

RESEARCH

Open Access



Introgression among North American wild grapes (*Vitis*) fuels biotic and abiotic adaptation

Abraham Morales-Cruz¹, Jonas A. Aguirre-Liguori¹, Yongfeng Zhou¹, Andrea Minio², Summaira Riaz², Andrew M. Walker², Dario Cantu² and Brandon S. Gaut^{1*} 

* Correspondence: bgaut@uci.edu

¹Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, CA, USA
Full list of author information is available at the end of the article

Abstract

Background: Introgressive hybridization can reassort genetic variants into beneficial combinations, permitting adaptation to new ecological niches. To evaluate evolutionary patterns and dynamics that contribute to introgression, we investigate six wild *Vitis* species that are native to the Southwestern United States and useful for breeding grapevine (*V. vinifera*) rootstocks.

Results: By creating a reference genome assembly from one wild species, *V. arizonica*, and by resequencing 130 accessions, we focus on identifying putatively introgressed regions (pIRs) between species. We find six species pairs with signals of introgression between them, comprising up to ~ 8% of the extant genome for some pairs. The pIRs tend to be gene poor, located in regions of high recombination and enriched for genes implicated in disease resistance functions. To assess potential pIR function, we explore SNP associations to bioclimatic variables and to bacterial levels after infection with the causative agent of Pierce's disease (*Xylella fastidiosa*). pIRs are enriched for SNPs associated with both climate and bacterial levels, suggesting that introgression is driven by adaptation to biotic and abiotic stressors.

Conclusions: Altogether, this study yields insights into the genomic extent of introgression, potential pressures that shape adaptive introgression, and the evolutionary history of economically important wild relatives of a critical crop.

Keywords: Adaptive introgression, Pierce's disease, Grapevines, Climate

Background

Species emerge from complex interactions among evolutionary processes. For example, genetic drift and local adaptation drive divergence between populations, which ultimately leads to genetic isolation and eventual speciation [1, 2]. The process of divergence can be slowed, in turn, by gene flow between populations, which maintains genetic similarity. There is growing evidence, however, that introgressive hybridization between populations and species does more than homogenize gene pools. It may also be a source of novelty that reassorts genetic variants into beneficial combinations,



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

permitting adaptation to new ecological niches [3]. This ability to reassort genetic variation may partially explain the highly reticulated evolutionary history of adaptive radiations like *Heliconius* butterflies [4], tomatoes [5], Darwin's finches [6], and African cichlids [7]. Introgression has also played a major role in the diversification and speciation of angiosperms [8], because hybridization affects an estimated ~ 25% of flowering plant species [9].

It is generally not known how frequently introgression occurs between species, whether introgression events are adaptive, and, if so, the traits that have been affected. Fortunately, genomic approaches have begun to provide some insights into these central questions. For example, the analysis of *Heliconius* genomes suggests that a large inversion was transferred between species and that this event was adaptive because the inversion contains a color pattern locus that controls mimicry and crypsis [4]. Similarly, a large chromosomal region was exchanged between distinct sunflower subspecies, likely facilitating genetic adaptation to xeric environments [10]. Recent work in cypress [11], oaks [12, 13], maize [14], and cultivated date palms [15] also suggest that introgression between plant species facilitates adaptation to local environments. The work in maize and date palms further highlights the importance of studying the wild relatives of crop species, because they are potential sources of traits for agronomic improvement [16]. Indeed, numerous studies have documented introgression between a crop and its wild relative, leading to a growing understanding of how introgression contributes to important agronomic traits like highland adaptation in maize [14], stress tolerance in potatoes [17], and perhaps fruit quality in apple [18]. We nonetheless emphasize that for many genomic studies of introgression among wild species, the potential phenotypic basis for adaptation has been either unclear or unstudied.

Here we explore the genomic extent and evolutionary dynamics of introgression among wild relatives of cultivated grapevines (*Vitis vinifera* ssp. *sativa*). The genus *Vitis* likely originated in North America ~ 45 my [19] and contains two subgenera [*Muscadinia* ($2n = 40$) and *Vitis* ($2n = 38$)] that encompass ~ 70 species across varied environments. The subgenus *Vitis* has a disjunct distribution across North America and Eurasia [19], with the ~ 25 North American species [20] distributed broadly across the continent, including the American Southwest, where extreme temperature changes and drought are pervasive abiotic stressors. All species within the subgenus are dioecious, interfertile, and often sympatric [21], suggesting the possibility of an extensive history of introgression among species [22, 23]. However, the extent and genomic location of introgressed regions remain unexplored among *Vitis* species, as do the potential functions and evolutionary forces that may shape successful introgression events.

Vitis is also an important study system because cultivated *V. vinifera* (hereafter *vinifera*) is the most valuable horticultural crop in the world [24] and also because it is a model for the study of perennial fruit crops [25]. It is not always appreciated, however, that the cultivation, sustainability, and security of grapevine cultivation relies on North American (NA) *Vitis* species as rootstocks that provide resistance to abiotic and biotic stress [21, 26, 27]. There is a need to identify additional sources of resistance to biotic and abiotic stress, however, because the major rootstock cultivars currently utilized represent a narrow genetic foundation [28]. One biotic stress is Pierce's disease (PD), which is a global threat to the sustainability of wine production [29]. PD is caused by a bacterium (*Xylella fastidiosa*) that spreads from plant to plant by xylem-feeding insect

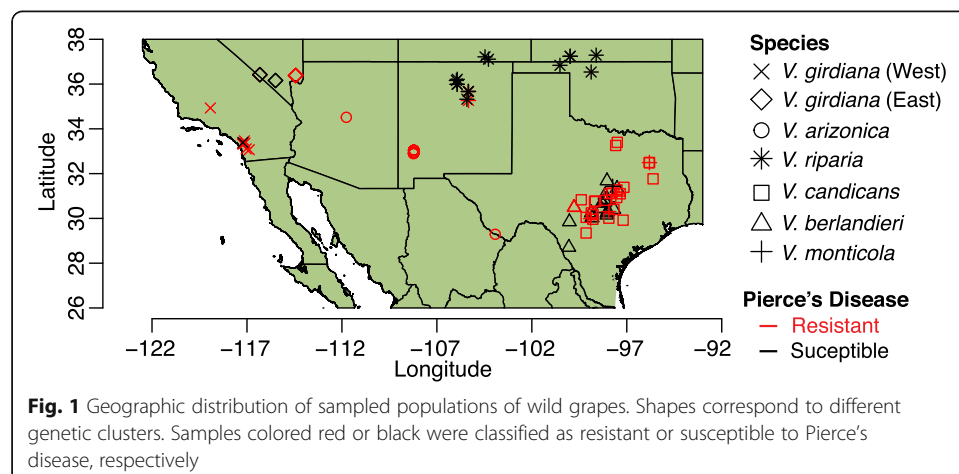
vectors. Several *Vitis* species are polymorphic for resistance to PD, including *V. arizonica* [30] and other species native to the Southwestern United States [31]. These observations open interesting questions about the potential introgression of pathogen resistance loci among wild grape species.

In this study, our goal is to characterize the genomic extent of introgression among wild *Vitis* species from the American Southwest. To do so, we have assembled a reference genome from one species (*V. arizonica*) and generated whole-genome resequencing data from 130 accessions representing six *Vitis* species from portions of their native ranges (Fig. 1). Some of these species have largely overlapping distributions (e.g., *V. candicans* and *V. berlandieri* in Texas), others have disjunct distributions (e.g., *V. arizonica*), and still another (*V. riparia*) has populations in the Southwest at the edge of a broader continental distribution. To complement our genetic data, we have also assessed PD resistance for each accession and gathered bioclimatic data from their location of origin. Given this multifaceted dataset, we address four sets of questions. First, given that species are interfertile and can overlap substantively in geographic range, are they genetically distinct? Second, if they are distinct, is there nonetheless genetic evidence for introgression? Third, if there is evidence for introgression, what are the genomic characteristics of introgressed regions, in terms of locations, size, and gene content? Finally, is there evidence that introgression events have played an adaptive role, as evidenced by genetic associations with either disease resistance or bioclimatic variables? In addressing these questions, this study provides novel insights into the evolutionary dynamics that shape the radiation of *Vitis* and identifies potentially useful genomic targets for breeding.

Results

Population structure and phylogeny

We generated a reference genome of *V. arizonica* from previously reported long-read sequences [32] that we assembled with the aid of a new optical map. The reference assembly contained 19 anchored pseudomolecules, an N50 of 25.9 Mb, a size of 503 Mb, and a BUSCO score of 96.4% (Additional File 1: Table S1 & Table S2). We generated short- and long-read RNAseq data to annotate genes within the genome, ultimately

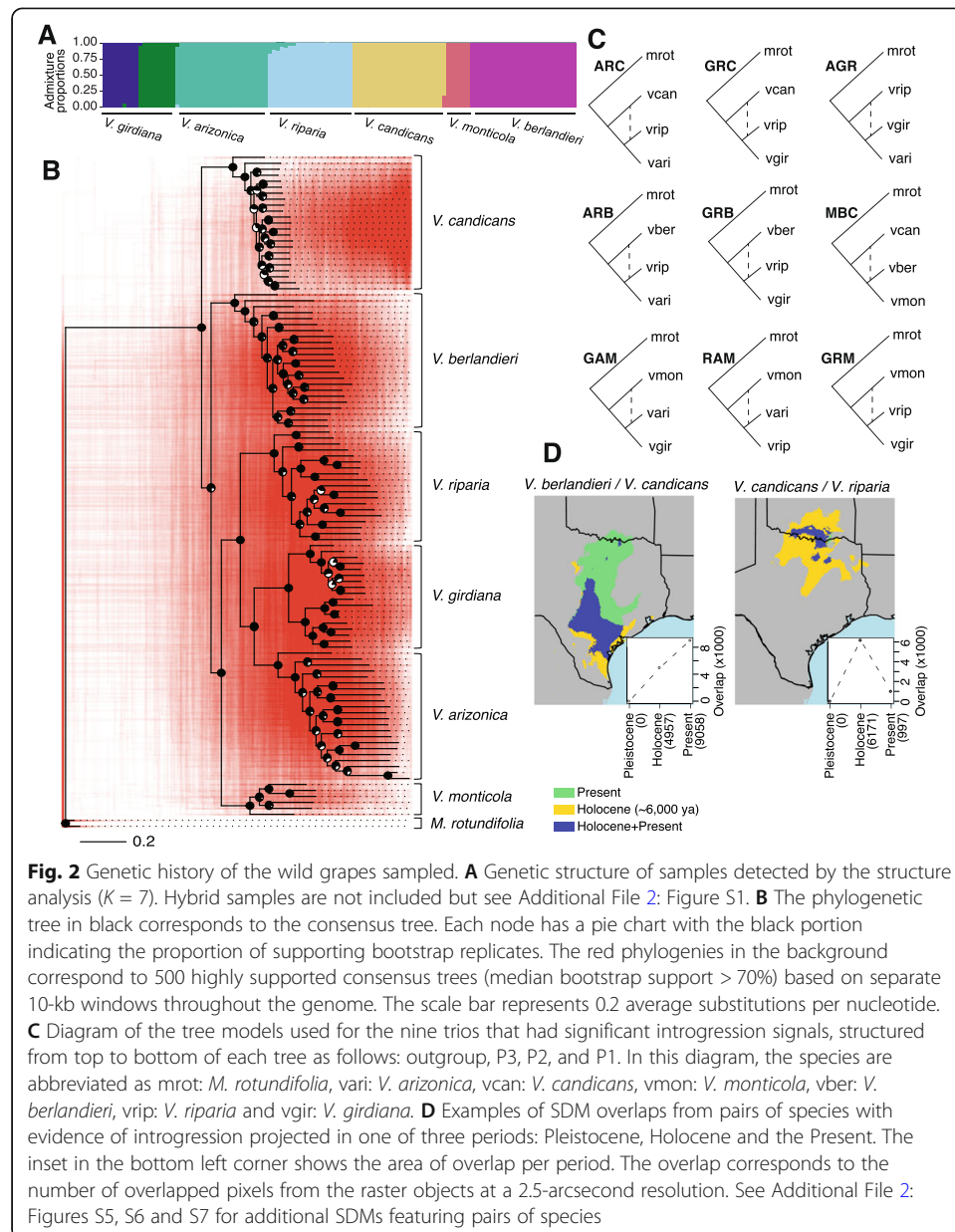


predicting 28,259 gene models. We then resequenced the genome of 130 *Vitis* samples from throughout the native range of six species (*V. arizonica*, *V. berlandieri*, *V. candicans*, *V. girdiana*, *V. monticola*, and *V. riparia*) (Fig. 1, Additional File 1: Table S3). After mapping resequencing data to the reference, we identified ~20 million SNPs among all samples (Additional File 1: Table S4) and used them to assess genetic structure using NGSadmix [33] with $K = 2$ to 10 clusters. The highest support was for $K = 7$ clusters, corresponding to one per species, except for *V. girdiana*, which had two disjunct groups from different geographical locations (Figs. 1 and 2A, Additional File 2: Figure S1). Based on NGSadmix results, we found and removed 19 hybrid individuals that had < 80% of the admixture proportion assigned to a single cluster [34, 35], leaving a final dataset of 111 accessions. Of these, four individuals (“vber09,” “vrip15,” “vcan26,” and “vcan27”) did not fit neatly into their initially proposed species group based on our genetic clustering; we treated each of the four as mis-named and assigned them to a species based on genetic grouping. Altogether, the 111 samples included accessions from *V. arizonica* ($n = 22$), *V. candicans* ($n = 24$), *V. berlandieri* ($n = 22$), *V. girdiana* ($n = 18$), *V. riparia* ($n = 19$), and *V. monticola* ($n = 6$) (Fig. 2A). Genome-wide nucleotide diversity per base pair (π) averaged 0.00284 across species and ranged from 0.00211 in *V. berlandieri* to 0.00353 in *V. monticola* (Additional File 1: Table S5).

We created a consensus phylogenetic tree based on a reduced number of SNPs that limit the effects of linkage disequilibrium (see “Methods”). The phylogeny had median bootstrap support of 88.5% for all nodes and strong support (> 76%) for nodes that separated species groups (Fig. 2B, Additional File 2: Figure S2). In addition, the accessions from each species formed a well-supported monophyletic clade, demonstrating that species are genetically identifiable and also justifying treating each named species as a separate group. Given the phylogeny, we calculated divergence times using a calibration point of 28.32 million years [22] for the separation of *M. rotundifolia* from the *Vitis* subgenus (Additional File 2: Figure S3). Divergence time estimates indicated that the deepest node of individual species often dated to ~10 million years or older, but the divergence between species typically exceeded 20 million years (Additional File 2: Figure S3). Despite strong support for the consensus tree, phylogenies based on individual 10Kb genomic regions from throughout the genome were highly discordant, representing “clouds” around species (Fig. 2B). These clouds illustrate potentially reticulate lineages and suggest the possibility of incomplete lineage sorting and/or a history of introgression among species.

Tests for introgression and geographic overlap among species

To formally test for introgression, we calculated the D statistic [36–38]. D is a genome-wide statistic that measures the excess of shared ancestral alleles based on a tree model of four populations, designated as (((P1,P2),P3),O), where P1, P2, and P3 are ingroups and O is the outgroup. D is expected to be zero under the null hypothesis of incomplete lineage sorting but deviates from zero when there is introgression between P2 and P3 [36, 37, 39]. We calculated D for all combinations of three *Vitis* ingroup species (hereafter called “trios”) that had an appropriate topology for the test according to the consensus tree (Fig. 2B, Additional File 1: Table S6). We used *Muscadinia rotundifolia* as the outgroup for all trios and estimated significance using a block jackknife approach



[37]. Of 11 trios tested, nine had a significant D value at $p < 0.0031$ (Additional File 1: Table S6), representing a total of six P2-P3 species pairs (Fig. 2C, Table 1).

The D statistic is useful for detecting introgression but a poor estimator of the introgressed proportion of genome [39, 40], so we applied the f_4 -ratio [37] to estimate the proportion of the genome with the signal of introgression. The values ranged from ~ 2.3% of the genome in comparisons between *V. riparia* and *V. candicans*, to ~ 8.0% between *V. arizonica* and *V. monticola* (Table 1). For some comparisons, we were able to test the same species pairs with different “control” P1 species. In two cases, the same species pairs yielded similar estimates (Table 1) providing some reassurance about the results. In contrast, the f_4 -ratio estimate varied widely, from 2.1 to 8.0%, for tests between *V. monticola* and *V. arizonica* depending on the P1 control species; notably, however, D was significant with either control. Following calculation of the f_4 -ratio, we

Table 1 Sets of taxa with significant D statistics and properties of the pIRs

Trio ^a	P2	P3	<i>f_d</i> -ratio	Mean pIRs (Kb) ^b	No. genes	Gene density ^c	Sweeps ^d	cM/kb ^e	Resistance genes ^f	PD ^g	Climate ^h
GAM	<i>arizonica</i>	<i>monticola</i>	803%	26936	2282	0.96	0.27	2.27	1.1	2.19*	1.27*
RAM	<i>arizonica</i>	<i>monticola</i>	2.07%	21044	597	0.92	0	1.1	0.56	1.12	1.51*
MBC	<i>berlandieri</i>	<i>candicans</i>	3.24%	16323	746	0.82*	0.39	1.52	1.12	0.59	2.02*
AGR	<i>girdiana</i>	<i>riparia</i>	3.32%	15449	955	0.99	0.32	1.61	0.52	0.80	1.42*
GRM	<i>riparia</i>	<i>monticola</i>	6.08%	27068	1325	0.79*	0.57	1.21	1.56*	0.81	1.13
GRB	<i>riparia</i>	<i>berlandieri</i>	7.47%	23329	1538	0.75*	0.93	1.43	2.00*	1.03	1.13*
ARB	<i>riparia</i>	<i>berlandieri</i>	7.40%	20007	1607	0.75*	0.77	1.8	1.73*	1.09	1.28*
ARC	<i>riparia</i>	<i>candicans</i>	2.43%	22625	480	0.77*	1.41*	1.23	3.55*	2.47*	1.14*
GRC	<i>riparia</i>	<i>candicans</i>	2.30%	21732	455	0.76*	0.43	1.45	3.01*	2.47*	1.33*

^aThe three letters in a trio represent three species that act as P1, P2, and P3, respectively, in introgression tests. The species are A: *V. arizonica*, B: *V. berlandieri*, C: *V. candicans*, G: *V. girdiana*, M: *V. monticola*, and R: *V. riparia*. For each trio, *M. rotundifolia* was used as the outgroup for introgression tests

^bThe mean length of pIRs after merging windows separated by < 1 kb in distance

^cGene density inside pIRs relative to the average gene density of the whole genome. Asterisks denote trios with significantly lower values of gene density in pIRs (*p* < 0.05)

^dSweeps refers to the relative enrichment of selective sweeps in pIR regions. Significant enrichment is denoted by an asterisk, with significance (*p* < 0.05) assessed by permutation

^ecM/kb refers to the ratio of the average recombination rate within pIR regions, relative to the genomic background

^fResistance genes refers to the enrichment of annotated disease resistance genes. Asterisks denote trios with significant enrichment (*p* < 0.05)

^gPD refers to the enrichment of Pierce's disease associated SNPs within pIRs. Significant enrichment is denoted by an asterisk (*p* < 0.05), with significance assessed by permutation

^hClimate refers to the enrichment of bioclimate associated SNPs within pIRs. Values correspond to the average relative enrichment of the top three bioclimatic variables. Significant enrichment is denoted by an asterisk, with significance assessed by permutation (*p* < 0.05)

also calculated the f -branch statistic [41], which recognizes that trios are not independent because they share branches. The f -branch statistic did not yield significant values for any of the internal branch nodes (Additional File 2: Figure S4), suggesting that introgression signals are due to multiple recurrent events instead of a single ancestral event. Taken together, these results show that (i) D and f_4 -ratio support historical introgression among species, (ii) the history of introgression is complex, potentially involving multiple species pairs and multiple events, and (iii) *V. riparia* was most commonly implicated in introgression events with other species (Fig. 2C).

A puzzling feature of these results is that some P2-P3 species pairs currently have few or no regions of geographic overlap (Figure S5; Additional File 2: Figure S5), suggesting limited opportunities for hybridization. To explore the potential for sympatry between species, we performed species distribution modeling (SDM). SDMs identify climatic factors that define the geographic distribution of a species and predict the change in species' distribution over time, given climate prediction models [42–44]. We constructed SDMs based on bioclimatic data from the present, the Holocene (~ 6000 ya), and the Pleistocene (~ 18,000 ya) (see “Methods”) (Fig. 2D). Some of the species pairs had no predicted geographic overlap in any of the three periods (Additional File 2: Figure S6), and none of these yielded genetic evidence for introgression. Others had overlapping distributions but without detected introgression events (Additional File 2: Figure S7). Finally, all species pairs with genomic evidence for introgression had some geographic overlap (Additional File 2: Figure S5). In some cases, however, the predicted geographic overlap between species was higher in the past. For example, *V. candicans* and *V. riparia* had little predicted overlap in the present and in the Pleistocene, but substantial predicted overlap in the Holocene (Fig. 2D). These SDMs support current or historical sympatry of species that yielded evidence for introgression, and they also suggest that detected introgression events are unlikely to have occurred very recently.

Introgression at chromosomal scales

For each of the nine trios with significant D values, we identified putative introgressed chromosomal regions by calculating f_d [39] and f_{dM} [41] in non-overlapping windows of 1000 SNPs along chromosomes (Additional File 2: Figures S8–S11). The two metrics were significantly correlated ($R > 0.75$, $p < 2.2e-16$) along the genome (Additional File 2: Figure S12) and gave qualitatively similar results; we focused on f_{dM} because positive values are interpretable as exchange of derived alleles between P2 and P3 [41]. For each trio, we defined putative introgressed regions (pIRs) (Dataset S1) [45] as the top f_{dM} windows that summed to the genomic proportion estimated by the f_4 -ratio (Table 1).

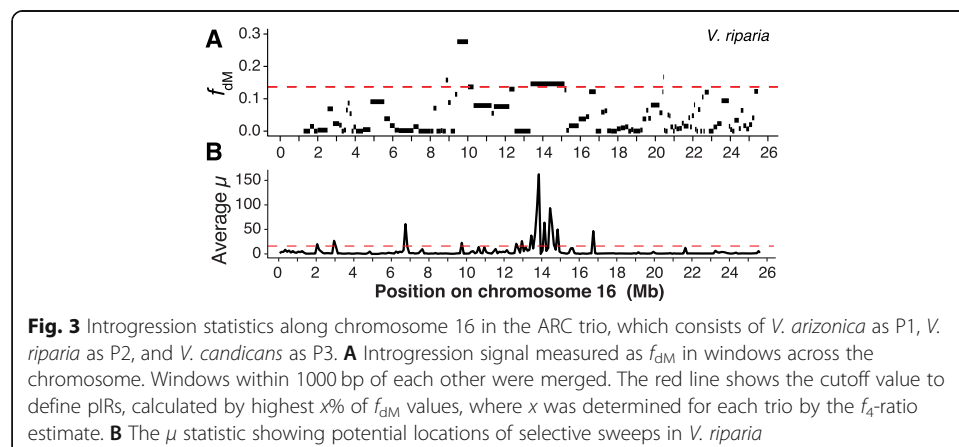
With pIRs identified, we evaluated their basic characteristics. For example, the mean length of pIRs ranged from 154 kb in the AGR trio (see Table 1 and Fig. 2C for trio definitions) to 271 kb in GRM, with the number of pIRs ranging from 54 to 269 (Table 1). The total complement of pIRs contained from 455 to 2282 genes across trios, representing an estimated 1.6 to 8.1% of all genes. We evaluated GO enrichment for the set of putatively introgressed genes and found 15 terms significantly enriched within pIRs (hypergeometric test, p value < 0.048 , Additional File 1: Table S7), including cell-cell signaling (GO:0007267), signaling receptor activity (GO:0038023), and nucleotide binding (GO:0000166). However, pIRs had significantly lower gene density per kilobase than

the genome-wide average in 6 of 9 trios (Table 1, Additional File 1: Table S8 & Additional File 2: Figure S13), and in no case were pIRs significantly enriched for gene density.

Finally, we assessed three additional features of pIRs. First, we reasoned that if pIRs result from introgression events, they should not follow the species consensus tree. We therefore investigated the phylogeny of pIRs for each trio using *twisst* [46], expecting P2 and P3 to be topologically reversed within introgressed regions. Indeed, for each trio the predominant phylogenetic weight was for the tree ((P1,P3),P2,O), as expected if there were introgression between P2 and P3 (Additional File 2: Figure S14). Second, some studies have suggested that introgressed regions should be in genomic regions of high recombination [47]. We used a genetic map to measure recombination in cM/kb [48]; in every trio, pIRs were in regions of higher recombination than the genomic average (Table 1, Additional File 1: Table S9). Third, we sought to determine if pIRs represent adaptive events by assessing their overlap with inferred selective sweeps. To do so, we evaluated sweeps in the P2 population (see “Methods”), focusing on 10-kp windows within the highest 5% μ statistic support. On average, 12.2% of pIRs had at least one putative sweep across the trios. However, sweeps were generally not enriched within pIRs; only one trio (ARC) had significantly more sweeps in pIRs than expected at random ($p < 2.0e^{-3}$, permutation test) (Table 1). This trio provided an interesting example of a potentially adaptive pIR between *V. riparia* and *V. candicans* on chromosome 16, where the pIR at 14–15 Mb contained multiple putative sweeps (Fig. 3).

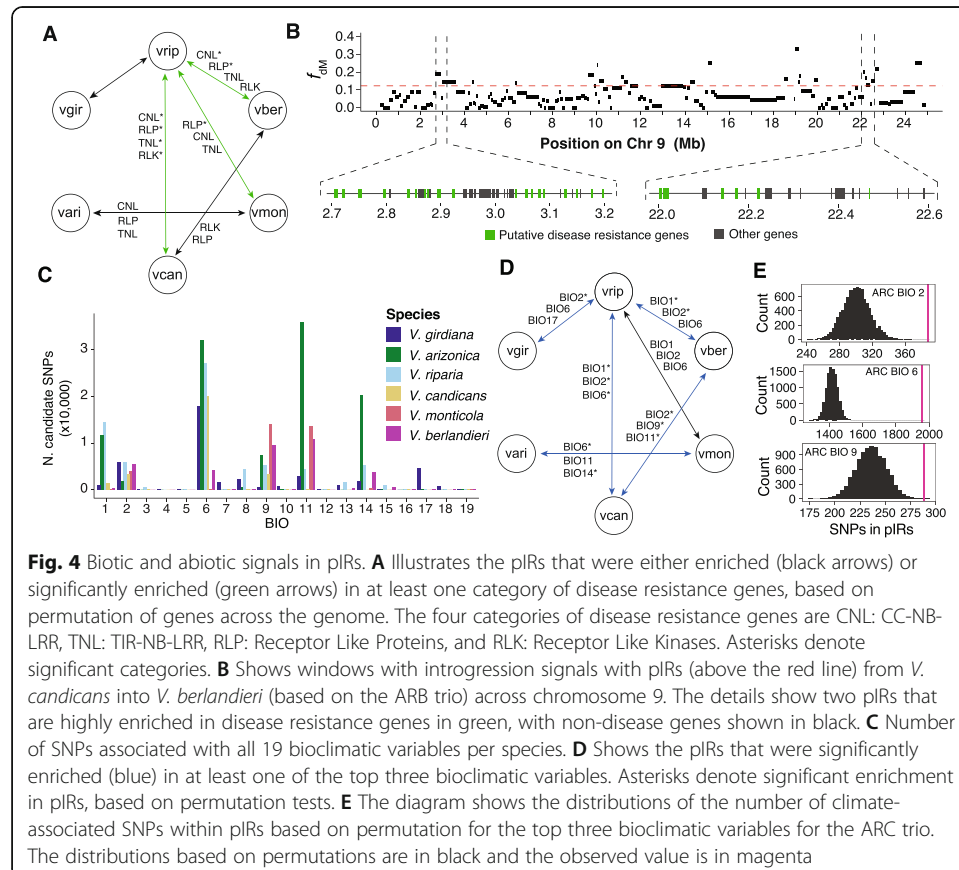
pIRs may be enriched for biotic associations

pIRs were not enriched for selective sweeps, which suggests superficially that they may not have resulted from recent adaptive events. However, another way to evaluate the potential for adaptation is to focus on function, especially potential roles of pIR genes in biotic resistance. We therefore subjected all 28,259 predicted genes to a pipeline designed to identify disease resistance genes [49] and tallied genes in the four most-studied types of pathogen recognition genes (PRGs)—i.e., CC-NB-LRR (CNL), TIR-NB-LRR (TNL), Receptor Like Proteins (RLP), and Receptor Like Kinases (RLK) genes. We found 208 CNLs, 55 TNLs, 258 RLKs, and 336 RLPs in the *V. arizonica* reference (Dataset S2) [45], representing 3% of all genes. We then assessed whether pIRs were



functionally enriched in PRGs among genes. Notably, PRGs were enriched within the pIRs for seven trios, with five significantly enriched (hypergeometric test, p value < 3.66×10^{-4}) (Fig. 4A, Table 1). Regions introgressed between *V. candicans* and *V. riparia* were especially noteworthy, because all four PRG types were significantly enriched. Illustrative examples include the pIRs at 3 Mb and 22 Mb on chromosome nine, which had 53.7% and 33.3% of their genes annotated as PRGs, respectively (Fig. 4B). The pIR at 14 Mb in chromosome 16 of the ARC trio was another region of interest (Fig. 3), because it contained multiple selective sweeps and had 39.1% of genes annotated as PRGs. Finally, we note that the enrichment of PRGs with pIRs is not due to the confounding effects of recombination. We separated the genome into four quadrants based on recombination rate (high, medium, low, and no recombination) and found that genome-wide PRGs are enriched only in low recombination genomic regions ($p = 0.0004$). Hence, pIRs are rare among high recombination genomic regions because they are enriched for PRGs.

The enrichment for PRGs suggests that pIRs could provide an adaptive benefit to biotic challenges. To pursue this idea further, we used quantitative measurements of *X. fastidiosa* bacterial levels in plant stems after manual infection (see “Methods”). We gathered *X. fastidiosa* data for 108 of the 111 accessions [30, 31] (Fig. 1) and performed genome-wide associations studies (GWAS) using bacterial levels as a quantitative phenotype (Additional File 1: Table S10). Our dataset was not ideal for GWAS because: (i) individual species had small sample sizes, which limits statistical power within



species and (ii) the entire dataset, while much larger, contained accessions from multiple species that reflect strong genetic structure (Fig. 2A). We nonetheless performed GWAS on the entire dataset, with a specific focus on whether pIR regions have enriched numbers of SNPs that associate with variation in bacterial levels.

To analyze the dataset, we applied Latent Factor Mixed Models (LFFM2) [50] while correcting for seven discrete genetic clusters (Fig. 2A) with seven Latent Factors (LFs) (Fig. 2A, Additional File 2: Figure S15A) and using a genome inflation factor to correct for population structure (Additional File 2: Figure S15B-C and Additional File 2: Figure S16). We also used a reduced set of 397,723 SNP sites that had no missing data across all 108 accessions and were polymorphic in at least one species. With this approach, we found bacterial levels were significantly associated (FDR adjusted $p < 0.05$) with 261 SNPs across the genome (Dataset S3, Additional File 2: Figure S17) [45], of which 25% (64/261) were located in pIRs of at least one of the nine trios. We then measured, for each trio, the relative enrichment of SNPs per kilobase within pIRs compared to the remainder of the genome; pIRs had higher densities of associated SNPs in six of nine trios (Table 1). Finally, we assessed the significance of this enrichment using a randomization approach that takes into account SNP densities (see “Methods”); the level of enrichment was significantly higher than expected at random in three of the nine trios ($p < 3.32e^{-2}$; permutation test) (Table 1, Additional File 1: Table S11).

Because genetic structure can increase the number of false positives (Additional File 2: Figure S16), we also tested the effect of increasing the number of LFs to further control for population structure. Based on Q-Q plots, we found that increasing the number of LFs reduced the impact of genetic structure, but led to fewer significantly positive associated SNPs—particularly for increasing LF = 7 to LF = 8—but with similar overall patterns (Additional File 2: Figure S18). At LF = 8, there were 129 associated SNPs. Although up to 10% of these outlier SNPs were found within pIRs for some trios, statistical tests for enrichment within pIRs were no longer significant. We suspect that the increase of LFs may better control for genetic structure but likely at the cost of statistical power in both GWAS (see ref. [51]) and enrichment tests. We conclude that pIRs house SNPs associated with bacterial levels and tend to be enriched for such SNPs in some analyses.

pIRs are enriched for abiotic associations

Given evidence for pIR enrichment for both defense-related genes and for SNPs associated with one biotic interaction (with *X. fastidiosa*), we also performed genome environment associations (GEA) to assess abiotic associations. We used BayPass [52] to identify associations between SNPs and 19 bioclimatic variables for each *Vitis* species separately. Outlier SNPs were defined as having Bayes Factor (BF) > 10 [52, 53]. For each species, we then identified the three bioclimatic variables with the highest number of outlier SNPs (Fig. 4C, Additional File 2: Figure S19). For most species, outlier SNPs were especially associated with low temperatures during the coldest period (BIO11 and BIO6), the temperature of the driest months (BIO9), and/or mean diurnal range (BIO2). The number of candidate SNPs associated with each of the three strongest bioclimatic variables ranged from 3380 SNPs (for BIO9 in *V. candicans*) to 39,587 (for BIO11 in *V. arizonica*) (Dataset S4) [45].

We then assessed whether pIRs in the P2 species were enriched for outlier SNPs, doing so separately for each trio and for each of the three bioclimatic variables. Out of 27 potential combinations (= 3 bioclimatic variables $\times$ 9 trios), 16 combinations were enriched for candidate SNPs within pIRs across seven trios ($p < 2.65e^{-2}$; permutation test) (Fig. 4E, Additional File 1: Table S12, Dataset S5) [45]. Thus, despite the fact that they tend to represent gene-poor and high recombination regions of the genome, pIRs are enriched for SNPs associated with bioclimatic variables.

Discussion

Genomic analyses have fueled a growing realization that introgression is an important evolutionary process. These genomic studies have been enhanced by the application of population genomic approaches, like D statistics [36, 37] and related metrics. Nonetheless, these studies often suffer from one (or more) of three limitations. First, a surprising number have relied on reduced representation sequencing methods and poor reference genomes [54]. As a consequence, these studies often have the power to identify a signal of introgression but lack the genomic resolution to adequately characterize the location and content of introgressed regions. Second, many—and perhaps most—studies have focused on introgression between a single pair of taxa [54]. While this simplifies analysis and interpretation, it makes it difficult to infer general evolutionary patterns. Finally, many studies lack information about the potential phenotypic effects of introgressed regions [55], although introgression events into a crop from a wild relative constitute notable exceptions [56]. Ultimately, information about potential phenotypic effects are crucial for understanding the evolutionary processes that affect retention of introgressed regions.

In this study, we have used whole genome resequencing data to identify putatively introgress regions (pIRs) across six wild *Vitis* species. *Vitis* is an interesting system for studying introgression because it is an example of an adaptive radiation [19], because several species grow in sympatry (Figs. 1 and 2D & Additional File 2: Figure S5) and because all wild species have the potential to be (or already are) agronomically important. Moreover, all *Vitis* species are inter-fertile, with hybrid individuals found in nature [21]. To facilitate our study of historical introgression, we first generated a reference genome assembly for *V. arizonica*. To date, four genomes have been published from other wild *Vitis* species, two from *V. riparia* [57, 58], *Vitis labrusca* L. [59], and one genome from *V. amurensis* [60]. However, the *V. arizonica* genome is more contiguous than other North American *Vitis* genomes (e.g., scaffold N50 of ~ 1 Mb [57] vs. 25.9 Mb), which presents a clear advantage for its use. Nonetheless, the use of a single genome raises the specter of reference bias. Our use of high coverage ($\sim 25\times$ average coverage) whole-genome resequencing data should prevent substantial ascertainment biases [61], but reference bias may also influence estimates of heterozygosity and allele frequencies [62, 63] and cannot capture true variants that are not present in reference haplotypes [64]. Full amelioration of reference bias likely requires a genome from all six species in the study, which are not currently available. If they do become available, the optimal strategy for reducing reference bias is not yet clear, but strategies like aligning the resequencing data to multiple assemblies, e.g., [65, 66] or to a *Vitis* pangenome may prove fruitful. Fortunately, however, at least two studies have evaluated reference

bias in population genetics analyses, and they concluded that the effect of the reference bias is unlikely to bias broad demographic and evolutionary genomic analyses [62, 63].

Evidence for extensive effects of introgression on genomes

Using the *V. arizonica* reference, we have produced a dataset of ~20 million SNPs from high coverage (~25×) data representing 130 accessions. We first used the SNPs to investigate genetic clustering and phylogenetic relationships. After the removal of 19 inferred hybrid individuals, we found that individuals from each species formed a monophyletic clade and also that inter-species relationships had strong support (Fig. 2). This last point is interesting because phylogenetic treatments of *Vitis* have not yet reached a broad consensus about species' relationships [67]. The lack of phylogenetic resolution is not surprising given the rapid radiation and inter-fertility of *Vitis* species, but our results suggest that further phylogenetic analyses based on whole genome data and population-level sampling may clarify relationships within this model genus.

Despite the well-supported inter-species phylogeny, we have also detected introgression. *D* and *f* statistics detect introgression between six distinct pairs of species, among a total of eight tested pairs (Fig. 2C). These analyses also suggest that four of the six taxa (and most notably *V. riparia*) have complex histories of hybridization and introgression with potentially more than one species. The results also imply that 2 to 8% of the genomes of some species owe their origins to introgression (Fig. 2C and Table 1). For comparison, the lower value is similar to the 2% of the human genome inferred as having neanderthal origin [68], and the higher value exceeds the 5% percentage of genome introgressed between *Z. mays* ssp. *mays* and ssp. *mexicana* [69]. Overall, our work supports the idea that introgression shapes substantial regions of wild extant plant genomes.

It remains difficult, however, to infer the evolutionary forces that lead to the retention of introgressed regions within hybrid individuals [70]. One hypothesis is that they are retained because they reduce genetic load, particularly when deleterious alleles are recessive. Under this model, the introgressed region from a donor population contributes to lower load in a hybrid compared to non-introgressed members of the receptor population, thereby creating a fitness advantage. This scenario may be most likely for regions of the genome that (i) are unlikely to contribute numerous deleterious mutations (and hence likely to be gene poor); (ii) have high recombination rates, where interference among mutations is minimized [47]; and (iii) come from donor populations with higher effective population sizes (N_e) than recipient species. This last point reflects the fact that high N_e species are generally expected to have a lower deleterious load. In our species, however, we do not find substantial variation in N_e , because nucleotide diversity varies <2-fold, making it hard to assess any relationship between population size and hybridization outcomes. Our results are, however, consistent with the first two points, because pIRs are gene poor in 5 of 9 trios (relative to randomly chosen genomic regions of the same size) and also because the pIRs in all trios are in regions with higher recombination rates than the genome-wide average (Table 1). Here we must recognize an important caveat: our trios are neither evolutionary nor statistically independent. As a consequence, conclusions based solely on the proportion of trios may be misleading. Nonetheless, these two specific trends are clear, because none of the nine trios have

enriched gene density within the pIRs and because all nine trios have pIRs with higher recombination rates than the genomic background (Table 1).

Potential adaptive significance of introgressed regions

Another non-exclusive hypothesis is that introgression is fueled by adaptation. Indeed, a recent simulation study has found that adaptation (and not deleterious load by itself) may be necessary to produce positive introgression statistics, particularly for regions that have high recombination rates and low gene density [71], like our pIRs. To investigate potential signals and causes of adaptation in our data, we have performed two distinct types of analyses. The first is selective sweep mapping within recipient species. While there were some compelling examples of overlapping sweeps and pIRs (Fig. 3), pIRs are generally not enriched for selective sweeps (Table 1). The lack of enrichment does not completely nullify an adaptive explanation for introgression, however, both because selective sweep mapping is inherently noisy and because sweeps are detectable only over a finite time frame. To the latter point, we believe that some pIRs reflect old events, based on two sources of information. First, most species pairs that show evidence for introgression do not currently grow in sympatry; SDMs suggest they were sympatric only in the past (Fig. 2D & Additional File 2: Figure S5). Second, some pIRs overlapped among trios, suggesting that introgression events pre-date the origin of current species (despite the lack of significant *f*-branch statistics on internal branches). For example, we compared two trios that had *V. riparia* as the P2 species. The first (GRM) had *V. monticola* as P3 and the second (GRB) had *V. berlandieri* as P3. The two trios had 142 and 293 pIRs, with 73 overlapping. This observation suggests the possibility that the introgression events contributing to these 73 pIRs pre-date the speciation of *V. monticola* and *V. berlandieri*.

Another approach to assess adaptation is to investigate associations with potentially adaptive functions. Accordingly, we have assessed gene content and explored associations with biotic and abiotic variables. For gene content, the pIRs have a consistent over-enrichment of genes involved in defense functions; seven of nine trios have pIRs that have more than the expected number of disease resistance genes, with five of seven statistically significant (Table 1). This phenomenon is most pronounced for *V. riparia*, because it is the P2 species in all five of these trios. Nonetheless, when we extend our work to include a quantitative phenotype—i.e., an assay of *X. fastidiosa* quantity after infection—we find that six of nine trios have pIRs enriched for SNPs associated with bacterial level, with three significantly so (Table 1). Interestingly, one of the three involves *V. arizonica*, so that our joint approach based on significant enrichments in gene content and bacterial levels implicates pIRs in disease function from a total of six of the nine trios (Table 1). We caution that pIR enrichment for associated SNPs is less marked when the GWAS model includes more latent factors (Additional File 2: Figure S18), and we must again recognize inherent limitations of our GWAS analyses and the lack of independence among trios. Nonetheless, the data on bacterial level and disease resistance enrichment (Fig. 4A) suggest that pathogen interactions could be a major feature shaping adaptive retention of introgressed regions in *Vitis*. Our results complement findings that plant-pathogen interactions play dominant roles in shaping the genetic diversity [72–74] and evolutionary dynamics [74–76] of plant populations, but they

also extend these findings by suggesting that pathogen interactions shape the outcome of introgression events [77]. More empirical work is merited toward this end, both for *Vitis* species and plants more generally.

We have also assessed the potential for adaptive introgression by performing genome-environment associations, which show that climate-associated SNPs are highly significantly enriched within pIRs for at least one of the top three bioclimatic variables in eight of the nine trios (Table 1). One interesting pattern is that the results imply that cold-adapted alleles (e.g., as indicated by pIR associations with BIO6) have introgressed into *V. riparia*. This is puzzling, both because *V. riparia* is the most cold-hardy species among North America wild *Vitis* [78] and also because our sample represents a geographic region that is likely to experience warmer temperatures than most *V. riparia* populations. However, recent studies about the physiological processes involved in grape cold hardiness may provide clues to this puzzle [79, 80]. Perennial plants reduce seasonal cold hardiness by the process of deacclimation, and the rate of deacclimation varies among wild grape species [79]. Because deacclimation affects the time of budbreak, it is a critical process for reproduction. We speculate that historical introgressions into *V. riparia* may have facilitated adaptation to southern regions by contributing alleles that avoid premature budbreak during the midwinter temperature increases that can be common in the Southwest. Clearly, this speculation requires further study to be tested thoroughly.

There is, however, another intriguing possibility, which is that climate-associations do not directly reflect adaptation to climate but rather serve as indirect indicators of biotic stressors within a particular climate. Some evidence from this possibility comes from the observation that, on average, 17.8% of the genes annotated as plant resistance genes in pIRs had a candidate SNP associated with a bioclimatic variable. If true, this idea strengthens the hypothesis that the complex history of introgression among *Vitis* species is driven in part by adaptation to biotic stress, which parallels inferences based on introgression events between neanderthals and humans [81] and between *Populus* species [77]. Overall, these results suggest that evolution has promoted the exchange of resistance alleles and genes among species, which—somewhat paradoxically—is one of the major goals of modern rootstock breeders. Hence, the pIRs identified by this study may prove to be valuable regions for agronomic focus.

Conclusions

In this study, we show that six wild *Vitis* species are genetically distinct, but there is widespread evidence of introgression among species, encompassing up to 8.0% of the genome. We also found that all species-pairs with evidence for introgression had some geographic overlap either in the present or in the past. The identified pIR were characterized by lower gene density and high recombination rates, matching theoretical expectations for introgressed regions. Phylogenetic analysis of the pIRs was also consistent with a history of introgression. Finally, we found some evidence that mutations significantly associated with pathogenic bacterial levels and bioclimatic variables are overrepresented in the pIRs, suggesting that these genomic regions have been exchanged and retained because of their adaptive benefits. The putative adaptive benefits suggest a model—in *Vitis* and perhaps more generally—for which biotic interactions are a major factor shaping the outcome of hybridization.

Methods

V. arizonica genome

Genome assembly

The *V. arizonica* genome was assembled using a hybrid strategy combining Single Molecule Real-Time (SMRT; Pacific Biosciences) data, which were reported previously [32], with the addition of Bionano NGM maps (Bionano Genomics). Genome sequences were assembled using SMRT reads following the custom procedure reported in <https://github.com/andreaminio/FalconUnzip-DClab> [82]. The pipeline performs the marking of repetitive content in SMRT reads using the TANmask and REPmask modules from DAMasker suite [83], both on raw reads before error correction and on the corrected reads used by FALCON ver. 2017.06.28-18.01 [84] to assemble the genome. Multiple combinations of assembly parameters were tested to reduce the sequence fragmentation. We used the following parameters for the final assembly:

```
length_cutoff_pr = 7500
```

```
ovlp_DBSplit_option = - x 500
```

```
ovlp_HPCdaligner_option = -mtan-mrep2-v-B128-M60-t60-k20-h256-e.9-l1000-s100-T16
```

```
overlap_filtering_setting = -- max_diff 100-- max_cov 400-- min_cov 3
```

FALCON_unzip was applied with default parameters to produce the primary contigs and the associated contigs [84] that were then polished from sequence error using PacBio reads with Arrow (from ConsensusCore2 ver. 3.0.0). Primary assembly contiguity was improved by scaffolding with SSPACE-Longread ver. 1.1 [85], followed by a gap-closing procedure with PBJelly (PBSuite ver. 15.8.4) [86, 87].

An optical map was based on Bionano molecules (Bionano Genomics) that were generated and assembled at the genome center of the University of California, Davis (Ming-Cheng Luo). Ultra-high molecular weight DNA (> 500 kbp) was extracted from young leaves by Amplicon Express (Pullman, WA). DNA was then labeled with a DLE-1 non-nicking enzyme (CTTAAG) and stained according to the Bionano Prep™ Direct Label and Stain (DLS) Kit (Bionano Genomics, San Diego, CA) instructions. Labeled DNA was loaded onto the SaphyrChip nanochannel array for imaging on the Saphyr system (Bionano Genomics, San Diego, CA). Optical maps were assembled with BioNano Solve ver. 3.3 [88]. The optical maps obtained were then used to scaffold the PacBio assembled sequences using HybridScaffold ver. 04122018 [88]. The procedure was performed in four iterations. In the first iteration, both sequences and optical maps were broken (“-B 2 -N 2”) when in conflict one with the other. In the second iteration, the scaffolds produced as results of the first iteration were compared to the optical maps, again both sequences and optical maps were broken (“-B 2 -N 2”) when in conflict. In the third iteration, scaffolds resulting from the previous step were compared to the optical maps, conflicts were resolved by breaking nucleotide sequences (“-B 1 -N 2”). In the fourth iteration, results of the previous scaffolding were used, conflicts were again resolved by breaking nucleotide sequences (“-B 1 -N 2”). The genomic sequences obtained were organized and sorted into two sets of chromosomes by using HaploSync ver. 0.1beta (<https://github.com/andreaminio/HaploSync>) and based on synteny with *Vitis vinifera* PN40024 chromosomes.

cDNA library preparation and sequencing

To help annotate the genome, we performed RNAseq using both short- and long-read sequencing technologies. Total RNA from *V. arizonica* leaves was isolated using a Cetyltrimethyl Ammonium Bromide (CTAB)-based extraction protocol as described in Ref. [89]. A Nanodrop 2000 spectrophotometer (Thermo Scientific, Hanover Park, IL) was then used to evaluate RNA purity. The RNA quantity was evaluated with the RNA broad range kit of the Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA) and the integrity using electrophoresis and an Agilent 2100 Bioanalyzer (Agilent Technologies, CA). Total RNA (300 ng, RNA Integrity Number > 8.0) was used for cDNA synthesis and library construction. An RNA-Seq library was prepared using the Illumina TruSeq RNA sample preparation kit v.2 (Illumina, CA, USA) following Illumina's Low-throughput protocol. This library was evaluated for quantity and quality with the High Sensitivity chip in an Agilent 2100 Bioanalyzer (Agilent Technologies, CA) and was sequenced in 100 bp single-end reads, using an Illumina HiSeq4000 sequencer (DNA Technology Core Facility, University of California, Davis). To prepare a cDNA long-read SMRTbell library, first-strand synthesis and cDNA amplification were accomplished using the NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (New England, Ipswich, MA, USA). The obtained cDNAs were then purified with ProNex magnetic beads (Promega, WI) following the instructions in the Iso-Seq Express Template Preparation for Sequel and Sequel II Systems protocol (Pacific Biosciences, Menlo Park, CA). Amplified cDNA was size-selected with a mode of 2 kb using ProNex magnetic beads (86 µl). At least 80 ng of the size-selected, amplified cDNA was used to prepare the cDNA SMRTbell library using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA), following the manufacturer's protocol. One SMRT cell was sequenced on the PacBio Sequel I platform (DNA Technology Core Facility, University of California, Davis). RNAseq data were submitted to NCBI under accession BioProject: PRJNA705722.

Genome annotation

The structural annotation of the *V. arizonica* genome was performed with the pipeline described here: https://github.com/andreaminio/AnnotationPipeline-EVM_based-DClab, an adaptation of a pipeline used previously [90] to use IsoSeq data as primary experimental evidence. In brief, high-quality Iso-Seq data from *V. arizonica* were used in PASA ver. 2.3.3 [91] to generate a set of high-quality gene models for the training of the following ab initio predictions software: Augustus ver. 3.0.3 [92], GeneMark ver. 3.47 [93], and SNAP ver. 2006-07-28 [94]. Gene predictions were generated also using BUSCO ver. 3.0.2 [95] with OrthoDB ver. 9 Plant conserved proteins. Repeats were annotated using RepeatMasker ver. open-4.0.6 [96] with the *Vitis* custom repeat library reported in Ref. [97].

Transcriptomic data was based on new data from *V. arizonica*, data from previous work [97], *Vitis* ESTs (NCBI, download date: 2016.03.15) and *Vitis* mRNA excluding transposable element-related proteins (NCBI, download date: 2016.03.15). RNAseq data were processed using Stringtie ver. 1.3.4d [98] and Trinity ver. 2.6.5 [99] with both on-genome and de novo protocols. Transcriptomic data was mapped, along with transcriptome assemblies and from the Iso-Seq data described above, on the *V. arizonica*

genome using PASA ver. 2.3.3 [91] and MagicBLAST v.1.4.0 [100]. Protein evidence obtained from swissProt viridiplantae (download date: 2016.03.15) and Vitis proteins excluding transposable element-related proteins (NCBI, download date: 2016.03.15) were mapped on the genomic sequences using Exonerate ver. 2.2.0 [101]. Predictions and experimental evidence were then processed by EvidenceModeler ver. 1.1.1 [102] to generate consensus gene models and alternative splicing information integrated from the available transcriptomic data using PASA ver. 2.3.3 [91]. The final functional annotation was produced integrating into Blast2GO ver. 4.1.9 [103] hits from the blastp ver. 2.2.28 [104] results against the Refseq plant protein database (<ftp://ftp.ncbi.nlm.nih.gov/refseq>, retrieved January 17th, 2017) and InterProScan ver. 5.28-67.0 [105].

Disease-related gene functions were annotated using the HMM models from the Disease Resistance Analysis and Gene Orthology (DRAGO 2) database [49]. All the 28,259 predicted proteins from the primary reference chromosomes were evaluated in DRAGO2. Enrichment analysis of functional categories was tested using GeneMerge v1.4 [106].

Genetic diversity data and analyses

Plant material

We collected fresh leaf tissue from 130 individuals from six American *Vitis* species in the wild grape germplasm collection at Davis, California. The germplasm reported in this study was collected from 1997 to 2016 as cuttings across the southwestern states and maintained at the Department of Viticulture and Enology, University of California, Davis, CA. Additional File 1: Table S3 provides details of global positioning coordinates and species designation based on the morphological features of leaves and growth habit of the field grown plants.

Genome resequencing

For the 130 samples, genomic DNA was extracted from leaf samples with the Qiagen DNeasy plant kit. The sequencing libraries were constructed with an insert size of ~ 300 bp using Illumina library preparation kits and were sequenced using the Illumina HiSeq 2500 platform with 2 × 150 bp paired reads to a target coverage of 25× following ref. [107]. The raw sequencing data has been deposited in the Short Read Archive at NCBI under BioProject ID: PRJNA731597.

SNP call and filtering

We filtered and evaluated raw reads using Trimmomatic-0.36 [108] and FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Filtered reads were then mapped to the reference genome with the BWA-MEM algorithm [109] implemented in bwa-0.78 [110]. Joint SNP calling was conducted using the HaplotypeCaller in the GATK v.4.0 pipeline following ref. [107]. We then filtered raw SNPs with bcftools v1.9 (<https://samtools.github.io/bcftools/>) and vcftools v0.1.15 (<https://vcftools.github.io/>). We kept SNP sites for downstream analysis if they were biallelic, had quality higher than 30, had a depth of coverage higher than five reads, had no more than three times the median coverage depth across accessions, and also had less than 25% of missing data among all samples. Additionally, the following expression was applied under the

exclusion argument of the filter function in bcftools: “QD < 2.0 | FS > 60.0 | MQ < 40.0 | MQRankSum < - 12.5 | ReadPosRankSum < - 8.0 | SOR > 3.0”.

Population structure and phylogenetic analyses

We used NgsAdmix to evaluate the genetic structure and the *thetastat* program to measure genetic diversity (π) within species from the ANGSD package v0.915 [33]. We tested values of K from 1 to 10 and used a minMaf filter of 0.05. We then employed the Cluster Markov Packager Across K (Clumpak) software [111] to detect the best K number of clusters. For the overall consensus phylogenetic tree, we used the SNPhylo pipeline (v20180901) to reduce the number of SNPs [112], using a linkage disequilibrium threshold of 0.5 and resulting in 15,893 informative sites. We then created a maximum likelihood phylogeny with the sites from SNPhylo using IQ-TREE v1.6.12 [113]. We used the Model Finder algorithm [114] implemented in IQ-TREE to search for the substitution model best among 550 different combinations and ultimately applied the “VT + F + R5” model. We used the ultrafast bootstrap option with 1000 replicates to obtain support values for each node. For the densitree (in red, Fig. 1C), we divided the genome in 10 kb windows but focused on used windows with at least 800 variable sites, resulting in 3342 windows. We created a consensus tree for each of the 3342 windows with 1000 replicates. We then focused on trees with a median bootstrap value higher than 70% across all nodes, reducing the number of windows to 1172, and randomly chose 500 of the 1172 trees for plotting. We used the R packages ape v5.4 [115] and phangorn v2.5.5 [116] to create the densitree plot and calculate tree statistics.

We also created a clock calibrated tree with BEAST v2.6.4 [117] using the 15,893 informative sites found by SNPphylo. BEAST was run in a chain length of 10 million with sampling every 500. A 10% burn-in was used to create the final tree (Additional File 2: Figure S3). To calibrate the time estimates, we used the estimates from ref. [22] of separation of *M. rotundifolia* with the *Vitis* subgenera as a prior, with a log normal distribution. For the time estimation, we used the Gamma Site Model distributions with the JTT site substitution model, a strict clock model, and the calibrated Yule model. We plotted multiple tree from BEAST using densitree v2.6.4 [118].

Admixture proportions

To test for introgression, we used the ABBA-BABA test as part of the Dsuite software [40]. First, we used the program Dtrios to calculate the overall D statistic and perform a block jackknifing of the statistic to obtain an associated p value, treating non-hybrid samples of each species as populations. The trios that were concordant with the 4-taxon topology of the test and had a p value < 0.001 were then used for more detailed analysis. For the f_d and f_{dM} tests of introgression, we used the program Dinvestigate from Dsuite, choosing non-overlapping windows of 1000 bp SNPs throughout the whole genome. We defined a pIR as windows with highest $x\%$ of f_{dM} values, where x was determined for each trio by the f_4 -ratio estimate (Table 1). To test for potential false positives of the f_4 -ratio estimates in closely related species we also calculated the f-branch statistic using Dsuite. We evaluated the phylogenies in the pIRs by creating phylogenetic trees of 10 kb SNP windows using the “phymtl_sliding_windows.py” script and then

using the “twisst.py” script to calculate average topology weights from the software twisst (<https://github.com/simonhmartin/twissst>) [46].

Selective sweeps

To detect selective sweeps, we used RAISeD v2.8 [119] for each species separately, with default parameters. We removed the gaps of the reference genome from the analysis to avoid potential errors. We split the genome into non-overlapping 10 kb windows created by bedtools v2.27.1 [120] and focused on the windows with top 5% of the μ statistic in the corresponding P2 population and defined those as highly supported sweeps. We first calculated the number of top 5% μ windows from the P2 species overlapping the IRs of each trio, using bedtools with a requirement of at least 50% of the window overlapping with the IR (-f 0.5). For the enrichment analysis, we randomly chose the same number of 10 kb windows as the number of HSSs chosen from the whole genome with 10,000 replicates. We counted the number of the random windows overlapping IRs in the same way as described before. We then compared the distribution of the observed values with the randomly generated distribution and performed a *t*-test to evaluate statistical significance. For plotting, we calculated average μ values for a 100 kb window using the bedtools mapping function.

Recombination

To estimate the location of pIRs relative to recombination rates across the genome, we employed the genetic mapping data from a previous study [48]. Briefly, the map data defined the recombination rate between two markers, based on a consensus of four different mapping populations that used both wild and cultivated *Vitis* parents. The physical location of 1662 markers was provided on the PN40024 *V. vinifera* reference. Since the *V. arizonica* assembly was anchored on the PN40024 *V. vinifera* reference, we were able to transform the markers and recombination values [48] and then assign recombination values (in cM/kb) to discrete regions of the *V. arizonica* genome.

Phenotype data, climate data, and SNP associations

Pierce’s disease assays

Evaluations for Pierce’s disease (PD) resistance were carried out using a greenhouse-based screen [121, 122]. Accessions with strong and intermediate PD resistance and inoculated and un-inoculated susceptible *V. vinifera* cultivar Chardonnay were used in all experiments as reference control plants. Nineteen screens were carried out from 2011 to 2020, and a minimum of four biological replicates of each accession were tested. Disease severity was assessed 10 to 14 weeks post inoculation and ELISA was used to measure the *X. fastidiosa* levels in the stem. Statistical analysis was performed using JMP Pro14 software (Copyright 2020, SAS Institute Inc.) to determine the variability of ELISA for the reference control plants across 19 experiments. In the next step, ELISA values of wild accessions were analyzed with the inclusion of the reference plants to adjust for variation among the 19 screens. PD resistance data for 38 of the 108 accessions used in this study were from a previous study [122] (Additional File 1: Table S10).

Genome-wide association studies (GWAS)

We used LFFM2 to evaluate the association between SNPs and *X. fastidiosa* bacterial level [123]. To control for demographic structure in the across-species associations, we used 7 latent factors (LF) based on the genetic clusters inferred by Admixture and PCA analyses (Additional File 2: Figure S15). We first used the function *lfmm_ridge* and $K = 7$ to compute the regularized least squares using a ridge penalty. Next, we used the *lfmm_test* function to estimate the genomic inflation factor based on the median Z scores. We verified that the calibrated p values had a flat distribution with a slight peak near zero (Additional File 2: Figure S15B-C). Since genetic structure could increase the number of false positives, we also compared GWAS models based on a linear regression model that does not control for genetic structure (see [51, 124, 125]), and models using LFMM with LF values ranging from $K = 6$ to $K = 12$ (Additional File 2: Figure S18).

To avoid biases associated to imputation (LFMM needs non-missing data), all GWAS models were applied to the subset of 397,723 SNPs that had no missing data across the dataset of $n = 108$ accessions. We also used an FDR of 0.05 to detect candidate SNPs and applied GWAS to data with and without a minor allele frequency of $< 5\%$, with no change in results. Once candidate SNPs were identified, we tested for enrichments of SNP density within pIRs. To do so, we randomized the state (candidate-associated SNP or non-candidate SNP) of observed SNPs across the genome. For each randomization, we counted the total number (t_i) of randomly assigned candidate SNPs within pIRs. We performed 10,000 randomizations and compared the distribution of t_i to the observed value (t_{obs}) to compute a p value. Note that this approach inherently accounts for any differences in the density of SNPs across chromosomal regions.

Genome environment associations (GEA)

Genome environment association studies identify outlier loci that correlate with the environment while controlling for demographic structure [126]. We used BayPass version 2 [52] to identify outlier loci that correlated with 19 bioclimatic variables. BayPass first uses the entire set of SNPs to estimate the covariance between populations and identify populations that are genetically closer because of recent co-ancestry or lower genetic structure. In a second step, BayPass analyzes the correlation between the allelic frequency of each SNPs (response variable) and independent environmental, phenotypic, and/or categorical variables. SNPs are considered candidates if they show a strong correlation with the independent variable after controlling for the co-ancestry between populations.

We obtained 19 bioclimatic variables from Worldclim [127] for the location of each accession, using the *extract* function of the raster package [128] in R (R Core Team). For each species, we ran BayPass using all the default parameters, using all polymorphic SNPs without any missing data (Nloc_{arizonica} = 5,456,474 SNPs; Nloc_{berlandieri} = 3,859,330 SNPs; Nloc_{candicans} = 4,971,805 SNPs; Nloc_{girdiana} = 3,598,951; Nloc_{monticola} = 4,664,260 SNPs; Nloc_{riparia} = 5,630,334 SNPs) and the 19 bioclimatic variables. Outlier SNPs were defined using Jeffreys' rule [53], so that they were considered outliers if they had a BF > 10 . For each species, we identified three bioclimatic variables that showed the highest number of outlier loci and focused on outlier SNPs. We further analyzed

the set of outlier SNPs to see if they were enriched in pIRs. We used bedtools v2.27.1 [120] to obtain the overlap between the candidate SNPs of the top 3 bioclimatic variables of each species and the pIRs. For each trio showing evidence of introgression, we used the randomization approach described above for PD associations to test whether pIRs in the P2 species were enriched with outlier SNPs associated with a bioclimatic variable.

Species distribution models

We constructed species distribution models (SDMs) to identify whether species having evidence of introgression had overlapping distributions in the present, the Holocene (~ 6000 ya) and the Pleistocene (~ 18,000 ya). For each species, we obtained occurrence records by combining coordinates data from the Global Biodiversity Information Facility (GBIF.org, 7 August 2020). Most of the six species analyzed had adequate occurrence data (> 70 entries), except *V. monticola* with 20 entries. The low values for *V. monticola* are likely to lead to an overestimation of its geographic distribution. We also obtained 19 bioclimatic variables from WorldClim [127] at a 2.5 resolution for the present, Holocene, and Pleistocene layers. For the past layers, we obtained bioclimatic data projected with the general circulation model CCSM [127].

We employed Maxent V. 3.4.1 [129] to construct and train the SDM of each species and then projected them onto the landscape for each time period. For each species, we ran 30 bootstrap replicates and used 70% and 30% of data for training and validating the models, respectively. Additionally, we evaluated the fit of the models by analyzing the area under the curve (AUC) of the receiver-operating characteristic curve (ROC). For each species, we binarized each replicate by defining as the probability of existence all areas that had a probability value in which the omission rate of the training and testing was above the 10% of prevalence, and/or that predicted the accessions we sampled. For each species, we summed all the binarized models and defined as the final distribution all geographic areas that were predicted by > 60% of the bootstrap replicates.

We used SDM to predict areas of overlap between pairs of species at different time periods. To simplify the data, for each period, first, we summed the SDM of each pair of 15 species pairs retained areas that were predicted by both species. The present, Holocene, and Pleistocene overlapped regions were arbitrarily set at values of 2, 5, and 10. Second, we summed the three time periods to identify the areas where each pair of species were potentially sympatric at different periods. Areas with values 2, 5, and 10 correspond to areas where there have been overlaps only in the present, Holocene, or Pleistocene, respectively. Areas with values 7, 12, or 15 are areas where there have been overlaps in the present and Holocene, present and Pleistocene, or Holocene and Pleistocene, respectively. Areas with value 17 are areas where there has been a continuous overlap between the pair of species. For each pair of species, we obtained the geographic area (measured by the number of overlapping pixels) that were predicted to have an overlap in the SDMs in the different periods.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-021-02467-z>.

Additional file 1: Table S1. Statistics of the *V. arizonica* b40-14 genome assembly. **Table S2.** Summary table of the BUSCO genes of both haplotypes detected in the *V. arizonica* genome assembly. **Table S3.** Accession

identifiers and geographical coordinates of the 130 accessions used in this study. **Table S4.** Chromosome size, raw number of predicted SNPs and number of SNPs after filtering across all 130 samples. **Table S5.** Genome-wide calculation of nucleotide diversity (π) within per species. **Table S6.** Results from genome-wide estimates of introgressions across all combination of trios tested. **Table S7.** GO terms significantly enriched ($p < 0.05$) in pIRs. **Table S8.** Gene density averages per 10 kb and relative enrichment. **Table S9.** Recombination rates per trio and values relative to the genome-wide estimated rate of recombination. **Table S10.** Quantitative measurements of Pierce's Disease evaluations. **Table S11.** Enrichments of SNPs significantly associated with PD in pIRs. **Table S12.** Enrichments of SNPs associated to the top three bioclimatic variables per trio in pIRs.

Additional file 2: Figure S1. Genetic clustering of all samples (including hybrid samples) detected by the structure analyses from $K = 2$ to $K = 10$. **Figure S2.** Consensus maximum likelihood phylogenetic tree created with IQtree using 15,893 informative sites found by SNPphylo. **Figure S3.** Clock calibrated phylogenetic tree created with BEAST using 15,893 informative sites found by SNPphylo. **Figure S4.** Results of the Fbranch statistic using species tree models. **Figure S5.** Pairs of species that show evidence of introgression present evidence of overlap at three periods: present, Holocene, Pleistocene. **Figure S6.** Pairs of species that do not show evidence of introgression present no or low overlap at three periods: present, Holocene, Pleistocene. **Figure S7.** Pairs of species that do not show evidence of introgression but have overlapping distributions at three periods: present, Holocene, Pleistocene. **Figure S8.** Genomic size distribution of the windows used to calculate the ABBA-BABA statistics per trio. **Figure S9.** Distribution the f_{dM} statistic values across windows per trio. **Figure S10.** Distribution the f_d statistic values across windows per trio. **Figure S11.** Manhattan plot of all the middle point value all windows with f_{dM} values for all nine trios. **Figure S12.** Scatterplot showing correlations between the f_{dM} and f_d statistics across all windows in the genome. **Figure S13.** Mean gene density of across the whole genome and within the pIRs of all the trios. **Figure S14.** Average weights of phylogenetic relationships in the pIRs. SNPs from pIRs for each trio were split in 10Kb windows to create a phylogenetic tree for each window. **Figure S15.** Principal component variance LFMM association analyses. **Figure S16.** QQ plots showing the logarithm of the p-values for a linear regression model, the LFMM model using 7 LF and without controlling for a genome inflation factor and controlling for a genome inflation factor. **Figure S17.** Manhattan plot of LFMM association results for the concentration of *X. fastidiosa*. The red lines indicate the threshold at an FDR < 0.05 . **Figure S18.** QQ plots showing the logarithm of the p-values for a linear regression model (gray), and LFMM models using 6-12 LF and controlling for a genome inflation factor. **Figure S19.** Number of associated candidate SNPs per bioclimatic variables per species.

Additional file 3. Review history.

Additional file 4: Dataset S1. Genomic windows identified as putative introgressed regions (pIRs) across nine trios.

Additional file 5: Dataset S2. Gene functional annotation for the genome reference of *Vitis arizonica* b40-14 version 1.

Additional file 6: Dataset S3. Significant SNPs (FDR adjusted $p < 0.05$) from GWAS with bacterial levels of the causative agent of Pierce's Disease.

Additional file 7: Dataset S4. Significant SNPs (Bayes' Factor > 10) from GEA with the top three bioclimatic variables per species.

Additional file 8: Dataset S5. Subset of significant SNPs (Bayes' Factor > 10) from GEA with the top three bioclimatic variables in the P2 species that were in putative introgressed regions (pIRs).

Acknowledgements

The authors thank R. Gaut for generating resequencing data, R. Figueroa-Balderas and O. Nguyen (Genome Center, UC Davis) for technical assistance, and B. Forrestal and J. Ross-Ibarra for providing comments.

Review history

The review history is available as Additional file 3.

Peer review information

Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

A.M.-C., J.A.-L., and B.S.G. designed the scope of the study. A.M. and D.C. generated, analyzed, and annotated the genome. S.R. and A.M.W. performed and supervised the gathering of phenotypic data. A.M.-C., J.A.-L., and Y.Z. performed population and evolutionary analyses. A.M.-C., J.A.-L., and B.S.G. wrote the paper. The author(s) read and approved the final manuscript.

Author's information

Twitter handle: @abramacruz (Abraham Morales-Cruz).

Funding

This work was funded by the NSF grant #1741627 to B.S.G., M.A.W., and D.C., and partially supported by funds to D.C. from the E.J. Gallo Winery and the Louis P. Martini Endowment in Viticulture. J.A.-L. was supported by a UC-Mexico postdoctoral fellowship.

Availability of data and materials

Raw genome resequencing data of the 130 wild grape accessions are available at NCBI under BioProject ID PRJNA731597 [130]. Transcriptome sequencing data of *V. arizonica* b40-14 used for gene model predictions are available at NCBI under BioProject ID: PRJNA705722 [131]. The published [32] long-read *V. arizonica* sequencing data were taken from NCBI BioProject ID PRJNA593045. Genome assembly files and gene annotation files are available in Zenodo [132]. Source code for the genome assembly is available in Github [133]. The genome files and a genome browser can also be found at www.grapegenomics.com/pages/Vari/. The Datasets S1-S5 generated during this study are also available at Figshare [45].

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, CA, USA. ²Department of Viticulture and Enology, University of California, Davis, Davis, CA, USA.

Received: 11 March 2021 Accepted: 12 August 2021

Published online: 03 September 2021

References

- Rundle HD, Nosil P. Ecological speciation: ecological speciation. *Ecol Lett*. 2005;8(3):336–52. <https://doi.org/10.1111/j.1461-0248.2004.00715.x>.
- Wu C-I. The genic view of the process of speciation: genic view of the process of speciation. *J Evol Biol*. 2001;14(6):851–65. <https://doi.org/10.1046/j.1420-9101.2001.00335.x>.
- Marques DA, Meier JL, Seehausen O. A combinatorial view on speciation and adaptive radiation. *Trends Ecol Evol*. 2019;34(6):531–44. <https://doi.org/10.1016/j.tree.2019.02.008>.
- Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, et al. Genomic architecture and introgression shape a butterfly radiation. *Science*. 2019;366(6465):594–9. <https://doi.org/10.1126/science.aaw2090>.
- Pease JB, Haak DC, Hahn MW, Moyle LC. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. Penny D, editor. *PLOS Biol*. 2016;14:e1002379.
- Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, et al. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*. 2015;518(7539):371–5. <https://doi.org/10.1038/nature14181>.
- Meier JL, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat Commun*. 2017;8(1):14363. <https://doi.org/10.1038/ncomms14363>.
- Anderson E, Stebbins GL. Hybridization as an evolutionary stimulus. *Evolution*. 1954;8(4):378–88. <https://doi.org/10.1111/j.1558-5646.1954.tb01504.x>.
- Mallet J. Hybridization as an invasion of the genome. *Trends Ecol Evol*. 2005;20(5):229–37. <https://doi.org/10.1016/j.tree.2005.02.010>.
- Todesco M, Owens GL, Bercovich N, Légaré J-S, Soudi S, Burge DO, et al. Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*. 2020;584(7822):602–7. <https://doi.org/10.1038/s41586-020-2467-6>.
- Ma Y, Wang J, Hu Q, Li J, Sun Y, Zhang L, et al. Ancient introgression drives adaptation to cooler and drier mountain habitats in a cypress species complex. *Commun Biol*. 2019;2(1):213. <https://doi.org/10.1038/s42003-019-0445-z>.
- Leroy T, Louvet J, Lalanne C, Le Provost G, Labadie K, Aury J, et al. Adaptive introgression as a driver of local adaptation to climate in European white oaks. *New Phytol*. 2020;226(4):1171–82. <https://doi.org/10.1111/nph.16095>.
- Nagamitsu T, Uchiyama K, Izuno A, Shimizu H, Nakanishi A. Environment-dependent introgression from *Quercus dentata* to a coastal ecotype of *Quercus mongolica* var. *crispula* in northern Japan. *New Phytol*. 2020;226(4):1018–28. <https://doi.org/10.1111/nph.16131>.
- Hufford MB, Lubinsky P, Pyhäjärvi T, Devengeno MT, Ellstrand NC, Ross-Ibarra J. The genomic signature of crop-wild introgression in maize. Mauricio R, editor. *PLoS Genet*. 2013;9:e1003477.
- Flowers JM, Hazzouri KM, Gros-Balthazard M, Mo Z, Koutroumpa K, Perrakis A, et al. Cross-species hybridization and the origin of North African date palms. *Proc Natl Acad Sci*. 2019;116(5):1651–8. <https://doi.org/10.1073/pnas.1817453116>.
- Burgarella C, Barnaud A, Kane NA, Jankowski F, Scarcelli N, Billot C, et al. Adaptive introgression: an untapped evolutionary mechanism for crop adaptation. *Front Plant Sci*. 2019;10:4. <https://doi.org/10.3389/fpls.2019.00004>.
- Hardigan MA, Laimbeer FPE, Newton L, Crisovan E, Hamilton JP, Vaillancourt B, et al. Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proc Natl Acad Sci*. 2017;114(46):E9999–10008. <https://doi.org/10.1073/pnas.1714380114>.
- Duan N, Bai Y, Sun H, Wang N, Ma Y, Li M, et al. Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. *Nat Commun*. 2017;8(1):249. <https://doi.org/10.1038/s41467-017-00336-7>.
- Ma Z-Y, Wen J, Ickert-Bond SM, Nie Z-L, Chen L-Q, Liu X-Q. Phylogenomics, biogeography, and adaptive radiation of grapes. *Mol Phylogenet Evol*. 2018;129:258–67. <https://doi.org/10.1016/j.ympev.2018.08.021>.
- Wan Y, Schwaninger HR, Baldo AM, Labate JA, Zhong G-Y, Simon CJ. A phylogenetic analysis of the grape genus (*Vitis* L.) reveals broad reticulation and concurrent diversification during neogene and quaternary climate change. *BMC Evol Biol*. 2013;13(1):141. <https://doi.org/10.1186/1471-2148-13-141>.
- Heinitz CC, Uretsky J, Dodson Peterson JC, Huerta-Acosta KG, Walker MA. Crop wild relatives of grape (*Vitis vinifera* L.) throughout North America. In: Greene SL, Williams KA, Khoury CK, Kantar MB, Marek LF, editors. *North Am Crop Wild*

- Relat Vol 2. Cham: Springer International Publishing; 2019 [cited 2020 Jul 28]. p. 329–51. Available from: http://link.springer.com/10.1007/978-3-319-97121-6_10
22. Aradhya M, Wang Y, Walker MA, Prins BH, Koehmstedt AM, Velasco D, et al. Genetic diversity, structure, and patterns of differentiation in the genus *Vitis*. *Plant Syst Evol*. 2013;299(2):317–30. <https://doi.org/10.1007/s00606-012-0723-4>.
 23. Ma Z-Y, Wen J, Tian J-P, Jamal A, Chen L-Q, Liu X-Q. Testing reticulate evolution of four *Vitis* species from East Asia using restriction-site associated DNA sequencing: Reticulate evolution of four *Vitis* species. *J Syst Evol*. 2018;56(4):331–9. <https://doi.org/10.1111/jse.12444>.
 24. Alston JM, Sambucci O. Grapes in the World Economy. In: Cantu D, Walker MA, editors. *Grape Genome*. Cham: Springer International Publishing; 2019 [cited 2021 Feb 8]. p. 1–24. Available from: http://link.springer.com/10.1007/978-3-030-18601-2_1
 25. Minio A, Lin J, Gaut BS, Cantu D. How single molecule real-time sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. *Front Plant Sci*. 2017;8:826. <https://doi.org/10.3389/fpls.2017.00826>.
 26. Walker MA, Lund K, Agüero C, Riaz S, Fort K, Heinitz C, et al. Breeding grape rootstocks for resistance to phylloxera and nematodes - it's not always easy. *Acta Hortic*. 2014;(1045):89–97. <https://doi.org/10.17660/ActaHortic.2014.1045.12>.
 27. Warschefsky EJ, Klein LL, Frank MH, Chitwood DH, Londo JP, von Wettberg EJB, et al. Rootstocks: diversity, domestication, and impacts on shoot phenotypes. *Trends Plant Sci*. 2016;21(5):418–37. <https://doi.org/10.1016/j.tplants.2015.11.008>.
 28. Riaz S, Pap D, Uretsky J, Laucou V, Boursiquot J-M, Kocsis L, et al. Genetic diversity and parentage analysis of grape rootstocks. *Theor Appl Genet*. 2019;132(6):1847–60. <https://doi.org/10.1007/s00122-019-03320-5>.
 29. Purcell AH, Saunders SR. Fate of Pierce's Disease Strains of *Xylella fastidiosa* in Common Riparian Plants in California. *Plant Dis*. 1999;83(9):825–30. <https://doi.org/10.1094/PDIS.1999.83.9.825>.
 30. Krivanek AF, Riaz S, Walker MA. Identification and molecular mapping of PdR1, a primary resistance gene to Pierce's disease in *Vitis*. *Theor Appl Genet*. 2006;112(6):1125–31. <https://doi.org/10.1007/s00122-006-0214-5>.
 31. Riaz S, Huerta-Acosta K, Tenschler AC, Walker MA. Genetic characterization of *Vitis* germplasm collected from the southwestern US and Mexico to expedite Pierce's disease-resistance breeding. *Theor Appl Genet*. 2018;131(7):1589–602. <https://doi.org/10.1007/s00122-018-3100-z>.
 32. Massonnet M, Cochetel N, Minio A, Vondras AM, Lin J, Muyle A, et al. The genetic basis of sex determination in grapes. *Nat Commun*. 2020;11(1):2902. <https://doi.org/10.1038/s41467-020-16700-z>.
 33. Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. *Genetics*. 2013;195(3):693–702. <https://doi.org/10.1534/genetics.113.154138>.
 34. Castillo A, Dorado G, Feuillet C, Sourdille P, Hernandez P. Genetic structure and ecogeographical adaptation in wild barley (*Hordeum chilense* Roemer et Schultes) as revealed by microsatellite markers. *BMC Plant Biol*. 2010;10(1):266. <https://doi.org/10.1186/1471-2229-10-266>.
 35. Vigouroux Y, Glaubitz JC, Matsuoka Y, Goodman MM, Sanchez G, J, Doebley J. Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. *Am J Bot*. 2008;95:1240–1253, 10, DOI: <https://doi.org/10.3732/ajb.0800097>.
 36. Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. *Mol Biol Evol*. 2011;28(8):2239–52. <https://doi.org/10.1093/molbev/msr048>.
 37. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. *Genetics*. 2012;192(3):1065–93. <https://doi.org/10.1534/genetics.112.145037>.
 38. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. *Science*. 2010;328(5979):710–22. <https://doi.org/10.1126/science.1188021>.
 39. Martin SH, Davey JW, Jiggins CD. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol Biol Evol*. 2015;32:244–57.
 40. Malinsky M, Matschiner M, Svoldal H. Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Mol Ecol Resour*. 2020;1755–0998.13265.
 41. Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, et al. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*. 2015;350:1493–8.
 42. Araújo MB, Peterson AT. Uses and misuses of bioclimatic envelope modeling. *Ecology*. 2012;93(7):1527–39. <https://doi.org/10.1890/11-1930.1>.
 43. Fourcade Y, Besnard AG, Secondi J. Evaluating interspecific niche overlaps in environmental and geographic spaces to assess the value of umbrella species. *J Avian Biol*. 2017;48(12):1563–74. <https://doi.org/10.1111/jav.01153>.
 44. Peterson AT, Soberón J, Pearson RG, Anderson RP, Martínez-Meyer E, Nakamura M, et al. *Ecological Niches and Geographic Distributions* (MPB-49). Princeton University Press; 2011. Available from: <https://books.google.com/books?id=wyFn9x9ViOC>
 45. Morales-Cruz A, Aguirre-Liguori J, Zhou Y, Minio A, Riaz S, Walker AM, Cantu D, Gaut BS. Introgression among North American wild grapes (*Vitis*) fuels biotic and abiotic adaptation. *Figshare*. <https://doi.org/10.6084/m9.figshare.13912178> (2021).
 46. Martin SH, Van Belleghem SM. Exploring evolutionary relationships across the genome using topology weighting. *Genetics*. 2017;206(1):429–38. <https://doi.org/10.1534/genetics.116.194720>.
 47. Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, et al. Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science*. 2018;360(6389):656–60. <https://doi.org/10.1126/science.aar3684>.
 48. Zou C, Karn A, Reisch B, Nguyen A, Sun Y, Bao Y, et al. Haplotyping the *Vitis* collinear core genome with rhAmpSeq improves marker transferability in a diverse genus. *Nat Commun*. 2020;11(1):413. <https://doi.org/10.1038/s41467-019-14280-1>.
 49. Osuna-Cruz CM, Paytavi-Gallart A, Di Donato A, Sundesha V, Andolfo G, Aiese Cigliano R, et al. PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Res*. 2018;46(D1):D1197–201. <https://doi.org/10.1093/nar/gkx1119>.
 50. Caye K, Jumentier B, Lepeule J, François O. LFMM 2: fast and accurate inference of gene-environment associations in genome-wide studies. *Kelley J, editor. Mol Biol Evol*. 2019;36:852–60.

51. Luo L, Tang Z, Schoville SD, Zhu J. A comprehensive analysis comparing linear and generalized linear models in detecting adaptive SNPs. *Mol Ecol Resour.* 2021;21(3):733–44. <https://doi.org/10.1111/1755-0998.13298>.
52. Gautier M. Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics.* 2015;201(4):1555–79. <https://doi.org/10.1534/genetics.115.181453>.
53. Jeffreys H. *Theory of probability.* 3rd ed. Oxford [Oxfordshire]: New York: Clarendon Press ; Oxford University Press; 1998.
54. Payseur BA, Rieseberg LH. A genomic perspective on hybridization and speciation. *Mol Ecol.* 2016;25(11):2337–60. <https://doi.org/10.1111/mec.13557>.
55. Suarez-Gonzalez A, Lexer C, Cronk QCB. Adaptive introgression: a plant perspective. *Biol Lett.* 2018;14(3):20170688. <https://doi.org/10.1098/rsbl.2017.0688>.
56. Janzen GM, Wang L, Hufford MB. The extent of adaptive wild introgression in crops. *New Phytol.* 2019;221(3):1279–88. <https://doi.org/10.1111/nph.15457>.
57. Girollet N, Rubio B, Lopez-Roques C, Valière S, Ollat N, Bert P-F. De novo phased assembly of the *Vitis riparia* grape genome. *Sci Data.* 2019;6(1):127. <https://doi.org/10.1038/s41597-019-0133-3>.
58. Patel S, Robben M, Fennell A, Londo JP, Alahakoon D, Villegas-Diaz R, et al. Draft genome of the Native American cold hardy grapevine *Vitis riparia* Michx. 'Manitoba 37'. *Hortic Res.* 2020;7:92.
59. Yang Y, Cuenca J, Wang N, Liang Z, Sun H, Gutierrez B, Xi X, Arro J, Wang Y, Fan P, Londo J, Cousins P, Li S, Fei Z, Zhong GY. A key 'foxy' aroma gene is regulated by homology-induced promoter indels in the iconic juice grape 'Concord'. *Hortic Res.* 2020;7:67. A key 'foxy' aroma gene is regulated by homology-induced promoter indels in the iconic juice grape 'Concord', 1, DOI: <https://doi.org/10.1038/s41438-020-0304-6>.
60. Wang Y, Xin H, Fan P, Zhang J, Liu Y, Dong Y, et al. The genome of *Shanputao* (*Vitis amurensis*) provides a new insight into cold tolerance of grapevine. *Plant J.* 2021;105(6):1495–506. <https://doi.org/10.1111/tpl.15127>.
61. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it: Prospects & Overviews. *BioEssays.* 2013;35(9):780–6. <https://doi.org/10.1002/bies.201300014>.
62. Günther T, Nettelblad C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *Di Rienzo A, editor. PLOS Genet.* 2019;15:e1008302.
63. Gopalakrishnan S, Samaniego Castruita JA, Sinding M-HS, Kuderna LFK, Räikkönen J, Petersen B, et al. The wolf reference genome sequence (*Canis lupus lupus*) and its implications for *Canis* spp. population genomics. *BMC Genomics.* 2017;18:495.
64. Brandt DYC, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. Mapping bias overestimates reference allele frequencies at the *HLA* genes in the 1000 Genomes Project Phase I Data. *G3 GenesGenomesGenetics.* 2015;5:931–41.
65. Chen N-C, Solomon B, Mun T, Iyer S, Langmead B. Reference flow: reducing reference bias using multiple population genomes. *Genome Biol.* 2021;22(1):8. <https://doi.org/10.1186/s13059-020-02229-3>.
66. Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gessing S, Kohlbacher O, et al. Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* 2009;10(9):R98. <https://doi.org/10.1186/gb-2009-10-9-r98>.
67. Miller AJ, Matasci N, Schwaninger H, Aradhya MK, Prins B, Zhong G-Y, et al. *Vitis* phylogenomics: hybridization intensities from a SNP array outperform genotype calls. Wang T, editor. *PLoS ONE.* 2013;8:e78680.
68. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature.* 2014;505(7481):43–9. <https://doi.org/10.1038/nature12886>.
69. Gonzalez-Segovia E, Pérez-Limon S, Cintora-Martínez GC, Guerrero-Zavala A, Janzen GM, Hufford MB, et al. Characterization of introgression from the teosinte *Zea mays* ssp. *mexicana* to Mexican highland maize. *PeerJ.* 2019;7:e6815.
70. Barton NH, Hewitt GM. Analysis of Hybrid Zones. *Annu Rev Ecol Syst.* Annual Reviews. 1985;16(1):113–48. <https://doi.org/10.1146/annurev.es.16.110185.000553>.
71. Zhang X, Kim B, Lohmueller KE, Huerta-Sánchez E. The impact of recessive deleterious variation on signals of adaptive introgression in human populations. *Genetics.* 2020;215(3):799–812. <https://doi.org/10.1534/genetics.120.303081>.
72. Laine A-L, Burdon JJ, Dodds PN, Thrall PH. Spatial variation in disease resistance: from molecules to metapopulations: Spatial variation in disease resistance. *J Ecol.* 2011;99(1):96–112. <https://doi.org/10.1111/j.1365-2745.2010.01738.x>.
73. Marden JH, Mangan SA, Peterson MP, Wafula E, Fescemyer HW, Der JP, et al. Ecological genomics of tropical trees: how local population size and allelic diversity of resistance genes relate to immune responses, susceptibility to pathogens, and negative density dependence. *Mol Ecol.* 2017;26(9):2498–513. <https://doi.org/10.1111/mec.13999>.
74. Stump SM, Marden JH, Beckman NG, Mangan SA, Comita LS. Resistance genes affect how pathogens maintain plant abundance and diversity. *Am Nat.* 2020;196(4):472–86. <https://doi.org/10.1086/710486>.
75. Bever JD, Mangan SA, Alexander HM. Maintenance of plant species diversity by pathogens. *Annu Rev Ecol Evol Syst.* 2015;46(1):305–25. <https://doi.org/10.1146/annurev-ecolsys-112414-054306>.
76. Burdon JJ, Thrall PH, Ericson L. The current and future dynamics of disease in plant communities. *Annu Rev Phytopathol.* 2006;44:19–39.
77. Suarez-Gonzalez A, Hefer CA, Lexer C, Cronk QCB, Douglas CJ. Scale and direction of adaptive introgression between black cottonwood (*Populus trichocarpa*) and balsam poplar (*P. balsamifera*). *Mol Ecol.* 2018;27(7):1667–80. <https://doi.org/10.1111/mec.14561>.
78. Pierquet P, Stushnoff C. Relationship of low temperature exotherms to cold injury in *Vitis Riparia* Michx. *Am J Enol Vitic.* 1980;31:1–6.
79. Kovaleski AP, Reisch BI, Londo JP. Deacclimation kinetics as a quantitative phenotype for delineating the dormancy transition and thermal efficiency for budbreak in *Vitis* species. *AOB PLANTS.* 2018 [cited 2021 Jan 21]; Available from: <https://academic.oup.com/aobpla/advance-article/doi/10.1093/aobpla/ply066/5127118>
80. Londo JP, Kovaleski AP. Deconstructing cold hardiness: variation in supercooling ability and chilling requirements in the wild grapevine *Vitis riparia*. *Aust J Grape Wine Res.* 2019;25(3):276–85. <https://doi.org/10.1111/ajgw.12389>.
81. Enard D, Petrov DA. Evidence that RNA viruses drove adaptive introgression between Neanderthals and modern humans. *Cell.* 2018;175:360–371.e13.
82. Minio A, Massonnet M, Figueroa-Balderas R, Castro A, Cantu D. Diploid genome assembly of the wine grape *Carménère*. *G3amp58 GenesGenomesGenetics.* 2019;9:1331–7.

83. Myers G. Efficient local alignment discovery amongst noisy long reads. In: Brown D, Morgenstern B, editors. *Algorithms Bioinforma*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014 [cited 2020 Nov 2]. p. 52–67. Available from: http://link.springer.com/10.1007/978-3-662-44753-6_5
84. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*. 2016;13(12):1050–4. <https://doi.org/10.1038/nmeth.4035>.
85. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics*. 2014;15:211.
86. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with pacific biosciences RS long-read sequencing technology. *PLoS ONE*. 2012;7:e47768.
87. English AC, Salerno WJ, Reid JG. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*. 2014;15(1):180. <https://doi.org/10.1186/1471-2105-15-180>.
88. Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat Biotechnol*. 2012;30(8):771–6. <https://doi.org/10.1038/nbt.2303>.
89. Blanco-Ulate B, Vincenti E, Powell ALT, 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*. 2013 [cited 2020 Nov 2];4. Available from: <http://journal.frontiersin.org/article/10.3389/fpls.2013.00142/abstract>
90. Vondras AM, Minio A, Blanco-Ulate B, Figueroa-Balderas R, Penn MA, Zhou Y, et al. The genomic diversification of grapevine clones. *BMC Genomics*. 2019;20:972.
91. Haas BJ. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*. 2003;31(19):5654–66. <https://doi.org/10.1093/nar/gkg770>.
92. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*. 2006;34(Web Server):W435–9. <https://doi.org/10.1093/nar/gkl200>.
93. Lomsadze A. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*. 2005;33: 6494–506.
94. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 2004;5:59.
95. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31(19):3210–2. <https://doi.org/10.1093/bioinformatics/btv351>.
96. Smit, AFA, Hubley, R, Green, P. RepeatMasker Open-4.0. 2015. Available from: <http://www.repeatmasker.org>
97. Minio A, Massonnet M, Figueroa-Balderas R, Vondras AM, Blanco-Ulate B, Cantu D. Iso-Seq allows genome-independent transcriptome profiling of grape berry development. *G3*. 2019;9(3):201008.2018.
98. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015;33(3):290–5. <https://doi.org/10.1038/nbt.3122>.
99. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644–52. <https://doi.org/10.1038/nbt.1883>.
100. Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinformatics*. 2019;20(1):405. <https://doi.org/10.1186/s12859-019-2996-x>.
101. Slater G, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005; 6(1):31. <https://doi.org/10.1186/1471-2105-6-31>.
102. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments. *Genome Biol*. 2008;9(1):R7. <https://doi.org/10.1186/gb-2008-9-1-r7>.
103. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005;21(18):3674–6. <https://doi.org/10.1093/bioinformatics/bti610>.
104. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
105. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30(9):1236–40. <https://doi.org/10.1093/bioinformatics/btu031>.
106. Castillo-Davis CI, Hartl DL. GeneMerge—post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*. 2003; 19:891–2.
107. Zhou Y, Massonnet M, Sanjak JS, Cantu D, Gaut BS. Evolutionary genomics of grape (*Vitis vinifera* ssp. *vinifera*) domestication. *Proc Natl Acad Sci*. 2017;114:11715–20.
108. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15): 2114–20. <https://doi.org/10.1093/bioinformatics/btu170>.
109. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio*. 2013 [cited 2020 Nov 2]; Available from: <http://arxiv.org/abs/1303.3997>
110. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*. 2014;30(20): 2843–51. <https://doi.org/10.1093/bioinformatics/btu356>.
111. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour*. 2015;15:1179–91.
112. Lee T-H, Guo H, Wang X, Kim C, Paterson AH. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics*. 2014;15(1):162. <https://doi.org/10.1186/1471-2164-15-162>.
113. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32(1):268–74. <https://doi.org/10.1093/molbev/msu300>.
114. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermini LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14(6):587–9. <https://doi.org/10.1038/nmeth.4285>.
115. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Schwartz R, editor. *Bioinformatics*. 2019;35:526–8.
116. Schliep K, Potts AJ, Morrison DA, Grimm GW. Intertwining phylogenetic trees and networks. Fitzjohn R, editor. *Methods Ecol Evol*. 2017;8(10):1212–20. <https://doi.org/10.1111/2041-210X.12760>.

117. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. Perteau M, editor. PLOS Comput Biol. 2019;15:e1006650.
118. Bouckaert RR. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics. 2010;26:1372–3.
119. Alachiotis N, Pavlidis P. RAISD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. Commun Biol. 2018;1(1):79. <https://doi.org/10.1038/s42003-018-0085-8>.
120. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2. <https://doi.org/10.1093/bioinformatics/btq033>.
121. Krivanek AF, Walker MA. *Vitis* resistance to Pierce's disease is characterized by differential *Xylella fastidiosa* populations in stems and leaves. Phytopathology®. 2005;95:44–52.
122. Riaz S, Tenschler AC, Heinritz CC, Huerta-Acosta KG, Walker MA. Genetic analysis reveals an east-west divide within North American *Vitis* species that mirrors their resistance to Pierce's disease. Chiang T-Y, editor. PLOS ONE. 2020;15:e0243445.
123. Frichot E, Schoville SD, Bouchard G, François O. Testing for associations between loci and environmental gradients using latent factor mixed models. Mol Biol Evol. 2013;30(7):1687–99. <https://doi.org/10.1093/molbev/mst063>.
124. Cubry P, Pidón H, Ta KN, Tranchant-Dubreuil C, Thuillet A-C, Holzinger M, et al. Genome wide association study pinpoints key agronomic QTLs in African Rice *Oryza glaberrima*. Rice. 2020;13(1):66. <https://doi.org/10.1186/s12284-020-00424-1>.
125. Rhoné B, Defrance D, Berthouly-Salazar C, Mariac C, Cubry P, Couderc M, et al. Pearl millet genomic vulnerability to climate change in West Africa highlights the need for regional collaboration. Nat Commun. 2020;11:5274.
126. Tiffin P, Ross-Ibarra J. Advances and limits of using population genetics to understand local adaptation. Trends Ecol Evol. 2014;29(12):673–80. <https://doi.org/10.1016/j.tree.2014.10.004>.
127. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. 2005;25(15):1965–78. <https://doi.org/10.1002/joc.1276>.
128. Hijmans R, van Etten J. Raster: Geographic data analysis and modeling. 2020. Available from: <https://cran.r-project.org/web/packages/raster>
129. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecol Model. 2006;190(3–4):231–59. <https://doi.org/10.1016/j.ecolmodel.2005.03.026>.
130. Morales-Cruz A, Aguirre-Liguori J, Zhou Y, Minio A, Riaz S, Walker AM, Cantu D, Gaut BS. Introgression among North American wild grapes (*Vitis*) fuels biotic and abiotic adaptation. Whole genome resequencing data for all 130 accessions. NCBI BioProject. <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA731597> (2021)
131. Morales-Cruz A, Aguirre-Liguori J, Zhou Y, Minio A, Riaz S, Walker AM, Cantu D, Gaut BS. Introgression among North American wild grapes (*Vitis*) fuels biotic and abiotic adaptation. Transcriptomic data used for gene annotation. NCBI BioProject. <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA705722> (2021)
132. Morales-Cruz A, Aguirre-Liguori J, Zhou Y, Minio A, Riaz S, Walker AM, Cantu D, Gaut BS. Introgression among North American wild grapes (*Vitis*) fuels biotic and abiotic adaptation. Zenodo. <https://doi.org/10.5281/zenodo.4977234> (2021)
133. Minio A, Cantu D. FALCON Unzip based pipeline integrating DAmasker and boosting Unzip speed. Github. <https://github.com/andreaminio/FalconUnzip-DCLab> (2021).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.