A reference-unbiased super-pangenome of the North American wild grape species (*Vitis* spp.) reveals genus-wide association with adaptive traits

Noé Cochetel[1], Andrea Minio[1], Andrea Guarracino[2,3], Jadran F. Garcia[1], Rosa Figueroa-Balderas[1], Mélanie Massonnet[1], Takao Kasuga[4], Jason Londo[5], Erik Garrison[2], Brandon Gaut[6], Dario Cantu[1,7*]

[1]Department of Viticulture and Enology, University of California Davis, Davis, CA, USA.
[2]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA.
[3]Genomics Research Centre, Human Technopole, Viale Rita Levi-Montalcini, Milan, Italy.
[4]Crops Pathology and Genetics Research Unit, United States Department of Agriculture–Agricultural Research Service, Davis, CA, USA.
[5]Horticulture Section, School of Integrative Plant Science, Cornell AgriTech, Cornell University, Geneva, NY, USA.
[6]Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, CA, USA.
[7]Genome Center, University of California Davis, Davis, CA, USA.

*Email: dacantu@ucdavis.edu

Abstract
Capturing the genetic diversity of wild relatives is crucial for improving crops because wild species are valuable sources of agronomic traits that are essential to enhance the sustainability and adaptability of domesticated cultivars. Genetic diversity across a genus can be captured in super-pangenomes, which provide a framework for interpreting genomic variations. Here we report the sequencing, assembly, and annotation of nine wild North American grape genomes, which were phased and scaffolded at chromosome scale. We generated a reference-unbiased super-pangenome using pairwise whole-genome alignment methods, revealing the full resolution of genomic diversity among wild grape species from sequence to gene level. The pangenome graph captured genomic variation between haplotypes within a species and across the different species, and it accurately assessed the similarity of hybrids to their parents. The species selected to build the pangenome ensured a comprehensive representation of the genus, as illustrated by capturing known allelic variants in the sex-determining region and for Pierce's disease resistance loci. Using pangenome-wide association analysis (pan-GWAS), we demonstrated the utility of the super-pangenome by effectively mapping short-reads from genus-wide samples and identifying loci associated with salt tolerance in natural populations of grapes. This study highlights how a reference-unbiased super-pangenome can reveal the genetic basis of the adaptive traits from wild relatives, potentially accelerating crop breeding research.

Introduction

The grapevine (*Vitis vinifera*) is an economically important crop of global significance, cultivated across the world to produce grape-based commodities such as wine, raisins, and grape juice. Historically, its productivity has been endangered by its high susceptibility to biotic and abiotic stresses, leading to the widespread practice of grafting *V. vinifera* cultivars on rootstocks derived from a handful of wild grape species[1,2]. The *Vitis* genus is composed of ~70 interfertile species[3] with the North American wild species holding particular significance due to their extensive genetic diversity and historical contribution to the development of disease-resistant cultivars and rootstocks[2]. The geographical distribution of the grape species in North America contributes to their diverse array of stress tolerance traits (e.g., salt and drought tolerance, resistance to bacterial and fungal diseases, etc). Dioecious and interfertile, natural populations of wild grapes have evolved and adapted to the environmental conditions of their respective habitats and through interspecific introgression[4]. Their tolerance to biotic and abiotic stress enables them to thrive in a variety of climates and ecological niches, making them valuable genetic resources for breeding programs.

Although several grape genomes have been sequenced (e.g., [5–7]), working separately with distinct references impedes the comprehensive understanding of the species' or genus' genetic architecture. Numerous reports indicate that the use of a single reference genome directly impacts the effectiveness of identifying intraspecific genetic variations because reference mapping often fails to capture novel or highly divergent sequences within a species[8,9]. These limitations have motivated the development of pangenomes, which are transformative because they provide a comprehensive view of the genetic landscape that may be otherwise overlooked[9]. The desire to capture an even broader spectrum of genetic diversity, encompassing wild relatives and diverse cultivars, further led to the generation of super-pangenomes, surpassing traditional pangenomes by incorporating genus-level diversity[8]. Super-pangenomes enable the identification of rare or unique genetic variations, population-specific alleles, and adaptive traits, providing a rich resource for crop improvement. Additionally, genus-level pangenomes can shed light on evolutionary history, domestication processes, and genetic relationships within a genus, contributing to a deeper comprehension of their genetic potential[8]. While pangenomes can have multiple definitions, graph-based methods are generally acknowledged to be the most comprehensive for eukaryotic genomes, because they include non-coding regions, regulatory elements, and intergenic regions, thereby enabling a more holistic understanding of genomic architecture and functional elements[10]. To date, however, super-pangenomes in plants have been limited by reducing diploid genomes to a haploid representation, by a constrained input order or phylogenetic dependencies, or by cataloging genetic variants against a single reference. None of these approaches overcome reference biases, and they often do not provide nucleotide-level resolution. Here we overcome these limitations by sequencing and assembling phased diploid genomes and building a super pangenome using all-vs-all whole-genome alignments. This approach is, by definition, reference-unbiased. It therefore overcomes a reliance on reference or tree-guided approaches, in an effort to better capture structural variations between distant species at the nucleotide-level resolution[11].

Here, we describe a reference-unbiased super-pangenome of North American wild grapes in the genus *Vitis.* To build the super-pangenome, we first sequenced and assembled diploid, chromosome-scaled genomes from nine representative species: *V. acerifolia*, *V. aestivalis*, *V. arizonica*, *V. berlandieri*, *V. girdiana*, *V. monticola*, *V. mustangensis*, *V. riparia*, and *V. rupestris*.

From these, we built a super-pangenome graph from all-vs-all chromosome alignments. From the generated graph, we inferred properties of the genic pangenome and extracted inter-species sequence variations. Our work has captured an unprecedented view of genetic diversity among North American wild grape species and effectively identified significant associations with salt tolerance using pangenome-wide association analysis (pan-GWAS).

Results

Assembly, annotation, and phasing of nine North American *Vitis* genomes

To construct the most comprehensive representation of the North American *Vitis* genus, nine accessions were selected to be characteristic of their respective group as defined previously[3] and for their potential in breeding programs (Fig. 1a-j, Supplementary Table 1).

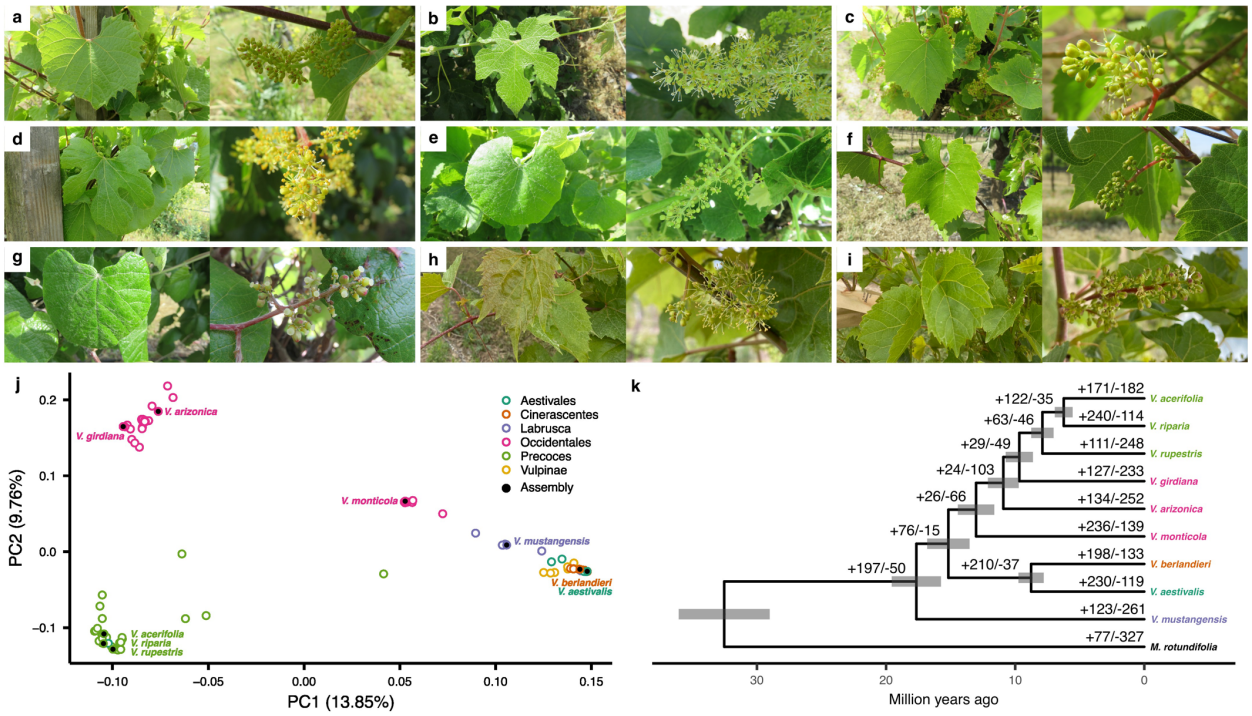

Fig. 1: Selection of nine wild North American *Vitis* species. a, Pictures from a representative leaf (left) and flower (right) of *V. acerifolia* (**a**), *V. aestivalis* (**b**), *V. arizonica* (**c**), *V. berlandieri* (**d**), *V. girdiana* (**e**), *V. monticola* (**f**), *V. mustangensis* (**g**), *V. riparia* (**h**), and *V. rupestris* (**i**), respectively. **j**, Principal component analysis results on SNP data derived from samples of natural populations (n = 90) aligned on *Vitis vinifera* Cabernet Sauvignon. Colors divide the *Vitis* genus into six groups. The accessions selected for assembly are represented by black-colored points. **k**, Phylogenetic relationships among the selected species. The phylogenetic tree was constructed based on 2,338 single-copy gene orthogroups and the branches represent divergence times in million years. *M. rotundifolia* was used as an outgroup to root the tree. Positive and negative numbers indicated on the branches represent expansions and contractions, respectively, as determined by gene family evolution analysis.

Genomes were sequenced using single-molecule real-time sequencing (SMRT; Pacific Biosciences) with 114-255X depth (Supplementary Table 2). Optical mapping (Bionano) data were generated to produce consensus genome maps for seven of the genomes with 791-3,074 depth (Supplementary Table 3). The Falcon-Unzip contigs (N50 = 0.31-1.14 Mb; Table 1) were scaffolded using the genome maps to produce hybrid assemblies (N50 = 1.13-9.22 Mb; Table 1). Protein-coding gene loci were annotated through an extensive *ab-initio* prediction pipeline relying on both Iso-Seq full-length transcripts and RNA sequencing (RNA-Seq) short reads

(Supplementary Table 4). After the generation of consensus models, isoforms prediction, filtering, and functional annotation, 57,003-74,142 genes were annotated in the nine genomes (Table 1), covering on average 33% of the genome. Each species showed a BUSCO score exceeding 95% with over 50% duplication supporting the diploid representation of the assemblies.

Table 1. Statistics for genome assembly and annotation for the nine selected *Vitis* species

Species	Accession	Sequence length (Mb)[1]	Anchored sequence (%)	Contig N50 (Mb)	Scaffold N50 (Mb)	Gene count[1]	Anchored gene (%)	Repeats (%)[1]	BUSCO (%)[1]
V. acerifolia	9018	486|482|55	94.6	0.31	1.74	31,605|31,193|2,751	95.8	44.7|44.8|56.7	90.8|90.5| 5.5
V. aestivalis	T52	548|545|97	91.8	0.96	2.45	32,795|33,222|5,373	92.5	47.8|47.9|53.4	94.6|93.8|10.1
V. arizonica	B40-14	466|423|66	93.1	0.68	9.21	28,043|25,925|3,035	94.7	47.3|47.0|59.8	89.9|83.9| 6.1
V. berlandieri	9031	554|549|114	90.6	1.06	2.68	33,866|33,472|6,804	90.8	47.7|47.5|52.1	95.3|95.0|12.6
V. girdiana	SC2	464|476|65	93.5	1.14	4.08	28,765|29,882|2,884	95.3	47.2|46.4|58.9	92.9|81.7| 5.6
V. monticola	TX6703	521|505|99	91.2	0.57	-	33,281|32,299|5,941	91.7	45.9|45.3|51.3	93.7|88.2|16.9
V. mustangensis	C114-96	454|432|71	92.6	0.49	1.13	27,803|26,177|3,510	93.9	44.5|44.3|56.3	92.8|85.3|11.0
V. riparia	HP-1	520|506|113	90.1	1.09	-	32,717|31,793|6,778	90.5	46.9|46.8|52.6	95.9|91.4|15.5
V. rupestris	B38	449|448|41	95.6	0.80	9.22	29,038|28,798|1,621	97.3	45.4|45.3|58.8	93.2|92.6| 4.6

[1] For the categories of sequence length, gene count, repeats (%), and BUSCO scores, three values are reported. The values represent haplotype 1, haplotype 2, and unplaced sequences.

Nine diploid sets of nineteen chromosomes were reconstructed using HaploSync[12] and a high-density *Vitis* genetic map[13] with a target of at least 90% of anchored sequences and protein-coding gene loci (Table 1). Both the allelic gene content and DNA sequencing (DNA-Seq) short-read coverage (Supplementary Table 5) were used to correct the diploid phasing. The best assembly performance reached 95.6% of sequence anchorage to the chromosomes for *V. rupestris*, representing 97.3% of its gene content. For all the genomes, the unplaced sequences were composed of short sequences with an N50 = 40,653-63,859 bp (Supplementary Table 6), mostly enriched in repeats (Table 1). As expected, the nine *Vitis* genomes presented high levels of heterozygosity, ranging from 16.79 to 23.77% (Supplementary Table 7). Despite such variations at the sequence level, gene content was well conserved between haplotypes, with genic-hemizygosity rates ranging from 2.09 to 4.95%, except for *V. arizonica* (9.35%) (Supplementary Table 8). The complete diploid representation of the genome assemblies was further supported by their length; representing approximately twice the haploid sizes estimated by flow cytometry, but also by the presence of 1.90-2.27X the haploid gene content of PN40024[5] (Supplementary Table 9). Altogether, the assembly of these nine high-quality genomes provided the foundation to build a comprehensive super-pangenome using comparable assembly/scaffolding qualities.

Given the nine reference genomes, we aligned in a pairwise configuration the nine haplotype 1 proteomes, resulting in the identification of 2,338 single-copy orthogroups (SCO). We gathered the SCO alignments into a supermatrix and used it to produce a species-wide phylogeny that matched previous inferences[14,15] and separated species by known *Vitis* groups[3] (Fig. 1k, Supplementary Fig. 1). We also estimated divergence times at each node, using clock-calibration (Fig. 1k). Finally, after the selection of one representative isoform per gene, we used the same proteomes to evaluate gene families across the North American *Vitis* phylogeny, identifying 450 rapidly evolving orthogroups (Fig. 1k). Interestingly, the gene families involved in the

contraction/expansion events were significantly enriched for gene ontology (GO) terms such as response to stress, signaling, response to stimulus, and cell communication, suggesting that they might be involved in the adaptation of each species (Supplementary Table 10).

The nucleotide-level super-pangenome graph of the North American *Vitis* genus

The 18 haplotype genomes (9 genomes x 2 haplotypes) were considered as distinct inputs for the super-pangenome construction. After pairwise comparisons of assemblies, we inferred that from 27.10 to 34.20% of the genome from each species was affected by genetic variants (Supplementary Table 11), values that greatly exceeded heterozygosity. We used nucleotide variants to estimate an average rate of $2.4 \times 10^{-8} \pm 6 \times 10^{-9}$ variants per nucleotide per year between *Vitis* genomes using the divergence times inferred in the clock-calibrated phylogeny (Fig. 1k). To avoid any bias towards a specific reference, the pipeline to construct the pangenome relied on all pairwise sequence alignments of the 18 genomes[16]. The workflow followed the PanGenome Graph Builder (PGGB) pipeline[11] which involves three main steps; i) wfmash to generate all the pairwise alignments, ii) seqwish to induce the pangenome graph from the alignments, and iii) smoothxg to polish the pangenome graph.

The generated pangenome graph had a compression ratio per chromosome of 81% on average (Fig. 2a) and contained 342 paths, each representing a chromosome of the input haplotype genomes. The graph structure was composed of ~ 200 million nodes (Supplementary Fig. 2) for a total length of ~ 1.7 Gbp (Fig. 2b) and representing ~1.5X the size of the diploid genome for a single *Vitis* species. Nodes were categorized into three different classes according to the number of species containing them: i) core genome; nodes present in the nine species, ii) dispensable genome; nodes present in more than one species but not all, and iii) private genome; species-specific nodes. The nodes in the private genome covered the largest portion of the pangenome (1.1 Gb). However, most of the nodes (> 100 million) were in the dispensable genome while both the core and the private genomes were composed of about 50 million nodes each (Supplementary Fig. 2). At the haplotype level, the core genome represented about half the genome size (48% on average), while the dispensable and the private genomes contained on average 36% and 16% of the genome, respectively (Fig. 2c).

One interesting feature of these analyses was that the private genome was composed of a significantly higher proportion of repeats compared to the other classes (Kruskal-Wallis, P value < 0.05). The private genome was ~ 40% more repetitive than the core genome (Fig. 2d). Class I transposable elements (TEs) represented the majority (80.3%) of the TEs in the pangenome (Supplementary Fig. 3). Remarkably, 56% of the private TEs were long terminal repeats (LTR) *Gypsy* retrotransposons, which represented only 24% of the core TEs (Fig. 2e). The over-representation of Gypsy elements in the private genome suggests they are strongly associated with genome divergence between species and may play a pivotal role in grape genome evolution.

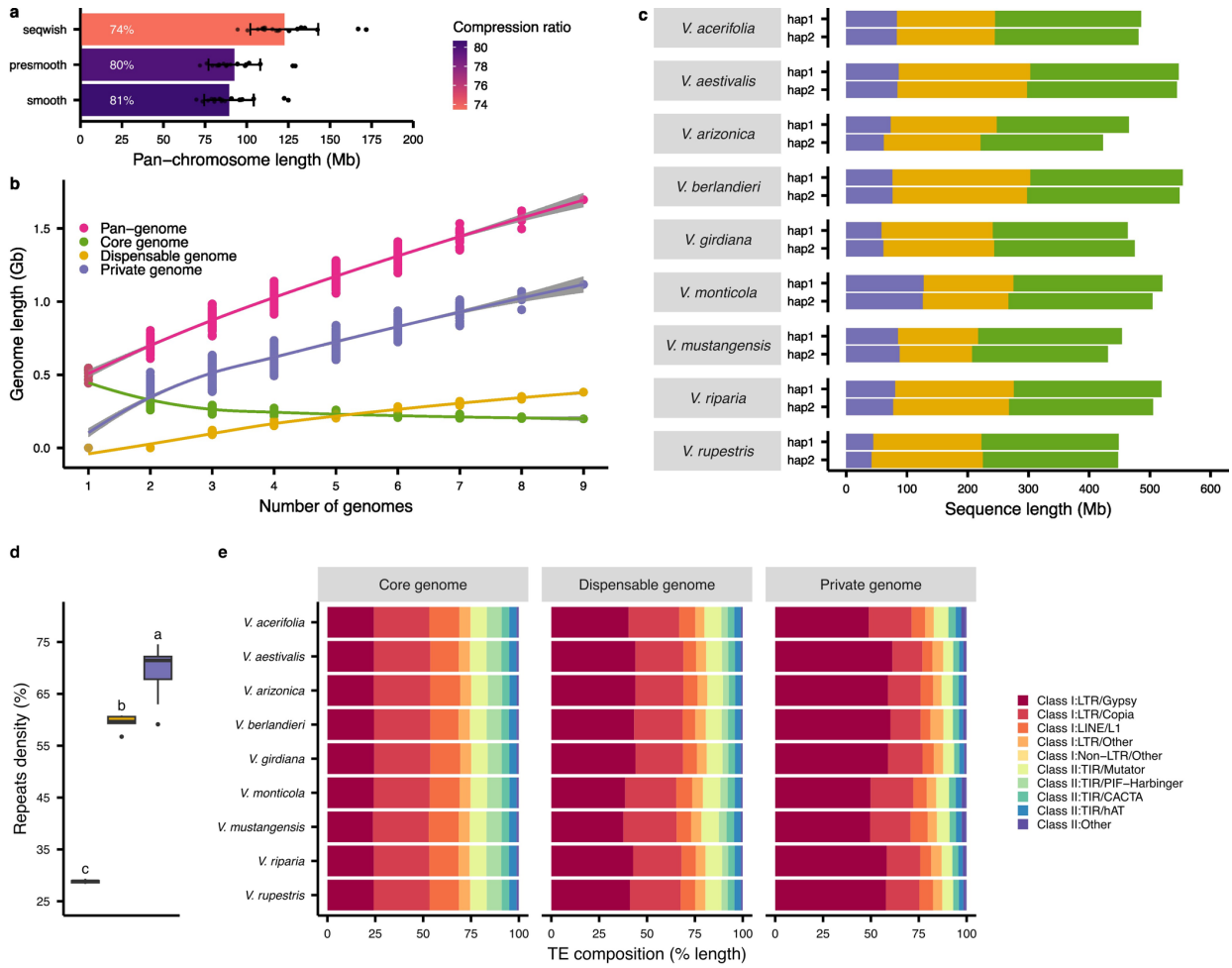

Fig. 2: Properties of the nucleotide-level super-pangenome graph. a, Compression ratio of the pangenome. For each step of the pangenome construction, raw (seqwish) and post-polishing (smoothxg), the mean ± sd (n = 19 chromosomes) of the total length of the unique nodes representing each chromosome is represented. The compression is expressed as a ratio of the total size of a chromosome in the graph compared with the sum of the lengths of the same chromosome in the 18 haplotypes used as input. **b**, Graph-based pangenome modeling. For every combination of 1-9 genomes, the total length of the unique nodes is represented per class. The line represents smoothed conditional means with a 0.95 confidence interval. **c**, Pangenome class composition per haplotype for each genome. **d**, Repeat density observed in the different node classes of the pangenome. Significant groups were determined using a multiple comparison test after Kruskal-Wallis with a *P* value < 0.05. The middle bars represent the median while the bottom and top of each box represent the 25th and 75th percentiles, respectively. The whiskers extend to 1.5 times the interquartile range and data beyond the end of the whiskers are plotted individually as outlying points. Panels **b**, **c**, and **d** share the same color legend. **e**, Transposable elements composition in the different pangenome classes. Categories with unspecified repeats or representing < 2% of the repeat content were classed as "Other".

Intra and inter-specific relatedness is captured in the super-pangenome

To evaluate the sensitivity of our pangenome, we added the previously published diploid genomes[7] of three hybrids to the pangenome, all derived from the cross of North American species included in our pangenome: 101-14 Millardet et de Grasset (101-14 Mgt: *V. riparia* x *V. rupestris*), Richter 110 (110R: *V. berlandieri* x *V. rupestris*), and Kober 5BB (*V. berlandieri* x *V. riparia*). We incorporated these hybrid genomes to evaluate whether we could accurately assess their origins and to characterize their effects on genomic components. As expected, the inclusion of the hybrid

genomes led to a decrease in the relative private and core genome sizes in favor of a larger dispensable genome, both in terms of length and node number (Supplementary Fig. 4).

To evaluate genome relatedness, we focused on the three hybrids and the three parental species from which the hybrids were derived. Three levels of comparisons were used: intra-genome by the comparison of the two haplotypes within each genome, inter-haplotype by the comparison of the hybrids' haplotypes to their parents, and inter-genome comparing haplotypes of different species. For each of the pairwise comparisons, we used the core and the private genome size to assess the haplotype similarity (there is no dispensable genome in such a pairwise configuration). First, intra-genomic comparisons were performed between the two haplotypes of each genome; on average, 57.2% was shared between the two haplotype sequences of each species (Supplementary Fig. 5a). Notably, the hybrids having *V. berlandieri* as a parental species showed the highest percentage of private nodes (50.1% on average), probably because this species is genetically distant from the two other parents *V. riparia* and *V. rupestris* (Supplementary Fig. 5b, Fig. 1k). Secondly, we evaluated if the pangenome could detect the relatedness of the hybrids with their parental species. Each hybrid haplotype was derived from a different species, as determined during their assembly[7]. When we compared each hybrid haplotype (e.g. 110R haplotype berlandieri) with the haplotypes of its parental genome (i.e. *V. berlandieri* haplotype 1 or 2), we found in average 55.8% of their sequences (nodes) in common (Supplementary Fig. 5c,d). Lastly, the smallest core genome sizes were detected when the genomes were compared in an inter-specific configuration, making the private genome the largest portion of the pangenome as expected when comparing distant species (Supplementary Fig. 5e-h). Altogether, these results illustrate the importance of haplotype-resolved genomes, demonstrate that the super-pangenome graph captures both intra and inter-specific variants, and accurately represents the relationship between parental species and hybrids, thus reinforcing the high sensitivity of the super-pangenome.

Gene-based pangenome inferred from the pangenome graph

We used the super-pangenome to study the distribution and characteristics of genes among the nine *Vitis* species. Genes were categorized into three different genome classes according to their node composition: i) core genes, when the gene sequence was composed of >80% of core nodes, ii) dispensable genes, when the sequence was composed of >80% dispensable nodes, and iii) private genes when the sequence was composed of >80% private nodes. The genes not falling in any of the above categories were classified as ambiguous; these represented on average only 2.6% of the gene content and were discarded for subsequent analyses (Supplementary Table 12). We also examined these categories as a function of species sampling: as genomes were added to the pangenome, the number of private genes continuously increased, inflating the overall gene pangenome size (Fig. 3a). In contrast, the number of core genes tended to slowly stabilize around 40,000 genes (Fig. 3a, Supplementary Table 13), representing more than 67% of gene content per species on average (Fig. 3b, Supplementary Table 13). Although the number of private genes continually increased with the addition of species, private genes represented were less numerous than core ranging from 2,580 to 9,145 genes per species.

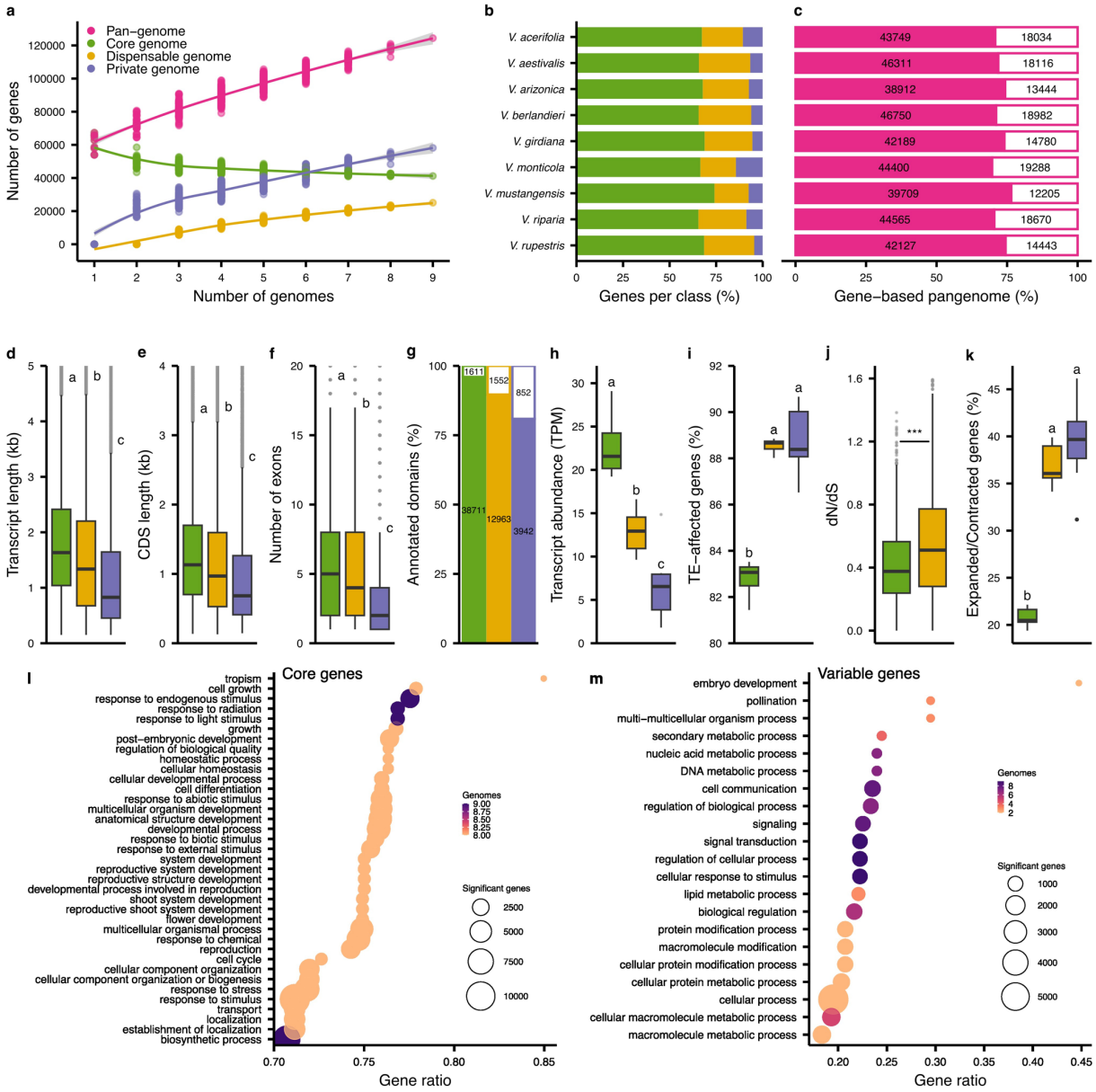

Fig. 3: Characterization of genes in the super-pangenome. a, Gene-based pangenome modeling. For each combination of 1-9 genomes, the number of genes is represented per class. The lines represent smoothed conditional means with a 0.95 confidence interval. **b**, Gene class composition for each genome. **c**, Percentage of genes attributed to the same class in both the graph-derived and orthology-based pangenomes. The number of consistent genes between both approaches is represented on the left of the bar while the different are on the right. **d**, Transcript length (kb); **e**, CDS length (kb); **f**, Number of exons; **g**, Percentage of annotated domains; **h**, Transcript abundance (TPM); **i,** TE-affected genes; **j,** dN/dS ratios, and **k**, Expanded/Contracted gene families per class of genomes. *P* values were determined using ANOVA and significant groups were assigned following Tukey's "Honest Significant Difference" method. The middle bars represent the median while the bottom and top of each box represent the 25th and 75th percentiles, respectively. The whiskers extend to 1.5 times the interquartile range and data beyond the end of the whiskers are plotted individually as outlying points. For the dN/dS ratios (**j**), *P* values were determined using a two-tailed Student's t-test. To prevent the compaction of the y-axis by extreme outlying points, the upper limit was capped. Panels **a**-**k** share the same color legend. Enriched gene ontology terms for the core genes (**l**) and the variable (dispensable + private) genes (**m**) are represented as circles with their size depending on the number of genes involved and their color based on the number of genomes having the term detected as significant. For each GO term, the number of significant genes relative to the total number of annotated genes for the term is represented as a gene ratio on the x-axis.

These results were supported by a classic orthology analysis based on colinearity. On average, 72.6 ± 2.3% of the genes were consistently categorized as core, dispensable, or private genes between the graph-inferred and the orthology-based gene pangenomes (Fig. 3c, Supplementary Table 14). The core genome was the most conserved class, with 85.8 ± 1.76% consistency. The three classes of genes also exhibited remarkable structural differences. Dispensable and private genes were significantly shorter than the core genes, with correspondingly smaller transcript and CDS lengths and fewer exons (Fig. 3d-f). The longer core genes had more annotated domains (Fig. 3g). A striking association was also observed between gene expression and gene conservation within the genus. Transcript abundance was significantly higher for core genes, suggesting functional conservation relative to dispensable and private gene sets (Fig. 3h). The variable (i.e., non-core) genes had higher TE density around them (Fig. 3i), which may suggest that TEs contribute to their lower expression.

We also investigated the evolutionary properties of genes. For example, we calculated non-synonymous/synonymous substitution (dN/dS) ratios (Fig. 3j) for each gene class. On average, dispensable genes had higher average dN/dS values, reflecting higher sequence conservation among core genes. The proportion of rapidly evolving gene families identified during the phylogeny analysis was twice as high in variable genes (dispensable/private) (Fig. 3k) compared to core genes (Kruskal-Wallis, P value < 0.05). This high proportion further suggests the possibility that dispensable genes contribute to the local adaptation of each species (Supplementary Table 15). Gene ontology (GO) enrichments supported this conjecture. For example, core genes were mainly associated with essential developmental functions such as the shoot system, flower development, and responses to stimulus. In contrast, variable genes were enriched in metabolic processes, cellular processes, and signaling (Fig. 3l,m). We focused particularly on nucleotide-binding site leucine-rich repeat (NLR) genes because they have been heavily studied in grapes for their association with disease resistance. Interestingly, NLRs possessing a toll/interleukin-1 receptor-like (TIR) domain, the TIR subclasses such as TIR-NBS-LRR genes, tended to be more abundant in the dispensable genome (Supplementary Fig. 6). Our results consistently supported the strong, conservative selection of core genes, presumably due to their essential functions. The pool of private and dispensable genes does not seem to be indispensable, at least under certain environmental conditions, meaning either that they are completely disposable or, more likely, provide benefits for local adaptation to specific abiotic and/or biotic stresses. They may thus be responsible for significant phenotypic variation among species.

Structural variants derived from the pangenome graph topology

By definition, the pangenome graph embeds polymorphisms and structural variants for all the pairs of genomes used as input. To extract this level of information from the graph, we employed the module "deconstruct" from the tool vg v1.48.0[17], selecting iteratively each genome in the graph as a reference. A variant analysis performed independently between genomes using NUCmer[18] provided similar results; the pangenome and NUCmer-based inferences were consistent across presenting 90% and 86% of SNPs and INDELs, respectively (Supplementary Table 16). SNPs were the most abundant variant type in the pangenome (Fig. 4a). Within individual genomes, SNPs remained the most numerous variant type, with insertion events being the least abundant (Fig. 4b). We classified based on size, showing that 1-bp variants (most SNPs) were most abundant and also that most of the remaining variants were shorter than 10 bp (Fig. 4c).

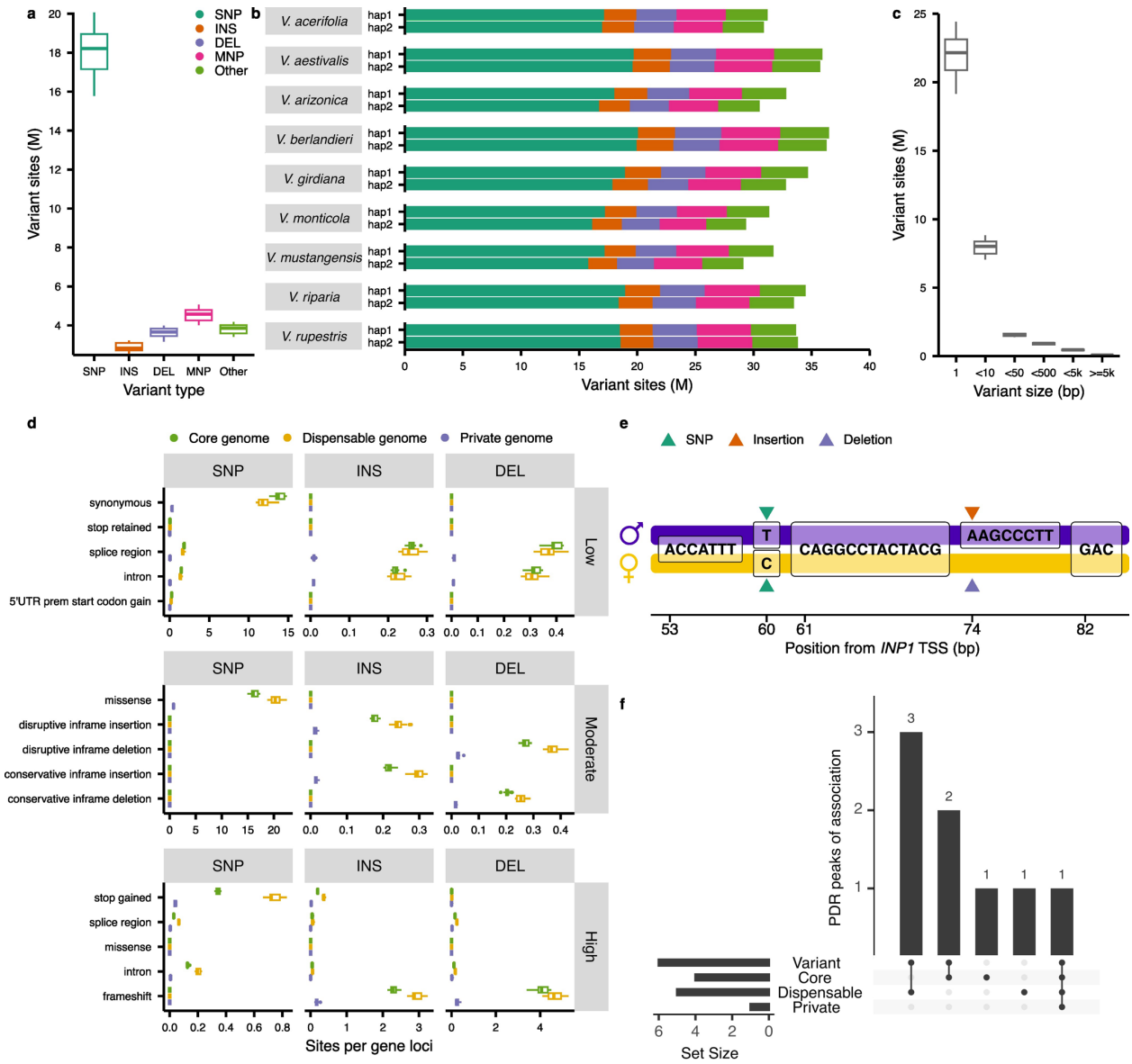

Fig. 4: Genetic variants from the graph topology. a, Total number of variants per class (n = 18 haplotype genomes; 9 genomes x 2 haplotypes). **b**, Number of variant sites in each haplotype per genome. Panels **a** and **b** share the same color legend. **c**, Distribution of variants by size. **d**, Functional impacts of the different variant types after annotation using SnpEff. The middle bars represent the median while the bottom and top of each box represent the 25th and 75th percentiles, respectively. The whiskers extend to 1.5 times the interquartile range and data beyond the end of the whiskers are plotted individually as outlying points. **e**, Structural variants at the sex-determining region within the *VviINP1* gene. The x-axis represents the position from the transcriptional start site (TSS) of *VviINP1*. The two horizontal bars represent the alleles of the male (purple) and the female (yellow) *VviINP1* coding sequence. Common nucleotides are presented in the middle of the two bars while allele-specific bases are indicated on the corresponding horizontal bar. Variants are indicated as triangles, SNPs in the two references are colored in green, the insertion in the male is in orange, and the deletion in the female is in purple. **f**, Consistency of the detected variants in the pangenome graph with the previously PDR-associated regions. The vertical bars represent the number of PDR-associated regions per genome class and detected or not as variants in the pangenome graph. The horizontal bars represent the total number of peaks corresponding to the variables indicated in the different rows.

We annotated the variants using snpEff v5.1[19] to estimate their impact on the protein sequence and noticed an interesting pattern. Generally, low-impact variants tended to be similarly distributed between the core and the dispensable genomes (Fig. 4d, except for the multiple nucleotide

polymorphisms (MNP); Supplementary Fig. 7a). However, when the impact on the protein sequence was moderate or high, most of the variants were found in the dispensable genome rather than in the core genome (Fig. 4d). The most striking evidence was illustrated by the SNPs. While they were similarly abundant in both the core and the dispensable genomes for synonymous polymorphisms (low impact), missense polymorphisms (moderate impact) were more abundant in the dispensable genome. The difference was even larger for the number of stop gains detected in genes (high impact variant), with ~2x more variants in the dispensable genome (Fig. 4d). The private genome was particularly less impacted by variants when compared with the core or dispensable genomes, however, it contained a similar number of modifier sites per gene loci than the core genome specifically in intergenic regions (Supplementary Fig. 7b).

We validated the nucleotide-level variants derived from the pangenome graph topology by investigating the structure of a gene from the grape sex-determining region[6], *VviINP1* (Fig. 4e). Remarkably, the super-pangenome graph captured the known variants distinguishing the *INP1* male and the female alleles, including an allele-specific SNP at 60 bp from the *VviINP1* transcriptional start site (TSS) and the 8-bp INDEL starting at 74 bp from the TSS. Moreover, the flower sexes of all the accessions used to build the pangenome were predicted correctly by their alleles in the pangenome and confirmed with phenotyping data from the field (Fig. 1a-i). To further explore the variants embedded in the pangenome graph, we also investigated polymorphisms associated with Pierce's disease resistance. Eight peaks of association were previously detected in *V. arizonica*[20] (one of the input genomes integrated into the pangenome) using genome-wide association studies (GWAS). The nodes of the pangenome graph corresponding to those regions were extracted (Fig. 4f). For 7 of the 8 association peaks, SNPs were detected in nodes that were either considered as variants from the graph topology or in the variable genome (private or dispensable). In other words, almost the entirety of the regions previously detected *de novo* by variant calling was represented in the pangenome graph structure. As a result, mapping the reads used in the previous GWAS should in theory only require genotyping to detect associated polymorphisms, thereby bypassing the variant calling process.

Pan-GWAS revealed loci associated with salt tolerance

In a previous study, 325 accessions from natural populations of 14 *Vitis* species were collected and evaluated for chloride exclusion, a process that confers salt tolerance in grapes[21]. Among the accessions used to build the pangenome, five were included in this earlier work and presented pronounced differences regarding salt accumulation in the roots and the leaves. In particular, *V. girdiana* accession SC2 was a high salt excluder. We sequenced the DNA of 153 samples from 12 species of the collection panel to perform a pangenome-wide association analysis (pan-GWAS). The variants extracted from the graph were used to construct a simplified version of the pangenome using haplotype 1 from *V. girdiana* as a reference (Fig. 5a). In this version of the graph, the representation of the sequence variations within *Vitis* was relative to *V. girdiana*. The reduction of the number of paths in the graph from 342 to 19 allowed us to prepare the pangenome for read mapping by significantly reducing the computational requirements. We indexed this reduced pangenome and mapped reads against it using vg. For each sample, >95% of the paired-end reads mapped onto the pangenome (Supplementary Table 17). After computing the snarls and the pack files from the graph, vg call was used to genotype the variants present in the graph (Fig. 5a).

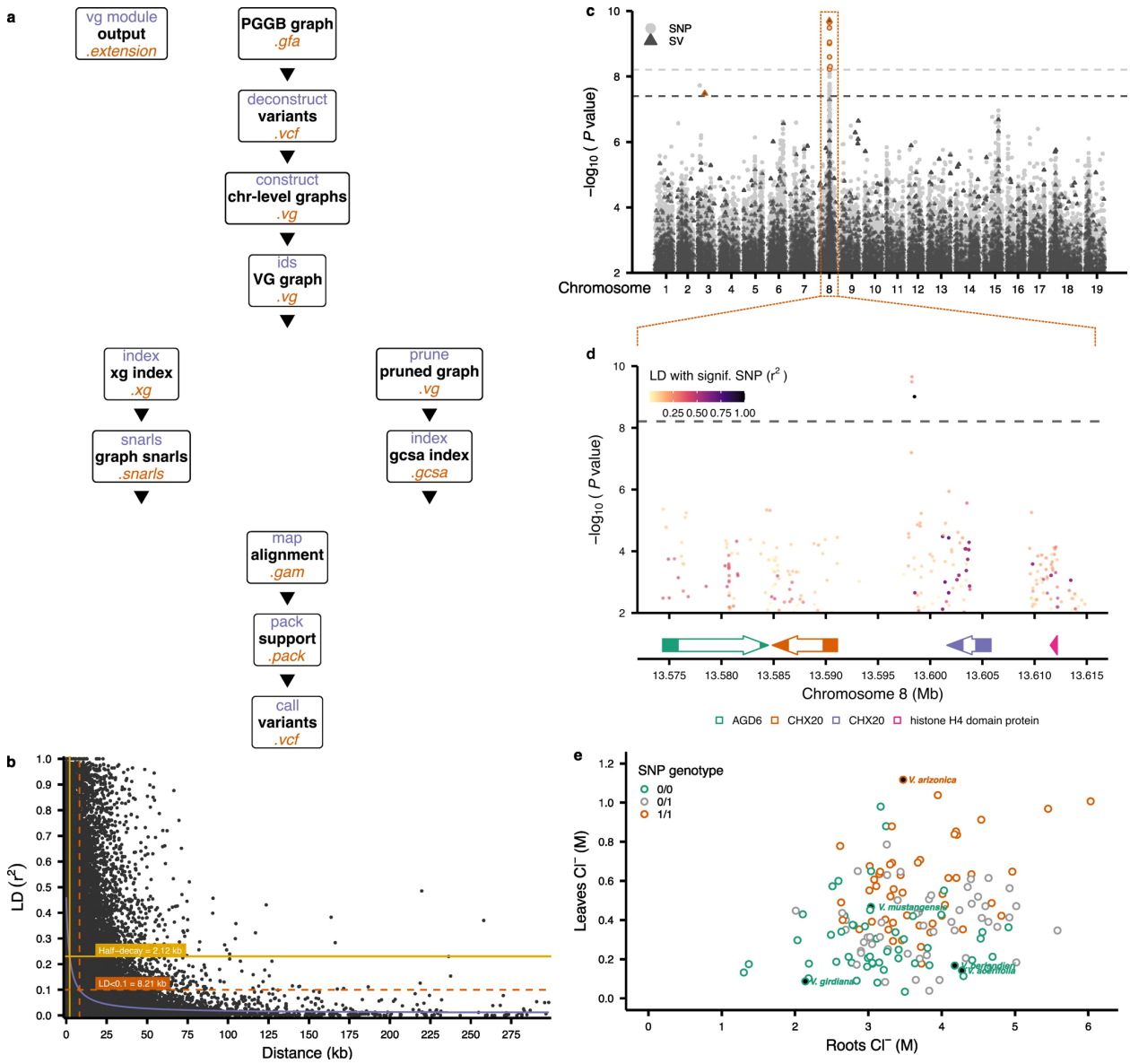

Fig. 5: Pan-GWAS revealed significant associations with chloride concentration in the leaves. a, Workflow to generate the variant file required for the pan-GWAS. The name of the vg module used for each step is indicated in purple, the output type is in black, and the file extension is orange. The PGGB graph is deconstructed to a variant file using *V. girdiana* haplotype 1 as a reference. Vg construct is used to construct a pangenome from the VCF and the reference fasta. After merging the IDs, indices are created to proceed with the alignments with vg map. The read support (pack) is extracted from the gam files and the vg call is finally performed. **b**, Linkage disequilibrium decay in the SNP dataset used for the pan-GWAS. Half-decay is represented in yellow, while minimum decay with a threshold set to LD < 0.1 is colored in orange. The number of points was randomly down-sampled to 50k for visual representation. **c**, Manhattan plot representing significant association with the chloride concentration in the leaves. Significant associations are detected using a Bonferroni threshold set to -log10(0.05/n) and colored in orange. The SNPs are represented as light grey points, and the SVs as dark grey triangles. **d**, Manhattan plot focused on the significant region in chromosome 8 outlined by the orange box in **c** for the association with the leaf chloride concentration. The filling color is based on the pairwise linkage disequilibrium with the SNP located at chr08:13598495. The gene structural annotation of this region is represented at the bottom. **e**, Chloride concentrations (M) in the roots (x-axis) and in the leaves (y-axis) from the different samples used for the GWAS. Samples corresponding to the accessions assembled and used in the pangenome are filled in black. The colors represent the genotype at the significant SNP, homozygous reference (0/0) in green, heterozygous (0/1) in grey, and homozygous alternative (1/1) in orange.

After filtering, we examined the distribution of SNPs in the 19 paths. The SNPs exhibited rapid decay of linkage disequilibrium (LD), reaching half the maximum average r2 at 2.12 kb (Fig. 5b), consistent with previous reports in grapes[14,22]. We also performed pan-GWAS separately using 8,091,983 SNPs and 1,261,953 SVs, which should provide sufficient statistical power[23]. For both SNPs and SVs, a significant peak of association was detected on chromosome 8, and a single 1-bp deletion was found on chromosome 3 (Fig. 5c). Those genotyping results were supported by variant calling from a standard GWAS approach based on the *V. girdiana* reference (Supplementary Fig. 8a).

The significant SNP detected in both pan- and standard GWAS was located in an intergenic region at chr08:13598495 (Fig. 5d) and was tightly associated with chloride concentration across samples (Fig. 5e). The SNPs in strong linkage disequilibrium with the significantly salt exclusion-associated SNP were located downstream, overlapping a gene locus spanning chr08:13601610-13605816 (Fig. 5d). Interestingly, the predicted protein from this region (VITVgdSC2_v1.0.hap1.chr08.ver1.0.g144890.t01) is a homolog of the *Arabidopsis thaliana* cation/H^+ exchanger AtCHX20 (AT3G53720). The putative function of this gene was further supported by the presence of two notable InterPro domains: the Cation/H+ exchanger (IPR006153) and the Sodium/solute symporter superfamily (IPR038770). In *V. vinifera* PN40024, the closest homolog is *Vitvi08g01174*, which is known to be induced by salt treatment[24]. In soybean (*Glycine max*), the closest homolog of VITVgdSC2_v1.0.hap1.chr08.ver1.0.g144890.t01 is GmNcl1 (Glyma.03g171500), which is also known to be induced by salinity[25]. Interestingly, the closest soybean homolog to AtCHX20 is GmSALT3 (Glyma.03g32900), the major salt tolerance gene in *G. max*[26]. The gene upstream of the significant SNP was also found to be homolog to *AtCHX20* (Fig. 5d). Altogether, these results show that our super-pangenome allows the detection of base-level associations.

Discussion

Wild *Vitis* species represent a valuable genetic pool for grape cultivars and rootstock improvement. Despite an increasing number of *Vitis* genomes being published, no integrative analysis has been performed yet, limiting the exploration of such a diverse repertoire. Here we assembled the genomes of nine North American wild *Vitis* species representative of the diversity of their genus (Fig. 1), scaffolded them at chromosome level, and then phased haplotype (Table 1) to comprehensively assess the structural variations present in complex and repetitive regions in diploid genomes[27]. These accessions included species that contain characteristic agronomic traits, such as disease resistance and drought tolerance. The accessions also include *V. berlandieri*, *V. riparia,* and *V. rupestris*, the wild species trio responsible for the production of most of the grape rootstocks[28].

In recent years, pangenomes have been generated for numerous crop species. However, the concept of super-pangenomes is still in its early stages of development, although notable examples can be found in tomato[29], rice[30], sorghum[31], and potato[32]. Our super-pangenome differs from these previous efforts by using a reference-unbiased, graph-based approach. Its construction relied on all-vs-all sequence alignments of chromosome-level diploid assemblies thereby providing access with a nucleotide resolution to intra- and inter-specific genetic variants, beyond the limited gene space and overcoming reference-bias (Fig. 2). Our efforts provide an exemplar for future plant super-pangenomes, but they also pave the way for future work in *Vitis*. For example, this super-

pangenome will be useful for additional comparative pangenomics analyses, and it will provide a backbone for including the wine-producing cultivars of *V. vinifera,* helping to identify and study variants of potential agronomic utility.

Our super-pangenome has provided insights into repetitive and genic content throughout the genus. For example, our characterization of the repetitive elements has shown that LTR Gypsy retrotransposons are enriched within private genomes. They are thus a predominant driving force behind the variable portion of the pangenome. In addition to their impact on genome size, TEs are known for impacting gene expression (Fig. 3h,i). With respect to genes, we have found (unsurprisingly) that most genes are in the core genome. These genes are enriched for essential functions. In contrast to the core genome, the variable genome was significantly more impacted by TEs (Fig. 3) and SVs (Fig. 4), and contained a higher proportion of expanded/contracted gene families, potentially comprising the adaptative gene repertoire from the wild species. Including genomes from domesticated cultivars is likely to further expand the size of the dispensable genome, for two reasons. First, the dispensable genome tends to increase with the addition of more samples (Fig 2a, 3a), and this trend is likely to continue. Second, domesticated cultivars, which are clonally propagated, have much higher rates of genic hemizygosity (typically 10 to 15%)[33,34] compared to the < 5% found in the outcrossing species in this study (except for *V. arizonica*). Thus, cultivars may have a larger proportion of their genes captured in the dispensable genome.

We have demonstrated the utility and accuracy of the super-pangenome in three ways. First, by investigating hybrid genomes, we have shown that the super-pangenome accurately identified their parental origins and accurately characterized origin at the haplotype level. Second, we have recapitulated valuable genetic variants at the nucleotide level, specifically variants in the sex-determining region (Fig. 4e) and that are associated with Pierce's Disease (PD) (Fig. 4f). The first led to the prediction of the flower sex phenotype of each accession by the presence/absence of nodes from the graph in the *INP1* gene, bypassing the need to perform manual gene annotation. For the latter, an interesting feature of the PD-associated variants is that they were more numerous in the variable (dispensable or private) genome compared to the core genome (Fig. 4f). This observation supports the notion that the variable genome contributes to species-specific adaptations, particularly those associated with biotic and abiotic stress, and it also consistent with the previous conjecture that the genetic basis for resistance may vary among *Vitis* species[20]. In the future, a pan-GWAS of PD-resistance across the genus may help reveal additional variants that contribute to this important trait.

Finally, we have demonstrated the utility of the super-pangenome by performing a pan-GWAS on chloride exclusion, leading to the identification of strong candidate genes on chromosome 8 (Fig. 5c,d). Thus, the pangenome has provided insights into a trait, salt tolerance, that is not only agronomically important but that is growing importance in the context of climate change and current agricultural uses. We note, however, that when we investigated another related trait, root-chloride concentration, no significant association was detected with pan-GWAS (or, for that matter, with standard GWAS) (Supplementary Fig. 8b,c). The current graph can be further augmented to embed the novel variants included within resequencing data, and it can be used as a reference for pan-GWAS based on germplasms that encompass multiple interfertile species. We, therefore, anticipate that the pangenome will be the backbone for many future applications, and particularly for unraveling the genetic basis of critical agronomic traits across *Vitis*.

Methods

Plant material and tissue collection

Several plant tissues were collected for sequencing of DNA, RNA, and full-length transcripts, Bionano next-generation mapping, and flow cytometry. All plant material intended for nucleic acid extraction was immediately frozen upon collection and ground in liquid nitrogen. Most of the samples were collected from accessions maintained by the Foundation Plant Services at the University of California, Davis (Supplementary Table 1).

Library preparation and sequencing

High-molecular-weight genomic DNA (gDNA) was isolated from young leaves using the method described in Chin et al. 2016[35]. DNA purity was evaluated with a Nanodrop 2000 spectrophotometer (Thermo Scientific, IL, USA), DNA quantity with the DNA High Sensitivity kit on a Qubit 2.0 Fluorometer (Life Technologies, CA, USA), and DNA integrity by pulsed-field gel electrophoresis. gDNA was cleaned with 0.45x AMPure PB beads (Pacific Biosciences, CA, USA) before library preparation. SMRTbell template was prepared with 15 µg of sheared DNA using SMRTbell Express Template Prep Kit (Pacific Biosciences, CA, USA) following the manufacturer's instructions. SMRTbell template was size selected using the Blue Pippin instrument (Sage Science, MA, USA) using a cutoff size of 25-80 kb. The size-selected library was cleaned with 1x AMPure PB beads. The SMRTbell library was sequenced on a PacBio Sequel II platform using a V2.0 chemistry (DNA Technology Core Facility, University of California, Davis, CA, USA) (Supplementary Table 2).
Ultra-high molecular weight DNA (> 500 kb) was also extracted from young fresh leaves using Amplicon Express (Pullman, WA, USA). After labeling with a DLE-1 non-nicking enzyme (CTTAAG) and staining following the instructions of the Bionano PrepTM Direct Label and Stain Kit (Bionano Genomics, CA, USA), DNA was loaded onto a SaphyrChip nanochannel array for imaging with the Saphyr system (Bionano Genomics, CA, USA) (Supplementary Table 3).

For *V. acerifolia*, *V. arizonica*, *V. monticola,* and *V. riparia*, DNA-Seq libraries were prepared from 1 µg of DNA extracted from young leaves using the Kapa LTP library prep kit (Kapa Biosystems, MA, USA). After quantity and quality evaluation with the High Sensitivity chip of a Bioanalyzer 2100 (Agilent Technologies, CA, USA), libraries were sequenced in 150-bp-long paired-end reads on an Illumina HiSeq 4000 (DNA Technology Core Facility, University of California, Davis, CA, USA) (Supplementary Table 5). For *V. rupestris*, the reads were produced using a HiSeq X Ten instrument (IDSeq, CA, USA). For the four other genotypes, the sequencing data were retrieved from the BioProject PRJNA731597[4].

Total RNA was extracted from young leaves as described in Cochetel et al. (2021)[36] using the Cetyltrimethyl Ammonium Bromide (CTAB)-based protocol from Blanco-Ulate et al. (2013)[37]. RNA purity was evaluated using a Nanodrop 2000 spectrophotometer (Thermo Scientific, IL, USA). DNA quantity was verified using a Qubit 2.0 Fluorometer and a broad range RNA kit (Life Technologies, CA, USA). Finally, the RNA integrity was evaluated through electrophoresis and with an Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA). Two types of cDNA libraries were produced. For short reads, the Illumina TruSeq RNA sample preparation kit v.2 (Illumina, CA, USA) was used for library preparation following the low-throughput protocol. Library quantity and quality were evaluated with the High Sensitivity chip in an Agilent Bioanalyzer 2100 (Agilent Technologies, CA, USA). The sequencing was performed using an Illumina HiSeq 4000

(DNA Technology Core Facility, University of California, Davis, CA, USA) to produce 100-bp-long paired-end reads. For full-length cDNA sequencing, cDNA SMRTbell libraries were prepared. The first-strand synthesis and cDNA amplification were performed with the NEB Next Single Cell/Low Input cDNA Synthesis & Amplification Module (New England, MA, USA). The cDNA was then purified with ProNex magnetic beads (Promega, WI, USA) according to the Iso-Seq Express Template Preparation for Sequel and Sequel II Systems protocol (Pacific Biosciences, CA, USA) (Supplementary Table 4). Long cDNA (> 2kb) fragments were selected with ProNex magnetic beads (Promega, WI, USA), and at least 80 ng were used to prepare the cDNA SMRTbell library. DNA damage repair and SMRTbell ligation were performed using the SMRTbell Express Template Prep kit 2.0 (Pacific Biosciences, CA, USA). One SMRT cell per genome was sequenced on the Sequel I platform (DNA Technology Core Facility, University of California, Davis, CA, USA) (Supplementary Table 4).

Genome assembly, phasing, and chromosome-scaling

For each genome, SMRT sequences (Pacific Biosciences, CA, USA) were assembled into contigs using Falcon-Unzip[35], as described in Massonnet et al., 2020[6] (Supplementary Table 2), and polished using Arrow from ConsensusCore2 v.3.0.0 (https://github.com/PacificBiosciences/ccs). Long imaged molecules (>150 kb) from the next-generation mapping (Bionano Genomics, CA, USA) were assembled using Bionano Solve v.3.3[38] with the parameters described in Cochetel et al., 2021[36]. Using HybridScaffold v.04122018[38] with the conflict resolution parameters "-B2 -N1", hybrid assemblies were generated through the scaffolding of the Arrow-polished contigs with the consensus genome maps. Despite four next-generation mappings for *V. monticola*, we were not able to generate a consensus genome map for this genome. Scaffolds were simultaneously phased and separated into chromosomes using the tool suite HaploSync[12] and the ~2,000 rhAmpSeq *Vitis* markers[13] (Supplementary Table 18). Illumina DNA-Seq reads were used during the chromosome reconstruction for coverage analysis (Supplementary Table 5). Gene space completeness was evaluated using BUSCO v.3.0.2[39]. In addition, CDS from PN40024[5] were mapped on each genome using pblat v.36x2[40] with the parameters "maxIntron=35000 -minIdentity=0". Before the alignments, the set of PN40024 CDS used was filtered to keep only the sequences uniquely mapping on PN40024 itself with a threshold fixed at 80% for identity and coverage. After the alignments on the nine species, hits were filtered with 50% identity and coverage, but also a minimum overlap of 80% on the reference gene loci (Supplementary Table 9). To evaluate heterozygosity, NUCmer from the tool MUMmer v.4.0.0beta5[18] was used with the parameter "--mum" to align the alternative haplotype of each genome against each other. Heterozygosity levels were estimated as the proportion of the genome covered by variants (Supplementary Table 7). Inter-species variations were also computed aligning the haplotypes of the different species against each other and using the same metrics as for the heterozygosity estimation (Supplementary Table 11).

Structural and functional annotation

The structural annotation of the protein-coding gene loci was performed using an extensive ab-initio prediction pipeline described in Cochetel et al. (2021)[36], available at https://github.com/andreaminio/AnnotationPipeline-EVM_based-DClab. Repeats were annotated using RepeatMasker v.open-4.0.6[41]. RNA-Seq reads were filtered based on quality and adapters were trimmed using Trimmomatic v.0.36[42]. Quality controls pre- and post-trimming were performed with fastQC[43]. Filtered reads were used to produce transcriptome assemblies through

Stringtie v.1.3.4d[44] and Trinity v.2.6.5[45]. For the IsoSeq reads, IsoSeq v.3 (https://github.com/PacificBiosciences/IsoSeq) was used for the extraction, and LSC v.2.0[46] for the polishing of the low-quality isoforms. High-quality gene models were then generated using PASA v.2.3.3[47] from the transcript evidence and external databases. The same input was used to produce genome alignments with Exonerate v.2.2.0[48], PASA, and MagicBLAST v.1.4.0[49]. Ab initio predictions were also generated using Augustus v.3.0.3[50], BUSCO v.3.0.2[39], GeneMark v.3.47[51], and SNAP v.2006-07-28[52]. Based on all the above evidence, EvidenceModeler v.1.1.1[53] was used to generate consensus gene models. The functional annotation was produced with Blast2GO v.4.1.9[54] combining blastp v.2.2.28[55] or diamond v.2.0.8[56] hits against the Refseq plant protein database (https:// ftp.ncbi.nlm.nih.gov/refseq/, retrieved January 17th, 2019) and InterProScan v.5.28-67.0[57] results. Hemizygosity was estimated through the alignments of the CDS of each haplotype per genome against each other using GMAP v.2019.09.12[58]. Results were filtered with a threshold of 80% for identity and coverage. To prevent assembly bias, CDS were also aligned on the unplaced sequences (Supplementary Table 8). Hmmsearch from HMMER v.3.3 (http://hmmer. org/) was used to scan the proteomes using NLR-specific Pfam domains[59]; PF00931, PF01582, PF00560, PF07723, PF07725, and PF12799. For proteins containing coiled-coil (CC) domains, annotations were extracted from the InterProScan results. NLR-annotator[60] was used as an additional source of evidence to identify NLR genes. After identifying NBS-LRR genes missing the NB-ARC domain (Pfam PF00931) as TIR-X or CC-X, eight classes were defined with CC-NBS-LRR, CC-NBS, TIR-NBS-LRR, TIR-NBS, NBS-LRR, and NBS. For gene ontology enrichment analysis, the R package topGO v2.48.0[61] was used with an FDR-adjusted *P* value threshold set to 0.01.

Flow cytometry

DNA content was estimated using flow cytometry as described in Cochetel et al. (2021)[36]. The nuclei extraction was performed using the Cystain PI absolute P kit (Sysmex America Inc., IL, USA). For the internal reference standard, *Lycopersicon esculentum* cv. Stupické polní tyčkové rané was selected, with a known genome size of 2 C = 1.96 pg; 1 C = 958 Mbp[62]. Young leaves (~ 5 mg) from grape and tomato were finely cut with a razor blade in a Petri dish containing 500 mL of extraction buffer and filtered through a 50 mm filter (CellTrics, Sysmex America Inc., IL, USA). Propidium iodide staining solution (2 mL) was then added to the nuclei suspension[63,64]. Measurements were acquired with a Becton Dickinson FACScan (Franklin Lakes, New Jersey) equipped with a 488 nm laser. Raw data were imported and analyzed with the R package flowPloidy v.1.22.0[65] (Supplementary Table 9).

Pangenome graphs construction

A nucleotide-level super-pangenome graph was built using the tools from the PGGB pipeline (https://github.com/pangenome/pggb)[11,16]. To test different configurations of the aligner (wfmash) and the graph builder (seqwish), chromosome 1 of the first haplotype from *V. acerifolia* and *V. berlandieri* were selected arbitrarily (Supplementary Figure 9). For wfmash, the segment length (1, 5, 10, 20, 50, 100kb) and the identity (70, 75, 80, 85, 90, 95%) were evaluated and alignments were converted to a graph using a fixed kmer size = 19 bp for seqwish. After the final selection of wfmash parameters "-s 10000 -p 85", the kmer size was tested from 15 to 405 bp and the default 49 bp was selected. The homologous chromosome sequences of each haplotype were aligned against each other between (e.g. chromosome 1 of species A haplotypes aligned against chromosome 1 of species B haplotypes) and within (e.g. chromosome 1 of species A haplotype 1

aligned against chromosome 1 of species A haplotype 2) species using wfmash (https://github.com/waveygang/wfmash) with the parameters "-p 85 -s 10000 -n 1". The unanchored sequences from each genome were not included. The alignments were merged per chromosome and processed with seqwish v0.7.3[16] to produce the pangenome graphs with "-k 49". Finally, two passes of smoothxg (https://github.com/pangenome/smoothxg) were performed to polish the pangenome. The parameters for the first pass were the following: "-w 68017 -K -X 100 -I 0.85 -R 0 -j 0 -e 0 -l 4001 -P "1,4,6,2,26,1" -O 0.03 -Y 1700 -d 0 -D 0 -V" and the parameters for the second were: "-w 76619 -K -X 100 -I 0.85 -R 0 -j 0 -e 0 -l 4507 -P "1,4,6,2,26,1" -O 0.03 -Y 1700 -d 0 -D 0 -V". These options were derived from the all-in-one PGGB command (https://github.com/pangenome/pggb) using target poa lengths (-l) of 4001 and 4507 for the first and second smoothxg pass, respectively. The chromosome-level graphs were joined using the module "ids" from vg[17] v.1.48.0 to produce a non-overlapping node id space. In the resulting pangenome, each chromosome from the 18 genomes was represented as a path. For each path, node ids and sequences were extracted and categorized in the core, dispensable, or private genomes based on their presence among species. A second pangenome was built adding to the later the genomes of three hybrid species; 101-14 Millardet et de Grasset (101-14 Mgt: *V. riparia* x *V. rupestris*), Richter 110 (110R: *V. berlandieri* x *V. rupestris*), and Kober 5BB (*V. berlandieri* x *V. riparia*)[7]. Wfmash and seqwish parameters were the same. For smoothxg, the parameters were modified to take into account the high number of genomes; for the first pass "-w 92023 -K -X 100 -I 0.85 -R 0 -j 0 -e 0 -l 4001 -P "1,4,6,2,26,1" -O 0.03 -Y 2300 -d 0 -D 0 -V" and the second pass "-w 103661 -K -X 100 -I 0.85 -R 0 -j 0 -e 0 -l 4507 -P "1,4,6,2,26,1" -O 0.03 -Y 2300 -d 0 -D 0 -V".

Orthology analysis

Transcripts from the nine species were aligned on the genome of each other using pblat v.36x2[40] with the parameters "maxIntron=35000 -minIdentity=0". After filtering with a threshold of 80% for identity and coverage, hits were summarized at the gene level. To evaluate colinearity between genomes, the above pblat hits were further filtered to cover at least 80% of the reference gene loci. Colinear blocks were then detected using McScanX_h from McScanX v.11.Nov.2013[66] with default parameters.

Transcript level quantification

Salmon v.1.10.1[67] was used for transcript-level quantification. To get high-accuracy alignments, decoys were generated using the genome sequences in addition to the CDS (https://combine-lab.github.io/alevin-tutorial/2019/selective-alignment/). The indexes were created using a kmer size of 31 bp. Transcript levels were then quantified with the parameters "--seqBias --gcBias --validateMappings" with the paired-end RNA-Seq data used for the gene annotation. Quantification results were imported and TPM values were extracted using the R package tximport[68] v.1.24.0.

Population structure

DNA-Seq samples were sequenced as described previously[4] and quality-based trimmed using trimmomatic. Cleaned DNA-Seq reads were mapped against the haplotype 1 of an equidistant reference genome from the *Vitis vinifera* clade, Cabernet Sauvignon[6], using bwa v.0.7.17-r1188[69] and the parameter "mem". Variant calling per sample was performed using the bcftools v1.17[70] commands mpileup and call with the parameters "--keep-alts --gvcf 0 --multiallelic-caller". The VCF files were merged using bcftools merge "--gvcf". Variant sites and samples were filtered

using bcftools filter "-e 'F_MISSING > 0.25 || MAF <= 0.01'", discarding 19 samples (153 remaining). Samples were further filtered to be representative of their groups and avoid any over-representation of a species/group overall. The maximum number of representative species per group was set to 10 with a selection based on coverage. The inclusion was ensured for the samples of the assembled genomes, a filter on coverage >= 10x was used, and this led to the final selection of 90 samples. Bedtools v.2.30.0[71] intersect was used to discard variant sites in repetitive regions, further filtered with bcftools filter "-e 'AC==0 || AC==AN' --SnpGap 10", bcftools view "-m2 -M2 -v snps", bcftools filter "-i 'QUAL>=30'", bcftools view "-e 'QUAL < 30 || MQBZ < -3 || RPBZ < -3 || RPBZ > 3 || SCBZ > 3 || INFO/DP<5 || INFO/DP>2000'" and pruned with plink v.1.90b6.21[72] "--indep-pairwise 50 5 0.2". The population stratification was finally evaluated by dimension reduction using plink "--pca".

Phylogenetic analysis

Proteomes from the representative haplotype 1 of each genome were compared to each other using OrthoFinder v.2.5.4[73]. The sequences of each single-copy gene orthogroups were aligned using MUSCLE v.5.1[74]. Alignments were concatenated and parsed with Gblocks v.0.91b[75] including positions with gap allowed in < 50% of the sequences. The evolutionary model was selected using the results from ModelTest-NG v.0.1.7[76]. The phylogenetic tree was generated using RAxML-NG v.1.1[77] (Supplementary Fig. 1). The Maximum Likelihood (ML) method was selected with the optimized evolutionary model JTT+I+G4+F, using ten parsimony starting trees, and a bootstrapping of 1,000 replicates. For clock calibration, the single copy orthologs were used as input for BEAUti v.1.10.4[78] to generate the preliminary XML file for BEAST v1.10.4[78] analysis. A monophyletic partition was set for the North American species in the subgenus *Vitis*. A calibration point was established at the crown age of this subgenus which was estimated to be 17.8 million years ago[79], with a normal distribution and standard deviation of 1. Ten independent Markov chain Monte Carlo (MCMC) chains, each consisting of 1,000,000 generations, were run using BEAST. The JTT substitution model, with four Gamma categories, a strict clock, and the Birth-Death Model with a random starting tree were used. Sampling was performed every 1,000 generations. The resulting log and trees files were combined using LogCombiner v.1.10.4[78], and the maximum clade credibility tree was generated using TreeAnnotator v.1.10.4[78] with a burn-in of 10,000 generations (Fig. 1k). For the dN/dS estimations, proteins from each haplotype used to construct the pangenome were separated into core and dispensable proteins selecting the longest isoform as representative. The orthogroups identification and their sequence alignments were performed as described above for the whole proteomes and converted to CDS using PAL2NAL v.13[80]. For the core proteins, the tree computed earlier with the representative haplotype 1 was used. For the dispensable proteins, a tree was computed for each orthogroup using RAxML-NG with the same parameters as for the proteomes. Using the CDS alignments and the phylogenetic trees, the dN/dS ratio was estimated for every core and dispensable orthogroups using codeml from PAML v.4.10.0[81] following the best practice guidelines[82]. For every gene of haplotype 1, a representative protein isoform was selected based on pfam domain coverage to study gene family evolution. They were aligned between genomes using all-vs-all diamond blastp with parameters "--sensitive --evalue 0.000001". The gene family clusters were defined using Markov clustering (mcl)[83] with an inflation value set to 3 resulting in the identification of 9,220 gene families. Using the previously clock-calibrated phylogenetic tree, a computational analysis of gene family evolution was performed using CAFE v.5.0[84] using default settings and a p-value threshold of

0.01. The lambda parameter specific to this dataset was 0.0103. After filtering for significant expansion or contraction, 450 gene families were retained.

Structural variant analysis

Single-nucleotide polymorphisms and structural variants were extracted from the pangenome graph using vg deconstruct and filtered to only consider top-level variants (LV=0). Multi-allelic sites present in the VCFs were normalized to bi-allelic sites using "norm" from bcftools with the parameter "-m-any". Files were divided into SNPs, INDELs, MNPs, and others using bcftools view and the parameter "--type". Insertions and deletions were extracted from the INDELs files using bcftools filter with "--include 'strlen(REF)<strlen(ALT)'" and "--include 'strlen(REF)>strlen(ALT)'", respectively. SnpEff v.5.1[19] was used to annotate the functional impact of the different types of variants. The unfiltered VCFs were compared with the inter-genomic pairwise comparisons made with NUCmer. For every reference genome, the insertion sites and deletions were concatenated into INDELs. The SNPs and INDELs detected from the different query genomes were combined per reference genome into a non-overlapping bed file using bcftools merge summarizing the entire proportion of the reference genome impacted by variants. The variant sites identified from vg deconstruct were compared with these SNPs and INDELs reference datasets using bedtools intersect.

Pan- and classic GWAS

The unfiltered VCFs from vg deconstruct were used with the reference genome fasta of the haplotype 1 of *V. girdiana* SC2 as inputs to the command "construct" from vg. One graph was constructed per chromosome and the id space of the 19 graphs was merged using vg ids. The gcsa index and the xg index were computed with vg index to proceed with the alignments using vg map. The read support was then extracted from the alignment files using vg pack. Finally, the variant genotyping was performed using the following command structure per sample: "vg call --threads number_of_cores -k sample.on.graph.pack -r graph.snarls -s sample -a graph.xg > sample.on.graph.vcf". The flag -a is essential to merge all the samples' VCFs as described in https://github.com/vgteam/vg#calling-variants-using-read-support. The 153 samples were merged into a single VCF using the command "merge --merge all" from bcftools. An initial filter was applied to discard variant sites from the VCF if they were presenting more than 25% of missing genotypes or a minor allele frequency lower than or equal to 1% using bcftools filter. SNPs sites were extracted using bcftools view and the parameter "-v snps" while the SVs were obtained using bcftools view "--exclude-types snps". The different chromosome-level VCF files were concatenated together using bcftools concat. After quality assessment, VCF files were further filtered with bcftools filter "-e 'MAF <= 0.05 || QUAL < 30 || DP < 153 || DP > 5000'". Each variant site was renamed with bcftools annotate using the parameter "--set-id +'%CHROM_%POS'". For linkage disequilibrium (LD) analysis, the following command was used to prune the dataset towards a subset of markers in approximate linkage equilibrium using plink: "plink --vcf name.vcf --double-id --allow-extra-chr -r2 gz --maf 0.1 --geno 0 --mind 0.1 --ld-window 10 --ld-window-kb 300 --ld-window-r2 0 --out name --thin-count 500000". LD decay was evaluated using the model from Hill and Weir (1988)[85] as previously described in grapes[86]. For GWAS, plink was used to prune the files using "--indep-pairwise 50 5 0.2" required to generate the standardized relatedness matrix to account for population structure with gemma v.0.98.3[87]. The GWAS was performed using the matrix and the unpruned variant datasets with linear mixed models (LMMs) to test for association using gemma.

A classic GWAS was also performed using the haplotype 1 of *V. girdiana* SC2 as a reference. The DNA-Seq reads were mapped using bwa and the parameter "mem". The alignments were prepared for GATK using several commands from the Picard toolkit (https://github.com/broadinstitute/picard) starting with the validation of the files using "ValidateSamFile". A sequence dictionary for the reference genome was obtained using "CreateSequenceDictionary". The commands "MarkDuplicates", "AddOrReplaceReadGroups" and "BuildBamIndex" were used to complete the file preparation. The variant calling was performed using GATK v.4.0.12.0[88] with the following command: HaplotypeCaller -I aln.bam --base-quality-score-threshold 20 --sample-ploidy 2 --native-pair-hmm-threads 4 -R ref.fasta -ERC GVCF --disable-read-filter GoodCigarReadFilter –output aln.g.vcf -L chr_name". The single sample gVCFs were imported into a GenomicsDB before the joint genotyping using "GenomicsDBImport" and the final VCF was generated using "GenotypeGVCFs". The VCF was filtered following GATK's guidelines using "bcftools filter -e 'QD < 2 || FS > 60 || SOR > 3 || MQ < 40 || MQRankSum < -12.5 || ReadPosRankSum < -8'" (https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants). The separation into SNPs and SVs, the last filterings and the GWAS followed the same steps as for the pan-GWAS.

Data visualization and statistical analysis

RStudio v.2023.03.1.446[89] running R v.4.2.1[90] was used to run the R packages cited above and all the figures were generated mostly using the package tidyverse v.2.0.0[91]. Genomics data were parsed using GenomicFeatures v.1.48.4[92], phylogenetic trees were drawn using ggtree v.3.4.4[93] after importing the data with treeio v.1.20.2[94], the upset plot was generated using UpSetR v.1.4.0[95], and the gene representations were obtained using gggenes v.0.5.0[96].

Data availability

All the raw sequencing data are available as an NCBI BioProject under the accession number PRJNA984685. Some of the DNA-seq data from the natural populations of wild grape species previously published were accessed from the BioProjects PRJNA731597 and PRJNA842753. The genome files associated with *V. arizonica* were already released in a previous work[20]. Genome assemblies, annotations, and the associated genome browser are accessible at http://www.grapegenomics.com/, files released in previous work[36] for *Muscadinia rotundifolia* cv. Trayshed were retrieved from the same source. Optical maps, genome files, and pangenome graphs were uploaded on Zenodo (10.5281/zenodo.8065808). Code and scripts used in the construction and analysis of the pangenome are available at https://github.com/noecochetel/North_American_Vitis_Pangenome.

Acknowledgments

This work was supported by the NSF grant #1741627 attributed to B.S.G. and D.C., and funds to D.C. from the E.&J. Gallo Winery, and the Louis P. Martini Endowment in Viticulture. Sequencing was performed by the DNA Technology Core, Genome Center at UC Davis. The authors acknowledge Victor Llaca (Corteva Agriscience), Lance Cadle-Davidson (USDA-ARS), and Bruce Reisch (Cornell University) for the *V. rupestris* optical maps.

References

1. Alston, J. M. & Sambucci, O. Grapes in the World Economy. in *The Grape Genome* (eds. Cantu, D. & Walker, M. A.) 1–24 (Springer International Publishing, 2019).
2. Rahemi, A., Dodson Peterson, J. C. & Lund, K. T. *Grape Rootstocks and Related Species*. (Springer International Publishing, 2022).
3. Walker, M. A., Heinitz, C., Riaz, S. & Uretsky, J. Grape Taxonomy and Germplasm. in *The Grape Genome* (eds. Cantu, D. & Walker, M. A.) 25–38 (Springer International Publishing, 2019).
4. Morales-Cruz, A. *et al.* Introgression among North American wild grapes (Vitis) fuels biotic and abiotic adaptation. *Genome Biol* **22**, 254 (2021).
5. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–468 (2007).
6. Massonnet, M. *et al.* The genetic basis of sex determination in grapes. *Nat Commun* **11**, 2902 (2020).
7. Minio, A., Cochetel, N., Massonnet, M., Figueroa-Balderas, R. & Cantu, D. HiFi chromosome-scale diploid assemblies of the grape rootstocks 110R, Kober 5BB, and 101–14 Mgt. *Sci Data* **9**, 660 (2022).
8. Khan, A. W. *et al.* Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. *Trends in Plant Science* **25**, 148–158 (2020).
9. Lei, L. *et al.* Plant Pan-Genomics Comes of Age. *Annu. Rev. Plant Biol.* **72**, 411–435 (2021).
10. Wang, S., Qian, Y.-Q., Zhao, R.-P., Chen, L.-L. & Song, J.-M. Graph-based pan-genomes: increased opportunities in plant genomics. *Journal of Experimental Botany* **74**, 24–39 (2023).
11. Garrison, E. *et al.* Building pangenome graphs. Preprint at https://www.biorxiv.org/content/10.1101/2023.04.05.535718v1 (2023).
12. Minio, A., Cochetel, N., Vondras, A. M., Massonnet, M. & Cantu, D. Assembly of complete diploid-phased chromosomes from draft genome sequences. *G3 Genes|Genomes|Genetics* **12**, jkac143 (2022).
13. Zou, C. *et al.* Haplotyping the Vitis collinear core genome with rhAmpSeq improves marker transferability in a diverse genus. *Nat Commun* **11**, 413 (2020).
14. Zecca, G., Labra, M. & Grassi, F. Untangling the Evolution of American Wild Grapes: Admixed Species and How to Find Them. *Front. Plant Sci.* **10**, 1814 (2020).
15. Ma, Z.-Y. *et al.* Phylogenomics, biogeography, and adaptive radiation of grapes. *Mol Phylogenet Evol* **129**, 258–267 (2018).
16. Garrison, E. & Guarracino, A. Unbiased pangenome graphs. *Bioinformatics* **39**, btac743 (2023).
17. Hickey, G. *et al.* Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol* **21**, 35 (2020).
18. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**, e1005944 (2018).
19. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w ^{1118} ; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
20. Morales-Cruz, A. *et al.* Multigenic resistance to Xylella fastidiosa in wild grapes (Vitis sps.) and its implications within a changing climate. *Commun Biol* **6**, 1–15 (2023).

21. Heinitz, C. C., Riaz, S., Tenscher, A. C., Romero, N. & Walker, M. A. Survey of chloride exclusion in grape germplasm from the southwestern United States and Mexico. *Crop Sci.* **60**, 1946–1956 (2020).
22. Myles, S. *et al.* Genetic structure and domestication history of the grape. *Proc Natl Acad Sci USA* **108**, 3530–3535 (2011).
23. Tello, J. & Ibáñez, J. Review: Status and prospects of association mapping in grapevine. *Plant Science* **327**, 111539 (2023).
24. Carrasco, D. *et al.* Coastal Wild Grapevine Accession (Vitis vinifera L. ssp. sylvestris) Shows Distinct Late and Early Transcriptome Changes under Salt Stress in Comparison to Commercial Rootstock Richter 110. *Plants* **11**, 2688 (2022).
25. Ning, L., Kan, G., Shao, H. & Yu, D. Physiological and transcriptional responses to salt stress in salt-tolerant and salt-sensitive soybean (Glycine max [L.] Merr.) seedlings. *Land Degradation & Development* **29**, 2707–2719 (2018).
26. Guan, R. *et al.* Salinity tolerance in soybean is modulated by natural variation in GmSALT3. *The Plant Journal* **80**, 937–950 (2014).
27. Garg, S., Balboa, R. & Kuja, J. Chromosome-scale haplotype-resolved pangenomics. *Trends in Genetics* **38**, 1103–1107 (2022).
28. Riaz, S. *et al.* Genetic diversity and parentage analysis of grape rootstocks. *Theor Appl Genet* **132**, 1847–1860 (2019).
29. Li, N. *et al.* Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat Genet* **55**, 852–860 (2023).
30. Shang, L. *et al.* A super pan-genomic landscape of rice. *Cell Res* **32**, 878–896 (2022).
31. Tao, Y. *et al.* Extensive variation within the pan-genome of cultivated and wild sorghum. *Nat. Plants* **7**, 766–773 (2021).
32. Tang, D. *et al.* Genome evolution and diversity of wild and cultivated potatoes. *Nature* **606**, 535–541 (2022).
33. Vondras, A. M. *et al.* The genomic diversification of grapevine clones. *BMC Genomics* **20**, 972 (2019).
34. Zhou, Y. *et al.* The population genetics of structural variants in grapevine domestication. *Nat. Plants* **5**, 965–979 (2019).
35. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, 1050–1054 (2016).
36. Cochetel, N. *et al.* Diploid chromosome-scale assembly of the *Muscadinia rotundifolia* genome supports chromosome fusion and disease resistance gene expansion during *Vitis* and *Muscadinia* divergence. *G3 Genes|Genomes|Genetics* **11**, jkab033 (2021).
37. Blanco-Ulate, B., Vincenti, E., Powell, A. L. T. & Cantu, D. Tomato transcriptome and mutant analyses suggest a role for plant stress hormones in the interaction between fruit and Botrytis cinerea. *Front. Plant Sci.* **4**, 142 (2013).
38. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat Biotechnol* **30**, 771–776 (2012).
39. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution* **35**, 543–548 (2018).
40. Wang, M. & Kong, L. pblat: a multithread blat algorithm speeding up aligning sequences to genomes. *BMC Bioinformatics* **20**, 28 (2019).
41. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013).

42. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
43. Andrews, S. FastQC: A Quality Control tool for High Throughput Sequence Data. (2014).
44. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
45. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
46. Au, K. F., Underwood, J. G., Lee, L. & Wong, W. H. Improving PacBio Long Read Accuracy by Short Read Alignment. *PLoS ONE* **7**, e46679 (2012).
47. Haas, B. J. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
48. Slater, G. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
49. Boratyn, G. M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B. & Madden, T. L. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinformatics* **20**, 405 (2019).
50. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
51. Lomsadze, A. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).
52. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
53. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
54. Gotz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).
55. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
56. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59 (2014).
57. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
58. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
59. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
60. Steuernagel, B. *et al.* The NLR-Annotator Tool Enables Annotation of the Intracellular Immune Receptor Repertoire. *Plant Physiol.* **183**, 468–482 (2020).
61. Alexa, A. & Rahnenfuhrer, J. topGO: Enrichment Analysis for Gene Ontology. (2016).
62. Doležel, J., Sgorbati, S. & Lucretti, S. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiol. Plant.* **85**, 625–631 (1992).
63. Doležel, J. & Bartoš, J. Plant DNA Flow Cytometry and Estimation of Nuclear Genome Size. *Ann. Bot.* **95**, 99–110 (2005).
64. Bertier, L., Leus, L., D'hondt, L., de Cock, A. W. A. M. & Höfte, M. Host Adaptation and Speciation through Hybridization and Polyploidy in Phytophthora. *PLOS ONE* **8**, e85385 (2013).

65. Smith, T. W., Kron, P. & Martin, S. L. flowPloidy: An R package for genome size and ploidy assessment of flow cytometry data. *Appl Plant Sci* **6**, e01164 (2018).
66. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
67. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
68. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, (2016).
69. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
70. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
71. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
72. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci* **4**, 7 (2015).
73. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).
74. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
75. Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution* **17**, 540–552 (2000).
76. Darriba, D. *et al.* ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution* **37**, 291–294 (2020).
77. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
78. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* **4**, vey016 (2018).
79. Wan, Y. *et al.* A phylogenetic analysis of the grape genus (Vitis L.) reveals broad reticulation and concurrent diversification during neogene and quaternary climate change. *BMC Evol Biol* **13**, 141 (2013).
80. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* **34**, W609–W612 (2006).
81. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* **24**, 1586–1591 (2007).
82. Álvarez-Carretero, S., Kapli, P. & Yang, Z. Beginner's Guide on the Use of PAML to Detect Positive Selection. *Molecular Biology and Evolution* **40**, msad041 (2023).
83. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**, 1575–1584 (2002).
84. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2020).
85. Hill, W. G. & Weir, B. S. Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology* **33**, 54–78 (1988).

86. Nicolas, S. D. *et al.* Genetic diversity, linkage disequilibrium and power of a large grapevine (Vitis vinifera L) diversity panel newly designed for association studies. *BMC Plant Biol* **16**, 74 (2016).
87. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821–824 (2012).
88. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
89. RStudio Team. RStudio: Integrated Development Environment for R. (2022).
90. R Core Team. R: A language and environment for statistical computing. (2022).
91. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).
92. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol* **9**, e1003118 (2013).
93. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics* **69**, e96 (2020).
94. Wang, L.-G. *et al.* Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Molecular Biology and Evolution* **37**, 599–603 (2020).
95. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
96. Wilkins, D. gggenes: Draw Gene Arrow Maps in 'ggplot2'. (2023).