The genomic and epigenomic dynamics of hemizygous genes across crops with contrasting mating systems

Yanling Peng[1], Yiwen Wang[1], Yuting Liu[1], Xinyue Fang[1], Lin Cheng[1], Qiming Long[1], Qi Xu[1], Nan Wang[1], Fan Zhang[1], Zhongjie Liu[1], Hua Xiao[1], Wei Hu[2], Songbi Chen[2], Sanwen Huang[1, 2], Brandon S. Gaut[3], Yongfeng Zhou[1, 2, *]

[1]National Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

[2]National Key Laboratory of Tropical Crop Breeding, Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou, China

[3]Department of Ecology and Evolutionary Biology, University of California, UC Irvine, CA, USA.

***Corresponding author(s): Yongfeng Zhou**

E-mail(s): zhouyongfeng@caas.cn

Author Contributions: Y.Z., B.S.G., and S.H. designed the research. Y.P., B.S.G., and Y.Z. wrote the manuscript. Y.P. performed the analyses. Y.W., L.C., Q.X., N.W., F.Z., Z.L., H.X., B.S.G., Y.Z., W.H., S.C., and S.H. revised the paper. Y.L., X.F., Q.L., and Y.P collected the data.

Competing Interest Statement: The authors declare no competing interests.

Abstract

Hemizygous genes are present in one of the two sister chromatids of diploid organisms. It comes to be known for their prevalent occurrence and vital roles in sex chromosome. However, hemizygous genes in genomes of diploid plants remain largely unexplored. In this study, we investigated the features, genetic, cis- and epigenetic regulations of hemizygous genes in seven crops. These crops represent three clonal lineages, one outcrossing species, and three putatively homozygous (selfed or doubled haploid) genomes. By remapping long reads to the primary genome assembly, we identified structural variants that included annotated genes. We found 3,399-5,610 hemizygous genes (10.1%-15.1%) in the three clonal plants. As expected, very few genes (0.003%-0.007%) were hemizygous in the three homozygous genomes, representing negative controls. The genome from an outcrossing species was intermediate between the two extremes. Hemizygous genes experienced a more recent origin and stronger selection pressure than diploid genes. We also found reduced expression of hemizygous genes compared to diploid genes, with ~20% expression levels on average, which violated the evolutionary model of dosage compensation. Furthermore, we detected higher DNA methylation levels on average in hemizygous genes and transposable elements, which may contribute to the reduced hemizygous gene expression levels. Finally, expression profiles showed that hemizygous genes were more tissue/treatment-specific expressed than diploid genes in fruit development, organ differentiation, and responses to abiotic and biotic stresses. Overall, hemizygous genes displayed distinct genomic, genetic and epigenomic features compared to diploid genes, providing new insights for the genetics and breeding of crops with heterozygous genomes.

Keywords: Clonal propagation, Structural variation, Integrative genomics, Heterozygous, Cis-regulation

Introduction

Hemizygous genes are present on only one chromatid of a diploid organism (1, 2). The most prominent examples of hemizygous genes are on the sex chromosomes of X/Y mammalian males or Z/W avian females (2, 3). Similar sets of hemizygous genes are present in sex-linked regions of X/Y dioecious plants, such as palms (4, 5) and grapevines (6-8), asparagus (9, 10), kiwifruit (11, 12) and strawberries (13). Numerous studies have focused on the evolution, gene expression and epigenetic regulation of sex-linked hemizygous genes compared to diploid genes (14-21). For example, genomic studies in mammalian males have consistently revealed lower mutation rates and more efficient selection in sex-linked hemizygous genes than diploid genes (14), due in part to the fact that hemizygosity uncovers recessive alleles and makes them visible to selection (22). Similar studies have generally shown that the ratio of sex-linked to diploid (X:AA) gene expression is ~0.5 in animals and plants (16, 23). A X:AA ratio of ~0.5 ratio is inconsistent with the hypothesis of dosage compensation, a mechanism that re-equalizes male and female expression and/or brings XY male expression back to its ancestral level (23). Interestingly, estimated expression levels of sex chromosome alleles (XY) in males show an overall trend of reduced expression of Y-linked alleles relative to X-linked alleles in both animals and plants (17-20). Some of these expression effects are mediated by epigenetic marks, such as histone markers and DNA methylation(18, 24-27). For example, the male-specific region of the papaya Y chromosome is associated with knob-like heterochromatic structures that are heavily methylated, which suggests that DNA methylation has played a role in the evolution of this Y chromosome (24). These observations indicate that sex-linked hemizygous genes often have unique epigenetic and regulatory features. In addition to sex-linked regions, it is worth noting that the absence of one paired allele is frequently observed in non-sex-linked regions of homologous chromosome in diploid plants, resulting in a significant presence of hemizygous genes (28). Yet, the extent and function of these genes remain largely uncharacterized. Here, we study hemizygous genes across a sample of plant genomes.

The identification of hemizygous genes in diploid plant genomes has become feasible with the emergence of long-read sequencing technologies and the advance of assembly

algorithms. Precise genome assemblies facilitate the identification of structural variants (SVs) in heterozygous diploid plant genomes (28-30), and thus permit genome-wide identification of hemizygous genes caused by SVs (28-30). For example, by remapping long-reads to a reference genome assembly, it has been inferred that ~13.5% and ~15% of genes are hemizygous in two clonal grapevine (*Vitis vinifera* ssp. *vinifera*) cultivars (28). While this high value may reflect, in part, unique features of long-term clonal lineages (28, 31), hemizygosity is not limited solely to clonal lineages, because ~8.89% and ~4% of genes are estimated to be hemizygous in an outcrossing wild rice species (*Oryza longistaminata*) and in avocado (*Persea americana*), respectively (32, 33). In contrast, as expected, only a few genes have been inferred to be hemizygous in inbred, selfed accessions. For example, only 0.73% and 0.35% of genes have been inferred to be hemizygous in rice cultivars Nipponbare and 93-11, respectively (33). The extent of hemizygosity in plant genomes has only begun to be appreciated, largely because genome projects have historically focused on selfed or homozygous materials (28). As a result, there is currently little information about natural variation in the number of hemizygous genes, about potential correlations between hemizygosity and life history traits (such as mating systems and historical population sizes), and about the evolutionary dynamics and putative functions of hemizygous genes. We do know, however, that hemizygosity can affect function. For example, white berry color in grapevines is related to a complex series of mutations that includes hemizygosity of a large genomic region (34). In this case, the key feature of hemizygosity is that it uncovers a recessive, non-functional allele, that interrupts anthocyanin biosynthesis.

Thus far, the extent of hemizygosity has been estimated in only a handful of plant genomes, and there have been no accompanying genome-wide analyses of hemizygous gene function and epigenetics. In this study, we build or gather chromosome-level assemblies for seven plant genomes and then identify SVs that define hemizygous genes. We have chosen seven samples that represent a range of species' mating systems and that also have available gene expression and epigenetic data. Our sample includes three clonal plants: two grapevine clones (Chardonnay and Pinot Noir) and one cassava (*Manihot esculenta*) accession (35). Since extensive hemizygosity may be elevated in clonal lineages, we have included comparatives from one outcrossing sample, Black Cottonwood (*Populus trichocarpa*) (36),

and three inbred/doubled haploid cultivars, including rice (*Oryza sativa* ssp. *indica*) (37), tomato (*Solanum lycopersicum*) (30) and sweet orange (*Citrus sinensis*) (38). To compare results across samples, we have assembled Pacific Biosciences (PacBio) long-read sequencing data for each genome and identified SVs that defined hemizygous genes.

Given the identification of hemizygous genes and the availability of transcriptomic and epigenomic data, we ask the following questions: (1) Are hemizygous genes widespread in diploid plant genomes, or are they particularly abundant in clonal lineages? (2) Do hemizygous genes have distinct sequences and evolutionary features compared to diploid genes? For example, are they enriched for specific biological processes? If they are, do they have distinct expression levels and patterns compared to diploid genes? (3) Hemizygous regions can include genes but also other sequence features, like transposable elements (TEs). How extensive are hemizygous TEs, and do they have detectable correlates with the expression of nearby diploid genes? And, finally, (4) Do hemizygous genes exhibit distinct epigenetic patterns relative to diploid genes? If they do, how is expression related to these epigenetic effects? By addressing these questions, our goal is to further understand the evolutionary and functional consequences of hemizygosity. Ultimately this knowledge will be beneficial for understanding the genetics, breeding and evolution of plants with heterozygous genomes.

Results

Prevalent hemizygous genes in clonal plant genomes

To identify hemizygous genes caused by SVs, we first built or gathered chromosome-level genome assemblies based on PacBio circular consensus sequencing (CCS) or continuous long read (CLR) reads for three clonal, one outcrossing, and three inbred/doubled haploid plant samples. The three clonal plants included the Pinot Noir and Chardonnay cultivars of grapevine (28) and the cassava variety TME204 (35). The outcrossing sample was from Black Cottonwood (36). The remaining three samples were either inbred of doubled haploids, and so should be nearly completely homozygous. The three included inbred rice cultivar MH63 (37), the genetically manipulated inbred tomato cultivar Heinz 1706 (30), and one manipulated doubled haploid sweet orange cultivar Valencia (38). Among these

seven genome assemblies, the Pinot Noir data were generated and assembled based on long PacBio CCS, ONT, and Hi-C reads for this study, while data and reference assemblies for the other samples were retrieved from public resources. The new Pinot Noir genome assembly was highly contiguous, with a scaffold N50 value of 27.1 Mb and a Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness score of 98.3%, and anchored to 19 chromosomes (Table 1). The genome assemblies for the additional species were built based on PacBio CCS/CLR reads, anchored to chromosome level, had scaffold N50 values of 13.2-67.6 Mb and BUSCO completeness scores of 96.4%-99.0% (Table 1).

Given these high-quality references, we identified hemizygous regions by remapping long-read PacBio data to genome assemblies, focusing on SVs >50 bp. The SVs were identified using the Sniffles pipeline (39), followed by several filtering steps, including thresholds for quality and coverage (see Methods). In addition, the raw data were downsampled to ~30× coverage before mapping to facilitate fair comparisons across samples (*SI Appendix* Fig. S1). This approach yielded from 20,822 to 46,418 heterozygous SVs (hSVs) in the three clonal samples. Of these, more than one-third (i.e, 10,165-29,614, 48.8%-63.8% of total hSVs) were heterozygous deletions (hDELs) relative to the reference (Table 1). The hDEL SV class was the focus for our studies, because for these SV events the reference contains information about content within the SV.

Focusing on hDELs as hemizygous regions, we found that they were extensive within clonal plants. In the three clonal genomes, hDELs encompassed from 83.7 to 90.1 Mb in length, corresponding to 11.5% to 16.9% of the total genome size (Table 1). Notably, the number of hSVs in cassava was demonstrably higher than the two grapevine varietals, likely reflecting a larger genome (~770Mb vs. ~500 Mb), features of the mating system and life history (such as the duration of the clonal lineage) or some combination. As a comparison, we identified SVs in outcrossing Black Cottonwood and the inbred samples. Black Cottonwood had 17,708 hSVs and 9,949 hDELs, with the latter representing 6.4% of the ~400Mb genome. In contrast, as we expected, the three homozygous (inbred or doubled haploid) samples contained far less evidence for hSVs. For example, the *indica* rice accession housed 133 hSVs and only 18 hDELs; similarly, the tomato and orange

samples had 28 and 191 hSVs with 10 and 52 hDELs. The total length of hDELs were negligible, ranging from 7 kb to 42 kb, representing only 0.001% to 0.013% of the total genome size (Table 1). The low values for the inbred and doubled haploid samples suggest that our methods have low false positive errors – i.e., representing < 0.013% of the genome.

We characterized heterozygous deletions further as to whether they contained complete genes or TEs, which we defined as hemizygous genes or TEs (Fig. 1A). In the clonal and outcrossing samples, we detected from 3,399 to 5,610 hemizygous genes (representing from 10.1% to 15.1% of total genes), with Chardonnay exhibiting the highest proportion (Table 1). These genomes also contained numerous (from 119,980 to 226,295) hemizygous TEs, representing 13.0% to 24.7% of annotated TEs (Table 1), with Chardonnay again exhibiting the highest proportion. In contrast to the clonal samples, the Black Cottonwood genome was less replete with hemizygous genes and TEs, with 1,570 hemizygous genes representing 4.5% of total genes and 39,773 hemizygous TEs (6.3% of total TEs) (Table 1). As expected, the three homozygous samples had far fewer hemizygous genes and TEs; each sample had one or two hemizygous genes (i.e., 0.003%-0.007% of annotated genes) with 8 to 46 hemizygous TEs (0.001%-0.01% of total TEs). These patterns mirrored previous findings that estimated that: i) ~15% and 13.5% of genes are hemizygous in clonal grapevine cultivars (28, 31); ii) there is a lower but notable percentage of hemizygous genes in outcrossing plants (i.e., 8.89% hemizygous genes in *O. longistaminata* (33), 4% in avocado (32), and now 4.5% in Black Cottonwood) and iii) as expected, there are consistently low rates of genic hemizygosity (<1%) in putatively homozygous materials (33). Overall, these results are consistent with the accumulation of hemizygous regions especially in clonal lineages.

We anticipate that our observations represent *bona fide* SVs, but for downstream characterization we thought it was critical to define a subset of genes with additional evidence for hemizygosity. We opted to focus on heterozygosity as an additional filter. Since hemizygous genes occur in one only sister chromatid, truly hemizygous genes should lack heterozygosity entirely. It is, however, difficult to give a proper cutoff value for heterozygosity, because sequencing error can contribute to a low level of apparent heterozygosity. In this case, we opted to use the conservative heterozygosity value of zero

to define reliably hemizygous genes. This additional filter reduced the hemizygous gene set to 1,403 from 3,399 genes in Pinot Noir, to 3,007 from 5,610 in Chardonnay, to 1,598 from 4,242 in cassava and to 309 out of 1,570 genes in Black Cottonwood. We focused on these genic sets for downstream analyses, and contrasted them to diploid genes – i.e., genes without any evidence of an overlapping hSV. We did not include the inbred/doubled haploid samples in subsequent analyses, because they contain so few hemizygous genes.

Given this heavily filtered set of hemizygous genes, we examined their length and coverage statistics. As expected, the average length of hemizygous genes was significantly shorter than that of diploid genes in the clonal samples (p < 2e-16, Fig. 1B), because one naively expects that shorter genes have a higher probability of being encompassed by an SV event and of lacking heterozygosity. Also as expected, the coverage of hemizygous genes was significantly lower than that of diploid genes (Wilcoxon rank-sum test, p < 2e-16, or p < 0.01; Fig. 1C). Altogether, the coverage and length of genes tend to confirm the hemizygosity of our gene sets.

Evolutionary and functional properties of hemizygous genes

We explored evolutionary and functional features of putatively hemizygous genes relative to diploid genes for the four samples with substantive hemizygosity. For example, we measured nonsynonymous (*Ka*) and synonymous (*Ks*) substitution rates for each gene by comparing sequences to a paired outgroup (e.g., *Muscadinia rotundifolia* for Chardonnay and Pinot Noir; *Ricinus communis* for cassava; and *Salix brachista* for Black Cottonwood). The median *Ks* value in hemizygous genes was significantly lower than diploid genes in Pinot Noir and cassava (Wilcoxon rank-sum test, p < 0.05); similar patterns were found in Chardonnay and Black Cottonwood but without significant values (Fig. 1D). Perhaps as a consequence of lower *Ks*, hemizygous genes have correspondingly higher median *Ka/Ks* values relative to diploid genes in Pinot Noir, Chardonnay and cassava (Wilcoxon rank-sum test, p < 0.05; Fig. 1E); similar patterns were found in Chardonnay and Black Cottonwood but without significant values (Fig. 1E). These *Ka/Ks* patterns are consistent with relaxed selection, lower mutation rates (as implied by lower *Ks* values) or some combination of these two processes in hemizygous genes relative to diploid genes.

We also investigated the proportion of single copy and multi-copy genes in both hemizygous genes and diploid genes. We hypothesized that hemizygous genes were more likely to belong to a multi-gene family, because gene family membership can provide functional redundancies that make hemizygosity potentially less detrimental. To measure membership in gene families, we implemented a BLASTP-based pipeline (see Methods). Our results supported our hypothesis, because we detected a lower proportion of single copy genes in hemizygous genes compared to diploid genes in all four plants (38.1% vs. 72.7% in Pinot Noir, p = 1.0e-06; 70.4% vs. 81.7% in Chardonnay, p = 0.07; 34.1% vs. 57.2% in cassava, p = 0.002; 44% vs. 72.8% in Black Cottonwood, p = 5.2e-05; Fisher's Exact Test) (Fig. 1F). We also applied the Orthofinder pipeline and found similar results – i.e., a lower proportion of single copy genes in hemizygous genes (2.9% vs. 32.7% in Pinot Noir, p = 1.6e-08; 3.8% vs. 27.3% in Chardonnay, p = 7.5e-06; 15.7% vs. 25.8% in cassava, p = 0.11; 14.2% vs. 24.8% in Black Cottonwood, p = 0.07; Fisher's Exact Test) (Fig. 1G). These results support the idea that hemizygous genes are more often members of multi-gene families.

Finally, we investigated the possible biological processes of hemizygous genes in the four plants (*SI Appendix* Fig. S2). In Pinot Noir, the top 10 enriched Gene Ontology (GO) terms were involved in biological processes such as regulation of reproductive processes, mitotic related processes, and responses to stimuli. In Chardonnay, the top 10 enriched GO terms were related to detection of other organisms, biotic stimuli and bacteria. In cassava, the top 10 enriched GO terms were related to responses to stress and environmental stimuli. In Black Cottonwood, the top 10 enriched GO terms were involved in regulation of signaling, regulation of cell communication, and response to stimulus. The results suggest that hemizygous genes can be involved in fundamental processes like reproduction and mitosis, but they were also consistently enriched for responses to biotic and abiotic stress.

Unique expression patterns of hemizygous genes

To explore the expression patterns of hemizygous and diploid genes, we amassed RNA-seq datasets from 60 samples belonging to 21 different tissues/treatments in Pinot Noir and 26 samples belonging to 10 different tissues/treatments in Chardonnay. The samples were

generated across different tissues, developmental stages (e.g., fruit development) and experimental regimes, such as stress treatments (*SI Appendix* Table S1). We also retrieved three leaf samples from one experiment without treatment in cassava, and three leaf samples from one experiment without treatment and nine leaf samples from abiotic stress experiments in Black Cottonwood (*SI Appendix* Table S1). In total, we recovered 313 Gb paired-end and 74 Gb single-end Illumina reads. Our goals with these data were: i) to investigate the level of expression in hemizygous genes relative to diploid genes and ii) to determine whether hemizygous genes had patterns of expression consistent with contributions to development and other processes.

We first assessed whether hemizygous genes were expressed. Across taxa and individual RNA-seq samples, a significantly higher proportion of hemizygous genes had no evidence for expression relative to diploid genes (Fig. 2A and *SI Appendix* Fig. S3). For example, across all samples in Pinot Noir, 61.2% (858 out of 1,403 genes) of hemizygous genes had evidence for expression, but that proportion was 89.9% (27,338 out of 30,403 genes) for diploid genes (Chi-sq=1099.9, df=1; $p<0.001$). This trend held true in each tissue/treatment for all taxa, such that of the total number of 21 RNA-seq tissues/treatments in Pinot Noir, hemizygous genes were expressed in lower proportions for all tissues/treatments. These results strongly suggest that hemizygous genes are enriched for pseudogenes or for a subset of genes that are expressed under fewer experimental and developmental conditions. On the flip-side, however, not all hemizygous genes were pseudogenes; across all data samples, we detected expression for 61.2%, 81.3%, 52.7% and 76.4% of hemizygous genes in Pinot Noir, Chardonnay, cassava, and Black Cottonwood, respectively (Fig. 2A).

For each RNA-seq sample, we next investigated average levels of expression for genes with evidence for expression. Hemizygous genes were consistently expressed at significantly lower levels than diploid genes based on average expression across all tissues/treatments (Fig. 2B) and within each tissue/treatment (*SI Appendix* Fig. S4). The hemizygous: diploid ratio of average expression ranged from 0.03 to 0.24 across Pinot Noir tissues/treatments, with median of 0.05 (Fig. 2C). These values were higher in Chardonnay (ranging from 0.18 to 0.31 with a median of 0.21). The hemizygous: diploid expression ratio was similar in Black Cottonwood, ranging from 0.19 to 0.34 with a median of 0.27.

All of these values are substantially lower than the 50% expected if hemizygous alleles were expressed at similar levels to diploid alleles. These results imply that there is a diminution of gene expression associated with hemizygosity, and this diminution typically results in < 50% of the expression of diploid genes, on average (see Discussion).

We then elucidated patterns of hemizygous gene expression across scenarios like fruit development, organ differentiation, and biotic and abiotic stress stimuli. First, we estimated whether there were significant expression changes for hemizygous and diploid genes between contrasting treatments (or control vs. treatment) within a single study. For example, for 12 paired comparisons in Pinot Noir, we found that the level of median hemizygous gene changed significantly for 10 (83.2%, $p < 0.05$, Wilcoxon test for paired samples) of contrasts (*SI Appendix* Fig. S5). Similarly, median diploid gene expression changed significantly in 10 out of 12 (83.2%, $p < 0.05$, Wilcoxon test for paired samples) paired comparisons (*SI Appendix* Fig. S5). Similar pattern was found in Chardonnay (*SI Appendix* Fig. S6). These observations suggest that hemizygous genes are not generally dissimilar to diploid genes with respect to differential expression across treatments and conditions. Second, we also assessed the proportion of hemizygous genes that were differentially expressed across comparisons. Hemizygous genes generally had a lower proportion of differentially expressed genes than diploid genes (Fig. 2D and *SI Appendix* Fig. S7). For example, in Pinot Noir, 19.1% (268 out of 1,403) of hemizygous genes and 72.4% (22,024 out of 30,403) of diploid genes were differentially expressed in at least one paired comparison (Chi-sq=1817.4, df=1; p<0.001). The corresponding values were 35.2% (1,057 out of 3,007) versus 54.4% (17,222 out of 31,634) for Chardonnay (Chi-sq=409.2, df=1; p<0.001) and 37.5% (116 out of 309) versus 55.6% (18,425 out of 33,130) (Chi-sq=39.754, df=1; p<0.001) in Black Cottonwood. These results indicate that statistically fewer expressed hemizygous genes responded to differences in fruit development, organ differentiation and abiotic and biotic stresses.

To further explore the expression pattern of hemizygous genes, we detected genes that were expressed in specific tissues or treatments. That is, we counted the number of genes that had significant evidence for expression in only one tissue/treatment of a paired comparison. Altogether, hemizygous genes had a higher proportion of tissue/treatment-specific genes

than the diploid genes (*SI Appendix* Fig. S8). Across 12 tissue/treatment comparisons in Pinot Noir, 29.3% to 79.5% of hemizygous genes were expressed in only one of the paired tissues or treatments, but these values were substantially lower for diploid genes (ranging from a low of 7.5% to a high of 25.6%). Similar patterns – i.e., more tissue/treatment-specific expression for hemizygous genes – were found in Chardonnay and Black Cottonwood.

We also classified pairwise comparisons of Pinot Noir and Chardonnay into three categories: fruit development, organ differentiation and abiotic and biotic stresses experiments (Fig. 2E). We then analyzed the overlap of differentially expressed genes among these three categories. For example, of 268 differentially expressed hemizygous genes in Pinot Noir, 19 (7.1%), 46 (17.2%), and 94 (35.1%) were found only in fruit development, in organ differentiation and in abiotic and biotic stresses processes, respectively, while 109 (40.6%) were shared among two or three processes. In contrast, of 22,024 differentially expressed diploid genes, 577 (2.6%), 1,671 (7.6%), and 2,895 (13.1%) were found only in fruit developmental, in organ differentiation and in abiotic and biotic stresses processes, respectively, while 16,881 (76.6%) were shared among two or three processes. This pattern was also found in Chardonnay (Fig. 2E). Compared to diploid genes, hemizygous genes were more often differentially expressed in only one of three processes and have distinct expression patterns.

The cis-regulatory effects of hemizygous and diploid TEs on gene expression

We then explored the cis-regulatory effects of TEs on gene expression. We used the RepeatModeler pipeline to identify TEs for the clonal and outcrossing plant samples, detecting from 633,676 to 1,195,837 TEs across the four taxa (Table 1). We then classified genes into four categories based on their proximity to annotated TEs. The four categories were: i) hemizygous genes with nearby TEs (i.e., within 2kb of either the 5' or 3' ends of genes), ii) hemizygous genes without nearby TEs, iii) diploid genes with nearby TEs, iv) diploid genes without nearby TEs.

Focusing on diploid genes, the pattern was consistent and clear: among the group of genes without TEs, a higher percentage were expressed (Fig. 3A) and at higher levels (Fig. 3B)

than genes with nearby TEs. This observation held across the four taxa and across individual RNA-seq samples (Additional file 1: Figs. S9 and S10). The difference could be striking; for example, in one leaf sample of Pinot Noir, 71% of diploid genes without a nearby TE were expressed, while only 61% of diploid genes with a nearby TE were expressed. The pattern in diploid genes was consistent with the findings that: i) host silencing of TEs near genes often negatively affect expression of a neighboring gene (40) – e.g., siRNA-targeted TEs are associated with reduced gene expression(41); and ii) TEs close to genes may disrupt cis-regulatory element such as enhancers, silencers, thus affecting gene expression (42, 43).

However, these patterns were not as consistent with hemizygous genes (Fig 3A, Additional file 1: Figs. S9 and S10). Hemizygous genes near TEs were generally expressed more often and more highly in Pinot Noir, while the converse was true in Chardonnay (Fig. 3A). In fact, expression levels of hemizygous genes without nearby TEs were significantly lower than that of hemizygous genes with nearby TEs in Pinot Noir (Fig. 3B, Wilcoxon rank-sum test, $p < 0.001$). Two other taxa (cassava and Black Cottonwood) had similar patterns of lower expression in hemizygous genes without nearby TEs, but the contrast was not significant (Fig. 3B). Thus, the relationship between hemizygous genes and the presence of nearby TEs tended to contradict patterns for diploid genes.

Like genes, TEs can be diploid or hemizygous, so we also explored the effect of hemizygous and diploid TEs on gene expression. To do so, we classified genes with nearby TEs into four categories: i) hemizygous genes with hemizygous TEs, ii) diploid genes with hemizygous TEs, iii) hemizygous genes with diploid TEs, and iv) diploid genes with diploid TEs. Across four taxa, there was no clear pattern in terms of the proportion of genes that were expressed in each of the categories. The percentage of expressed diploid genes with nearby diploid TEs was generally higher, though not always, compared to the percentage of expressed diploid genes with nearby hemizygous TEs (Fig. 3C; *SI Appendix* Fig. S11), but results varied among taxa. Turning to expression levels, rather than the proportion of expressed genes, diploid genes near diploid TEs tended to be expressed at higher levels, on average, than diploid genes near hemizygous genes (Fig. 3D; *SI Appendix* Fig. S12). However, this contrast was significant only for Black Cottonwood (Fig. 3D).

Overall, the results generally suggest that SVs near genes (i.e., that result in hemizygous TEs) tend to reduce expression of diploid gene more than nearby diploid TEs.

Higher DNA methylation in hemizygous genes and TEs

Finally, we investigated DNA methylation patterns of hemizygous genes relative to diploid genes. We surveyed genome-wide levels of DNA methylation and gene expression from leaves of Pinot Noir, Chardonnay, cassava, and Black Cottonwood (Additional file 1: Table 2). For hemizygous genes in Pinot Noir leaves, we detected average weighted genomic DNA methylation levels of 53.8%, 24.7% and 2.3% in the CG, CHG and CHH sequence contexts, respectively (Fig. 4A and 4B). Like previous reports (44, 45), the genic methylation level was lower than the genome-wide methylation level; in hemizygous versus diploid genes, the average DNA methylation level was 42.1% vs. 39.5%, 30.2% vs. 7.3%, and 2.5% vs. 1% in the CG, CHG, and CHH contexts, respectively. These patterns were largely consistent across taxa and generally reflect higher methylation levels for hemizygous genes compared to diploid genes.

As expected, TEs were methylated at higher levels than genome-wide averages, but interestingly, hemizygous TEs tended to be methylated at higher levels than diploid TEs. For example, in the Pinot Noir sample, hemizygous and diploid TEs have methylation levels of 80.4% and 65.6% in the CG context, 58.6% and 42.8% in the CHG context and 4.2% and 3.3% in the CHH context. Similar patterns were found in Chardonnay, cassava, and Black Cottonwood (Fig. 4B), and hence hemizygous TEs generally have higher methylation levels than diploid TEs.

Hemizygous gene expression levels correlated with gene body methylation

We then turned to the methylation status of individual genes. We defined gene body-methylated genes (gbM) as genes with CG methylation but without CHG and CHH methylation. We also categorized CHG methylated genes (mCHG) as genes with CHG methylation, and unmethylated genes (UM) as genes without CG, CHG and CHH methylation. Across the clonal taxa, the hemizygous genes tended to harbor a lower proportion of gbM genes than diploid genes. For example, in Pinot Noir, 20.5% of

hemizygous genes (395 out of 1,403) and 35.1% of diploid genes (10,682 out of 30,404) on average were gbM (Fig. 5A). In Black Cottonwood, the proportion of gbM in hemizygous genes (4.2%, 13 out of 309) was slightly higher than that in diploid genes (3.3%, 1,084 out of 33,129) (Fig. 5A). Overall, the results indicated that hemizygous genes were less often gbM than diploid genes and that gbM was more extensive in the clonal samples compared to single outcrossing sample in our dataset (Fig. 5A). The difference between gbM proportions in hemizygous and diploid genes (i.e., hemizygous < diploid) and average CG genic methylation ratio pattern (hemizygous > diploid) can be explained by the higher proportion of hemizygous mCHG genes. For example, in Pinot Noir, 25.2% of hemizygous genes (353 out of 1,403) and 6.5% of diploid genes (1,991 out of 30,403) were mCHG.

After identifying gbM, mCHG and UM genes, we investigated their expression patterns. We identified a few distinct trends. First, a relatively small proportion of mCHG genes were expressed, no matter whether they were hemizygous or diploid genes. This pattern was consistent with previous findings that mCHG methylation decreases gene expression (45). The high proportion of hemizygous mCHG genes contributed to the overall lower expression levels of hemizygous vs. diploid genes (Fig. 2B). Second, a high (> 73.3%) proportion of gbM and UM genes were expressed for diploid genes, but this was not always the case for hemizygous genes (Fig. 5B). In Chardonnay and cassava, a high (>70.4%) proportion of hemizygous gbM and UM genes were expressed, but these proportion were as low as 20.5% and 30.8% in Pinot Noir and Black Cottonwood (Fig. 5B). Third, the patterns based on the proportion of expressed genes were largely reflected in expression levels. That is, mCHG genes were relatively lowly expressed, no matter if they were hemizygous or diploid (Fig. 5C); gbM and UM genes were expressed at higher levels than mCHG genes (*SI Appendix* Fig. S13); and hemizygous gbM and UM genes were consistently more lowly expressed than diploid genes (Fig. 5C).

Discussion

Hemizygous genes have been extensively studied in sex-linked regions, but they can also occur beyond sex-linked regions of homologous chromosomes due to SVs. Some SVs will

lead to the presence of a single allele on one sister chromatid of an otherwise diploid organism. The prevalence, expression and epigenetics of these hemizygous genes has rarely been investigated. Here we have integrated genomic, transcriptomic and epigenomic analyses to estimate the frequency of hemizygous genes and to characterize their features, expression, and epigenetic regulation.

Hemizygous genes are most common in clonal lineages

Consistent with previous work, we found that hemizygous genes are more common in clonal, as opposed to outcrossing lineages. Although hemizygosity has already been measured in a handful of plant taxa – i.e., primarily grape varieties (28, 31) and rice species (33) – we have extended observations to additional grape cultivars, including a new Pinot Noir assembly, clonally propagated cassava, an outcrossing species (Black Cottonwood) and three species expected to have completely homozygous genomes. Focusing on SVs that represent deletions relative to the reference assembly (hDELs), we have found, as expected, little evidence for hemizygosity in the homozygous samples. Across our three homozygous samples, two (orange and rice) had estimates of ~0.005% of the genome captured within hDELs of > 50 bp, with an even lower estimate in tomato (Table 1). While these results are not particularly surprising, the homozygous samples act as a negative control and show that we do not estimate high hemizygosity where there should be none.

In contrast to the homozygous samples, our work substantiates a growing consensus that outcrossing species can harbor a substantive portion of their genome as hemizygous. In Black Cottonwood, for example, we have estimated that 3.2% of the genome, containing 4.5% of the genes, is captured by hDELs, mimicking levels found in outcrossing rice (33) and avocado (32) (Avocado is clonally propagated in cultivation, but the investigated tree had been produced by a recent outcrossing event). In contrast, long-term clonal lineages consistently have a more substantial fraction of their genomes and genes captured in a hemizygous state. We caution that most of the observations to date have been based on grapevine clones, some of which have been propagated for 1000 or more years (46). However, by including cassava, we have not only shown that it is similar to grapes (with > 10% of the genome captured in hemizygous regions, Table 1), but also that the

phenomenon is not limited to grapevines. Moreover, the results accentuate how a traditional focus on inbreeding plants like *Arabidopsis thaliana* (47), rice (48) and tomato (49) has biased our understanding of genetic variation. Inbreeding plants are typically highly homozygous with few sequence variants (30, 37, 38), but the genome of clonal plants are highly heterozygous with genetic diversity that includes SVs and hemizygous genes (50).

High genetic variation in clonal lineages is not particularly unexpected, for two reasons. First, previous work on SVs has inferred, based on population samples, that they tend to be deleterious (28, 33). Second, forward simulations have consistently revealed that heterozygous, deleterious variants are expected to accumulate over time in clonal lineages, without the matching phenomenon in outcrossing plants (28, 51, 52). This accumulation reflects the fact that recessive deleterious alleles can hide as heterozygotes within a clonal lineage; in contrast, they are expected to occasionally become homozygous and thus visible to selection in outcrossing systems. This accumulation also reflects that recombination is limited (i.e., effectively zero) in strictly clonal lineages, meaning that deleterious mutations cannot recombine onto different genetic backgrounds. In this context, it is interesting to note that domesticated, clonally propagated cassava has a marked 26% higher genomic burden of putatively deleterious nucleotides compared to its wild congener (53).

Despite previous studies about the accumulation of deleterious variants in clonal lineages, the large number of hemizygous genes in clonal lineages is still somewhat surprising, because functionally hemizygous genes cannot (by definition) be recessive. Hence, the dynamics of the accumulation of hemizygous genes are likely to differ somewhat from the deleterious recessive case studied by forward simulation. Assuming that many (but not all; see below) of the SV events are deleterious, several functional and evolutionary processes likely contribute to the accumulation of hemizygous genes in clonal lineages. One is a ratchet mechanism – i.e., once an SV occurs in a clonal lineage, it has only one possible fate, so long as it is not lethal, which is to remain in the clonal lineage. By this process, clonal lineages are expected to accumulate SVs. In theory, this accumulation is more likely when the SV events do not severely affect fitness; for that reason, we expect deleterious SVs often have moderate functional effects.

Hemizygous gene expression is moderated by epigenetic effects

Indeed, we have accrued evidence that hemizygous genes have moderated effects, based on three pieces of evidence. First, hemizygous genes are more likely to be non-expressed than diploid genes in our sample species (Fig. 2A). That is, a higher proportion of hemizygous genes appear to be pseudogenes. Second, hemizygous genes are more likely to be members of gene families (Fig. 1F and Fig. 1G), implying that they are more likely to be functionally redundant. Thus, the loss of one copy of a multi-copy gene is likely to carry fewer fitness consequences than the loss of one allele of a critical single-copy gene. Finally, and somewhat surprisingly, as a group, hemizygous genes tend to be expressed at less than half the level of an average diploid genes, at about 20% (Fig. 2C). This value is substantially less than the 50% expected of a single gene. It is hard to know the cause of this low expression pattern. It is possible, for example, that hemizygous genes are a biased sample that were lowly expressed in their diploid state before the SV event. Another possibility is that epigenetic effects act especially strongly on hemizygous genes to moderate their expression (see below).

In this context, it is worth accentuating that our expression observations are unprecedented. The only other hemizygous genes studied intensively – i.e., sex-linked genes – tend to have a X:AA gene expression ratio of ~0.5 in human, mouse, and nematode (16, 54); Another possibility for sex-linked genes is dosage compensation, which hypothesizes that hemizygous X-linked genes are expressed at twice the level of diploid genes per active allele to balance the gene dose between the X chromosome and autosomes (21). For this case, the X:AA gene expression ratio would hover around 1.0. This upregulation may be sufficient to mitigate negative fitness effects, even if expression still falls significantly short of ancestral expression levels, and may mitigate the effects of aneuploidy (55-57); additional selection for compensatory up-regulation may be unnecessary for such loci (58). We see none of these effects. In our study, we do not see any overarching evidence of complete or even partial dosage compensation of hemizygous genes. Instead, the opposite is true: the expression of heterozygous alleles are dampened, on average, so that they are substantially less expressed than the average diploid allele.

We suspect that this dampening effect is at least partially due to epigenetic phenomena, for three reasons. First, in all four species investigated, hemizygous TEs have elevated levels of DNA methylation relative to diploid TEs (Fig. 4A). Several phenomena may contribute to this observation, including that hemizygous TEs may be more recent insertions (and therefore more recently targeted by host mechanisms). Whatever the cause, the data hint that hemizygous TEs differ quantitatively in their methylation effects. Second, hemizygous genes also exhibit higher levels of methylation than diploid genes, specifically a higher proportion of mCHG alleles (Fig. 5A), which are typically lowly expressed (Fig. 5C). Finally, we have shown that genes near TEs are consistently more lowly expressed than genes far from TEs (Fig. 3B), but this effect is more prominent for genes near hemizygous TEs (Fig. 3D). This may be a partial explanation as to why genes close to SVs are associated with reduced gene expression levels in tomato (59).

These observations have interesting parallels to previous studies that have suggested that DNA methylation is correlated with reduced gene expression levels for sex-limited genes on the Y or W chromosome (60, 61). High levels of DNA methylation have also been associated with sex chromosomes in sticklebacks and papaya (24, 25). In addition, DNA methylation is a key feature in X-chromosome inactivation (26). These results suggest some similar features of DNA methylation patterns between sex-linked and non-sex-linked hemizygous genes. Clearly, we cannot be certain what, if any, epigenetic mechanisms might be shared between sex-linked hemizygosity and that which we have studied here, but it is an interesting question for further research.

Are hemizygous genes merely functional remnants?

Given the evidence – i.e., that hemizygous genes tend to be shorter than diploid genes (Fig. 1B), expressed at lower levels (Fig. 2B), potentially subjected to lower levels of purifying selection (as measured by *Ka/Ks*; Fig. 1E), and more heavily methylated (Fig. 4A and 4B) – it is tempting to conclude that hemizygous genes are typically pseudogenes. Are they merely functional remnants of previously functional genes? While to this question is likely yes for most hemizygous genes, there is some tantalizing evidence that the answer is this question may often be ‘no’.

Evidence supporting functionality of some hemizygous genes comes in a few forms. For example, a reasonable proportion of hemizygous genes have gbM patterns of methylation (Fig. 5A). In both hemizygous and diploid genes, gbM has higher expression levels than UM genes (Fig. 5C, *SI Appendix* Fig. S13). Moreover, several studies have detected a correlation between the presence of gbM and the enhance of gene or allelic expression (62, 63), while others have found evidence that it is subject to natural selection (63) based on population genetic arguments. In short, although the functional role of gbM (if any) is debated, it typically is a mark deposited and maintained on active genes (45). The fact that some hemizygous gene bear this epigenetic mark superficially suggests that they can be easily dismissed as non-functional.

In addition, hemizygous genes (as a group) demonstrate patterns of tissue/treatment-specific expression that are similar to diploid genes. This pattern does not hold at the single gene level, but nonetheless up to 37.5% of hemizygous genes exhibit tissue/treatment-specific expression in Black Cottonwood (Fig. 2D). Of course, tissue/treatment-specific expression patterns are not proof of function, but it does indicate that some hemizygous are induced under different environmental and developmental conditions. Finally, there are some consistent patterns of GO enrichment, particularly for responses to biotic and abiotic stresses (*SI Appendix* Fig. S2). Again, GO enrichment is not proof of function, but all of this evidence combines to make it reasonable to hypothesize that not all hemizygous genes are functional 'junk'. Of course, the mere act of uncovering of a recessive allele can have important functional consequences; we invoke again the compelling case of hemizygosity and the white berry phenotype of grapes (34).

To sum: Our work has contributed to an emerging picture that clonal lineages are particularly replete with hemizygous genes but that outcrossing diploids still have substantial regions of hemizygosity. Many of the genes in these regions are not expressed, and the regions themselves appear to be prone to enhanced methylation levels. These enhanced methylation levels – particularly in hemizygous TEs – may affect patterns of cis-regulation, such that differences in hemizygosity between clonal lineages may contribute to phenotypic differences. These pervasive hemizygous genes may thus be more important

than previously thought for understanding the genetics, breeding and evolution of plants with heterozygous genomes.

Materials and Methods

Sample selection and genome assembly and annotation

We used genome assemblies based on long-reads sequencing data for seven diploid plant samples. Three were clonal samples, including two grapevine (*V. vinifera* ssp. *vinifera*) crop, varietals, heterozygous Pinot Noir and Chardonnay (28), and the TME204 varietal of cassava (*M. esculenta*) (35). One was outcrossing sample, Black Cottonwood (*P. trichocarpa*) (36). The remaining three included one naturally inbred rice cultivar (*O. sativa* ssp. *indica*), MH63 (37); one manipulated inbred tomato cultivar (*S. lycopersicum*), Heinz 1706 (30); and one manipulated doubled haploid sweet orange cultivar (*Citrus sinensis*), Valencia (38) (Table 1).

Among the seven genome assemblies, the data for the assembly of Pinot Noir were generated for this study; we focused on Pinot Noir because of the wealth of expression data available for the reference grapevine PN40024 that was based on a Pinot Noir-derived genotype (64), while these expression data were produced from Pinot Noir (*SI Appendix* Table S1). The plant material was grown at Agriculture Genomics Institute at Shenzhen (AGIS), Chinese Academy of Agriculture Science (CAAS). DNA extraction and the construction of SMRTbell libraries followed ref. (30). SMRTbell libraries were sequenced on the PacBio Sequel II platform in the CCS mode, generating a total of 33 Gb (66× genomic coverage). DNA extraction and the preparation of ultra-long Oxford Nanopore Technologies (ONT) libraries followed ref. (65). ONT libraries were sequenced on the ONT platform, generating an additional 14 Gb (28×) of extra-long reads with an average length of 99 kb. For Hi-C library construction, chromatin was digested with the restriction enzyme MboI using a previously described Hi-C library preparation protocol (66). The Hi-C libraries were sequenced on an Illumina HiSeq X Ten Platform, generating a total of 82 Gb (160×).

The Pinot Noir genome was assembled using hifiasm (v.0.13) (67) , which generated a primary assembly (p_ctg) and an alternative assembly (a_ctg), both of which consisted of gapless contigs. The p_ctg assembly consisted of 461 contigs with a contig N50 of 23.6 Mb, while the a_ctg assembly included 342 contigs with a contig N50 of 24.2 Mb. Subsequently, the contig-level assemblies were aligned to 19 chromosomes using the Cabernet Sauvignon genome (http://www.grapegenomics.com/pages/VvCabSauv/) (67) as a reference with the default parameters of RagTag (v2.1.0) (68). The primary and alternative contigs were then grouped and sorted using Juicer (v.1.6) (69) and 3D-DNA (v.180922) (70) software, and anchored to 19 chromosomes using Hi-C reads in Juicebox (71). The genome assembly was manually corrected in IGV (v2.12.3) (72) by remapping ultra-long ONT reads to the genome assembly. We also filled and closed gaps using selected and assigned contigs, achieving gap-free assemblies for Pinot Noir. Ultimately, two phased genomes were obtained, including the haplotig1 genome spanning 495.2 Mb sequences (scaffold N50 of 27.1 Mb), and the alternative genome (haplotig2) spanning 489.6 Mb. The haplotig1 genome assembly was used for downstream analyses; this genome scored 98.3% completeness in gene-space using Embryophyta_odb10 datasets based on BUSCO (v5.3.2) (73) (Table 1). The gene annotations of Pinot Noir were transferred from *Vitis vinifera cv.* PN40024 v4.2 (http://www.grapegenomics.com/pages/PN40024/blast.php) (74) using liftoff (v1.6.3) (72) with default parameters. Repetitive elements (TEs) of Pinot Noir were identified using RepeatModeler and masked using RepeatMasker (75) with default parameters.

The remaining six genome assemblies were retrieved from public resources. We downloaded Chardonnay genome assembly and annotation (VvChar04_v1) from Genome Database for Grapevine (http://www.grapegenomics.com/pages/VvChar/) (28), then used RagTag (v2.1.0) (68) to anchor and orient VvChar04_v1 to chromosome level based on the reference genome *Vitis vinifera* cv. PN40024 v4.2 (74), and updated the final gene annotation based on VvChar04_v1 gff file using (v1.6.3) (72). For cassava, we downloaded the genome assembly (hifiasm152_l3.hic.hap1.p_ctg) from NCBI (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/020/916/445/GCA_020916445.1_hifiasm152_l3.hic.hap1.p_ctg/), and updated the genome annotation from cassava AM560-2.v8 (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/659/605/GCF_001659605.2_M.escul

enta_v8/). For outcrossing Black Cottonwood, we downloaded the genome and annotation files Cottonwood from NCBI (https://www.ncbi.nlm.nih.gov/assembly/GCF_000002775.5) (36). For the three inbred and doubled haploid samples, rice, tomato and sweet orange, we downloaded: the genome assembly and annotation of rice (*O. sativa* ssp. *indica*) MH63RS3 from Rice Information Gateway (http://rice.hzau.edu.cn/rice_rs3/) (37); the genome assembly and annotation of tomato (*S. lycopersicum* Heinz 1706) SL5.0 from a tomato database (http://solomics.agis.org.cn/tomato/ftp/) (30); and the genome assembly of sweet orange (*C. sinensis*) Valencia from NCBI (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/018/104/345/GCA_018104345.1_ASM1810434v1/), and updated the genome annotation based on genome assembly and annotation of DVS_A1.0 (https://www.ncbi.nlm.nih.gov/assembly/GCA_022201045.1/) (38) using liftoff (v1.6.3) (72). Repetitive elements (TEs) of these six samples were identified for this study using the same Repeatmasker pipeline that was applied to Pinot Noir. Across these genomes, 26,874-59,903 protein-coding genes were annotated. The total gene lengths were in the 114.3 to 180.8 Mb range, occupying 15%~45.7% of their respective genome sizes (Table 1). We also identified from 458,740 to 1,195,837 TEs across species, occupying from 51.83% to 71.10% of each genome (Table 1).

Identification and characterization of hemizygous genes

To identify hemizygous genes, we retrieved the raw long-read PacBio CCS/CLR data of the seven genome assemblies publicly, except for Pinot Noir: (1) for two clonal propagated samples, we retrieved the PacBio CLR data of Chardonnay and PacBio CCS data of cassava from the NCBI Short Read Archive under accession PRJNA550461 and PRJEB43673, respectively; (2) for one outcrossing samples, we retrieved the PacBio CLR data of Black Cottonwood from NCBI SRA under accession PRJNA791651; (3) we recovered the PacBio CLR data of rice, PacBio CCS of tomato, PacBio CLR of sweet orange from NCBI SRA under accession PRJNA302543, PRJNA733299, and PRJNA347609, respectively.

We then remapped the corresponding long-read PacBio CCS or CLR data to each of the seven surveyed plant genome assemblies to call SVs using a Sniffles pipeline. In this pipeline, PacBio CCS or CLR reads were normalized to 30 × sequencing depth for each sample using seqkit (v 2.2.0) (76), the depth of normalized sequencing reads were calculated using bedtools (v2.30.0) (77) coverage with default option, in which the windows were made as 1000 bp using bedtools makewindows, the fourth line of the output files was sequencing depth. The distribution of sequencing depth was plotted for each taxon (*SI Appendix* Fig. S1). The remaining PacBio reads were mapped onto genome assemblies using Minimap2 (2.24-r1122) (78) with the MD flag, and variant callings were performed using Sniffles (v2.0.6) (39). SV analysis outputs (VCF files) were filtered based on the following three steps: (1) we removed SVs that had ambiguous breakpoints (flag: IMPRECISE) and also low-quality SVs that did not pass quality requirements of Sniffles (flag: UNRESOLVED); (2) we removed SV calls shorter than 50 bp; (3) we removed SVs with less than four supporting reads. Hemizygous regions were defined as deletion regions with 0/1 flags based on SV inferences, and genes that 100% overlapped hemizygous regions were defined as hemizygous genes. Genes were extracted from hemizygous regions of the genome with bedtools (v2.30.0) (77) intersect with command "bedtools intersect -wo -a hemizygous_regions.bed -b gene.bed -F 1". The remaining genes were termed diploid genes.

To help identify reliable hemizygous genes, we calculated SNP heterozygosity. SNPs were called based on PacBio CCS reads for Pinot Noir and cassava, and Illumina paired-end data for Chardonnay and Black Cottonwood. PacBio CCS reads were mapped onto the corresponding genome assembly using Minimap2 (2.24-r1122) (78) with the -ax map-hifi and MD flag. SAM format was converted to BAM format and sorted using Samtools (v1.9) (79). SNPs were called using Deepvariant (v1.4.0) (80) with the PacBio mode and default parameters. For Illumina paired-end data, the adapters of raw data were trimmed and low-quality data were discarded using Trimmomatic (v0.39) (81) with the options: LEADING:3 TRAILING:3 SLIIDINGWINDOW:10:20 MINLEN:36. Second, Illumina reads were mapped to Chardonnay using bwa (v0.7.17-r1188) (82). SNPs were called using Deepvariant (v1.4.0) (80) with the WGS mode and default parameters. Then, SNPs were filtered using with VCFtools (v0.1.16) (83): (1) SNPs with less than five supported reads

were removed, (2) SNP with "0/1" flag were selected. The Illumina data for Chardonnay from NCBI SRA under SRR5627799 (PRJNA388292); The Illumina data for Black Cottonwood under SRR17455010 (PRJNA791651). Once SNPs were identified, the heterozygosity of each gene was calculated as the total heterozygous SNP numbers divided by the length of each gene.

We then characterized sequence features, such as coverage, length, synonymous mutation rate (*Ks*), and non-synonymous/synonymous mutation ratio (*Ka/Ks*) of hemizygous and diploid genes, and the proportion of single copy genes of hemizygous and diploid genes. Gene length was calculated as the length between transcription start and end. We calculated the coverage of hemizygous vs. diploid genes. The sequence depth of each gene was detected using bedtools (v2.30.0) (77) with the coverage option, disregarding sequencing depths of >100 or < 3 for possible sequencing bias.

Ka and *Ks* values were estimated using MCScanX pipeline (https://github.com/wyp1125/MCScanX) based on Pinot Noir-*M. rotundifolia*, Chardonnay-*M. rotundifolia* and cassava-*R. communis*, Black Cottonwood-*S. brachista* genome sequence comparisons. We downloaded the genome fasta and gene annotation gff file of *M. rotundifolia* (http://www.grapegenomics.com/pages/Mrot/download.php), *R. communis* (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/019/578/655/GCF_019578655.1_ASM1957865v1/), and S. brachista (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/009/078/335/GCA_009078335.1_ASM907833v1/). The corresponding coding and protein sequences were converted from fasta and gff file using gffread (v0.12.7) (84). The BLASTP was performed using protein sequences (E-value < $1e^{-10}$, top 5 matches and outfmt 6) to search all possible homologous gene pairs between each species pair. The output files based on BLASTP analysis were used as inputs for MCScanX, and pairwise *Ka*, *Ks* values of syntenic homologous genes were estimated using the Perl script "add_ka_and_ks_to_collinearity.pl" in the MCscanX package, which implemented the *Nei-Gojobori* algorithm (85).

We used two methods to determine the single copy and multi-copy genes. Using Pinot Noir-*M. rotundifolia*, Chardonnay-*M. rotundifolia*, cassava-*R. communis*, and Black Cottonwood-*S. brachista*, we performed all-to-all BLASTP between each species and its respective outgroup. Single copy orthologous genes were determined when they met the criteria of an *e*-value $< 1 \times 10^{-10}$, with similarity >70% and coverage >70%; the top five matches were kept if more than five hits met the preset requirements. Second, the OrthoFinder (86) was also used to detect the single copy orthologous genes.

Finally, we characterized the biological function of hemizygous genes using eggnog-mapper (http://eggnog-mapper.embl.de/) (87). GO analysis was performed using the ClusterProfiler package (88) in R 4.1.0. We employed a P value < 0.05 to represent significantly enriched terms.

Dissection of hemizygous gene expression patterns

To understand how hemizygous genes responded to fruit development, organ differentiation, and biotic and abiotic stress stimulus, we downloaded publicly available RNA-seq data from NCBI (*SI Appendix* Table S1). For Pinot Noir, RNA-seq data represented fruits at two development stages in each of three projects (PRJNA260535, RJNA381300, PRJEB36552); RNA-seq data representing organ differentiation were recovered from leaves and fruits (PRJNA373967), leaves and immature/mature fruit (PRJNA381300); leaves and stem (PRJDB5807); and flowers and fruits (PRJEB39263). RNA-seq data regarding abiotic and biotic stress stimuli were generated from leaves under CO_2 treatment (PRJNA837346), leaves under drought treatment (PRJNA433817), fruits under water deficit treatment (PRJNA268857), and embryogenic callus under yeast treatment (PRJNA732451). For Chardonnay, we retrieved one dataset related to fruit development (PRJNA260535), one dataset related to organ differentiation (leaves and embryogenic callus, PRJNA691261), and three different datasets related to stress stimulus (PRJNA402079, PRJNA268857, PRJEB31325). The cassava RNA-seq data were generated from leaves (PRJNA787456). The Black Cottonwood RNA-seq data were generated from leaves (PRJNA549496) and abiotic stress stimulus experiments (PRJEB19784).

Raw RNA-seq reads were trimmed by quality using Trimmomatic (v0.39) (81) with the options: LEADING:3 TRAILING:3 SLIIDINGWINDOW:10:20 MINLEN:36. High-quality reads were mapped onto the primary genome assemblies using HISAT2 (v.2.2.1) (89) with default parameters. Raw count for each gene was calculated based on FeatureCounts (2.0.1) with the option: -p -B (paired-end reads, single-end reads without -B) -C -t transcript -g gene_id. Gene expression was quantified in normalized fpkm (fragment per kilobase per million) with a custom R script using the GenomicFeatures package (90) in R 4.1.0. In each tissue/treatment, gene expression was averaged over the biological replicates in each surveyed crop. Expressed genes were defined as those with fpkm > 0.

Exploration of cis-regulatory effects of TEs on gene expression

Based on the identification of repeat sequences, we explored the cis-regulatory effects of TEs on gene expression. For this purpose, we first assigned each TE to its closest gene when it was within 2 kb (the distance to either 5' or 3' end of gene with > 0 kb and <2 kb) using command "bedtools closest -wo -a gene.bed -b TE.bed", and thus genes were separated in four classes: hemizygous genes with nearby TEs, hemizygous genes without nearby TEs, diploid genes with nearby TEs, diploid genes without nearby TEs. We divided genes near TEs into four categories: hemizygous genes with either hemizygous or diploid TEs, and diploid genes with either hemizygous or diploid TEs.

Unveiling DNA methylation patterns of hemizygous genes

Bisulfite-seq (BS-seq) for four samples were either generated for this study or downloaded from public sources (*SI Appendix* Table S2). The Chardonnay clone chosen for BS-seq was FPS 04, a clone commonly grown in California and throughout the world. The reference plant is located at Foundation Plant Services, University of California. DNA was isolated with the Qiagen DNeasy Plant Mini kit, and bisulfite libraries were prepared as previously described (91). Libraries were pooled and sequenced (150bp paired-end) on the Illumina HiSeq2500. As a control for bisulfite conversion, lambda-DNA was spiked into each library preparation to measure the conversion rate of unmethylated cytosines (0.5% w/w). For publicly available datasets, BS-seq were generated from leaves in Pinot Noir

(PRJNA381300) (92), Chardonnay (PRJNA691261) (93), cassava (PRJNA793021) (35), and Black Cottonwood (PRJNA549497) (36).

BSseq reads were trimmed for quality and adapter sequences using Trimmomatic (v0.39) (81) with the options: LEADING:3 TRAILING:3 SLIIDINGWINDOW:10:20 MINLEN:36. Low quality reads and reads less than 36 bp were discarded. Bismark (v0.23.1) (94), in conjunction with bowtie2 (v2.1.0) (95) with default parameters were used to align trimmed reads to the respective genome reference.

The number of methylated and unmethylated reads per cytosine was calculated using the bismark_methylation_extractor in Bismark (v0.23.1) (94). Methylated cytosines were identified using a binomial test incorporating the estimated rates of bisulfite conversion errors (P<0.01 after Benjamini-Yekutieli FDR correction) (96). False methylation rates (FMR) for each library were estimated for each taxon as one previous study performed (91), FMRs were estimated using lambda-DNA or chloroplast DNA using MethylExtract (97). A minimum coverage of two was required at each cytosine to determine methylation status. DNA methylation distribution plots were performed with deepTools (98).

We defined body-methylated genes following the strategy of Ref. (99). Briefly, we quantified the level of DNA methylation for each protein-coding region for each context-CG, CHG, CHH. The protein-coding region was defined as the annotated translation start to the termination codon. Taking the CG context as an example, n_{CG} was the number of cytosine residues at CG sites with ≥2 coverage in the gene of interest, m_{CG} was the number of methylated cytosine residues at CG sites for the same gene, and p_{CG} was the average proportion of methylated cytosine residues across all genes. Assuming a binomial probability distribution, the one-tailed P value for the departure of CG methylation levels from average genic proportion of DNA methylation was calculated as:

$$P_{\mathrm{CG}} = \sum_{i=m_{\mathrm{CG}}}^{n_{\mathrm{CG}}} \left(\frac{n_{\mathrm{CG}}}{i}\right) {p_{\mathrm{CG}}}^{i} (1 - p_{\mathrm{CG}})^{n_{\mathrm{CG}}-i}$$

Where P_{CG} was a proxy of DNA methylation level. Using the same rationale, P_{CHG} and P_{CHH} were calculated for CHG and CHH context, respectively.

Given the binomial results, taking the similar strategies of Ref. (63), a gene was inferred to be gene body methylated (gbM) if CG methylation was significantly higher than the background (P_{CG} <=0.05), while CG and CHG methylation were not significantly higher than the background (P_{CHG} > 0.05 and P_{CHH} > 0.05). Similarly, a gene was inferred to be CHG methylated if CHG methylation was higher than the background (P_{CHG} <= 0.05) and CHH methylation was not significantly higher than the background (P_{CHH} > 0.05). CHG methylated genes also tended to be CG methylated, but CG methylation was not required in our categorization. A gene was inferred to be CHH methylated if CHH methylation was higher than the background (P_{CHH} <= 0.05). CHH methylated genes also tend to be CG and CHG methylated. Finally, a gene was inferred to be unmethylated (UM) if CG, CHG, and CHH methylation were not significantly higher than the background (P_{CG} > 0.05, P_{CHG} > 0.05, and P_{CHH} > 0.05). In any other case, the gene methylation state was not inferred.

Data availability

The PacBio CCS, ONT and Hi-C sequence data have been deposited to the NCBI short reads achieved under the project number: PRJNA951461 and the National Genomics Data Center (NGDC) Genome Sequence Archive (GSA) (https://ngdc.cncb.ac.cn/gsa/), with BioProject number PRJCA016741. The Bisulfite-seq data have been deposited to the NCBI short reads achieved under the project number: PRJNA987409. The genome assembly and annotation have been deposited to zenodo: https://zenodo.org/record/8089258. All data supporting the findings of this study are available within the manuscript and its supporting information are available from the corresponding author upon request.

Acknowledgments

We thank R. Gaut for generating the BS-seq data. This work was supported by the National Natural Science Foundation of China (No. 32372662), the Science Fund Program for Distinguished Young Scholars of the National Natural Science Foundation of China (Overseas) to Yongfeng Zhou, and the National Key Research and Development Program of China (No. 2023YFF1000100; 2023YFD2200700), the Agricultural Science and Technology Innovation Program (CAAS-ZDRW202101), the National Science Foundation grant (No.1741627) to Brandon S. Gaut.

References

1. D. Charlesworth, Young sex chromosomes in plants and animals. *New Phytol* **224**, 1095-1107 (2019).
2. D. Charlesworth, Plant sex determination and sex chromosomes. *Heredity (Edinb)* **88**, 94-101 (2002).
3. B. Charlesworth, J. L. Campos, B. C. Jackson, Faster-X evolution: Theory and evidence from Drosophila. *Mol Ecol* **27**, 3753-3771 (2018).
4. E. Cherif *et al.*, Male-specific DNA markers provide genetic evidence of an XY chromosome system, a recombination arrest and allow the tracing of paternal lineages in date palm. *New Phytol* **197**, 409-415 (2013).
5. M. F. Torres *et al.*, Genus-wide sequencing supports a two-locus model for sex-determination in Phoenix. *Nat Commun* **9**, 3969 (2018).
6. S. Picq *et al.*, A small XY chromosomal region explains sex determination in wild dioecious V. vinifera and the reversal to hermaphroditism in domesticated grapevines. *BMC Plant Biol* **14**, 229 (2014).
7. M. Massonnet *et al.*, The genetic basis of sex determination in grapes. *Nat Commun* **11**, 2902 (2020).
8. H. Badouin *et al.*, The wild grape genome sequence provides insights into the transition from dioecy to hermaphroditism during grape domestication. *Genome Biol* **21**, 223 (2020).
9. A. Harkess *et al.*, The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nat Commun* **8**, 1279 (2017).
10. K. Murase *et al.*, MYB transcription factor gene involved in sex determination in Asparagus officinalis. *Genes Cells* **22**, 115-123 (2017).
11. T. Akagi *et al.*, Two Y-chromosome-encoded genes determine sex in kiwifruit. *Nat Plants* **5**, 801-809 (2019).
12. T. Akagi *et al.*, A Y-encoded suppressor of feminization arose via lineage-specific duplication of a cytokinin response regulator in kiwifruit. *Plant Cell* **30**, 780-795 (2018).
13. J. A. Tennessen *et al.*, Repeated translocation of a gene cassette drives sex-chromosome turnover in strawberries. *PLoS Biol* **16**, e2006062 (2018).
14. B. Vicoso, B. Charlesworth, Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet* **7**, 645-653 (2006).
15. B. Charlesworth, D. Charlesworth, The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci* **355**, 1563-1572 (2000).
16. Y. Xiong *et al.*, RNA sequencing shows no dosage compensation of the active X-chromosome. *Nature Genetics* **42**, 1043-1047 (2010).
17. B. Charlesworth, The evolution of chromosomal sex determination and dosage compensation. *Curr Biol* **6**, 149-162 (1996).
18. R. Bergero, S. Qiu, D. Charlesworth, Gene loss from a plant sex chromosome system. *Curr Biol* **25**, 1234-1240 (2015).
19. Q. Zhou, D. Bachtrog, Chromosome-wide gene silencing initiates Y degeneration in Drosophila. *Curr Biol* **22**, 522-525 (2012).
20. D. Prentout *et al.*, Plant genera *Cannabis* and *Humulus* share the same pair of well-differentiated sex chromosomes. *New Phytol* **231**, 1599-1611 (2021).
21. S. Ohno, *Sex chromosomes and sex linked genes* (Springer Verlag, Berlin, Germany, 1967).
22. P. A. Eyer, A. J. Blumenfeld, E. L. Vargo, Sexually antagonistic selection promotes genetic divergence between males and females in an ant. *P Natl Acad Sci USA* **116**, 24157-24163 (2019).

23. A. Muyle, G. A. B. Marais, V. Bacovsky, R. Hobza, T. Lenormand, Dosage compensation evolution in plants: theories, controversies and mechanisms. *Philos Trans R Soc Lond B Biol Sci* **377**, 20210222 (2022).
24. W. Zhang, X. Wang, Q. Yu, R. Ming, J. Jiang, DNA methylation and heterochromatinization in the male-specific region of the primitive Y chromosome of papaya. *Genome Res* **18**, 1938-1943 (2008).
25. D. C. H. Metzger, P. M. Schulte, The DNA methylation landscape of stickleback reveals patterns of sex chromosome evolution and effects of environmental salinity. *Genome Biol Evol* **10**, 775-785 (2018).
26. W. Reik, A. Lewis, Co-evolution of X-chromosome inactivation and imprinting in mammals. *Nat Rev Genet* **6**, 403-410 (2005).
27. A. Muyle, D. Bachtrog, G. A. B. Marais, J. M. A. Turner, Epigenetics drive the evolution of sex chromosomes in animals and plants. *Philos Trans R Soc Lond B Biol Sci* **376**, 20200124 (2021).
28. Y. F. Zhou *et al.*, The population genetics of structural variants in grapevine domestication. *Nat Plants* **5**, 965-979 (2019).
29. D. Tang *et al.*, Genome evolution and diversity of wild and cultivated potatoes. *Nature* **606**, 535-541 (2022).
30. Y. Zhou *et al.*, Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527-534 (2022).
31. A. M. Vondras *et al.*, The genomic diversification of grapevine clones. *BMC Genomics* **20**, 972 (2019).
32. E. A.-O. Solares *et al.*, Insights into the domestication of avocado and potential genetic contributors to heterodichogamy. LID - 10.1093/g3journal/jkac323 [doi] LID - jkac323.
33. Y. Kou *et al.*, Evolutionary Genomics of Structural Variation in Asian Rice (Oryza sativa) Domestication. *Mol Biol Evol* **37**, 3507-3524 (2020).
34. P. Carbonell-Bejerano *et al.*, Catastrophic unbalanced genome rearrangements cause somatic loss of berry color in grapevine. *Plant Physiol* **175**, 786-801 (2017).
35. W. Qi *et al.*, The haplotype-resolved chromosome pairs of a heterozygous diploid African cassava cultivar reveal novel pan-genome and allele-specific transcriptome features. *Gigascience* **11** (2022).
36. B. T. Hofmeister *et al.*, A genome assembly and the somatic genetic and epigenetic mutation rate in a wild long-lived perennial Populus trichocarpa. *Genome Biol* **21**, 259 (2020).
37. J. M. Song *et al.*, Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol Plant* **14**, 1757-1767 (2021).
38. L. Wang *et al.*, Somatic variations led to the selection of acidic and acidless orange cultivars. *Nat Plants* **7**, 954-965 (2021).
39. F. J. Sedlazeck *et al.*, Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**, 461-468 (2018).
40. J. D. Hollister, B. S. Gaut, Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res* **19**, 1419-1428 (2009).
41. J. D. Hollister *et al.*, Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. *Proc Natl Acad Sci U S A* **108**, 2322-2327 (2011).
42. C. D. Hirsch, N. M. Springer, Transposable element influences on gene expression in plants. *Biochim Biophys Acta Gene Regul Mech* **1860**, 157-165 (2017).
43. J. M. Noshay *et al.*, Assessing the regulatory potential of transposable elements using chromatin accessibility profiles of maize transposons. *Genetics* **217**, 1-13 (2021).

44. K. Roessler *et al.*, The genome-wide dynamics of purging during selfing in maize. *Nat Plants* **5**, 980-990 (2019).
45. A. M. Muyle, D. K. Seymour, Y. Lv, B. Huettel, B. S. Gaut, Gene body methylation in plants: mechanisms, functions, and important implications for understanding evolutionary processes. *Genome Biol Evol* **14** (2022).
46. J. Ramos-Madrigal *et al.*, Palaeogenomic insights into the origins of French grapevine diversity. *Nat Plants* **5**, 595-603 (2019).
47. I. Arabidopsis Genome, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).
48. S. A. Goff *et al.*, A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296**, 92-100 (2002).
49. C. Tomato Genome, The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641 (2012).
50. B. S. Gaut, D. K. Seymour, Q. Liu, Y. Zhou, Demography and its effects on genomic variation in crop domestication. *Nat Plants* **4**, 512-520 (2018).
51. H. Xiao *et al.*, Adaptive and maladaptive introgression in grapevine domestication. *Proc Natl Acad Sci U S A* **120**, e2222041120 (2023).
52. Y. Zhou, M. Massonnet, J. S. Sanjak, D. Cantu, B. S. Gaut, Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. *Proc Natl Acad Sci U S A* **114**, 11715-11720 (2017).
53. P. Ramu *et al.*, Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat Genet* **49**, 959-963 (2017).
54. P. P. Khil, N. A. Smirnova, P. J. Romanienko, R. D. Camerini-Otero, The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nature Genetics* **36**, 642-646 (2004).
55. P. Stenberg, J. Larsson, Buffering and the evolution of chromosome-wide gene regulation. *Chromosoma* **120**, 213-225 (2011).
56. J. H. Malone *et al.*, Mediation of *Drosophila* autosomal dosage effects and compensation by network interactions. *Genome Biol* **13**, r28 (2012).
57. J. E. Mank, Sex chromosome dosage compensation: definitely not for everyone. *Trends Genet* **29**, 677-683 (2013).
58. L. Q. Gu, J. R. Walters, Evolution of sex chromosome dosage compensation in animals: a beautiful theory, undermined by facts and bedeviled by details (vol 9, pg 2461, 2017). *Genome Biology and Evolution* **11**, 1135-1135 (2019).
59. M. Alonge *et al.*, Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145-161 e123 (2020).
60. R. K. Slotkin, R. Martienssen, Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**, 272-285 (2007).
61. N. Zamudio *et al.*, DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes Dev* **29**, 1256-1270 (2015).
62. A. Muyle, B. S. Gaut, Loss of gene body methylation in eutrema salsugineum is associated with reduced gene expression. *Mol Biol Evol* **36**, 155-158 (2019).
63. A. Muyle, J. Ross-Ibarra, D. K. Seymour, B. S. Gaut, Gene body methylation is under selection in Arabidopsis thaliana. *Genetics* **218** (2021).
64. X. Shi *et al.*, The complete reference genome for grapevine (Vitis vinifera L.) genetics and breeding. LID - uhad061.
65. S. Deschamps *et al.*, A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat Commun* **9**, 4844 (2018).
66. J. M. Belton *et al.*, Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268-276 (2012).

67. H. Y. Cheng, G. T. Concepcion, X. W. Feng, H. W. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170-+ (2021).
68. M. Alonge *et al.*, RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* **20**, 224 (2019).
69. N. C. Durand *et al.*, Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**, 95-98 (2016).
70. O. Dudchenko *et al.*, De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92-95 (2017).
71. J. T. Robinson *et al.*, Juicebox.js Provides a cloud-based visualization system for Hi-C Data. *Cell Syst* **6**, 256-258 e251 (2018).
72. H. Thorvaldsdottir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192 (2013).
73. M. Seppey, M. Manni, E. M. Zdobnov, BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227-245 (2019).
74. D. Navarro-Paya *et al.*, The grape gene reference catalogue as a standard resource for gene selection and genetic improvement. *Front Plant Sci* **12** (2022).
75. J. M. Flynn *et al.*, RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* **117**, 9451-9457 (2020).
76. W. Shen, S. Le, Y. Li, F. Hu, SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962 (2016).
77. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
78. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
79. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
80. R. Poplin *et al.*, A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**, 983-+ (2018).
81. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
82. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
83. P. Danecek *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
84. G. Pertea, M. Pertea, GFF Utilities: GffRead and GffCompare. *F1000Res* **9** (2020).
85. M. Nei, T. Gojobori, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418-426 (1986).
86. D. M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20** (2019).
87. C. P. Cantalapiedra, A. Hernandez-Plaza, I. Letunic, P. Bork, J. Huerta-Cepas, eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic Scale. *Mol Biol Evol* **38**, 5825-5829 (2021).
88. T. Wu *et al.*, clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).
89. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360 (2015).
90. M. Lawrence *et al.*, Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118 (2013).
91. D. K. Seymour, B. S. Gaut, Phylogenetic shifts in gene body methylation correlate with gene expression and reflect trait conservation. *Mol Biol Evol* **37**, 31-43 (2020).

92. P. Lu *et al.*, Genome encode analyses reveal the basis of convergent evolution of fleshy fruit ripening. *Nat Plants* **4**, 784-791 (2018).
93. D. Lizamore, R. Bicknell, C. Winefield, Elevated transcription of transposable elements is accompanied by het-siRNA-driven de novo DNA methylation in grapevine embryogenic callus. *BMC Genomics* **22**, 676 (2021).
94. F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).
95. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
96. R. Lister *et al.*, Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523-536 (2008).
97. G. Barturen, A. Rueda, J. L. Oliver, M. Hackenberg, MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Res* **2**, 217 (2013).
98. F. Ramirez, F. Dundar, S. Diehl, B. A. Gruning, T. Manke, deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187-191 (2014).
99. S. Takuno, B. S. Gaut, Body-methylated genes in *Arabidopsis thaliana* are functionally important and evolve slowly. *Mol Biol Evol* **29**, 219-227 (2012).

Figures and Tables

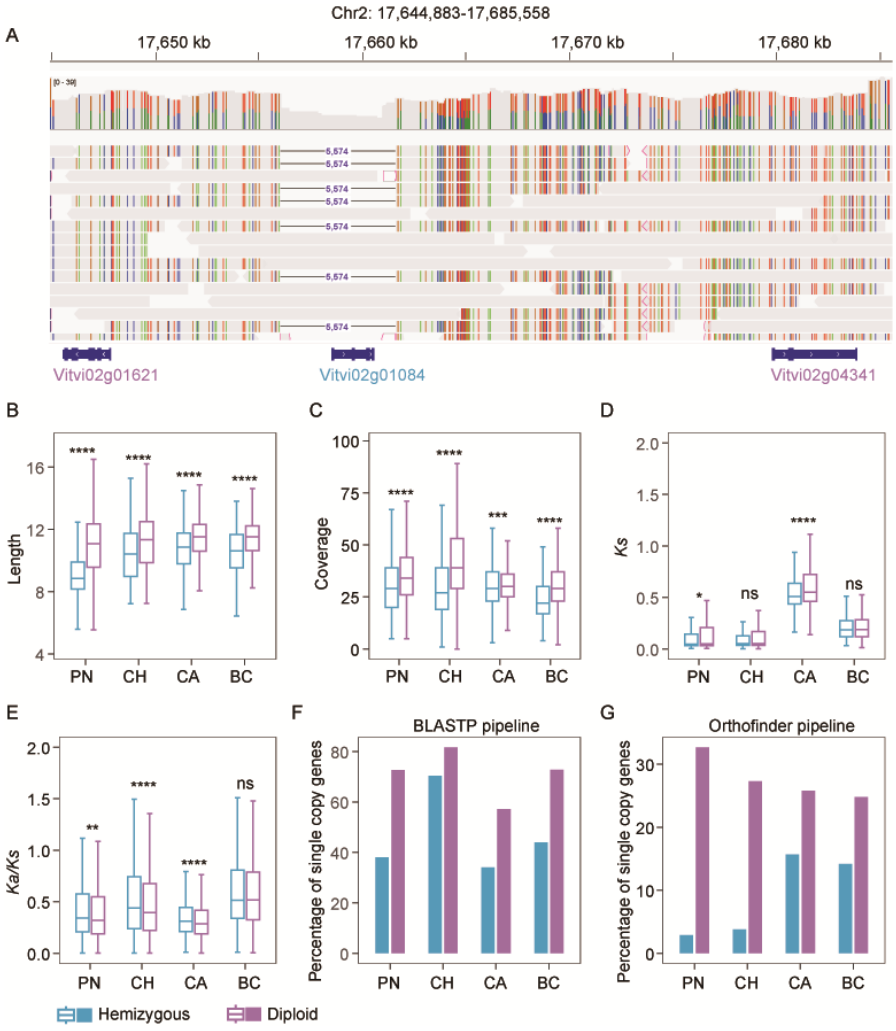

Fig. 1. The identification and characterization of hemizygous genes. (A) an example of a hemizygous gene, shown for Pinot Noir in IGV. (B-G) boxplots comparing genic statistics between hemizygous and diploid genes, including B) the gene length, shown as $\log_2$(length bp), C) the coverage, D) *Ks* and E) *Ka/Ks*. (F, G) the proportion of single-copy genes for hemizygous and diploid genes, respectively, based on the BLASTP and Orthofinder pipeline. In each boxplot, the line in the middle of the box is the median, the edges of the box represent first and 3rd quartile, and the whiskers represent the range. Abbreviations used: PN for Pinot Noir, CH for Chardonnay, CA for Cassava, and BC for Black Cottonwood.

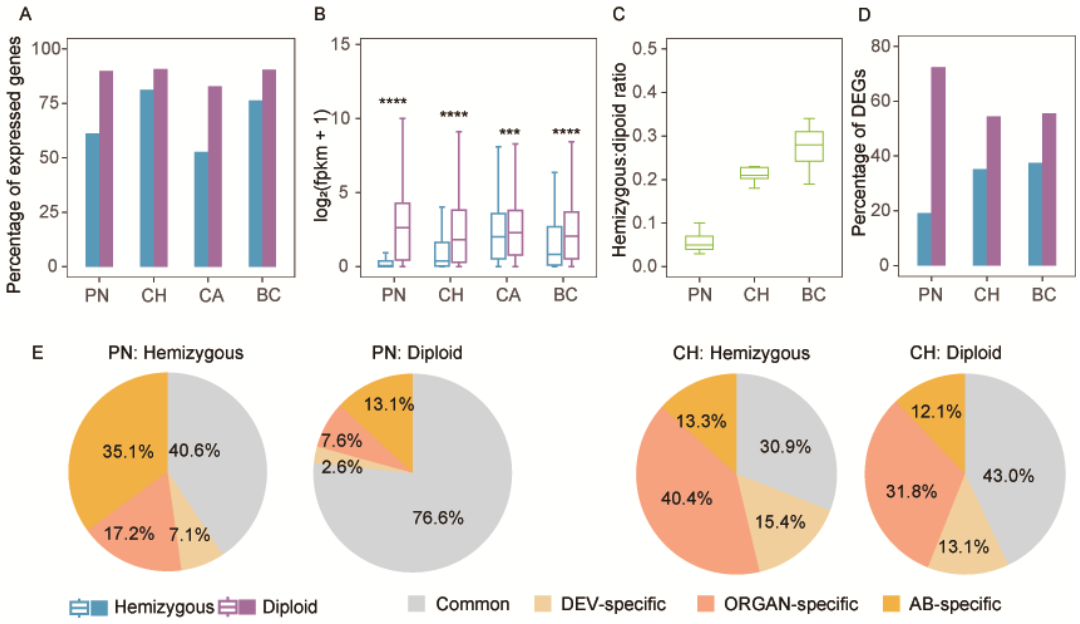

Fig. 2. Gene expression patterns of hemizygous and diploid genes across four taxa, namely Pinot Noir (PN) and Chardonnay (CH), cassava (CA), Black Cottonwood (BC). (A) The proportion of expressed hemizygous genes and diploid genes across all data from the four taxa. (B) The expression level, shown as $\log_2$(fpkm+1), for hemizygous and diploid genes. (C) The distribution of hemizygous: diploid expression ratios from each of the three taxa, cassava was discarded for analysis as it contained only one tissue. (D) The proportion of differentially expressed genes for hemizygous and diploid genes for the three taxa that allow control-treatment contrasts, cassava was discarded for analysis as it contained only one tissue. (E) Pie charts of common and unique differentially expressed hemizygous and diploid genes among three processes, including fruit development (DEV), organ differentiation (ORGAN), abiotic and biotic stress stimulus processes (AB). Data are shown only for Pinot Noir and Chardonnay because they are the only taxa that included all three processes, i.e., fruit development, organ differentiation, abiotic and biotic stress stimulus processes. In B and C, the line in the middle of the box is the median, the edges of the box represent first and 3rd quartile, and the whiskers represent the range.

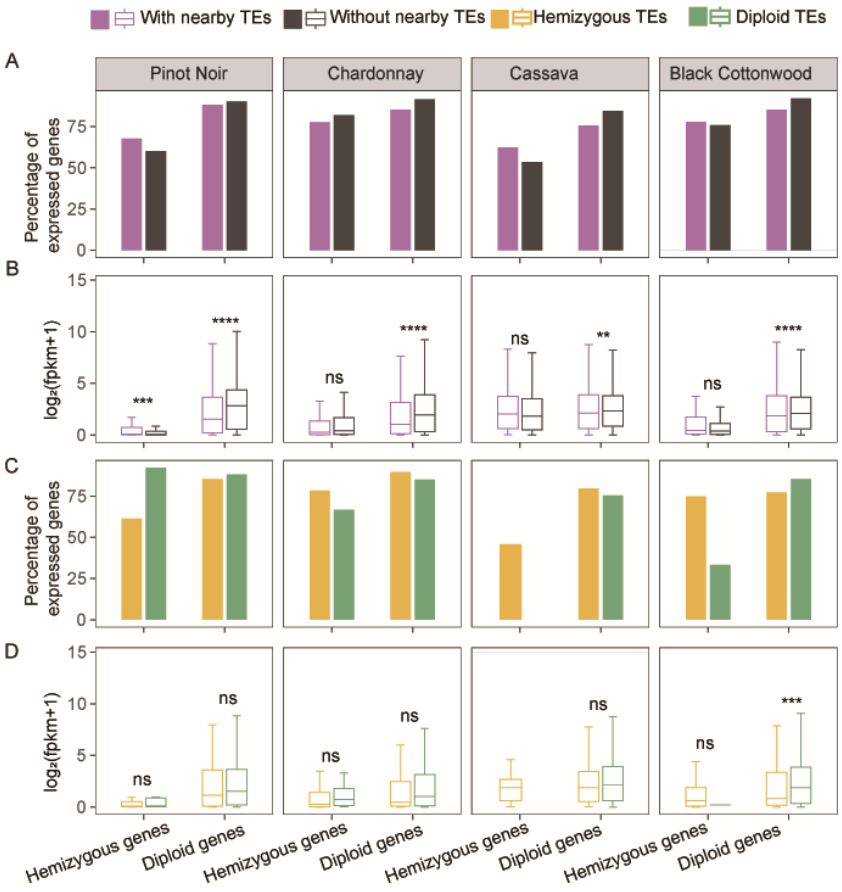

Fig. 3. The cis-regulation of TE effects on hemizygous and diploid genes for four taxa. (A) The proportion of expressed hemizygous and diploid genes with nearby TEs and without nearby TEs across four taxa. (B) Expression levels, shown as $\log_2$(fpkm+1), of expressed hemizygous and diploid genes without nearby TEs and with nearby TEs. (C) The proportion of expressed hemizygous and diploid genes with nearby hemizygous and diploid TEs. (D) Expression levels, shown as $\log_2$(fpkm+1), of hemizygous and diploid genes with nearby hemizygous and diploid TEs. In Figure B and D, asterisks indicate the results of Wilcoxon rank-sum comparisons, with ns: $p > 0.05$, *: $p <= 0.05$, **: $p <= 0.01$, ***: $p <= 0.001$, ****: $p <= 0.0001$; the line in the middle of the box is the median, the edges of the box represent first and 3rd quartile, and the whiskers represent the range.

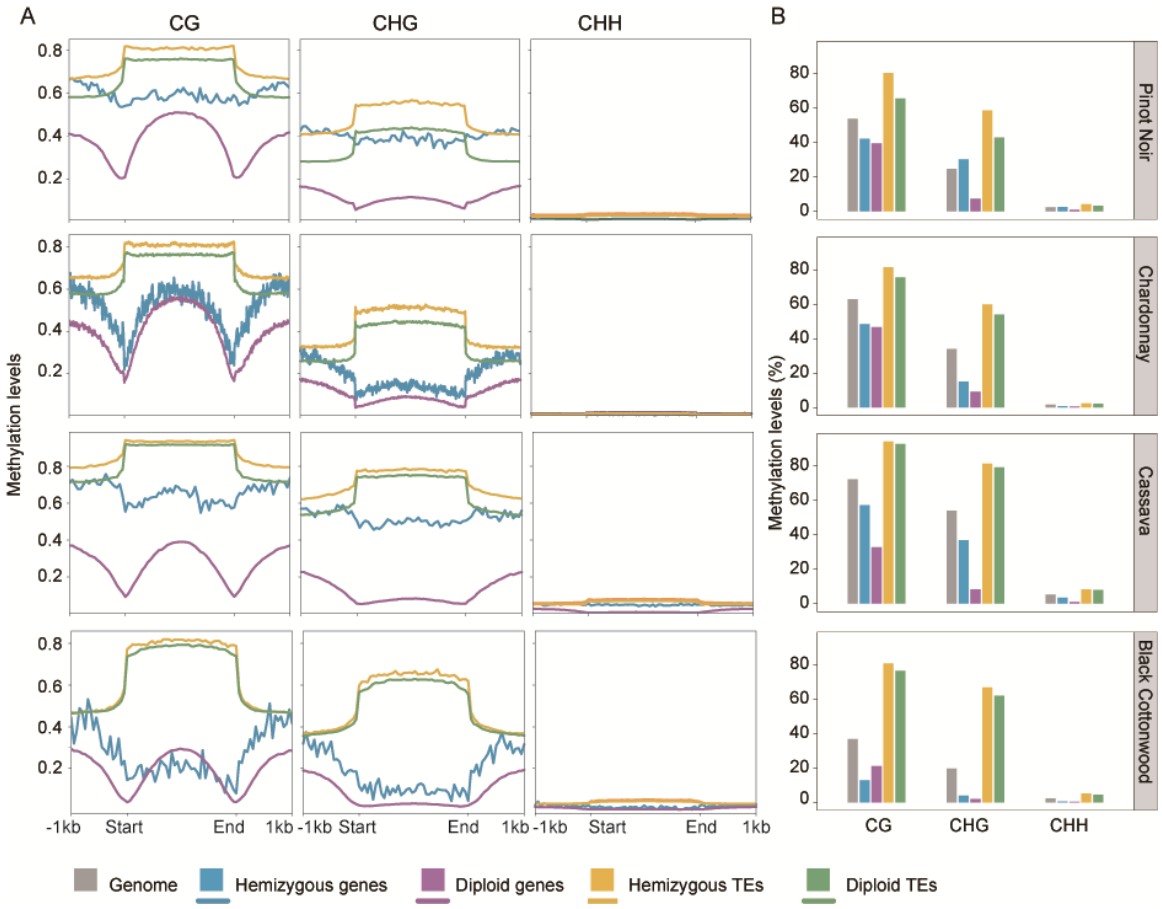

Fig. 4. Methylation patterns of hemizygous genes, diploid genes, hemizygous TEs and diploid TEs in each of four taxa. (A) The global distribution of DNA methylation levels at hemizygous genes, diploid genes, hemizygous TEs and diploid TEs. Start and end denote the transcription start and stop sites of genes or the beginning or end of the TE annotations. The graphs include a 1-kb window upstream and downstream of each feature. (B) The average DNA methylation level for each of the three methylation contexts (CG, CHG or CHH) in each sequence type within each taxon.

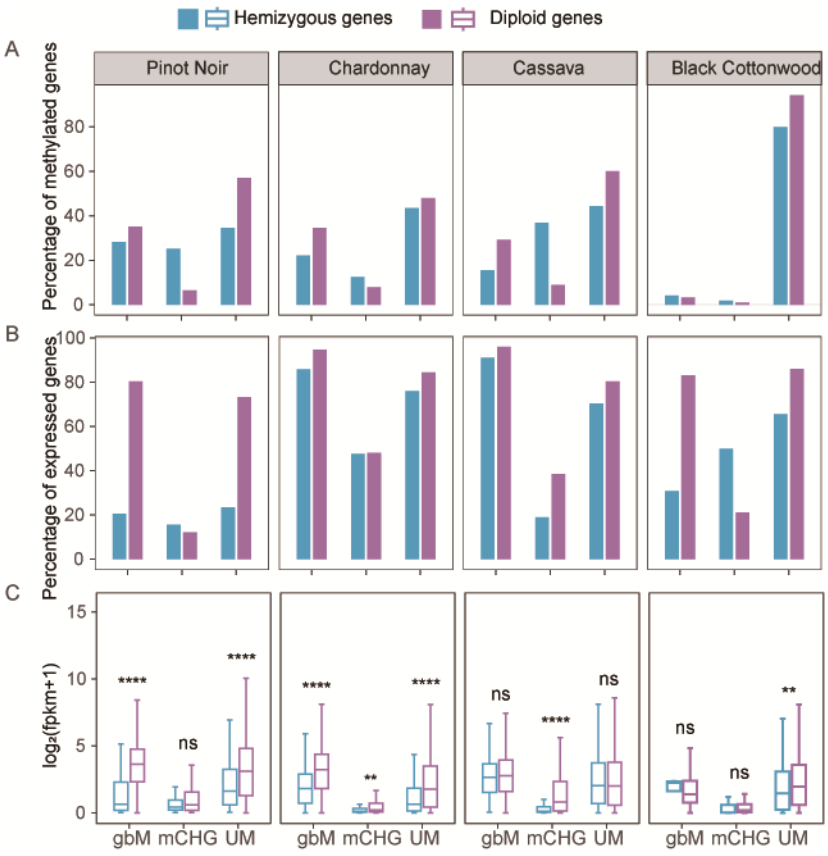

Fig. 5. Epigenetic effects on the expression of hemizygous genes in four taxa. (A) The proportion of body-methylated genes (gbM), CHG methylated genes (mCHG) and unmethylated genes (UM) for hemizygous and diploid genes across four taxa. (B) The proportion of expressed hemizygous and diploid gbM, mCHG, and UM across four taxa. (C) Expression levels, shown as $\log_2$(fpkm+1) of expressed gbM, mCHG and UM genes. The results of Wilcoxon rank-sum test are indicated by ns: $p > 0.05$, *: $p <= 0.05$, **: $p <= 0.01$, ***: $p <= 0.001$, ****: $p <= 0.0001$. In Figure C, the line in the middle of the box is the median, the edges of the box represent first and 3rd quartile, and the whiskers represent the range.

Table 1. Quality of genome assembly, characterization of SVs andidentification of hemizygous genes based on SV inferences in seven taxa.

Plants	**PN**	**CH**	**CA**	**BC**	**Rice**	**Tomato**	**Orange**
Mating system	Clonal	Clonal	Clonal	Outcross	Inbred	Inbred	DH
Genome sizes (Mb)	495.2	606	762.4	392.2	395.8	801.8	334.3
*N*chr	19	19	18	19	12	12	9
*N*scaffolds	0	175	1396	27	0	0	223
Busco(%)	98.3	96.4	99	98.8	98.7	98.6	98.6
50(Mb)	27.1	30.4	35.5	13.2	31.9	67.6	32.3
*N*hSVs	36,556	20,822	46,418	17,708	133	28	191
*N*hDELs	19,935	10,165	29,614	9,949	18	10	52
%*N*hDELs	54.5	48.8	63.8	56.2	13.5	35.7	27.2
hDEL length (Mb)	83.7	90.1	88	25.1	0.02	0.007	0.013
%hDEL length	10.4	14.9	11.5	3.2	0.005	0.001	0.004
*N*genes	33,803	37,243	32,659	34,699	59,903	36,648	26,893
*N*TEs	922,166	1,195,837	826,051	633,676	616,102	637,638	442,868
*N*hemigenes	3,399	5,610	4,242	1,570	2	1	2
%*N*hemigenes	10.1	15.1	13.0	4.5	0.003	0.003	0.007
*N*hemiTEs	119,980	226,295	203,804	39,773	9	8	46
%*N*hemiTEs	13.0	18.9	24.7	6.3	0.001	0.001	0.010

Abbreviations used: PN for Pinot Noir, CH for Chardonnay, CA for cassava, BC for Black Cottonwood, *N*chr for chromosomes numbers, DH for doubled haploid, *N*scaffolds for number of unplaced scaffolds, *N*hSVs for number of heterozygous SVs, *N*hDELs for number of heterozygous deletions, %*N*hDELs for the proportion of heterozygous deletions, %hDEL length for the proportion of heterozygous deletions length, *N*genes for total gene numbers, *N*TEs for total TE numbers, *N*hemigenes for Number of hemizygous genes, %*N*hemigenes for the proportion of hemizygous genes, *N*hemiTEs for numbers of hemizygous TEs, %*N*hemiTEs for the proportion of hemizygous TEs numbers.