Pyrocoelia pectoralis genome. GigaScience. 2017;6(12):1-7.
647 22. Khost DE, Eickbush DG, Larracuente AM. Single-molecule sequencing resolves the
648 detailed structure of complex satellite DNA loci in Drosophila melanogaster. Genome
649 Research. 2017;27(5):709-21.
650 23. Yoshida Y, Koutsovoulos G, Laetsch DR, Stevens L, Kumar S, Horikawa DD, et al.
651 Comparative genomics of the tardigrades Hypsibius dujardini and Ramazzottius
652 varieornatus. PLOS Biology. 2017;15(7):e2002266.
653 24. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore
654 sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology.
655 2018;36:338–45
656 25. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid,
657 finished microbial genome assemblies from long-read SMRT sequencing data. Nature
658 Methods. 2013;10:563-9.
659 26. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:
660 assessing genome assembly and annotation completeness with single-copy orthologs.
661 Bioinformatics. 2015;31(19):3210-2.
662 27. Vinson JP, Jaffe DB, O'Neill K, Karlsson EK, Stange-Thomann N, Anderson S, et al.
663 Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. Genome
664 Research. 2005;15(8):1127-35.
665 28. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de
666 novo assembly of highly heterozygous genomes from whole-genome shotgun short reads.
667 Genome Research. 2014;24(8):1384-95.
668 29. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: Synteny Reduction for
669 Third-gen Diploid Genome Assemblies. bioRxiv. 2018:286252.

670 30. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: An easy-
671 to-use annotation pipeline designed for emerging model organism genomes. Genome
672 Research. 2008;18(1):188-96.
673 31. Lacombe T, Boursiquot J-M, Laucou V, Di Vecchi-Staraz M, Péros J-P, This P.
674 Large-scale parentage analysis in an extended set of grapevine cultivars (Vitis vinifera L.).
675 Theoretical and Applied Genetics. 2013;126(2):401-14.
676 32. Marroni F, Scaglione D, Pinosio S, Policriti A, Miculan M, Di Gaspero G, et al.
677 Reduction of heterozygosity (ROH) as a method to detect mosaic structural variation. Plant
678 Biotechnology Journal. 2017;15(7):791-3.
679 33. Gómez-Romero L, Palacios-Flores K, Reyes J, García D, Boege M, Dávila G, et al.
680 Precise detection of de novo single nucleotide variants in human genomes. Proceedings of
681 the National Academy of Sciences. 2018.
682 34. Boss PK, Davies C, Robinson S P. Anthocyanin composition and anthocyanin
683 pathway gene expression in grapevine sports differing in berry skin colour. Australian
684 Journal of Grape and Wine Research. 1996;2(3):163-70.
685 35. Longbottom ML, Dry PR, Sedgley M. Observations on the morphology and
686 development of star flowers of *Vitis vinifera* L. cvs Chardonnay and Shiraz. Australian
687 Journal of Grape and Wine Research. 2008;14(3):203-10.
688 36. Emanuelli F, Battilana J, Costantini L, Le Cunff L, Boursiquot JM, This P, et al. A
689 candidate gene association study on muscat flavor in grapevine (*Vitis vinifera* L.). BMC plant
690 biology. 2010;10:241.
691 37. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants
692 from high-throughput sequencing data. Nucleic Acids Research. 2010;38(16):e164-e.
693 38. Cipriani G, Marrazzo MT, Di Gaspero G, Pfeiffer A, Morgante M, Testolin R. A set of
694 microsatellite markers with long core repeat optimized for grape (*Vitis*spp.) genotyping. BMC
695 plant biology. 2008;8(1):127.

696 39. Laucou V, Lacombe T, Dechesne F, Siret R, Bruno JP, Dessup M, et al. High
697 throughput analysis of grape genetic diversity as a tool for germplasm collection
698 management. Theoretical and Applied Genetics. 2011;122(6):1233-45.
699 40. Shajii A, Yorukoglu D, William Yu Y, Berger B. Fast genotyping of known SNPs
700 through approximate k-mer matching. Bioinformatics. 2016;32(17):i538-i44.
701 41. Cipriani G, Spadotto A, Jurman I, Di Gaspero G, Crespan M, Meneghetti S, et al. The
702 SSR-based molecular profile of 1005 grapevine (*Vitis vinifera* L.) accessions uncovers new
703 synonymy and parentages, and reveals a large admixture amongst varieties of different
704 geographic origin. Theoretical and Applied Genetics. 2010;121(8):1569-85.
705 42. Lorenzis Gd, Imazio S, Brancadoro L, Failla O, Scienza A. Evidence for a Sympatric
706 Origin of Ribolla gialla, Gouais Blanc and Schiava cultivars (*V. vinifera* L.). South African
707 Journal of Enology and Viticulture. 2014;35(1):149-56.
708 43. Maul E, Eibach R, Zyprian E, Töpfer R. The prolific grape variety (*Vitis vinifera* L.)
709 'Heunisch Weiss' (= 'Gouais blanc'): bud mutants, "colored" homonyms and further offspring.
710 Journal of Grapevine Research. 2015;54(2):79-86.
711 44. Lijavetzky D, Cabezas JA, Ibanez A, Rodriguez V, Martinez-Zapater JM. High
712 throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-
713 sequencing approach and SNPlex technology. BMC Genomics. 2007;8:424.
714 45. Myles S, Chia JM, Hurwitz B, Simon C, Zhong GY, Buckler E, et al. Rapid genomic
715 characterization of the genus vitis. PLoS One. 2010;5(1):e8219.
716 46. Laucou V, Launay A, Bacilieri R, Lacombe T, Adam-Blondon AF, Berard A, et al.
717 Extended diversity analysis of cultivated grapevine *Vitis vinifera* with 10K genome-wide
718 SNPs. PLoS One. 2018;13(2):e0192540.
719 47. This P, Jung A, Boccacci P, Borrego J, Botta R, Costantini L, et al. Development of a
720 standard set of microsatellite reference alleles for identification of grape cultivars. Theoretical
721 and Applied Genetics. 2004;109(7):1448-58.

722 48. Kunst L, Samuels AL. Biosynthesis and secretion of plant cuticular wax. Progress in
723 Lipid Research. 2003;42(1):51-80.
724 49. Cheng JB, Russell DW. Mammalian wax biosynthesis. I. Identification of two fatty
725 acyl-Coenzyme A reductases with different substrate specificities and tissue distributions.
726 Journal of biological Chemistry. 2004;279(36):37789-97.
727 50. Rowland O, Domergue F. Plant fatty acyl reductases: Enzymes generating fatty
728 alcohols for protective layers with potential for industrial applications. Plant Science.
729 2012;193-194:28-38.
730 51. Krauss P, MarkstÄDter C, Riederer M. Attenuation of UV radiation by plant cuticles
731 from woody species. Plant, Cell & Environment. 2005;20(8):1079-85.
732 52. Leide J, Hildebrandt U, Reussing K, Riederer M, Vogg G. The Developmental Pattern
733 of Tomato Fruit Wax Accumulation and Its Impact on Cuticular Transpiration Barrier
734 Properties: Effects of a Deficiency in a *β*-Ketoacyl-Coenzyme A Synthase (LeCER6). Plant
735 physiology. 2007;144(3):1667-79.
736 53. Isaacson T, Kosma DK, Matas AJ, Buda GJ, He Y, Yu B, et al. Cutin deficiency in the
737 tomato fruit cuticle consistently affects resistance to microbial infection and biomechanical
738 properties, but not transpirational water loss. the Plant Journal. 2009;60(2):363-77.
739 54. Yeats TH, Rose JKC. The Formation and Function of Plant Cuticles. Plant
740 physiology. 2013;163(1):5-20.
741 55. Xue D, Zhang X, Lu X, Chen G, Chen Z-H. Molecular and Evolutionary Mechanisms
742 of Cuticular Wax for Plant Drought Tolerance. Frontiers in Plant Science. 2017;8:621.
743 56. Konlechner C, Sauer U. Ultrastructural leaf features of grapevine cultivars (*Vitis*
744 *vinifera* L. ssp. *vinifera*). OENO One. 2016;50(4).
745 57. Dobritsa AA, Shrestha J, Morant M, Pinot F, Matsuno M, Swanson R, et al.
746 CYP704B1 is a long-chain fatty acid omega-hydroxylase essential for sporopollenin
747 synthesis in pollen of Arabidopsis. Plant physiology. 2009;151(2):574-89.

748 58. Carrier G, Le Cunff L, Dereeper A, Legrand D, Sabot F, Bouchez O, et al.
749 Transposable Elements Are a Major Cause of Somatic Polymorphism in *Vitis vinifera* L.
750 PLoS ONE. 2012;7(3):e32973.
751 59. Gambino G, Dal Molin A, Boccacci P, Minio A, Chitarra W, Avanzato CG, et al.
752 Whole-genome sequencing and SNV genotyping of 'Nebbiolo' (*Vitis vinifera* L.) clones.
753 Scientific Reports. 2017;7(1):17294.
754 60. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local
755 alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics.
756 2012;13(1):238.
757 61. Fallon TR, Lower SE, Chang C-H, Bessho-Uehara M, Martin GJ, Bewick AJ, et al.
758 Firefly genomes illuminate the origin and evolution of bioluminescence. bioRxiv.
759 2017:237586.
760 62. Chen J, Hu Q, Zhang Y, Lu C, Kuang H. P-MITE: a database for plant miniature
761 inverted-repeat transposable elements. Nucleic Acids Research. 2014;42(Database
762 issue):D1176-81.
763 63. Smit A, Hubley R. RepeatModeler Open-1.0. 2008-2015 [Available from:
764 http://www.repeatmasker.org.
765 64. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR:
766 ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
767 65. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al.
768 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and
769 isoform switching during cell differentiation. Nature biotechnology. 2010;28(5):511-5.
770 66. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method
771 employing protein multiple sequence alignments. Bioinformatics. 2011;27(6):757-63.
772 67. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for
773 Eukaryotic Genomes. Genome Research. 2003;13(9):2178-89.

774 68. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic
775 Acids Research. 2000;28(1):27-30.
776 69. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.
777 BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
778 70. Pundir S, Martin MJ, O'Donovan C. UniProt Protein Knowledgebase. In: Wu CH,
779 Arighi CN, Ross KE, editors. Protein Bioinformatics: From Protein Modifications and
780 Networks to Proteomics. New York, NY: Springer New York; 2017. p. 41-55.
781 71. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic
782 Acids Research. 2017;45(D1):D158-D69.
783 72. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al.
784 Versatile and open software for comparing large genomes. Genome biology. 2004;5(2):R12.
785 73. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-
786 MEM. arXiv:13033997v1. 2013.
787 74. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2:
788 Somatic mutation and copy number alteration discovery in cancer by exome sequencing.
789 Genome Research. 2012;22(3):568-76.
790 75. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of
791 occurrences of k-mers. Bioinformatics. 2011;27(6):764-70.
792 76. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and
793 space complexity. BMC Bioinformatics. 2004;5:113.
794 77. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large
795 datasets. Bioinformatics. 2014;30(22):3276-8.
796 78. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R
797 language. Bioinformatics. 2004;20(2):289-90.
798 79. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592-3.

799 80. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
800 sequence data. Bioinformatics. 2014;30(15):2114-20.

801 Supporting information

802 **S1_Table.pdf: BUSCO analysis of the Chardonnay FALCON Unzip assembly before**
803 **and after curation.**

804 **S2_Fig.pdf: Redundant primary contig reduction**. Circular representations of **A)** FALCON
805 Unzip Chardonnay assembly and **B)** the same assembly after curation. Tracks are: length-
806 ordered contigs **(i)**, read-depth of mapped Illumina paired end reads **(ii)** and heterozygous
807 SNPs density **(iii).**

808 **S3_Fig.pdf: Gene expansion of Chardonnay Chromosome 5 (primary contigs) region**
809 **containing FAR2-like genes.** Alignments are indicated as black lines (dotplot), the ORFs
810 for FAR2-like genes and pseudogenes are indicated for both Pinot noir and Chardonnay.

811 **S4_Fig.pdf: Distribution of clonal markers over the Chardonnay assembly**. Chardonnay
812 chromosome-ordered primary contigs **(i)** and clonal marker variants **(ii)**.

813 **S5_Dataset.xlsx: Chardonnay clonal-specific markers. Sheet 1)** Chardonnay clone-
814 specific markers with read-counts, **Sheet 2)** markers in gene models with Annovar and
815 Provean predictions, **Sheet 3)** Summaries of clonal marker detection screening against
816 validation datasets.

817 **S6_Archive.tar.gz: Scripts and workflows for data analysis.** Extract with tar for linux or
818 Mac, or 7zip (7zip.org) for Windows. Contents: **bin/**, All custom scripts used for analysis; **lib/**,
819 Custom Perl library for scripts; **src/**, Source code for window coverage masking program;
820 **workflows/**, Commands used with comments for all data analysis; **Makefile**, The GNU
821 Make pipeline for marker discovery.