The genomic diversification of clonally propagated grapevines

Amanda M. Vondras[1], Andrea Minio[1], Barbara Blanco-Ulate[2], Rosa Figueroa-Balderas[1], Michael A. Penn[1], Yongfeng Zhou[3], Danelle Seymour[3], Ye Zhou[1], Dingren Liang[1], Lucero K. Espinoza[1], Michael M. Anderson[1], M. Andrew Walker[1], Brandon Gaut[3], Dario Cantu[1]*

[1] Department of Viticulture and Enology, University of California Davis, Davis, CA 95616

[2] Department of Plant Sciences, University of California, Davis, CA 95616

[3] Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92617

*Corresponding author. Telephone: +1 530-752-2929 Email: dacantu@ucdavis.edu

Abstract

Vegetatively propagated clones accumulate somatic mutations. The purpose of this study was to better understand the consequences of clonal propagation and involved defining the nature of somatic mutations throughout the genome. Fifteen Zinfandel winegrape clone genomes were sequenced and compared to one another using a highly contiguous genome reference produced from one of the clones, Zinfandel 03.

Though most heterozygous variants were shared, somatic mutations accumulated in individual and subsets of clones. Overall, heterozygous mutations were most frequent in intergenic space and more frequent in introns than exons. A significantly larger percentage of CpG, CHG, and CHH sites in repetitive intergenic space experienced transition mutations than genic and non-repetitive intergenic spaces, likely because of higher levels of methylation in the region and the increased likelihood of methylated cytosines to spontaneously deaminate. Of the minority of mutations that occurred in exons, larger proportions of these were putatively deleterious when they occurred in relatively few clones.

These data support three major conclusions. First, repetitive intergenic space is a major driver of clone genome diversification. Second, clonal propagation is associated with the accumulation of putatively deleterious mutations. Third, the data suggest selection against deleterious variants in coding regions such that mutations are less frequent in coding than noncoding regions of the genome.

Keywords

Clonal propagation, DNA methylation, genome diversification, somatic mutations, structural variation, transposable elements

Introduction

Cultivated grapevines are clonally propagated. As a result, the genome of each cultivar is preserved, except for the accumulation of mutations that accumulate over time and can generate distinguishable clones [1-4]. Somatic mutations are responsible for several notable phenotypes. For example, a single, semi-dominant nucleotide polymorphism can affect hormone response [5] and recessive insertion of the *Gret1* retrotransposon in the promoter of the *VvmybA1* transcription factor inhibits anthocyanin accumulation in white varieties [6], as do additional mutations affecting the color locus [7-10]. The fleshless fruit of an Ugni Blanc clone and the reiterated reproductive meristems observed in a clone of Carignan are both caused by dominant transposon insertion mutations [11,12]. In citrus, undesirable mutations can be unknowingly propagated that render fruit highly acidic and inedible [13,14]. Interestingly, somatic mutations in plum are associated with a switch from climacteric to non-climacteric ripening behavior [15].

There is limited understanding and evidence of the extent, nature, and implications of the somatic mutations that accumulate in clonally propagated crops [16]. Genotyping approaches based on whole genome sequencing make it possible to identify genetic differences without predefined markers [17-19] and expedite learning the genetic basis of valuable traits and developmental processes [15,20]. Still, few previous studies have used genomic approaches to study somatic variations among clones [17-21]. The first to publish a genome-wide exploration of somatic variation in grapevine was Carrier *et al.* (2012), finding that transposable elements were the largest proportion of somatic mutation types affecting four Pinot Noir clones [18].

Whole genome sequencing was also used to study structural variations and complex chromosomal rearrangements in Tempranillo, comparing diverse accessions of phenotypically distinct Tempranillo Tinto and Tempranillo Blanco to better understand the basis of somatic mutations giving rise to red versus white fruit [20]. Genomic tools could be used to comprehensively describe the extent of somatic mutations and infer the processes affecting clone genomes.

Mutations occur in somatic cells that proliferate by mitosis. These can occur by a variety of means, including single base-pair mutations [22,23] that are more prevalent in repetitive regions because methylated cytosines passively deaminate to thymines [24-26], polymerase slippage that drives variable microsatellite insertions and deletions [27], and larger structural rearrangements and hemizygous deletions [10,20]. Transposable elements are also a major source of somatic mutations in grapevines [18], though transcriptional and post-transcriptional mechanisms exist to prevent transposition and maintain genome stability [28-31]. Notably, methylation of transposable elements is one specific mechanism that prevents transposition, which establishes a tradeoff, then, between methylation and the transposition of mobile elements.

At the cellular level, distinct clones can emerge following a mutation in a shoot apical meristem that spreads throughout a single cell layer, creating periclinal chimeras. This chimera is stable for Pinot Meunier, a clone of Pinot Noir with distinct L1 and L2 layers in shoots [3]. Each cell layer in a stratified apical meristem like that observed in grape [32] is developmentally distinct. The distinct cell layers will remain so provided cell divisions occur anticlinally. But, periclinal divisions and cellular rearrangements can result in the homogenization of a mutant genotype across cell layers [33]. This is the case for green-yellow bud sports of the grey-fruited Pinot Gris, wherein sub-epidermal white cells invaded and displaced epidermal pigmented cells

[9]. In contrast to replacement (L1 cells invade L2), displacement is likely more common because of the relative disorganization of the inner cell layers [32,33].

Meristem architecture is related to the fate of somatic mutations, as it influences the impact of these mutations and the likelihood of competition between cell lineages, also known as diplontic selection [34-36]. Provided each cellular layer is maintained by anticlinal divisions, deleterious mutations can be preserved in periclinal chimeras [35,37]. In addition, the predominance of "hidden", heterozygous recessive somatic mutations [2,37] may further shield somatic mutations from selective forces. These factors are permissive of the accumulation of somatic mutations. Diplontic selection could occur if periclinal cell divisions result in the invasion of one cell layer by cells from another [34,35]. This mechanism could oppose the accrual of deleterious mutations expected by Muller [38,39]. A recent study of the long-lived pedunculate oak described substantial intra-organismal genetic variation, but did not draw conclusions about the contribution of somatic variations to large-scale oak evolution [21]. Evidence of diplontic selection in plants is remarkably scarce [37], though its likelihood given different circumstances has been modeled [34,35,40]. Given the prevalence of chimerism and rearrangements documented in the model [9,33], grapevine is a suitable model for investigating the possibility of selection during vegetative propagation.

Zinfandel is the third-most cultivated wine grape in California [41,42] DNA profiling produced evidence that Zinfandel is synonymous with Primitivo grown in Italy [43] and Croatian Pribidrag and Crljenak Kastelanski [44]. Historical records plus the cultivation of closely related cultivars support Croatia as the likely origin of Zinfandel [44-47] and also that Primitivo was likely brought to the Gioia del Colle region in Italy by Benedictine monks in the 17th century [3,48]. The reported variability in Zinfandel [49-51], including subtle variability in phenolic

metabolites (Additional file 1), and its long history of cultivation make it a useful model for studying clonal variation in grapevine, specifically, and the nature of the accumulation of somatic mutations in clonally propagated crops, generally.

The purpose of this study was to better understand the nature of the somatic variations that occur during clonal propagation. Representatives from at least a portion of Zinfandel's history [44-47] from Croatia, Italy, and California were sequenced and compared using Zin03 as reference. First, we show that intergenic space drives clonal diversification. As previously reported, transposable element insertions varied among clones [18]. This report expands that understanding to implicate methylation as an indirect driver of clonal diversification; rare somatic heterozygous SNPs were most observed in the repetitive intergenic regions, likely because of the high levels of transposition-inhibiting methylation and associated transition mutations that are prevalent there. Second, the data support an important component of Muller's ratchet [38], that asexually propagated organisms accumulate deleterious mutations. Third, somatic mutations were relatively scarce in the coding regions of genes relative to introns and intergenic space, suggesting some degree of negative selection against deleterious mutations.

Results

Zinfandel genome assembly, annotation, and differences between haplotypes

The clone used for the genome assembly, Zinfandel 03 (Zin03), was acquired by FPS in 1964 from the Reutz Vineyard near Livermore, California that was planted during Prohibition (1920 – 1933) [52]. Zin03 was sequenced using Single Molecule Real-Time (SMRT; Pacific Biosciences) technology at ~98x coverage and assembled using FALCON-unzip [53], a diploid-aware assembly pipeline. The genome was assembled into 1,509 primary contigs (N50 = 1.1

Mbp) for a total assembly size of 591 Mbp, similar to the genome size of Cabernet Sauvignon (590 Mbp) [53] and larger than Chardonnay (490Mb) [19] and PN40024 (487 Mb) [54]. Fifty two percent of the genome was phased into 2,246 additional phased sequences (haplotigs) where the homologous chromosomes were distinguishable with an N50 of ~442 kbp (Table 2). A total of 53,560 complete protein-coding genes were annotated on the primary (33,523 genes) and haplotig (20,037 genes) assemblies (Table 2).

Of the 20,037 genes annotated on the haplotig assembly, 18,878 aligned to the primary assembly, leaving 1,159 genes that may exist hemizygously in the genome due to structural variation between homologous chromosomes or because of substantial divergence in sequence between haplotypes. These genes were annotated with a broad variety of putative functions, including biosynthetic processes, secondary metabolism, and stress responses. Long reads were mapped to both the primary and haplotig assemblies to evaluate the circumstances that explain the differences between haplotypes. Structural variants (SVs) between the haplotypes were examined by mapping long SMRT sequencing reads onto Zin03's primary and haplotig assemblies with NGMLR and calling SVs with Sniffles [55].

A total of 22,399 SVs accounted for 6.94% (41.0 / 591 Mbp) of the primary assembly's length and 6.02% (8.4 / 139 Mbp) of the primary assembly's gene-associated length (Fig. 1a, Table 3). SVs intersected 4,559 genes in the primary assembly (13.6% of primary assembly genes) and 390 SVs spanned more than one gene. Manual inspection of the long reads aligned to the primary assembly support that large, heterozygous deletions and inversions occurred in the Zin03 genome that were either inherited from different structurally distinct parents or arose during clonal propagation (Fig. 1b,c,d). Importantly, there was substantial hemizygosity in the

genome, with long reads supporting deletions affecting 2,521 genes and 4.56% of the primary assembly's length (Table 3).

Next, we considered whether specific structural variation could account for the 1,159 genes uniquely found in the haplotig assembly. Three hundred eighty-two genes of the previously mentioned 1,159 genes that uniquely exist within the haplotig assembly intersected structural variations. Two hundred ninety of these intersected deletions, accounting for the failure to identify them on the primary assembly. Some of the haplotig genes that failed to map to the primary assembly intersected additional types of SVs, including duplications (80 genes), insertions (89 genes), and inversions (16 genes).

These results reveal structural differences between Zinfandel's haplotypes. These differences could have been inherited and/or could have occurred during clonal propagation. Overall, these structural variations affected 4,559 primary assembly genes. Importantly, these data show that a notable portion of the primary assembly's length (4.56%) is hemizygous.

Differences in structure and gene content between Zinfandel and other grape genomes

The Zin03 genome was compared to PN40024 and Cabernet Sauvignon to identify cultivar-specific genes that may contribute to Zinfandel's characteristics. PN40024 is the inbred line derived from Pinot Noir used to develop the first grape genome reference [54] and Cabernet Sauvignon (CS08) was recently used to construct the first diploid, haplotype-resolved grape genome for which long reads are available [53]. Overall, 1,801 genes were not shared between all three genotypes (Zin03, Pinot Noir, and Cabernet Sauvignon; Fig. 2a). Three hundred nine protein coding genes were found uniquely in Zin03 relative to PN40024 and CS08; 223 were annotated on the primary assembly and 86 were annotated on the haplotigs (Fig. 2a, Additional

file 2). These genes had a panoply of functions that included but were not limited to nucleotide binding (60 genes), protein binding (58 genes), stress response (34 genes), and kinases (28), and were associated with membranes (48 genes), signal transduction (23 genes), carbohydrate metabolism (12 genes), and lipid metabolism (8 genes; Additional file 2).

Structural differences between Zin03 and CS08 were explored in more detail by mapping the long SMRT reads of CS08 onto Zin03's primary and haplotig assemblies with NGMLR and calling SVs with Sniffles (Fig. 2b, Table 3). Overall, these SVs corresponded to 17.74% (159/ 897 Mbp) of the Zin03 assembly's total length, 12.5% of its total protein-coding regions (28 / 223 Mbp), and 25.6% of all Zin03 genes. SVs affected 9,885 genes in the primary assembly and 3,804 genes in the haplotigs. Manual inspection of the alignment of long CS08 reads to Zin03's primary assembly support that large SVs exist between the two genotypes (Fig. 2c,d). Next, we considered whether specific structural variation called by Sniffles could account for the 576 Zin03 genes absent from CS08 according to the reciprocal mapping analysis (Fig. 2a). Of these 576 Zinfandel genes, 268 genes intersected 454 deletions supported by long CS08 reads aligned to Zin03.

Though Zinfandel had few unique genes, high levels of structural variation between Zinfandel (Zin03) and Cabernet Sauvignon (CS08) were observed and these affected considerable protein-coding regions of the genome. These results justify constructing a Zinfandel-specific reference to better capture genomic variability among Zinfandel clones that could otherwise be missed, particularly if an alternative reference lacks sequences present in Zinfandel.

Relatedness among Zinfandel clones

206 Fifteen Zinfandel clones, including Zin03, were sequenced using Illumina. The resulting
207 reads were aligned to the Zin03 primary assembly to characterize SNPs, small INDELs, variable
208 transposon insertions, and large structural variants. Principal Component Analysis (PCA) of
209 variants among the clones showed no clear pattern in their relationships to one another based on
210 their recorded origins prior to acquisition (Fig. 3a). The ambiguity surrounding the travels and
211 histories of these clones means that it should not be taken for granted that the Californian
212 selections, for example, ought to be more closely related to one another than to the Italian or
213 Croatian selections. Notably, Pribidrags 5 and 15, which have a known and close relationship, do
214 not co-localize in the PCA (Fig. 3a, Table 1).
215 A kinship analysis [56] was then used to quantitatively assess the relationships between
216 the Zinfandel selections. These values range from zero (unrelated) to 0.5 (self). Additional
217 cultivars were included in the analysis with known relationships to help contextualize the
218 differences between clones and the integrity of the analysis (Fig. 3b). Cabernet Franc and Merlot
219 have a parent - offspring relationship, as do Pinot Noir and Chardonnay [57,58]. These pairs had
220 kinship coefficients of 0.15 and 0.18, respectively (Fig. 3b). As a possible grandparent of
221 Sauvignon Blanc, Pinot Noir had a kinship coefficient of 0.05 with Sauvignon blanc [59,60].
222 Zinfandel selections had kinship coefficients between 0.42 and 0.45; this is likely because of the
223 accrual of somatic mutations among clones (Fig. 3b).
224 Across the Zinfandel clones, the median number of homozygous and heterozygous
225 variants called relative to Zin03 were 37,437 and 718,174, respectively. Between 10-fold and 27-
226 fold more heterozygous variants were called than homozygous variants in each clone, and less
227 than 10% of sites did not share the Zin03 reference allele (Additional file 3).
228

Clonal versus cultivar genetic variability

Overall, an average of 761,948 variant sites were identified in individual Zinfandel clones when short reads were mapped on the Zin03 primary assembly. On average, 6,153,830 variant sites were identified in other cultivars (Pinot noir, Chardonnay, Sauvignon Blanc, Merlot, Cabernet Franc) relative to Zin03 (Additional file 3). Both of these figures excluded heterozygous sites at which the diploid genotype called for a given sample was identical to that called for Zin03.

Variants were 7.9X more frequent in other cultivars relative to Zin03 than for Zinfandel clones; on average, mutations in clones occurred once every 723 bases and once every 92 bases in other cultivars (Additional file 3). However, the ratio of transitions to transversion mutations and the proportions of the severities of the predicted variant effects were similar for both groups (Additional file 3). The normalized count of variants differed between cultivars and Zinfandel clones on the basis of variants' location in the genome, the type of variant, and the zygosity of the variant (Fig. 4).

Variants in non-Zinfandel cultivars and heterozygous variants among Zinfandel clones were significantly more prevalent in intergenic space than introns and exons and significantly more prevalent in introns than exons (Tukey HSD, $p < 0.01$). Unlike homozygous variants between cultivars and as expected, homozygous variants were rare among clones (Fig. 4, Additional file 3). Still, the normalized count of homozygous INDELs in intergenic space, introns, and exons were significantly different among Zinfandel clones (Tukey HSD, $p < 0.01$), as were the normalized count of homozygous intergenic versus genic (exons and introns) SNPs (Tukey HSD, $p < 0.01$). The normalized count of homozygous SNPs in exons and introns were not significantly different in Zinfandel clones (Tukey HSD, $p > 0.01$). The accrual of

predominantly heterozygous and likely recessive variants [2] is consistent with what would be expected given physically separate homologous chromosomes and the absence of sexual reproduction. The differences in mutation abundances observed were initially surprising; if somatic mutations occurred randomly and absent mechanisms that make certain sites more or less susceptible to mutation, then different regions of the genome should have had equal levels of mutations. This was not the case (Figure 5).

The accrual of somatic mutations in Zinfandel clones

Heterozygous sites found among the 15 Zinfandel clones ought to be a mixture of sites inherited from their shared ancestral plant and somatic mutations that arose during clonal propagation. To better understand the nature of somatic mutations, the data were handled slightly differently than they were to construct Figure 4; all 15 Zinfandel clones were included and all heterozygous calls were considered, even if all genotypes were identically heterozygous. Thirty percent of heterozygous SNPs, 24% of heterozygous INDELs, and 47% of heterozygous structural positions were shared by all 15 Zinfandel clones (Fig. 5a). Because all clones are identically heterozygous at these loci, these variants are those inherited from Zinfandel's parents.

Individual and subsets of Zinfandel clones accumulated heterozygous mutations as clonal propagation occurred (Fig. 5a). Thirteen percent and 16% of heterozygous INDELs and SNPs, respectively, and 1% of large (>50 bp) structural variants occurred in only one or two clones (Fig. 5a). The distribution of SVs called by Delly is markedly different than those of SNPs and INDELs (Fig. 5a). For both SNPs and INDELs, there were 3 and 3.5-fold as many heterozygous variants shared by all 15 clones as there were uniquely occurring variants; there were 71.5-fold more structural variants shared by all clones than there were unique variants in individual clones

(Fig. 5a). This might imply that the mechanisms that give rise to small mutations are more common among clones than the large-scale changes associated with SVs.

The distribution of unique and shared heterozygous INDELs in exons, introns, repetitive, and non-repetitive intergenic spaces were not equal (Fig. 5b). The distribution of INDELs in exons was significantly different than the distributions of INDELs in each other feature considered (Kolmogorov-Smirnov Test, $p < 0.01$). Similarly, the distributions SNPs in genic (exons, introns) and intergenic (repetitive, non-repetitive) regions were not equal (Fig. 5b). Shared heterozygous SNPs were most common in intergenic non-repetitive regions and introns and least common in exons and repetitive intergenic regions (Fig. 5b). Interestingly, unique heterozygous SNPs occurred at high rates in repetitive intergenic regions (Fig. 5b).

That shared heterozygous sites are mostly in non-repetitive intergenic space and unique heterozygous sites are mostly in repetitive space may have to do with the increased likelihood that methylated cytosines spontaneously deaminate and the prevalence of methylated repetitive sequences in those regions [22,25,29,30]. This is also supported by the significantly higher ratio of transitions to transversions in repetitive intergenic regions than in exons, introns, and non-repetitive intergenic space (Fig. 5c). Furthermore, the mean percentage of CpG, CHG, and CHH sites affected by transition mutations was significantly higher in repetitive intergenic space than genic and non-repetitive intergenic spaces (Fig. 5d; Tukey HSD, $p < 0.01$). The mean percentage of CpG sites affected by transition mutations was also significantly higher in introns than exons (Tukey HSD, $p < 0.01$). Compatible with this hypothesis, INDELs, which should not increase in frequency due to methylation, did not occur preferentially in repeats (Fig. 5b).

The impact of specific variants also varied with their prevalence among the clones (Fig. 5e). “High impact” mutations were predicted by SNPEff [61]. The high impact mutations

identified in these data included exon losses, start and stop site gains and losses, frameshifts, gene fusions, splice acceptor mutations, and splice donor mutations. These mutations are predicted to be deleterious because of their disruptive effects on the coded protein. For these reasons, we designated such mutations as putatively deleterious in this manuscript. These were counted for each Zinfandel clone relative to Zin03. Relatively low proportions of heterozygous variants shared by all Zinfandel clones were putatively deleterious. In contrast, larger proportions of exonic SNPs and INDELs that occurred in individual or subsets of clones were putatively deleterious (Fig. 5e).

Together, these results show that mutations associated with clonal propagation are most numerous outside of coding regions of the genome, indicating that clone genomes diversify most rapidly in the intergenic space, particularly in repetitive and likely methylated regions (Fig. 5). Though a minority of somatic mutations occurred in exons, we show that exonic mutations that occur in few or individual clones are more often deleterious than exonic heterozygous variants shared by all or most clones. In other words, clonal propagation is associated with the accumulation of putatively deleterious heterozygous mutations.

Zinfandel clones incur unique transposon insertions

Transposable element insertions (TEI) contribute to somatic variation in grape [6,11,12,18]. Relative to Zin03, 1,473 TEI were identified among the Zinfandel clones. A large fraction of TEI (26.7%) occurred uniquely in individual clones (Fig. 6a) and included 325 retrotransposons, mostly Copia and Gypsy LTRs, and 69 DNA-transposons (Fig. 6b). Because uniform loci are excluded, in-common TEI were not captured when clones were compared to Zin03. Comparing the clones relative to PN40024, however, revealed that the majority (64.8%)

of TEI were shared among the 15 Zinfandel clones. Five hundred thirty TEI occurred in only one, two or three clones (Fig. 6a). This result supports the derivation of these selections from a common ancestral plant and the accumulation of somatic variations over time.

In addition to being suggestive of their shared heritage, the positions of these insertions and their proximity to coding genes were notable. Three-hundred forty-seven TEI occurred within 314 coding genes. The remaining 938 TEIs were in intergenic regions (Fig. 6c). The median upstream and downstream distance of intergenic TEs from the closest feature were 11,811 and 11,279 base-pairs, respectively, and 25% of TEI were less than 4,345 bases downstream of the closest feature and/or less than 3,826 bases upstream of the closest feature (Fig. 6c).

Discussion

Consideration of the genomic differences among Zinfandel clones revealed what is likely a complex history not easily reconstructed. Analyses of the relationships between clones did not reveal groupings of clones per their recorded countries of origin. Somatic mutations may help identify individual clones but could also blur the historical relationships between them. It is also plausible that pairs of clones from any given region are not direct cuttings of one another but of Zinfandels from another region; the clones now grown in California, for example, may have been imported on numerous independent occasions from various other regions, meaning some may indeed be more closely related to one of the Primitivo or Croatian clones than they are to other Californian clones. It would be unwise to assume a single migratory path radiating from an ancestral mother plant ought to be applicable to the clones.

343 Despite this ambiguity, the examination of SNPs, INDELs, transposable elements and
344 other structural variants all support the derivation of all but one of the clonal selections from a
345 common ancestral Zinfandel mother plant and show the accumulation of somatic mutations over
346 time (Figs. 5 and 6). The structure of the Zinfandel genome, location of mutations among clones,
347 their frequency and prevalence, and the relationship between these factors provides some insight
348 into the nature of mutations in clonally propagated plants. Mutations among clones were
349 predominantly heterozygous (Fig. 4) and uncommon heterozygous mutations shared by a subset
350 of or individual clones were increasingly deleterious when they occurred in exons (Fig. 5e).
351 There are costs and benefits associated with clonal propagation [16]. Among the benefits
352 are that the plants need not breed true-to-type; clonal propagation generally fixes heterozygous
353 loci and valuable phenotypes. However, the increase in the proportion of deleterious alleles
354 supports Muller's ratchet, which posits that sex is advantageous and that clonal propagation
355 increases mutational load [38]. Though these and previous data do not tell which mutations are
356 actually recessive or dominant, they could remain hidden if they are recessive or do not manifest
357 their deleterious effects [2,62]. However, even after taking into consideration the total length of
358 exons, introns, and intergenic space (repetitive and non-repetitive), heterozygous mutations
359 occurred at varying frequency in these regions and were least abundant in coding regions. The
360 rarity of mutations in exons and commonality of mutations in repetitive intergenic space may
361 have at least two components.
362 Mutations are likely more frequent in repetitive intergenic space as a result of the
363 regulation of transposition by DNA methylation. Repetitive intergenic space had the highest rate
364 of relatively unique SNPs and the ratio of transitions to transversions was significantly higher
365 there than in other regions. DNA methylation is an important epigenetic control and is one

366 mechanism that maintains genome stability and impairs the transposition of mobile elements
367 [29,63,64]. Methylated cytosines, however, spontaneously deaminate faster than unmethylated
368 cytosines [24,30]. Together, the expectations that intergenic regions are rich in transposable
369 elements, that these regions are typically highly methylated and as a result will experience
370 greater transition rates account for the high rates of SNPs in repetitive intergenic spaces among
371 Zinfandel clones. Also notable, these data show that some transposable elements are not entirely
372 silenced, with a substantial number inserting in genes or in close proximity to genes (Fig. 6c).
373 These insertions could be effectively inconsequential or not; transposable element insertions can
374 result in novel transcripts and affect gene expression regulation [11,65]. Gene body methylation
375 is appreciated as a mutagenic "double-edged sword" [66], with benefits coming at the price.
376 Recent work observed region-specific methylation in vegetatively propagated Sardinian white
377 poplar that may serve an advantageous function [67] and others have suggested that the
378 epigenome contributes to the success of vegetatively propagated plants [68]. Future work might
379 also consider the long-term price associated with intergenic mutagenesis and the potential loss of
380 methylation in vegetatively propagated plants.

381 The rarity of exonic mutations was surprising. After accounting for the length of these
382 spaces in the genome and their repetitiveness, we expected uniform rates of mutation in exons,
383 introns, and intergenic space. Instead, we observed that although rare somatic mutations in exons
384 were increasingly deleterious, they were relatively scarce. Some degree of negative selection
385 against deleterious variants in coding regions could explain why mutations were less frequent in
386 coding than noncoding regions of the genome. The possibility of diplontic, clonal selection or
387 competition between cell lineages that could purge otherwise consequential deleterious
388 mutations has been modeled, but evidence of its occurrence is sparse [16,34,39]. The structures

of apical meristems [35,69] and the tendency of somatic mutations to be heterozygous and recessive [2] place constraints on the likelihood that deleterious mutations would be subjected to negative selection. Periclinal divisions across cell layers could enhance diplontic selection [34] against dominant and/or hemizygous recessive alleles. Four and one half of Zinfandel's genome is hemizygous; structural variations identified within the Zinfandel genome and the rampant hemizygosity reported in Chardonnay [10] could also expose otherwise hidden somatic variations to selective pressure hostile to the accumulation of deleterious mutations. Additional work should explore to what degree each of these factors, or others not considered here, explain why somatic mutations in exons were relatively infrequent and characterize the realized long-term consequences of mutation accumulation versus selection for grapevine and other clonally propagated plants.

Conclusions

This study described the nature of the mutations causing the diversification of 15 clonally propagated grapevines and confirm their derivation from a single ancestral mother Zinfandel. The findings indicate that repetitive intergenic space, likely because of its higher rates of methylation in plants, is a significant contributor to the pool of mutations differentially observed among the clones. In addition, the analyses revealed that though relatively infrequent compared to intergenic mutations, mutations in exons were increasingly deleterious the less common they were among Zinfandel clones. This result is consistent with the expectation that vegetative propagation is associated with the accrual of mutations and adds that negative selection may simultaneously purge mutations from the genome. These findings add novel insight and nuance to our understanding of the nature and fates of mutations during vegetative propagation.

412

413 **Methods**

414 *Zinfandel plant material and additional accessions*

415 Fifteen Zinfandel clones were used for this study. Plants were confirmed to be clones of
416 Zinfandel using the following microsatellite markers: VVMD5, VVMD7, VVMD27, VVMD31,
417 VVMD32, VVMS2, VRZAG62, and VRZAG79 [44,70,71]. Fourteen of these clones are
418 available through Foundation Plant Services (FPS) at the University of California Davis. Nine of
419 the fifteen clones belong to the Zinfandel Heritage Vineyard Project, a collection of rare
420 Zinfandel vine cuttings grown in the same vineyard. The identification numbers, common
421 names, and source of the clones used in this study are listed in Table 1. An FPS identification
422 number suffix of ".1" indicates that the clone underwent microshoot tip tissue culture therapy,
423 with two exceptions. Pribidrag 13 and Pribidrag 15 are directly derived from the same plants as
424 Pribidrag 4 and Pribidrag 5, respectively, but did not undergo microshoot tip tissue culture
425 therapy. They are labeled with identical FPS numbers to make clear that the relationship between
426 them is known. In this manuscript, Zinfandel clones will be referred to by the clone numbers and
427 common names listed in Table 1.

428

429 *DNA extraction, library preparation, and sequencing*

430 High quality genomic DNA was isolated from grape leaves using the method described in
431 Chin *et al*. (2016) [53]. DNA purity was evaluated with a Nanodrop 2000 spectrophotometer
432 (Thermo Scientific, Hanover Park, IL), quantity with a Qubit 2.0 Fluorometer (Life
433 Technologies, Carlsbad, CA) and integrity by electrophoresis. For SMRT sequencing, SMRTbell
434 libraries for the Zinfandel reference FPS clone 03 (Zin03) were prepared as described by Chin *et*

435 *al*. (2016). For Illumina sequencing, DNA sequencing libraries for each of the fifteen Zinfandel
436 clones were prepared using the Kapa LTP library prep kit (Kapa Biosystems) as described by
437 Jones *et al*., (2014) [72]. Final libraries were evaluated for quantity and quality using a
438 Bioanalyzer 2100 (Agilent Technologies, CA). Zin03 SMRTbell libraries were sequenced on a
439 PacBio RS II and Illumina libraries were sequenced in 100 and 150 base-pair paired-end reads
440 on an Illumina HiSeq3000 sequencer (DNA Technology Core Facility, University of California,
441 Davis). Genome sequences of additional *V. vinifera* were used in this study, including long reads
442 from Cabernet sauvignon (NCBI BioProject PRJNA316730) and short reads from Cabernet
443 franc, Chardonnay, Merlot, Pinot Noir, and Sauvignon blanc (NCBI BioProject PRJNA527006).
444
445 *Zinfandel genome assembly and annotation*
446 *De novo* assembly of Zinfandel (Zin03) was performed at DNAnexus (Mountain View,
447 CA, USA) using PacBio RS II data and the FALCON-unzip (v. 1.7.7) pipeline [53]. FALCON-
448 unzip was used for its ability to assemble contiguous, phased diploid genomes with better
449 resolved heterozygosity [53,73]. Repetitive sequences were masked prior to error correction
450 using TANmask and REPmask modules in Damasker [74]. After error-correction (13,073 bp
451 length cut-off), a total of 1.68 million error-corrected reads (N50 15Kbp, 29-fold coverage of
452 expected genome size) were obtained and repeats were masked before overlap detection in the
453 FALCON pipeline (v. 1.7.7). PacBio reads were assembled after testing multiple parameters to
454 produce the least fragmented assembly. These conditions are listed in Additional file 4.
455 Haplotype reconstruction was performed with default parameters. Finally, contigs were polished
456 with Quiver (Pacific Biosciences, bundled with FALCON-unzip v. 1.7.7). Repeats were

457 annotated on the Zin03 assembly using RepeatMasker (v. open-4.0.6) [75] and a *V. vinifera*
458 repeat library [76].

459 The publicly available RNAseq datasets listed in Additional file 4 were used as
460 transcriptional evidence for gene prediction. Each RNAseq sample was trimmed with
461 Trimmomatic (v. 0.36; Additional file 4) and assembled with Stringtie (v. 1.3.3) [77] to
462 reconstruct variety-specific transcripts. A detailed list of all experimental data used for the
463 annotation procedure is in Additional file 4. This data was then mapped on the genome using
464 Exonerate (v. 2.2.0, transcripts and proteins) [78] and PASA (v. 2.1.0, transcripts) [79].
465 Alignments, and *ab initio* predictions generated with SNAP (v. 2006-07-28) [80], Augustus [81],
466 and GeneMark-ES [82] were used as input for EVidenceModeler (v. 1.1.1) [83].
467 EVidenceModeler was used to identify consensus gene structures using the weight reported in
468 Additional file 4. Functional annotation was performed using the RefSeq plant protein database
469 (ftp://ftp.ncbi.nlm.nih.gov/refseq, retrieved January 17th, 2017) and InteProScan (v. 5) as
470 previously described [76].
471
472 *Genetic variant calling*
473 Comparisons between Zinfandel clones and between Zin03 and other cultivars were
474 made using the Zin03 genome as reference. This pipeline is described in Additional file 5. Small
475 insertions and deletions (INDELs), single nucleotide polymorphisms (SNPs), and structural
476 variations (SVs) were analyzed. The short Illumina reads belonging to the fifteen Zinfandel
477 clones and additional cultivars were trimmed using Trimmomatic (v. 0.36; Additional file 4).
478 Quality filtered and trimmed paired-end reads were then randomly down-sampled to 84 million
479 (~14X coverage) in each library to mitigate the possibility of sequencing depth-dependent

480 outcomes. All libraries were aligned to Zin03 using bwa (v. 0.7.10) and the -M parameter [84].
481 For all genotypes, the median number of reads mapping to the Zinfandel reference genome was
482 97%. Next, Picard Tools (v. 2.12.1) were used to mark optical duplicates, build BAM indices,
483 and validate SAM files (http://broadinstitute.github.io/picard). Variants were called using
484 GATK's HaplotypeCaller (v. 3.5) [85]. Then, called variants were filtered and annotated (--
485 filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum
486 < -8.0"). Variant call files were combined using GATK's GenotypeGVCFs. Having mapped
487 Illumina reads corresponding to the Zinfandel reference onto itself, erroneous non-reference
488 Zin03 calls (8.1%) were removed. The variants called included SNPs and INDELs.
489 Next, large structural variations among clones, between Zin03 and other cultivars, and
490 between Zin03's haplotypes were studied. First, Zin03 genes were compared to PN40024 and
491 Cabernet Sauvignon (CS08) by mapping coding sequences on genome assemblies using Gmap
492 (v. 2015-09-29) and the following parameters: -K 20,000 -B 4 -f 2. Hits with at least 80%
493 identity and reciprocal coverage are reported. Genes annotated on Zin03's haplotig assembly
494 were also mapped to Zin03's primary assembly to assess differences in gene content between
495 Zin03's haplotypes. SMRT reads from Zin03 and CS08 were mapped to Zin03 using NGMLR
496 (v. 0.2.7) and structural differences were called with Sniffles (v.1.0.8) [55]. Zinfandel clones
497 were compared to one another using Illumina short reads and Delly (v. 0.7.8) with default
498 parameters [86]. The structural variations identified by Sniffles and Delly in Zin03 were
499 intersected. Several filters were applied to the results of SV analyses. Transversions, non-
500 reference Zin03 genotype calls, SVs annotated at the ends of contigs, and SVs that intersected
501 the repeat annotation were filtered from Delly output.
502

Transposon insertion analysis

PoPoolationTE2 (v. 1.10.04) [87] was used to identify transposon insertions in the Zinfandel clones; it was used following the workflow outlined in its software manual (https://sourceforge.net/p/popoolation-te2/wiki/Manual/). Insertions were called relative to Zin03 genome assembly and PN20024 [54]. As described in Kofler *et al.* (2016), PoPoolationTE2 analyses transposable element insertions and can identify novel and annotated TE insertions provided insertions fall within predefined families of TEs. The annotation produced by RepeatMasker was used for the analysis. In this manuscript, the TE insertions among the clones are reported using the classification system and nomenclature described by Wicker *et al.* (2007) [88]. In instances where the TE order and/or superfamily was not annotated, only the TE class and order, when available, are named in the associated figures and text.

Relationships between Zinfandel clones

The relationships between Zinfandel clones were visualized by Principal Component Analysis and their relatedness was quantified (VCFtools v. 0.1.15) based on the method described by Manichaikul *et al.* (2010) [56]. This approach gives information about the relationship of any pair of individuals (unrelated, 3rd degree relative, 2nd degree relative, full siblings, and self) by estimating their kinship coefficient, which ranges from zero (no relationship) to 0.50 (self). These analyses used SNPs outside of repetitive regions.

List of abbreviations

ZAP, Zinfandel Advocates and Producers; UC, University of California; FPS, Foundation Plant Services; SMRT, Single Molecule Real-Time; Zin03, Zinfandel 03; CS08, Cabernet Sauvignon

08; PCA, Principal component analysis; TEI, Transposable Element Insertions; SV, Structural variant; INDEL, Insertion/Deletion; SNP, Single Nucleotide Polymorphism

Funding

This work was partially supported by start-up funds from the College of Agricultural and Environmental Sciences (UC Davis) to DC, the Louis P. Martini Endowment in Viticulture to DC and the NSF PGRP grant #1741627 to DC, MAW, and BG.

Author contributions

AMV, MAW, BG, and DC designed the experiments. BBU, YZ, MMA, and RFB collected the biological material and generated the data. MAP carried out the chemical analysis of the clones. AMV, AM, YZ, DS, DL, and LKE analyzed the data. AMV and DC prepared the figures and wrote the manuscript. All authors contributed to the final version of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are grateful for the vision of the late James A. Wolpert, who established the original Zinfandel clone trials with the support of the Zinfandel Advocates and Producers (ZAP).

References

1. Riaz S, Garrison KE, Dangl GS, Boursiquot J-M, Meredith CP. Genetic divergence and chimerism within ancient asexually propagated winegrape cultivars. J Amer Soc Hort Sci. 2002;127:508–14.

549 2. Zhou Y, Massonnet M, Sanjak JS, Cantu D, Gaut BS. Evolutionary genomics of grape (*Vitis*
550 *vinifera* ssp. *vinifera*) domestication. PNAS. 2017;114:11715–20.

551 3. Franks T, Botta R, Thomas MR. Chimerism in grapevines: implications for cultivar identity,
552 ancestry and genetic improvement. Theor. Appl. Genet. 2002;104:192–9.

553 4. Ramu P, Esuma W, Kawuki R, Rabbi IY, Egesi C, Bredeson JV, *et al*. Cassava haplotype map
554 highlights fixation of deleterious mutations during clonal propagation. Nat Genet. Nature
555 Publishing Group; 2017;49:959–63.

556 5. Boss PK, Thomas MR. Association of dwarfism and floral induction with a grape "green
557 revolution" mutation. Nature. 2002;416:847–50.

558 6. Kobayashi S, Goto-Yamamoto N, Hirochika H. Retrotransposon-induced mutations in grape
559 skin color. Science. 2004;304:982.

560 7. Walker AR, Lee E, Robinson SP. Two new grape cultivars, bud sports of Cabernet Sauvignon
561 bearing pale-coloured berries, are the result of deletion of two regulatory genes of the berry
562 colour locus. Plant Mol Biol. Kluwer Academic Publishers; 2006;62:623–35.

563 8. Yakushiji H, Kobayashi S, Goto-Yamamoto N, Tae Jeong S, Sueta T, Mitani N, *et al.* A skin
564 color mutation of grapevine, from black-skinned Pinot Noir to white-skinned Pinot Blanc, is
565 caused by deletion of the functional VvmybA1 allele. Biosci. Biotechnol. Biochem.
566 2006;70:1506–8.

567 9. Pelsy F, Dumas V, Bévilacqua L, Hocquigny S, Merdinoglu D. Chromosome Replacement
568 and Deletion Lead to Clonal Polymorphism of Berry Color in Grapevine. PLoS Genet. 2015;11.

569 10. Zhou Y, Minio A, Massonnet M, Solares E, Lyu Y, Beridze T, *et al.* Structural variants,
570 clonal propagation, and genome evolution in grapevine (*Vitis vinifera*). bioRxiv. 2018. pp. 1–48.

571 11. Fernandez L, Torregrosa L, Segura V, Bouquet A, Martínez-Zapater JM. Transposon-
572 induced gene activation as a mechanism generating cluster shape somatic variation in grapevine.
573 The Plant Journal. 2010;61:545–57.

574 12. Fernandez L, Chaïb J, Zapater JMM, Thomas MR, Torregrosa L. Mis□expression of a
575 PISTILLATA□like MADS box gene prevents fruit development in grapevine. The Plant
576 Journal. 2013;73:918–28.

577 13. Whitham TG, Slobodchikoff CN. Evolution by Individuals, Plant-Herbivore Interactions, and
578 Mosaics of Genetic Variability: The Adaptive Significance of Somatic Mutations in Plants.
579 Oecologia. 1981;49:287–92.

580 14. Soost RK, Cameron JW, Bitters WP, Platt RG. Citrus bud variation, old and new. Calif
581 Citrograph. 1961;46:188–93.

582 15. Farcuh M, Li B, Rivero RM, Shlizerman L, Sadka A, Blumwald E. Sugar metabolism
583 reprogramming in a non-climacteric bud mutant of a climacteric plum fruit during development
584 on the tree. Journal of Experimental Botany. 2017;68:5813–28.

585 16. McKey D, Elias M, Pujol B, Duputié A. The evolutionary ecology of clonally propagated
586 domesticated plants. New Phytologist. 2010;186:318–32.

587 17. Gambino G, Molin AD, Boccacci P, Minio A, Chitarra W, Avanzato CG, *et al.* Whole-
588 genome sequencing and SNV genotyping of "Nebbiolo" (*Vitis vinifera* L.) clones. Scientific
589 Reports. 2017;7:1–15.

590 18. Carrier G, Le Cunff L, Dereeper A, Legrand D, Sabot F, Bouchez O, *et al.* Transposable
591 Elements Are a Major Cause of Somatic Polymorphism in *Vitis vinifera* L. PLoS ONE. 2012;7.

592 19. Roach MJ, Johnson DL, Bohlmann J, van Vuuren HJJ, Jones SJM, Pretorius IS, *et al.*
593 Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar
594 Chardonnay. PLoS Genet. 2018;14.

595 20. Carbonell-Bejerano P, Royo C, Torres-Pérez R, Grimplet J, Fernandez L, Franco-Zorrilla
596 JM, *et al.* Catastrophic unbalanced genome rearrangements cause somatic loss of berry color in
597 grapevine. Plant Physiology. 2017;175:786–801.

598 21. Plomion C, Aury J-M, Amselem J, Leroy T, Murat F, Duplessis S, *et al.* Oak genome reveals
599 facets of long lifespan. Nature Plants. 2018;4:440–52.

600 22. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, *et al.*
601 The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana. Science.
602 2010;327:92–4.

603 23. Hershberg R, Petrov DA. Evidence That Mutation Is Universally Biased towards AT in
604 Bacteria. PLoS Genet. 2010;6.

605 24. Selker EU. Premeiotic instability of repeated sequences in Neurospora crassa. Annu. Rev.
606 Genet. 1990;24:579–613.

607 25. Meunier J, Khelifi A, Navratil V, Duret L. Homology-Dependent Methylation in Primate
608 Repetitive DNA. PNAS. 2005;102:5471–6.

609 26. Mautino MR, Rosa AL. Analysis of Models Involving Enzymatic Activities for the
610 Occurrence of C-T Transition Mutations During Repeat-Induced Point Mutation (RIP) in
611 Neurospora crassa. J theor Biol. 1998;192:61–71.

612 27. Schlötterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic Acids
613 Research. 1992;20:211–5.

614 28. Qi Y, He X, Wang X-J, Kohany O, Jurka J, Hannon GJ. Distinct catalytic and non-catalytic
615 roles of ARGONAUTE4 in RNA-directed DNA methylation. Nature. 2006;443:1008–12.

616 29. Shen H, He H, Li J, Chen W, Wang X, Guo L, *et al.* Genome-Wide Analysis of DNA
617 Methylation and Gene Expression Changes in Two Arabidopsis Ecotypes and Their Reciprocal
618 Hybrids. The Plant Cell. 2012;24:875–92.

619 30. Cantu D, Vanzetti LS, Sumner A, Dubcovsky M, Matvienko M, Distelfeld A, *et al.* Small
620 RNAs, DNA methylation and transposable elements in wheat. BMC Genomics. 2010;11.

621 31. Chan SW-L, Henderson IR, Jacobsen SE. Gardening the genome: DNA methylation in
622 Arabidopsis thaliana. Nat Rev Genet. 2005;6:351–60.

623 32. Thompson MM, Olmo HP. Cytohistological Studies of Cytochimeric and Tetraploid Grapes.
624 American Journal of Botany. 1963;50:901–6.

625 33. Hocquigny S, Pelsy F, Dumas V, Kindt S, Heloir M-C, Merdinoglu D. Diversification within
626 grapevine cultivars goes through chimeric states. Genome. 2004;47:579–89.

627 34. Klekowski EJ. Plant clonality, mutation, diplontic selection and mutational meltdown.
628 Biological Journal of the Linnean Society. 2003;79:61–7.

629 35. Klekowski EJ, Kazarinova-Fukshansky N, Mohr H. Shoot Apical Meristems and Mutation -
630 Stratified Meristems and Angiosperm Evolution. American Journal of Botany. 1985;72:1788–
631 800.

632 36. Tilney-Bassett RAE. Plant chimeras. Edward Arnold (Publishers) Ltd.; 1986.

633 37. Klekowski EJ. Mutation rates in mangroves and other plants. Genetica 1998;102/103:325–
634 31.

635 38. Muller HJ. Some genetic aspects of sex. The American Naturalist. 1932;66:118–38.

636 39. Pineda-Krch M, Fagerström T. On the potential for evolutionary change in meristematic cell
637 lineages through intraorganismal selection. Journal of Evolutionary Biology. 1999;12:681–8.

638 40. Orive ME. Somatic Mutations in Organisms with Complex Life Histories. Theoretical
639 Population Biology. 2001;59:235–49.

640 41. CDFA. *Grape Crush Report, Final 2016 Crop*. 2016;1-5.

641 42. CDFA. *California Grape Crush Report Preliminary 2015*. 2016;1-141.

642 43. Bowers JE, Bandman EB, Meredith CP. DNA Fingerprint Characterization of Some Wine
643 Grape Cultivars. AJEV. 1993;44:266–74.

644 44. Maletic E, Pejic I, Karoglan Kontic J, Piljac J, Dangl G, Vokurka A, *et al.* The Identification
645 of Zinfandel on the Dalmatian Coast of Croatia. Acta Hort. 2003;603:251–4.

646 45. Mirošević N, Meredith CP. A review of research and literature related to the origin and
647 identity of the cultivars Plavac mali, Zinfandel and Primitivo (*Vitis vinifera* L.). Acta Hort.
648 2000;65:45–9.

649 46. Maletic E, Pejic I, Kontic JK, Piljac J, Dangl GS, Vokurka A, *et al.* Zinfandel, Dobricic, and
650 Plavac mali: The genetic relationship among three cultivars of the Dalmatian Coast of Croatia.
651 AJEV. 2004;55:174–80.

652 47. Fanizza G, Lamaj F, Ricciardi L, Resta P, Savino V. Grapevine cvs Primitivo, Zinfandel and
653 Crljenak kastelanski: Molecular analysis by AFLP. Vitis. 2005;44:147–8.

654 48. Russo G, Liuzzi V, D'Andrea L, Alviti G. Comparison among Five Clones of “Primitivo”
655 Vine in Southern Italy. Hajdu E, Borbas E, editors. Acta Hort. 2003;603:779–86.

656 49. Wolpert JA. Performance of Zinfandel and Primitivo Clones in a Warm Climate. AJEV.
657 1996;47:124–6.

658 50. Fidelibus MW, Christensen LP, Katayama DG, Verdenal P-T. Performance of Zinfandel and
659 Primitivo grapevine selections in the central San Joaquin Valley, California. AJEV.
660 2005;56:284–6.

661 51. Zdunić G, Simon S, Malenica N, Budić-Leto I, Maletic E, Karoglan Kontić J, *et al.*
662 Intravarietal variability of Crljenak Kastelanski' and Its Relationship with 'Zinfandel' and
663 ‘Primitivo’ Selections. Acta Hort. 2014;1046:573–80.

664 52. Sweet NL, Wolpert JA. The Zinfandels of FPS. FPS Grape Program Newsletter. 2007;10–9.

665 53. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, *et al.* Phased
666 diploid genome assembly with single-molecule real-time sequencing. Nat Meth. 2016;13:1050–
667 4.

668 54. Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, *et al.* The grapevine
669 genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature.
670 2007;449:463–7.

671 55. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Haeseler A, *et al.* Accurate
672 detection of complex structural variations using single-molecule sequencing. Nat Meth.
673 2018;15:461–8.

674 56. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship
675 inference in genome-wide association studies. Bioinformatics. 2010;26:2867–73.

676 57. Boursiquot J-M, Lacombe T, Laucou V, Julliard S, Perrin FX, Lanier N, *et al.* Parentage of
677 Merlot and related winegrape cultivars of southwestern France: discovery of the missing link.
678 Australian Journal of Grape and Wine Research. 2009;15:144–55.

679 58. Bowers J, Boursiquot J-M, This P, Chu K, Johansson K, Meredith C. Historical genetics: The
680 parentage of chardonnay, gamay, and other wine grapes of northeastern France. Science.
681 1999;285:1562–5.

682 59. Regner F, Stadlbauer A, Eisenheld C, Kaserer H. Genetic Relationships Among Pinots and
683 Related Cultivars. AJEV. 2000;51:7–14.

684 60. Imazio S, Labra M, Grassi F, Winfield M, Bardini M, Scienza A. Molecular tools for clone
685 identification: the case of the grapevine cultivar “Traminer.” Plant Breeding. 2002;121:531–5.

686 61. Cingolani P, Platts A, Le Lily Wang, Coon M, Nguyen T, Wang L, *et al.* A program for
687 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly.
688 2012;6:80–92.

689 62. Gaut BS, Seymour DK, Liu Q, Zhou Y. Demography and its effects on genomic variation in
690 crop domestication. Nature Plants. 2018; doi:0.1038/s41477-018-0210-1

691 63. Chen RZ, Pettersson U, Beard C, Jackson-Grusby L, Jaenisch R. DNA hypomethylation
692 leads to elevated mutation rates. Nature. 1998;395:89–93.

693 64. Hirochika H, Okamoto H, Kakutani T. Silencing of Retrotransposons in Arabidopsis and
694 Reactivation by the ddm1 Mutation. Plant Cell. 2000;12:357–68.

695 65. Hirsch CD, Springer NM. Transposable element influences on gene expression in plants.
696 Biochimica et Biophysica Acta. 2017;1860:157–65.

697 66. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-Wide Evolutionary Analysis of
698 Eukaryotic DNA Methylation. Science. 2010;328:916–9.

699 67. Guarino F, Cicatelli A, Brundu G, Heinze B, Castiglione S. Epigenetic Diversity of Clonal
700 White Poplar (*Populus alba* L.) Populations: Could Methylation Support the Success of
701 Vegetative Reproduction Strategy? PLoS ONE. 2015;10:e0131480–20.

702 68. Douhovnikoff V, Dodd RS. Epigenetics: a potential mechanism for clonal plant success.
703 Plant Ecol. 2014;216:227–33.

704 69. Klekowski EJ Jr., Kazarinova-Fukshansky N. Shoot Apical Meristems and Mutation:
705 Selective Loss of Disadvantageous Cell Genotypes. American Journal of Botany. 1984;71:28–
706 34.

707 70. Thomas MR, Cain P, Scott NS. DNA typing of grapevines: a universal methodology and
708 database for describing cultivars and evaluating genetic relatedness. Plant Mol Biol.
709 1994;25:939–49.

710 71. Sefc KM, Regner F, Turetschek E, Glössl J, Steinkellner H. Identification of microsatellite
711 sequences in Vitis riparia and their applicability for genotyping of different *Vitis* species.
712 Genome. 1999;42:367–73.

713 72. Jones L, Riaz S, Morales-Cruz A, Amrine KCH, McGuire B, Gubler WD, *et al.* Adaptive
714 genomic structural variation in the grape powdery mildew pathogen, Erysiphe necator. BMC
715 Genomics. 2014; doi:10.1186/1471-2164-15-1081

716 73. Minio A, Lin J, Gaut BS, Cantu D. How Single Molecule Real-Time Sequencing and
717 Haplotype Phasing Have Enabled Reference-Grade Diploid Genome Assembly of Wine Grapes.
718 Front Plant Sci. 2017;8:481–6.

719 74. Myers G. Efficient Local Alignment Discovery amongst Noisy Long Reads. Wroclaw,
720 Poland: Springer, Berlin, Heidelberg; 2014. 52–67.

721 75. Smit A, Hubley R, Green P. *RepeatMasker Open-4.0*. 2013. Available from:
722 http://www.repeatmasker.org

723 76. Minio A, Massonnet M, Figueroa-Balderas R, Vondras AM, Blanco-Ulate B, Cantu D. Iso-
724 Seq Allows Genome-Independent Transcriptome Profiling of Grape Berry Development. G3:
725 Genes, Genomes, Genetics; 2019;9:755–67.

726 77. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of
727 RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11:1650–67.

728 78. Slater GSC, Birney E. Automated generation of heuristics for biological sequence
729 comparison. BMC Bioinformatics. 2005; doi:10.1186/1471-2105-6-31

730 79. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, *et al*. Improving
731 the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic
732 Acids Research. 2003;31:5654–66.

733 80. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004; doi:10.1186/1471-2105-
734 5-59

735 81. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio
736 prediction of alternative transcripts. Nucleic Acids Research. 2006;34:W435–9.

737 82. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in
738 novel eukaryotic genomes by self-training algorithm. Biological Journal of the Linnean Society.
739 2005;33:6494–506.

740 83. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, *et al.* Automated eukaryotic gene
741 structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.
742 Genome Biol. 2008;9.

743 84. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
744 Bioinformatics. 2009;25:1754–60.

745 85. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, *et*
746 *al*. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices
747 pipeline. Current Protocols in Bioinformatics. 2013.

748 86. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant
749 discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28:333–9.

750 87. Kofler R, Gómez-Sánchez D, Schlötterer C. PoPoolationTE2: Comparative Population
751 Genomics of Transposable Elements Using Pool-Seq. Mol Biol Evol. 2016;33:2759–64.

752 88. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, *et al*. A unified
753 classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.

754
755
756
757
758
759
760
761
762
763

764 **Tables**

765

Table 1. Clone identifying information

Clone #	Common name	Origin	Foundation Plant Services
1	Primitivo	Bari, Italy	Primitivo FPS 03
2	Primitivo	Conegliano, Italy	Primitivo FPS 06
4	Pribidrag	Svinšće, Croatia	Zinfandel FPS 43.1
5	Pribidrag	Svinšće, Croatia	Zinfandel FPS 44.1
6	Zinfandel	California, USA	Zinfandel FPS 10
7	Zinfandel	California, USA	Zinfandel FPS 24
8	Zinfandel	California, USA	Zinfandel FPS 37
9	Zinfandel	California, USA	Zinfandel FPS 39
10	Zinfandel	California, USA	Zinfandel FPS 56.1
11	Zinfandel	California, USA	Zinfandel FPS 40
12	Pribidrag	Marušići, Croatia	In testing at FPS
13	Pribidrag	Svinšće, Croatia	Zinfandel FPS 43.1
14	Crljenak kaštelanski	University of Zagreb, Croatia	-
15	Pribidrag	Svinšće, Croatia	Zinfandel FPS 44.1
Zin03	Zinfandel	California, USA	Zinfandel FPS 03

766

Table 2. Summary statistics of the Zinfandel genome assembly and annotation.

	Primary	Haplotig
Total length	591,171,721	306,029,957
Number of contigs	1,509	2,246
N50	1,062,797	442,393
N75	366,308	185,785
L50	154	200
L75	395	463
Median contig length (bp)	161,249	37,307
Longest contig (bp)	7,901,503	2,609,171
Shortest contig (bp)	17,787	1,970
Average GC content (%)	34.45%	34.37%
Number of genes	33,523	20,037
	Total	*Average per gene*
Number of exons	244,880	4.57
Number of introns	191,320	3.57
	Average (bp)	*Maximum (bp)*
mRNA lengths	4,166	94,143
Exon lengths	245.79	7,992
Intron lengths	191,320	41,647
Intergenic distances	10,309	302,473

Table 3. Sniffles analysis of structural variation between cultivars and between Zinfandel parental haplotypes

	Cabernet Sauvignon vs. Zinfandel					Zinfandel haplotig vs. Zinfandel primary				
	Median Size (bp)	*Count*	*Genes*	*Total SV size (Mb)*	*% genome*	*Median Size (bp)*	*Count*	*Genes*	*Total SV size (Mb)*	*% genom*
Deletions	196	46,363	9,219	115.0	12.82	203	12,031	2,521	26,953,558	4.56
Duplications	5,518	2,884	3,286	48.7	5.43	1,966	553	535	7,604,041	1.29
Insertions	88	37,407	5,225	23.9	2.66	92	9,647	2,081	5,594,259	0.95
Inversions	6,037	607	1,440	20.6	2.30	3,592	111	391	5,521,214	0.93
Duplicated Insertions	339	9	2	0.0439	0.0049	385	3	2	6,861	0.001
Inverted Duplications	293	65	12	0.0418	0.0047	113	54	11	12,930	0.002

Figure legends

Figure 1. Structural variation between Zin03 haplotypes. **a**. Distribution of structural variation sizes. Boxplots show the 25th quartile, median, and 75th quartile for each type of SV. Whiskers are 1.5$^{\text{Inter-Quartile Range}}$. Diamonds indicate the mean $\log_{10}$(length) of each type of SV; **b,c,d.** Examples of heterozygous structural variants between haplotypes that intersect genes. For each reported structural variation, (from top to bottom) the coverage, haplotype-resolved alignment of reads, and the genes annotated in the region are shown; **b.** 4 kbp heterozygous deletion of two genes; **c.** 11 kbp heterozygous deletion of two genes; **d.** 22 kbp inversion that intersects a single gene. Triangles indicate boundaries of the inversion. A gap is shown rather than the center of the inverted region.

788 **Figure 2.** Gene content and structural variability between Zin03 and other *V. vinifera* genomes.
789 **a.** Uniquely occurring Zinfandel genes and the number of Zinfandel genes that align well to other
790 cultivars with >=80% identity and reciprocal coverage. The total number of hits (or total gene
791 content for Zin03) is indicated by the "Set Size" and the exclusive hits for each intersection is
792 indicated as the "Intersection Size"; **b. Boxplot shows the sizes of structural variations; c,d.**
793 Selected deletions in Cabernet sauvignon relative to Zin03 that intersect genes. For each reported
794 deletion, (from top to bottom) the coverage of reads over the region by long Zinfandel and
795 Cabernet Sauvignon reads, haplotype-resolved alignment of the reads, and the genes annotated in
796 the region are shown; **b.** Two genes are completely deleted in Cabernet Sauvignon relative to
797 Zinfandel and are deleted in one Zinfandel haplotype; **c.** One gene contains a homozygous partial
798 deletion in Cabernet Sauvignon.
799
800 **Figure 3.** The relationships between Zinfandel selections. **a.** Principal component analysis of
801 Zinfandel selections based on SNP data. Zin03 was not included in the analysis; **b.** Kinship
802 analysis of Zinfandel selections and other cultivars with known relationships based on SNP data
803 and outside of annotated repeats. The Kinship coefficient, PHI, is shown, as well as a
804 dendrogram constructed by hierarchically clustering genotypes using their kinship coefficients.
805
806 **Figure 4**. Characterization of variants and their frequency among Zinfandel selections and other
807 *vinifera* cultivars (Pinot Noir, Chardonnay, Merlot, Cabernet Franc, and Sauvignon Blanc). The
808 normalized rate of variants (number of variants divided by the total feature length in the genome
809 * 1k) by type (SNP, INDEL), feature (Intergenic, Intron, Exon), and genotype (Non-Zinfandel
810 Cultivars, Zinfandel selections). Boxplots show the 25^{th} quartile, median, and 75^{th} quartile.

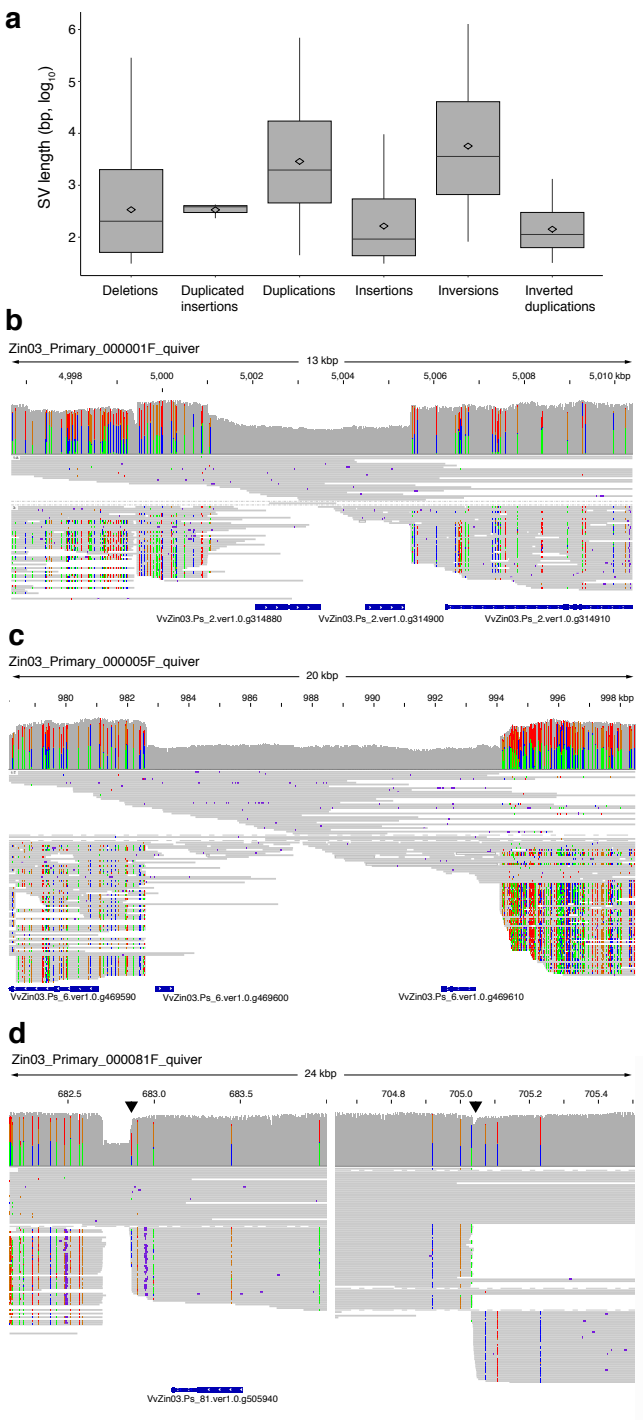
Figure 5. The abundance and impact of shared and unique heterozygous mutations among Zinfandel clones. **a.** The number of heterozygous SNPs, INDELs, and SVs are shared by N Zinfandel clones; **b.** The number of SNPs and INDELs shared by N clones in exons, introns, intergenic repeats ("Repeats"), and non-repetitive intergenic space; **c.** The ratio of transitions (Tr) to transversions (Tv) for heterozygous SNPs that uniquely occur in single Zinfandel clones and in different genome features. Different letters correspond to significant differences in Tr/Tv rates between features (ANOVA, Tukey HSD, $p < 0.01$); **d.** The percentage of CpG, CHG, and CHH in exons, introns, intergenic repeats ("Repeats"), and non-repetitive intergenic space that experiences transition mutations. Comparisons were made between features for each type of C-repeat separately. Different letters correspond to significant differences (Tukey HSD, $p < 0.01$); **e.** Proportion of exonic SNPs and INDELs that are deleterious and shared by N Zinfandel clones

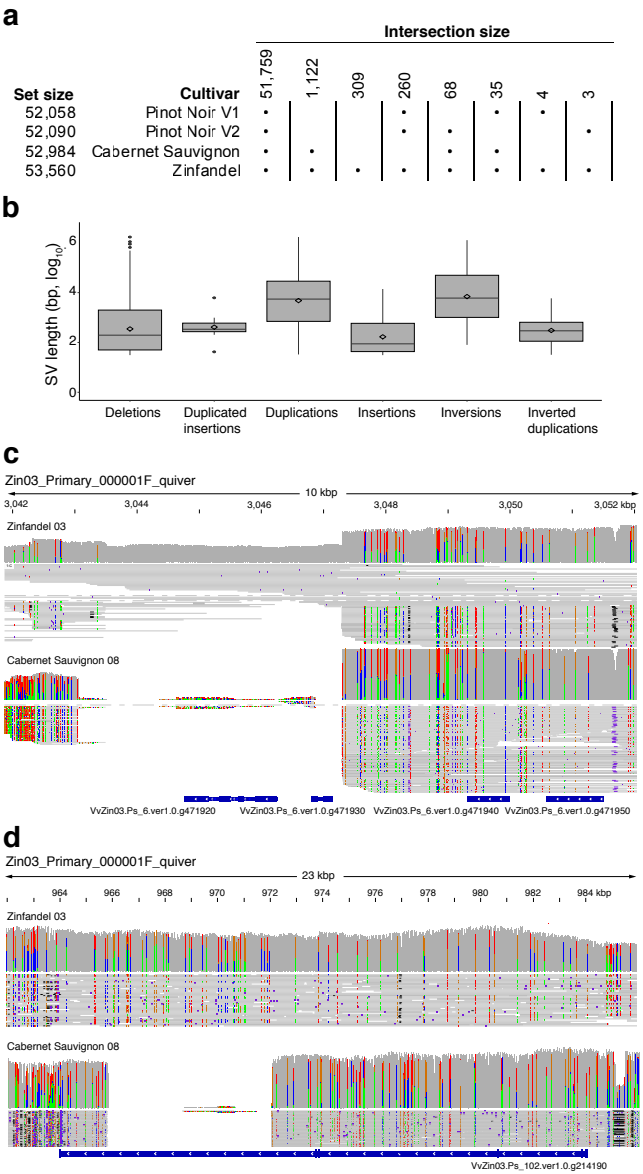
Figure 6. Transposable element insertions among Zinfandel selections. **a.** Transposable element insertions shared among N Zinfandel selections relative to Zin03 and PN40024; **b.** Types of transposable element insertions shared by N Zinfandel selections;**c.** The proximity of intergenic transposable element insertions to genes

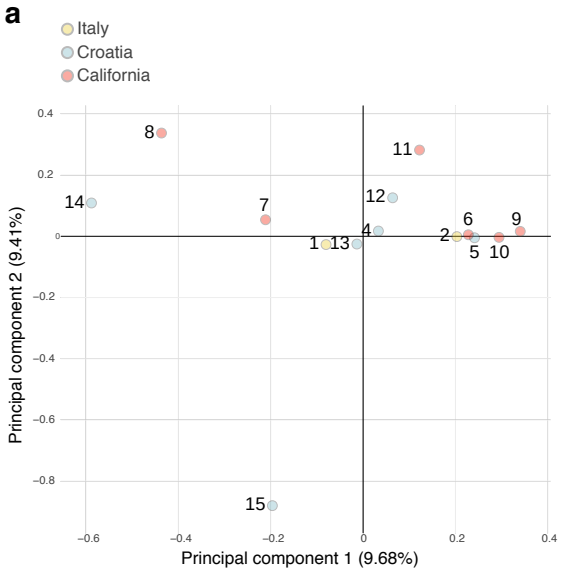
Additional files

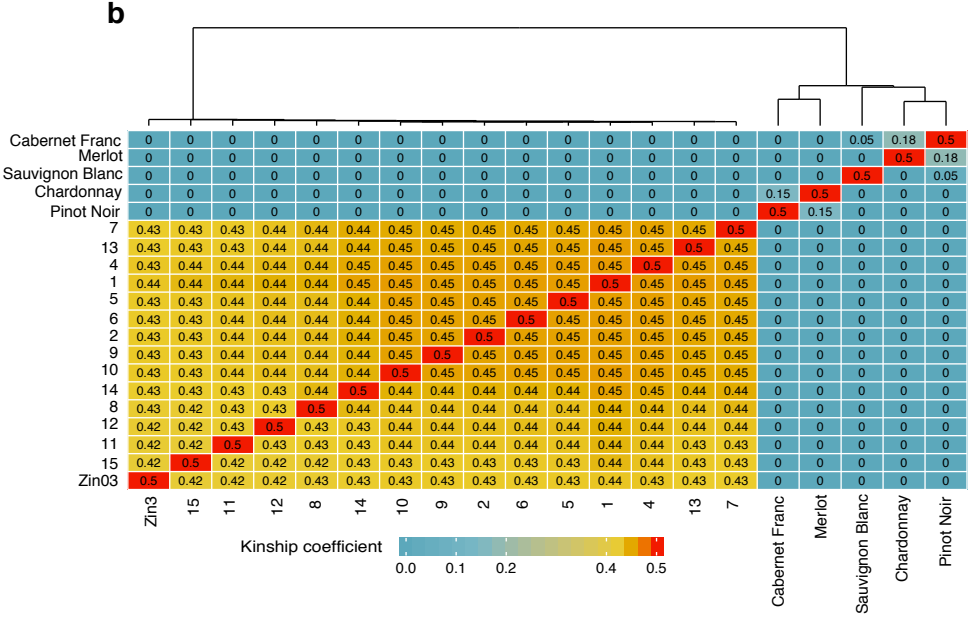
Additional file 1. .docx ; Method to extraction phenolic metabolites from Heritage Vineyard Zinfandel clones and discriminant analysis of Zinfandel clones based on their phenolic profiles.

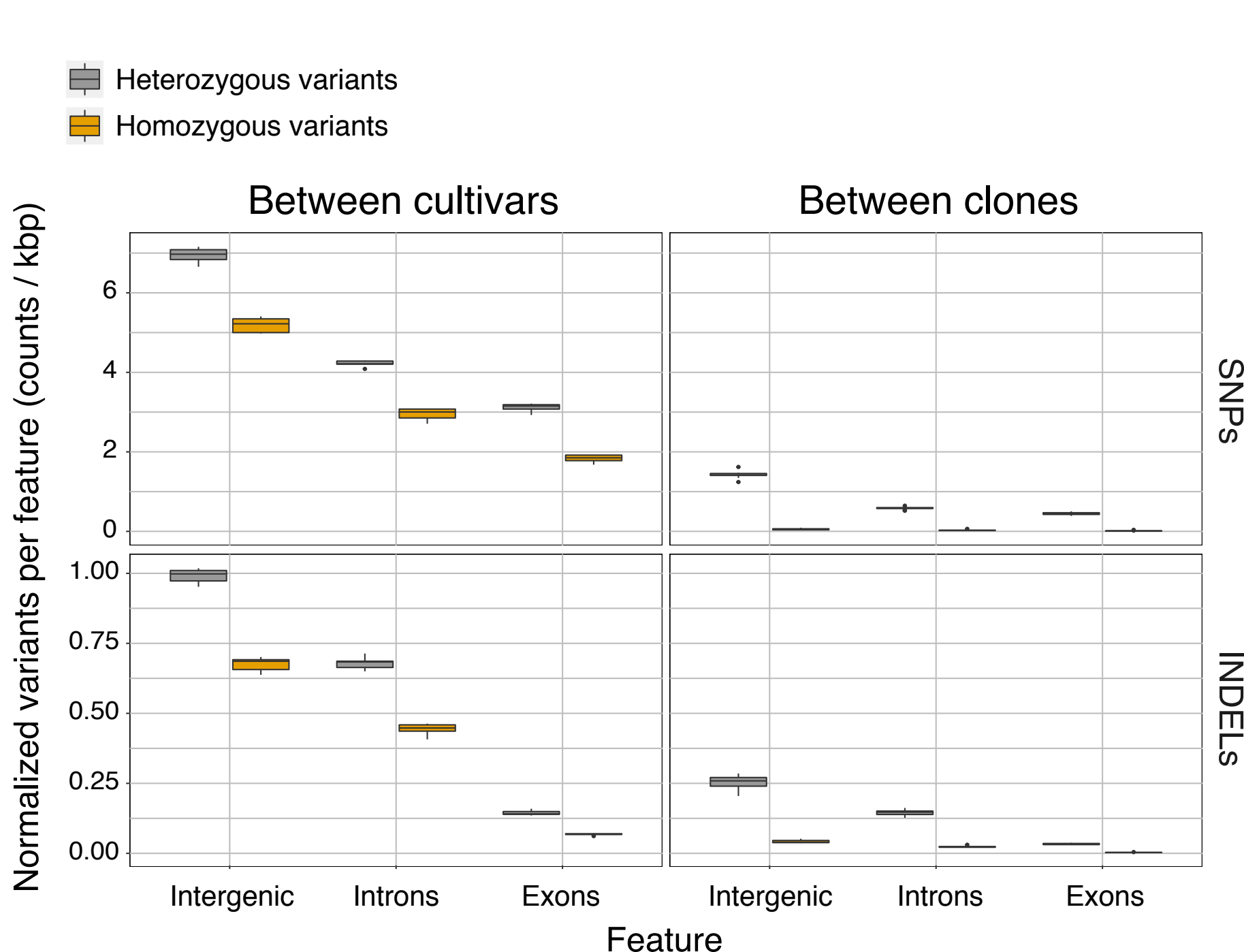
834 **Additional file 2.** .xlsx ; Unique genes identified in Zinfandel, not identified in Pinot Noir and
835 Cabernet Sauvignon (309), with associated Gene Ontology categories.
836
837 **Additional file 3.** .xlsx ; The first tab of this excel file is a summary of variants relative to the
838 Zinfandel reference genome and the second is a summary of the SnpEff analysis of variants, with
839 mean values ± SEM shown, and excluding sites where samples and Zin03 have identical
840 heterozygous genotypes at the locus.
841
842 **Additional file 4.** .txt ; Settings and data used for Zin03 genome assembly, annotation, and
843 variant calling.
844
845 **Additional file 5.** .sh ; Bioinformatic pipeline for SNP, INDEL, and SV calling.
846

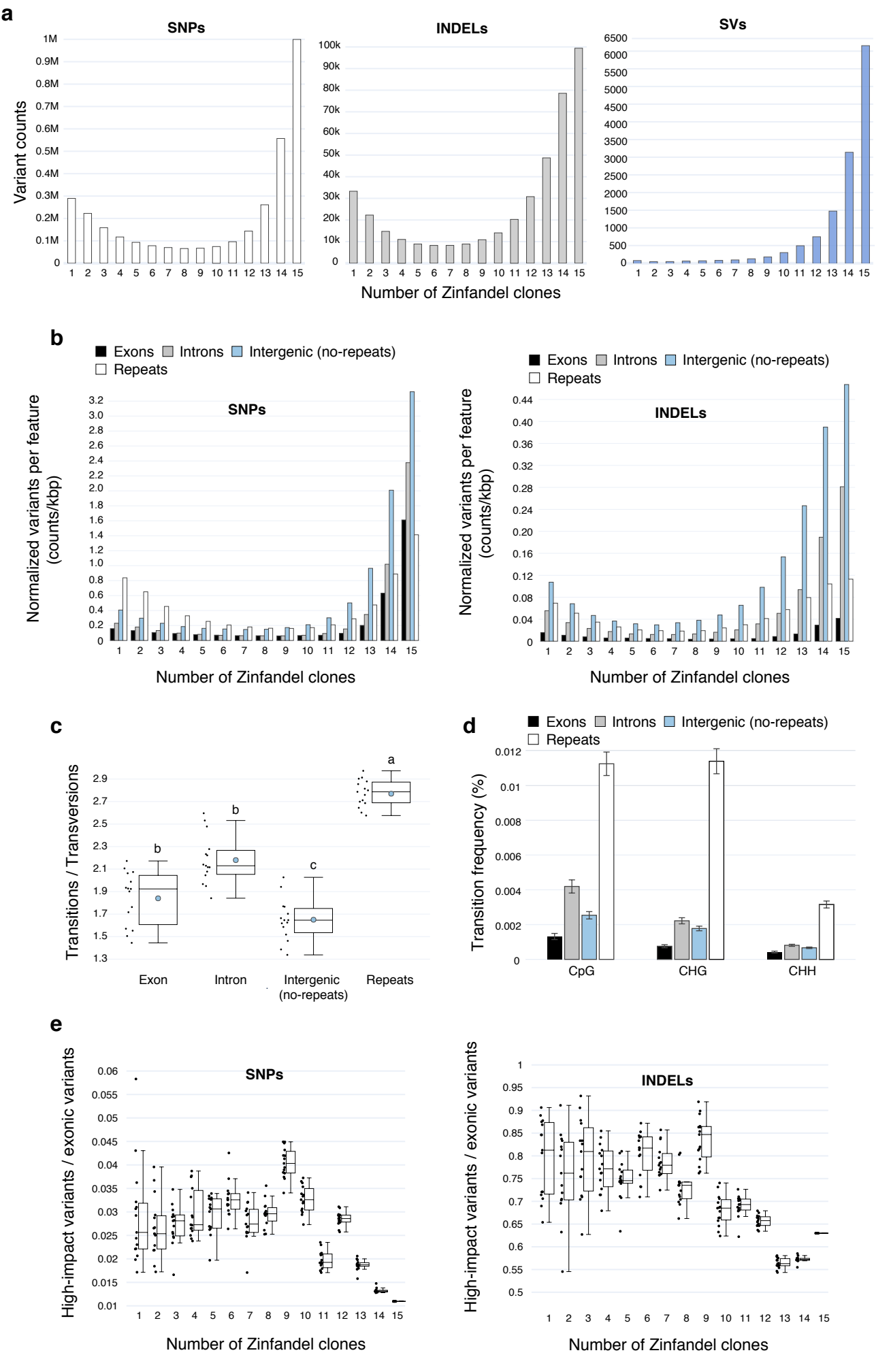

a
SV length (bp, log10)
6
5
4
3
2
Deletions
Duplicated insertions
Duplications
Insertions
Inversions
Inverted duplications
b
Zin03_Primary_000001F_quiver
13 kbp
4,998
5,000
5,002
5,004
5,006
5,008
5,010 kbp
VvZin03.Ps_2.ver1.0.g314880
VvZin03.Ps_2.ver1.0.g314900
VvZin03.Ps_2.ver1.0.g314910
c
Zin03_Primary_000005F_quiver
20 kbp
980
982
984
986
988
990
992
994
996
998 kbp
VvZin03.Ps_6.ver1.0.g469590
VvZin03.Ps_6.ver1.0.g469600
VvZin03.Ps_6.ver1.0.g469610
d
Zin03_Primary_000081F_quiver
24 kbp
682.5
683.0
683.5
704.8
705.0
705.2
705.4
VvZin03.Ps_81.ver1.0.g505940

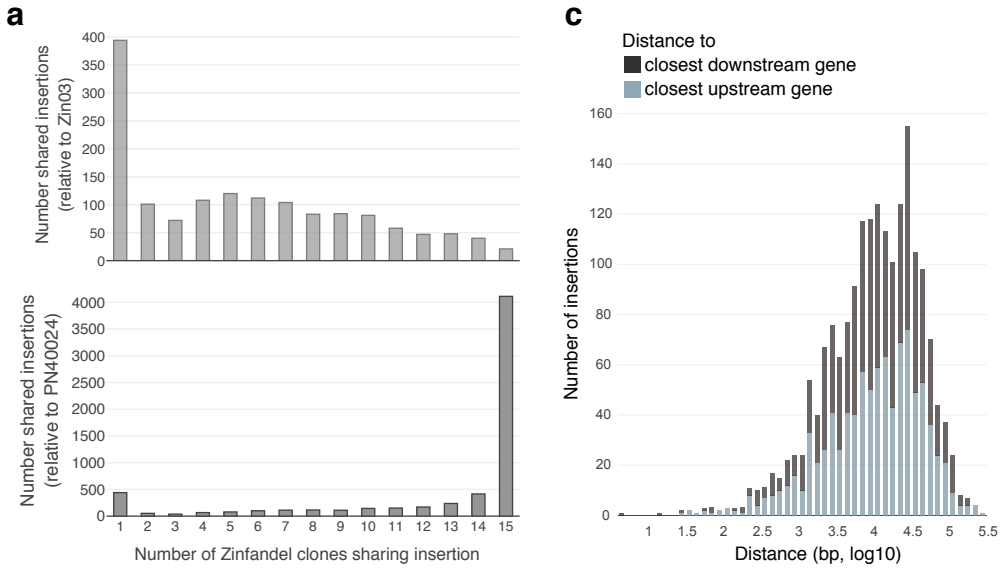

a
Intersection size
51,759
1,122
309
260
68
35
4
3
Set size
Cultivar
52,058 Pinot Noir V1
52,090 Pinot Noir V2
52,984 Cabernet Sauvignon
53,560 Zinfandel
b
SV length (bp, log10)
0
2
4
6
Deletions
Duplicated insertions
Duplications
Insertions
Inversions
Inverted duplications
c
Zin03_Primary_000001F_quiver
10 kbp
3,042
3,044
3,046
3,048
3,050
3,052 kbp
Zinfandel 03
Cabernet Sauvignon 08
VvZin03.Ps_6.ver1.0.g471920
VvZin03.Ps_6.ver1.0.g471930
VvZin03.Ps_6.ver1.0.g471940
VvZin03.Ps_6.ver1.0.g471950
d
Zin03_Primary_000001F_quiver
23 kbp
964
966
968
970
972
974
976
978
980
982
984 kbp
Zinfandel 03
Cabernet Sauvignon 08
VvZin03.Ps_102.ver1.0.g214190

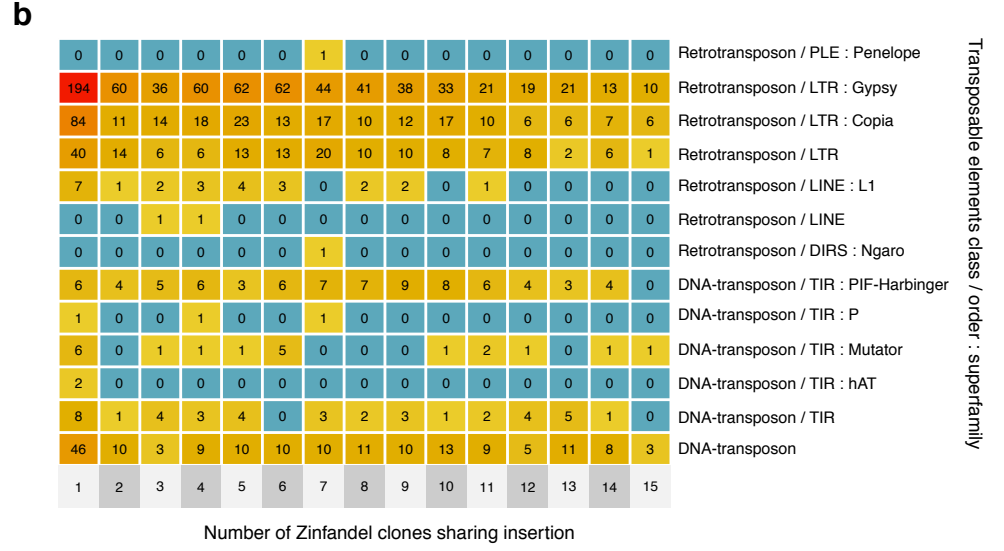

a
Italy
Croatia
California
Principal component 2 (9.41%)
Principal component 1 (9.68%)
8
11
14
7
12
6
9
4
2
1
13
5
10
15


b
Cabernet Franc
Merlot
Sauvignon Blanc
Chardonnay
Pinot Noir
7
13
4
1
5
6
2
9
10
14
8
12
11
15
Zin03
Zin3
15
11
12
8
14
10
9
2
6
5
1
4
13
7
Cabernet Franc
Merlot
Sauvignon Blanc
Chardonnay
Pinot Noir
Kinship coefficient
0.0
0.1
0.2
0.4
0.5


Heterozygous variants
Homozygous variants
Between cultivars
Between clones
SNPs
INDELs
Normalized variants per feature (counts / kbp)
0
2
4
6
0.00
0.25
0.50
0.75
1.00
Intergenic
Introns
Exons
Intergenic
Introns
Exons
Feature


a
SNPs
INDELs
SVs
Variant counts
Number of Zinfandel clones
b
Exons
Introns
Intergenic (no-repeats)
Repeats
SNPs
INDELs
Normalized variants per feature (counts/kbp)
Number of Zinfandel clones
c
Transitions / Transversions
Exon
Intron
Intergenic (no-repeats)
Repeats
d
Transition frequency (%)
CpG
CHG
CHH
e
SNPs
INDELs
High-impact variants / exonic variants
Number of Zinfandel clones


a
Number shared insertions (relative to Zin03)
400
350
300
250
200
150
100
50
0
Number shared insertions (relative to PN40024)
4000
3500
3000
2500
2000
1500
1000
500
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Number of Zinfandel clones sharing insertion
c
Distance to
closest downstream gene
closest upstream gene
Number of insertions
160
140
120
100
80
60
40
20
0
1 1.5 2 2.5 3 3.5 4 4.5 5 5.5
Distance (bp, log10)


b
0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 Retrotransposon / PLE : Penelope
194 60 36 60 62 62 44 41 38 33 21 19 21 13 10 Retrotransposon / LTR : Gypsy
84 11 14 18 23 13 17 10 12 17 10 6 6 7 6 Retrotransposon / LTR : Copia
40 14 6 6 13 13 20 10 10 8 7 8 2 6 1 Retrotransposon / LTR
7 1 2 3 4 3 0 2 2 0 1 0 0 0 0 Retrotransposon / LINE : L1
0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 Retrotransposon / LINE
0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 Retrotransposon / DIRS : Ngaro
6 4 5 6 3 6 7 7 9 8 6 4 3 4 0 DNA-transposon / TIR : PIF-Harbinger
1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 DNA-transposon / TIR : P
6 0 1 1 1 5 0 0 0 1 2 1 0 1 1 DNA-transposon / TIR : Mutator
2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 DNA-transposon / TIR : hAT
8 1 4 3 4 0 3 2 3 1 2 4 5 1 0 DNA-transposon / TIR
46 10 3 9 10 10 10 11 10 13 9 5 11 8 3 DNA-transposon
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Number of Zinfandel clones sharing insertion
Transposable elements class / order : superfamily