









ARTICLE

<https://doi.org/10.1038/s41467-019-14280-1>

OPEN

Haplotyping the *Vitis* collinear core genome with rhAmpSeq improves marker transferability in a diverse genus

Cheng Zou ^{1,11}, Avinash Karn ^{2,11}, Bruce Reisch ², Allen Nguyen³, Yongming Sun³, Yun Bao³, Michael S. Campbell⁴, Deanna Church ⁴, Stephen Williams⁴, Xia Xu⁵, Craig A. Ledbetter⁶, Sagar Patel⁷, Anne Fennell ⁷, Jeffrey C. Glaubitz¹, Matthew Clark ⁸, Doreen Ware ^{9,10}, Jason P. Londo⁵, Qi Sun¹ & Lance Cadle-Davidson ^{5*}

Transferable DNA markers are essential for breeding and genetics. Grapevine (*Vitis*) breeders utilize disease resistance alleles from congeneric species ~20 million years divergent, but existing *Vitis* marker platforms have cross-species transfer rates as low as 2%. Here, we apply a marker strategy targeting the inferred *Vitis* core genome. Incorporating seven linked-read de novo assemblies and three existing assemblies, the *Vitis* collinear core genome is estimated to converge at 39.8 Mb (8.67% of the genome). Adding shotgun genome sequences from 40 accessions enables identification of conserved core PCR primer binding sites flanking polymorphic haplotypes with high information content. From these target regions, we develop 2,000 rhAmpSeq markers as a PCR multiplex and validate the panel in four biparental populations spanning the diversity of the *Vitis* genus, showing transferability increases to 91.9%. This marker development strategy should be widely applicable for genetic studies in many taxa, particularly those ~20 million years divergent.

¹BRC Bioinformatics Facility, Institute of Biotechnology, Cornell University, Ithaca, NY 14853, USA. ²School of Integrative Plant Science, Cornell AgriTech, Cornell University, Geneva, NY 14456, USA. ³Integrated DNA Technologies, Redwood City, CA 94063, USA. ⁴10x Genomics, Inc., Pleasanton, CA 94566, USA. ⁵USDA-ARS, Grape Genetics Research Unit, Geneva, NY 14456, USA. ⁶USDA-ARS, Crop Diseases, Pests and Genetics Research, Parlier, CA 93648, USA. ⁷Agronomy, Horticulture and Plant Science Department, South Dakota State University, Brookings, SD 57007, USA. ⁸Department of Horticultural Science, University of Minnesota, Saint Paul, MN 55108, USA. ⁹Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. ¹⁰USDA-ARS, Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853, USA. ¹¹These authors contributed equally: Cheng Zou, Avinash Karn. *email: Lance.CadleDavidson@ars.usda.gov

Accelerated breeding is helping to meet the challenge of declining food security in the face of rapid population growth and environmental change. Molecular markers are widely deployed to accelerate crop and livestock breeding programs. These DNA markers are useful for germplasm characterization, marker-assisted selection, marker-assisted introgression, and genomic selection. Over the past four decades, DNA marker systems have evolved from interrogating small numbers of loci and individuals (e.g. Restriction Fragment Length Polymorphisms (RFLPs)¹, or simple sequence repeats (SSR)²) to tens of thousands of loci in large study populations (e.g. fluorescence hybridization-based microarray or next-generation sequencing based genotyping)³. However, for breeding applications, single nucleotide polymorphism (SNP) microarrays are constrained by high startup costs and can be affected by ascertainment bias. Alternatively, next-generation sequencing based marker platforms, such as restriction-site associated DNA⁴ and genotyping-by-sequencing (GBS)⁵, suffer from high missing data rates and heterozygote under-calling. This issue has been overcome in grapevine (*Vitis* spp.) through amplicon sequencing (AmpSeq); however, multiplexing of AmpSeq markers is limited to hundreds of loci per PCR reaction due to spurious primer-primer interactions and off-target amplification⁶.

In breeding programs involving highly diverse species and/or genera with rampant structural variation, an important additional concern is marker transferability. For example, *Eucalyptus* (*Eucalyptus* spp.) breeding efforts draw genes from species that diverged 2 to 5 million years ago (Mya)⁷ while grape breeding can include species that diverged up to 20 Mya⁸. Therefore, universal molecular marker panels are needed that can span the diversity present in broad gene pools. In this study, we use the grape genus (*Vitis*) as a model for the development of a pan-generic marker panel. Cultivated grape (*V. vinifera* subsp. *vinifera*) was domesticated from *V. vinifera* subsp. *sylvestris* around ~6000–8000 years ago^{9,10} and is among the most important horticultural crops in the world¹¹. Many grape breeders introgress desirable traits, including abiotic stress tolerance and disease resistance^{12–14}, from wild species within the genus. Furthermore, the *Vitis* genus displays a high degree of structural diversity, presenting a challenge for the development of transferable markers.

Multiple factors contribute to the marker transferability problem, including: (1) Null alleles due to local polymorphism. In this case, genetic variability in a PCR primer site or SNP chip probe site causes binding failure, or polymorphism in a restriction enzyme site causes a null allele in a GBS assay. In a study using three SNP chips (BovineSNP50, OvineSNP50, and EquineSNP50) to genotype species that split as long as 50 Mya, the genotyping failure rate increased by 1.5% per million years of divergence¹⁵. (2) Lack of polymorphism. For example, only 17–33% of the markers on the most widely used SNP genotyping array in maize, the Illumina MaizeSNP50 BeadChip, are polymorphic among European maize inbred lines¹⁶. Similarly, polymorphism of grape SNP markers drops to as low as 2.3% when applied to different *Vitis* species¹⁷. Moreover, only 2% of cattle SNPs are polymorphic in water buffalo (diverged ~12 Mya)^{18,19}. In a large, multispecies study in animals, Miller et al. showed that polymorphism retention decayed exponentially with divergence time¹⁵. Only 5% of markers were polymorphic in species that diverged 5 Mya. (3) Genomic structural variation. Many plant species display a high degree of structural variation between individual genomes^{20,21}. Markers that fall within the so-called dispensable genome²² are less likely to transfer to related species, or even, within the species. Furthermore, they will sometimes be located at different chromosomal positions in different individuals. For example, in *Vitis* we have shown that some markers that tag the flower sex locus in one grape species are located on a different linkage group in other

species, even though the flower sex locus itself remains in the same position⁶.

The aim of this study is to develop a low-cost marker system that is transferrable across the entire *Vitis* genus. Our strategy focuses on the colinear core genome and achieves a high level of multiplexing by incorporating RNase H2 enzyme-dependent amplicon sequencing (rhAmpSeq)²³, which improves the multiplex capacity of AmpSeq via improved amplification specificity and by minimizing spurious primer-primer interactions and off-target amplification. We first identify the *Vitis* core genome based on syntenic whole genome alignment of 10 independent de novo assemblies, consisting of three publicly available genomes (including the *Vitis* reference, *V. vinifera* cv PN40024 genome) and seven linked-read assemblies constructed specifically for this study. Similar to AmpSeq, Illumina sequencing of rhAmpSeq amplicons can capture haplotype allelic series comprised of multiple SNPs and/or Indels per amplicon, resulting in more informative markers than platforms designed for a specific bi-allelic SNP. To help identify highly informative genomic regions and minimize both ascertainment bias and the probability of primer mismatch, we incorporate additional shotgun whole genome sequence data from a *Vitis*-wide diversity panel of 40 accessions. In total, 2000 rhAmpSeq markers are developed, spanning all 19 chromosomes, with an average inter-marker distance of 200 kilobases (kb). The marker panel is validated in four families encompassing much of the genetic diversity in US breeding programs. This strategy generates a highly polymorphic marker set with a low missing data rate across diverse germplasm. The marker set can be used not just for genotyping known alleles, but also for the discovery of novel haplotypes. These short-range novel haplotypes help us to infer the four haplotypes in biparental families directly without any computation estimation. Our strategy should be applicable to many other crop, livestock, and wild species, with or without comprehensive prior genomic resources.

Results

De novo genome assemblies of seven *Vitis* accessions. For accurate definition of the core genome, genome assemblies are required for a representative panel that spans the genomic structural diversity present in the target taxon. The de novo assembly panel in this study included two accessions from wild *Vitis* species, two accessions of wild-wild interspecific hybrids, two accessions of interspecific hybrid grape cultivars (crossing of wild and domesticated species), and four accessions of widely-cultivated modern cultivars of *V. vinifera* subsp. *vinifera* including the reference genome PN40024 (Supplementary Table 1, Fig. 1a). The genomes of Sultanina (a seedless table grape)²⁴ and Cabernet Sauvignon²⁵ were obtained from the public database. In addition, we de novo assembled seven genomes using linked reads, with total raw sequencing read depths ranging from 39-fold to 66-fold. After assembly with the Supernova Assembler (version 2.0.1), contig N50s ranged from 43.9 to 56.86 kb, and the scaffold N50s ranged from 278 kb to 2.1 megabase pair (Mbp). For all nine genomes, more than 90% of the BUSCO genes were represented in full length, indicating good coverage of the gene space (Fig. 1b). Supernova 2.0.1 produces diploid assemblies comprised of two locally phased pseudo-haplotypes. We found that these two phased pseudo-haplotypes only contain one heterozygous site every 371–406 base pairs, which might be due to the under-calling after misassembly of one haplotype. Therefore, only one pseudohap1 assembly was used to represent each genome in the downstream analyses. The molecular length, effective depth, and assembly statistics for each genome are shown in Supplementary Table 1.

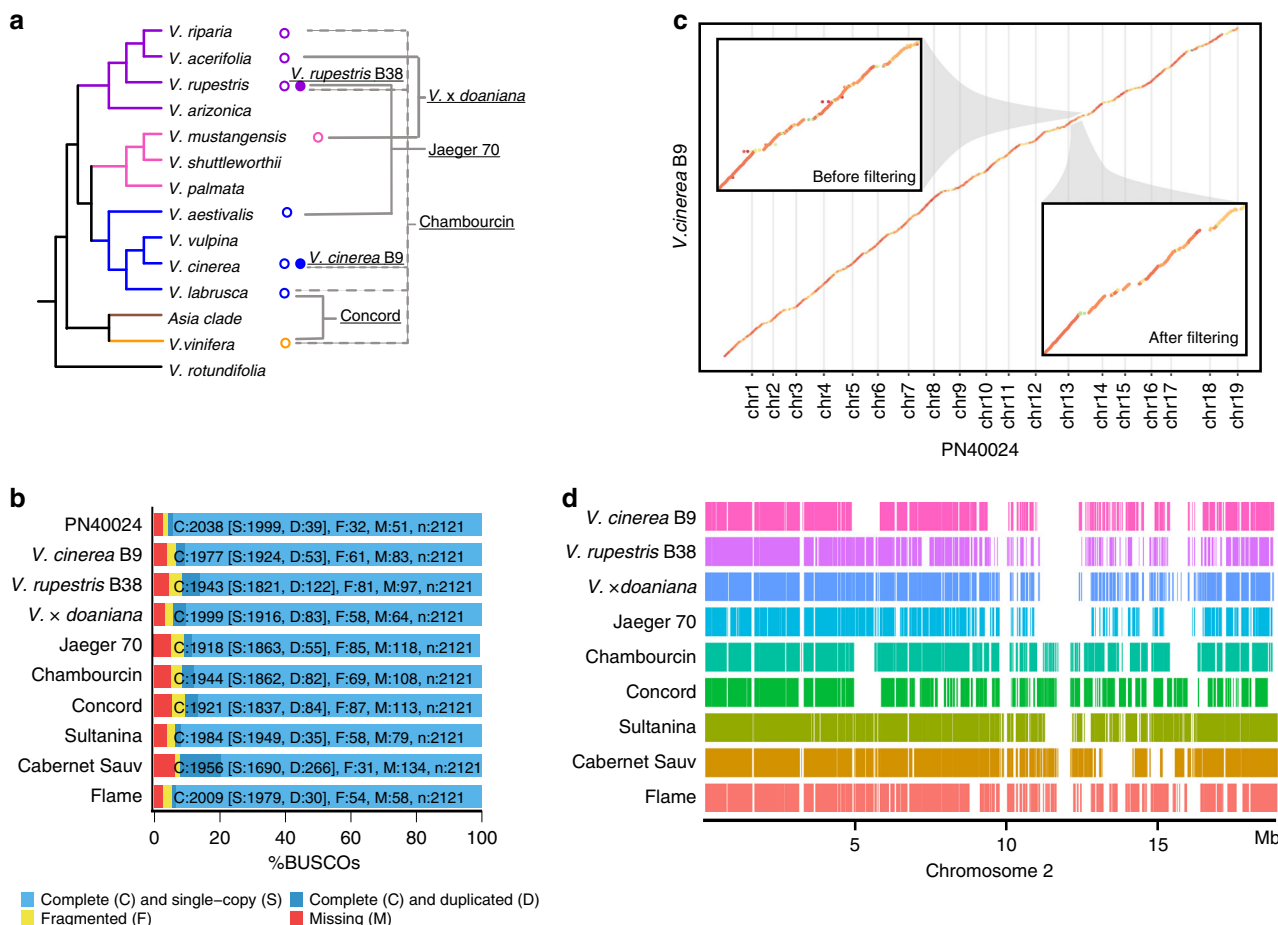


Fig. 1 Genome assembly and core genome construction. **a** Simplified illustration of the evolutionary history of the *Vitis* genus and the genomes included in this study. To enrich diversity in the primer design panel, interspecific hybrid grapes with diverse genetic backgrounds were sequenced. Open circle denote species represented by a sequenced hybrid accession; Closed circle denote one non-hybrid accession from this species was sequenced. Dotted lines denote more than two taxa were included in the accession. **b** BUSCO scores for the ten genomes examined in this study. **c** Syteny of the raw 10x assembly with the reference. The upper left panel is the raw pairwise genome alignment of a region on chromosome 14 constructed by minimap2; the lower right panel is filtered alignment of the same region requiring a syntenic one-to-one match. **d** Coverage of chromosome 2 across the nine genomes. Each line indicates a region that can be aligned to the reference.

Construction of genus-wide core genome. To construct the *Vitis* core genome, the nine assemblies were aligned to the grape reference genome (PN40024 version 12X.v2²⁶). As 41.4% of the grape genome is composed of transposable elements¹¹, we masked the repetitive genomic regions prior to alignment by kmer frequency. A high degree of collinearity between each assembly and the reference was observed, validating the assembly qualities in general. To further maximize one-to-one correspondence, the pairwise syntenic alignments were smoothed by collapsing small, local tandem duplicates, present either in the PN40024 reference or the assembly (Fig. 1c, Supplementary Fig. 1). The total length of one-to-one matched blocks after smoothing ranged from 160 to 226 Mbp and depended on genetic distance to *V. vinifera*. After the kmer-based repeat masking, 88% of the coding sequences were retained on the reference genome PN40024. For the other genomes aligned with PN40024, before smoothing 64 to 77% of the reference coding sequences were retained, and 55 to 39% were retained after smoothing (Supplementary Table 1). Core genomic regions were distributed across each chromosome, except for gaps that represent either structural variation (the dispensable genome) or mis-assembly. As expected, the genomes of the wild species display more structural variation relative to the reference genome than those of domesticated cultivars. For example, on average, only 37% of the reference genome was collinear with wild species

versus 54% with cultivars (Supplementary Table 1). An illustration of collinearity with chromosome 2 of each assembly is presented in Fig. 1d. The total length of the core genome decays exponentially as more assemblies are included in the core genome (Supplementary Fig. 2). By fitting an exponential decay function, the inferred core genome size is predicted to plateau at 39.8 Mbp with 17 assemblies in the model.

We define the coverage of the core genome presence as the number of times a reference PN40024 chromosomal region aligned with another genome assembly in a collinear, syntenic fashion. About 10% of the reference genome was present in all nine genome assemblies—we define these regions as the grape core genome. Sixty-four percent of the inferred core genome lies within gene regions. We consider these 9386 genes in the collinear core genome to be core genes, and they are significantly enriched in cellular metabolic processes, RNA binding function, and membrane localization (Supplementary Fig. 3). Certain genome features are more prevalent in the core genome (Supplementary Fig. 4). For example, gene density is higher in the core genome versus the remaining genome ($\rho = -0.75$, $P < 1E-13$ in two-sided Spearman's correlation test) and transposable elements (TEs) are depleted ($\rho = 0.76$, $P < 1E-13$ in two-sided Spearman's correlation test), in particular the most abundant TE family, Gypsy.

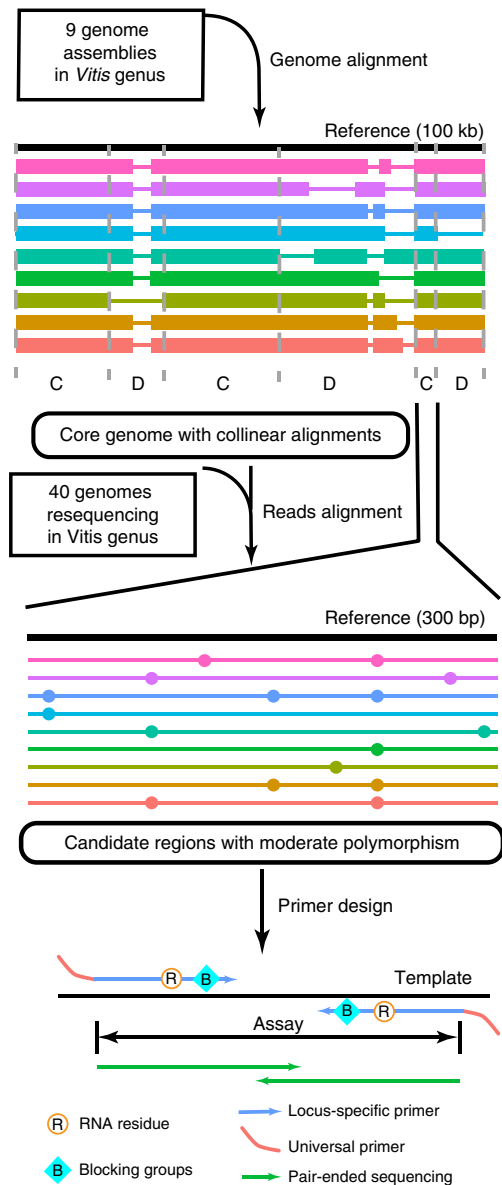


Fig. 2 Marker design pipeline based on genus-wide core genome and polymorphism. Colored blocks, alignment between the query genome and the reference genome. C, core genome; D, dispensable genome. Colored dots denote genomic variants. In the rhPCR, an RNA residue and several bases of blocking DNA are added to the allele-specific primer. The RNA residue together with the blocking DNA can only be cleaved by the RNase H2 enzyme when the match between the target and the primer is perfect. After the RNA residue and the blocking DNA are removed from the 3' end, the extension reaction will continue.

Marker design and statistics. To develop markers universally effective in diverse germplasm, rhAmpSeq markers were designed to target only the core *Vitis* genome, with multiple additional attributes taken into consideration (Fig. 2). First, to decrease off-target amplification, primer binding sites had to be unique and had to avoid sequence variation. Second, the polymorphism level of PCR products had to be moderate. Elevated polymorphism is both advantageous and disadvantageous in marker design. A higher level of polymorphism manifests as multiallelic haplotype markers, which are more informative than biallelic ones; however, it also increases the risk of null alleles or off-target amplification from its paralogs in the genome. In our genotyping pipeline, instead of calling multiple independent SNPs separately, the

entire amplicon of 200–300 bp is used as a haplotype allele tag. To extensively characterize genus level polymorphism in these targeted regions, we called variants based on grape whole genome sequencing data from two sources: (1) 47 *Vitis* accessions with at least three-fold sequencing depth retrieved from the NCBI SRA, and (2) seven *Vitis* accessions shotgun sequenced specifically for this study (these are different from the seven de novo assembled accessions) (Supplementary Table 2). Principal component analysis of the resulting genotypes indicated that the wild species are substantially more genetically diverse than the cultivated lines (Supplementary Fig. 5). To balance the composition of wild species and cultivars in this diversity panel, we randomly selected twenty accessions from each group. Across the 40 accessions, SNP density in the core genome was 0.032 (i.e., 32 SNPs per kilobase), and was very similar in the core genes (0.031) (Supplementary Fig. 6). To balance primer transferability against information content, we focused on moderately polymorphic regions by discarding loci outside the 25th and 75th percentiles. As expected, the missing genotype rate in the *Vitis* diversity panel decreases as an exponential decay as more assembled genomes are included in the core genome construction (Supplementary Fig. 7). The final consideration for marker design was physical distribution across the genome. Initially, candidate regions were randomly chosen to obtain one marker per ~200 kb of reference genome. Subsequently, to improve efficiency in gene mapping, we included more gene-rich regions, where the recombination rate is typically elevated²⁷. Successful primer designs were obtained for 99.6% of the candidate regions, with amplicon size ranging from 270 to 330 bp (Supplementary Fig. 8). Of these, 98% were predicted to be multiplex-competent in a single reaction. In total, 2000 rhAmpSeq markers were designed and synthesized (Supplementary Data 1).

Marker validation in four grapevine families. The 2000 rhAmpSeq marker panel was then evaluated in the four grapevine breeding families representing a wide range of genetic diversity in US breeding programs, including wine grapes, table grapes, wild species, and interspecific hybrids; hereafter, each family is referred to using the two-letter initials preceding its description: (1) HC: “Horizon” × *V. cinerea* B9 (PI588154), a complex F₁ family maintained at Cornell University in Geneva, New York and including *V. vinifera*, *V. cinerea*, *V. aestivalis* var. *lincecumii*, and *V. rupestris*^{6,28}. (2) MN: MN1264 × MN1246, a complex F₁ family maintained at the University of Minnesota Horticultural Research Center in St. Paul, Minnesota and including *V. vinifera*, *V. riparia*, *V. rupestris*, *V. labrusca*, *V. aestivalis*, and *V. cinerea*^{6,29}. (3) RS: *V. riparia* 37 (PI588259) × “Seyval blanc”, an F₂ family maintained at South Dakota State University in Brookings, South Dakota and including *V. riparia*, *V. rupestris*, and *V. aestivalis*³⁰. (4) BC: B37-28 × C56-11, a modified backcross (mBC₁) family maintained at USDA-ARS in Parlier, California and including *V. vinifera* and *V. aestivalis*³¹.

Firstly, to examine amplification and sequencing bias in the rhPCR, we calculated the average read depth for each marker (Fig. 3a). After log (base 10) transformation, sequencing depth was nearly normal in distribution, and 90% of markers ranged from 1- to 100-fold in depth, indicating that the amplification was efficient for most markers, and depth was sufficient for genotyping. Secondly, we checked the reproducibility of the rhAmpSeq platform in generating similar data quantities among 96-well plates of samples, using different DNA extraction protocols and Illumina sequencers. The average sequencing depth per sample was greater than ten for all 96-well plates (Fig. 3b). The number of individuals with depth <10 in the MN family was significantly higher than in other families

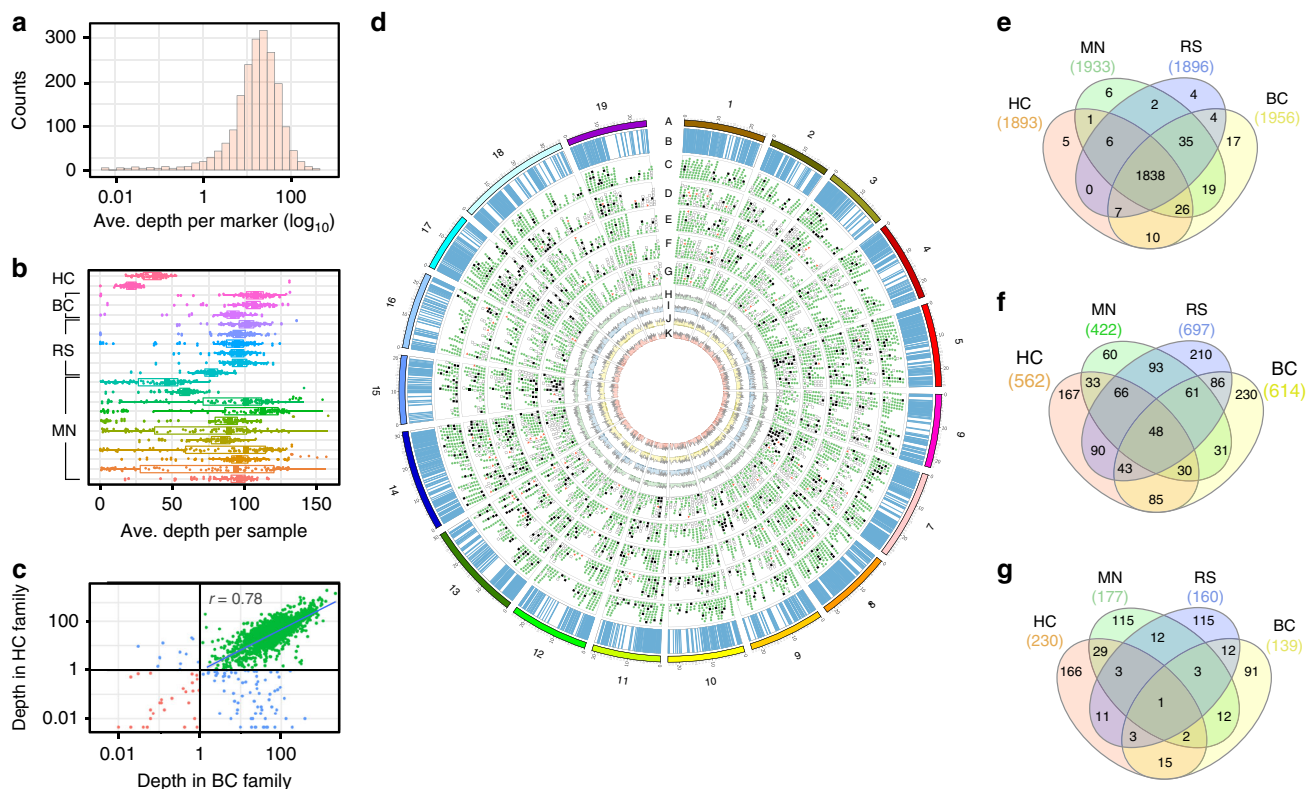


Fig. 3 Performance of 2,000 rhAmpSeq markers in four families. **a** Distribution of average depth of markers on a log₁₀ scale. **b** Average depth per sample in all plates of four families. Each point represents one sample in the plate. The bound box represents the 25th or 75th percentiles and the whiskers extends to 1.5 times of the interquartile range. **c** Correlation of read depth in two families. The Pearson correlation coefficient was calculated for the markers with average depth larger than one in both families. **d** Circos plot of marker performance in four families and the metapopulation. **a** denotes Chromosomes; **b** denotes physical locations of rhAmpSeq core genome markers; **c-g** denote rhAmpSeq markers that failed to return data (red triangle), distorted (black square), monomorphic (white square) and on linkage map (green circle) for consensus genetic map, HC, MN, BC and RS families; **h-k** denote read depth across for each marker in HC, MN, BC, RS families, respectively. **e** Venn diagram of the markers with average read coverage greater than 1 across the four families. **f** Venn diagram of the monomorphic markers in four families. **g** Venn diagram of the distorted markers in four families.

(χ^2 , two-tailed chi-squared test, $P < 1E-6$), as was the variance of read depth, likely due to different DNA extraction protocols. The DNA of the MN family was extracted using an automated magnetic bead pipeline, while the other families were manually extracted using commercial kits that included a column-based purification step. The HC family had less depth than the others, as it was sequenced on a MiSeq, which has a lower output than the NextSeq500 used for the other families. Thirdly, we examined the correlation of marker sequencing depth between two families. After excluding markers with average depth less than one, the Pearson correlation coefficient (r) was 0.78, indicating that sequencing depth is mainly influenced by the composition of the probe and the target sequence, and less so by the genetic background (Fig. 3c).

The performance of each marker is illustrated in a circos plot (Fig. 3d) and summarized in Table 1. Despite no previous testing or troubleshooting of amplification conditions, very few markers displayed null alleles, and 91.9% (1838) of the markers had a mean read depth of at least one in all four families (Fig. 3e). In addition, markers classified as missing, monomorphic, or segregation-distorted were distinct for each family (Fig. 3e-g), and 97.1% of the markers amplified in at least one of the families. As expected, more markers were monomorphic in the F₂ RS family (697) and in the modified backcross BC family (614) with narrow genetic distances shown in Supplementary Fig. 9, than in the wider, multi-species F₁ crosses HC (562) and MN (422). The rhAmpSeq markers had a mean of 5.7 alleles per marker across seven genotyped parental accessions (Supplementary Fig. 10).

Core-genome mapping indicates possible genome divergence.

We constructed parent-specific genetic maps based on the segregating markers in each of the four families. The total genetic distances ranged from 1333.4 to 2224.2 cM (Table 1). The average Pearson’s correlation (r) between physical and genetic positions ranged from 0.79 to 0.93 genome-wide, and the genetic maps covered 95.9–98.2% of the reference genome (Supplementary Data 2). In general, the parental genetic maps were highly similar among parents, but there were some parent- or family-specific anomalies. For example, the distal 30.1% of chromosome 19 was monomorphic (non-segregating) for both parents of HC. Other regions failed to recombine in a specific parent, with the most extreme case being no recombination over the distal 64.4% of chromosome 17 in the female parental map of the MN family (Fig. 4). Parent-specific repression of recombination indicates candidate regions of structural variation in the genome revealed by the core markers, for further exploration.

We analyzed markers excluded from the linkage maps to determine whether they perform poorly in all families. Almost all excluded markers were monomorphic or displayed segregation distortion (Table 1). Most monomorphic markers or distorted markers (47.7% or 77.3%, respectively) were specific to one family, and only 18.6% of the monomorphic markers and 5.5% of the distorted markers were in problematic three or four families (Fig. 3). We found that distorted markers were enriched in minor allele frequency surrounding 0.33, most likely due to the hemizyosity, or null alleles, in one of the parental genomes. In a recently published genome of a highly heterozygous cultivated

Table 1 Summary statistics of the genetic maps of each family and the consensus genetic map.

	HC	MN	RS	BC	Consensus Map
Number of vines	157	1007	504	260	600
Cross type	F ₁	F ₁	F ₂	F ₁	-
rhAmpSeq core genome markers	2000	2000	2000	2000	2000
Markers in linkage groups	1153	1387	1113	1222	1661
Markers that failed to return data	55	14	30	25	31
Monomorphic markers ^a	562(34)	422(48)	697(111)	614(42)	111
Distorted markers ^a	230(47)	187(93)	160(11)	139(50)	228
Male genetic map size (cM)	2224.2	1760.8	1415.8	1787.7	-
Female genetic map size (cM)	1333.4	1519.0	1569.5	1502.9	-
Sex-averaged genetic map size (cM)	-	-	-	-	1198.1
Male genome-wide recombination rate (cM/Mbp)	4.9	3.8	3.1	3.9	-
Female genome-wide recombination rate (cM/Mbp)	2.9	3.3	3.4	3.3	-
Sex-averaged genome-wide recombination rate (cM/Mbp)	-	-	-	-	2.61
Male genome-wide correlation of the genetic and physical map (r)	0.79	0.93	0.79	0.86	-
Female genome-wide correlation of the genetic and physical map (r)	0.86	0.90	0.80	0.86	-
Sex-averaged genome-wide correlation of the genetic and physical map (r)	-	-	-	-	0.95
Genome ^b coverage (%)	95.9%	98.2%	96.9%	96.4%	98.6%

Abbreviation: HC, Horizon × *V. cinerea* B9; MN, MN 1264 × MN 1246; RS, *V. riparia* 37 × Seyval; BC B37-28 × C56-11

^aThe number in the parentheses indicates the number of markers with a null allele segregating according to Mendelian expectations

^bRelative to 12X.v2 version of the PN40024 reference genome

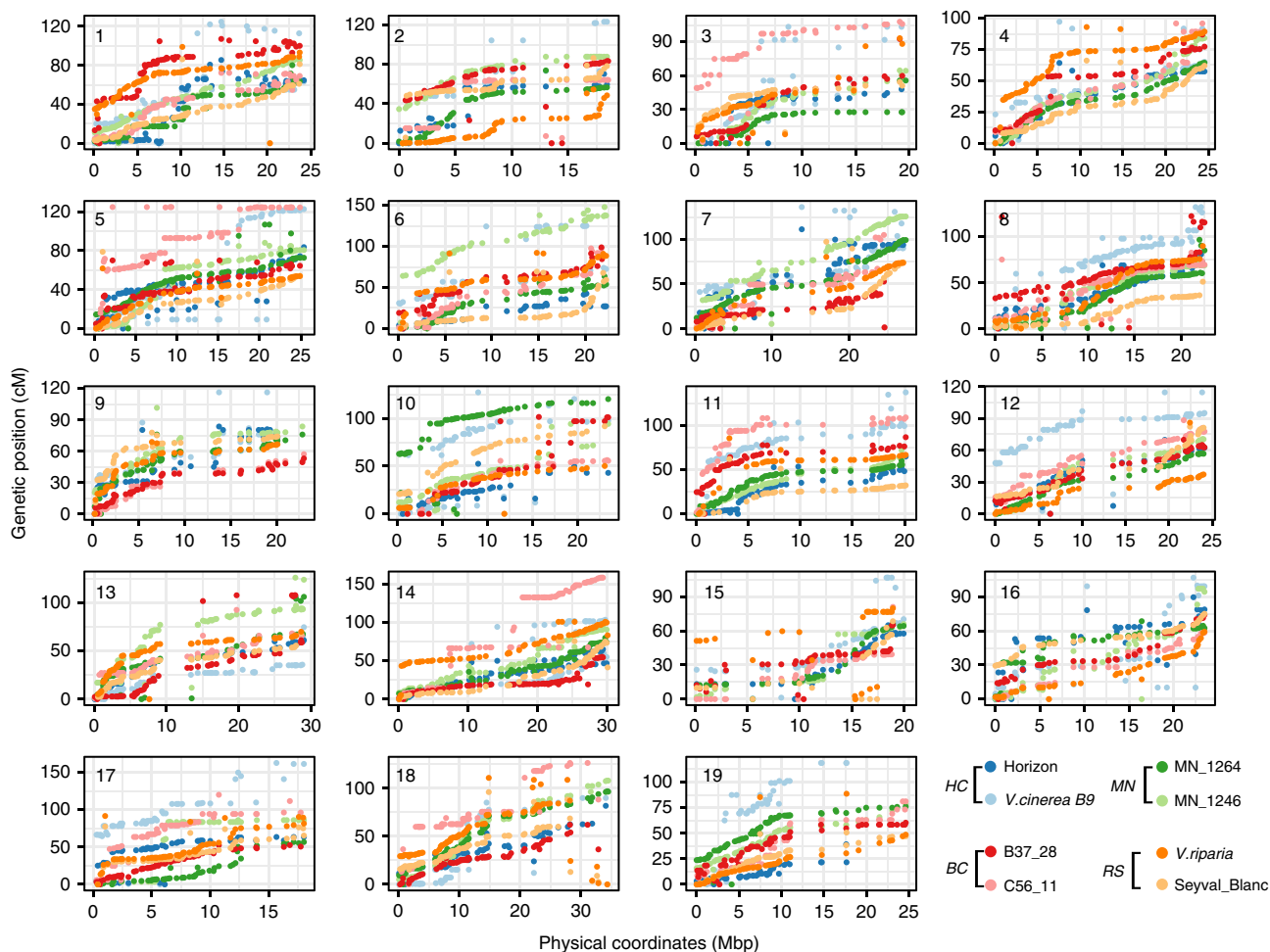


Fig. 4 Relationship between the genetic and physical maps for the nineteen chromosomes of the parents in four families. Genetic distance: cM; Physical maps: Mbp. The genetic distances of the markers were derived from the genetic map developed for each family and physical distances are from version 12X.2 of the PN40024 reference genome.

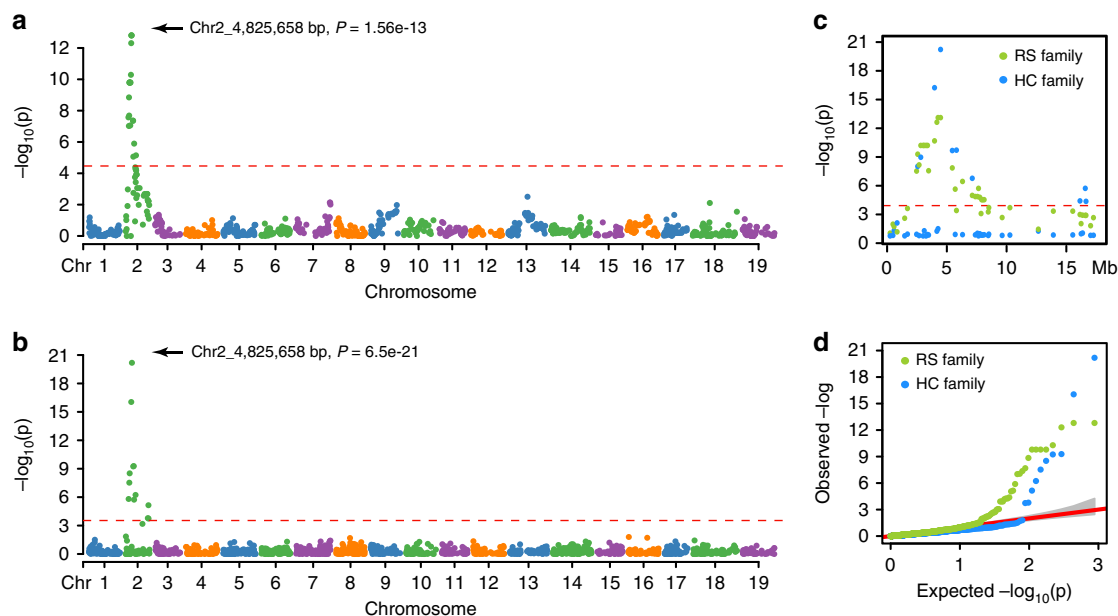


Fig. 5 Association mapping of flower sex loci in the RS and HC families. **a, b** Manhattan plot of GWAS results for flower sex in the RS and HC families, respectively. Bonferroni threshold at 4.5 ($\alpha = 0.05$) is shown in dotted red line. **c** The p -values on chromosome 2 only. **d** QQ-plot for the flower sex GWAS analysis in the RS and HC families. Source data are provided as a Source Data file.

grape, 14.2% of the genes were hemizygous²¹. We tested if the distorted markers segregated in a Mendelian pattern when allowing hemizygosity among the alleles. We found that up to 52% of distorted markers and up to 15% of monoallelic markers followed Mendelian laws of segregation if assuming one or two alleles were hemizygous in the parents (Table 1). While missing one copy of the allele could have a biological basis (deletion in the genome), it could also be attributed to mismatches in the primers resulting in failed PCR amplification. By inferring the genetic variances at the primer regions using the whole genome shotgun sequences of four parents, we found that 83% of the markers with a potential hemizygous allele had at least one mismatch in the primers, which is two times higher than markers with no hemizygous allele. Thus, our primer design pipeline successfully targeted the core genome, avoiding biologically hemizygous sites, but returned null alleles in about 5% of markers per parent. Nevertheless, we were able to place 83% of the 2000 markers in this panel on a consensus genetic map using a joint-linkage mapping population of 600 vines (150 vines from each family), with an average Pearson's correlation (r) between physical and genetic positions of 0.95 across all chromosomes and 98.6% coverage of the physical genome (Supplementary Data 1 and Supplementary Fig. 11). This high genetic mapping rate indicates that most markers should behave as true Mendelian markers in most *Vitis* taxa.

A transferable flower sex marker. In the *Vitis* genus, all of the wild species are dioecious while the domesticated grapevine *V. v. ssp. vinifera* is hermaphroditic³². The region around 5Mbp on chromosome 2 has been identified in several linkage mapping and population genetic studies, and the boundary of the sex locus and the sex determining genes were proposed but are still under debate^{32–36}. Here, we used the sex locus to assess the performance of our rhAmpSeq markers in genetic mapping. A total of 1712 and 1784 post-imputed, filtered markers were analyzed for association with the flower sex trait measured in 146 and 106 vines from the HC and RS families, respectively. After Bonferroni correction for multiple comparisons, 13 and 17 markers, respectively, were significantly associated with flower sex. Marker

chr2_4825658 (chromosome 2 at position 4,825,658 bp) was the most significant marker in both the HC ($P = 6.5E-21$, male (M) allele dominant over female (f)) and RS ($P = 1.56E-13$, hermaphroditic (H) allele dominant over female (f)) families, and was concordant with flower sex phenotype for 143/146 HC progeny (97.9%) and for 100/102 RS progeny (95.3%) (Fig. 5). In our previous study of the sex locus using genotyping-by-sequencing markers, the distinct genetic markers associated with flower sex inconsistently spanned a 1 Mb region in different mapping families⁶. In contrast, here the same marker was most significant in both HH \times Mf and Hf-selfed families, which emphasizes the transferability of these core genome markers.

Discussion

A set of universal genetic markers that work for related taxa is desired in many genetic studies. In marker-assisted breeding, universal markers can be used in crosses between distant relatives to generate heterosis or introgress useful alleles^{37–39}. In molecular ecology and evolutionary studies, universal markers allow comparison of genetic characters among related species^{40,41}. In genera or families containing many economically important species, transferable, universal markers can decrease the time and effort required for marker development^{42,43}. While the transferability for low-throughput microsatellite (SSR) markers is relatively good, ranging from 27% to 77% in the different taxa of plants and animals⁴⁴, the transferability of high-throughput SNP genetic markers have been as low as 2%^{17,45}.

In this study, we developed and validated a pipeline for designing universal markers that work across the diverse *Vitis* genus, which were diverged 20 Mya. Using the rhAmpSeq targeted sequencing platform, 93% of markers returned data for all four families tested, and around 70% of markers were polymorphic in every family. All parental genetic maps were highly correlated with physical position in the PN40024 reference genome ($r = 0.86$ to 0.95). Although 10 to 20% of the markers in each family deviated from expected Mendelian segregation ratios, these markers were family-specific and were clustered on particular chromosomes. The vast majority of markers were informative for consensus genetic map construction, indicating high

marker transferability. Furthermore, in two families where the sex trait was analyzed, the same marker explained the most phenotypic variation and was the most significantly associated. This result suggests that not only are random markers transferable, but functional markers are also transferable. Thus, it appears that markers designed to target a genus-wide core genome are transferable in all key aspects, including amplification, polymorphism, segregation, and marker-trait association.

The design of transferable markers depended on the construction of a genus-wide core genome comprised of collinear, syntenic blocks. Previously, markers designed from shotgun resequencing had limited transferability because only local genetic variation could be assessed, and large and complex structural variation was often overlooked. Any long collinear block conserved within a structurally diverse taxon is suggestive of strong selection against structural variation within the block, and markers designed within such blocks are more likely to consistently occur in different species with consistent segregation patterns. For this study, linked-read scaffolding of de novo assembly enabled the identification of collinear blocks at a relatively inexpensive price point (about \$3000 per 475 Mbp genome). By using genus-wide sequence data to design primers targeting conserved sequences flanking regions of moderate polymorphism in the inferred core genome, we obtained markers that reliably returned informative data in most cases.

Previously, we found that the AmpSeq genotyping platform outperforms GBS for highly diverse and heterozygous species, due to reduced missing data, increased coverage and increased accuracy at heterozygote sites, as well as elevated transferability among species⁶. In contrast to SNP arrays or Kompetitive allele-specific PCR (KASP), which typically target two alleles per marker, or site, the AmpSeq genotyping platform allows identification of numerous alleles as short haploblocks because the entire amplified target (typically 200–250 bp) is sequenced via NGS. The rhAmpSeq markers developed in this study had a mean of 5.7 alleles per marker across seven genotyped parental accessions (Supplementary Fig. 10). Compared to biallelic SNPs, these multiallelic haplotype markers simplify phasing along the chromosome and provide more information about ancestry. The high information content, even coverage, and unbiased sequencing of rhAmpSeq amplicons make this platform applicable for population genetics and ecology studies. Relative to AmpSeq, the rhAmpSeq technology simply adds an RNA base and blocker DNA at the 3' end of each primer. When the match is perfect between the primers and template, this RNA-base and blocker are cleaved by RNase H2 enzyme⁴⁶. This step increases the genotyping specificity and increases the multiplexing capacity up to 5000 markers per reaction⁴⁶.

Here we developed a strategy for genus-wide haplotype marker design considering the syntenic core genome and genus-wide polymorphism. In combination with the rhAmpSeq platform, this genotyping pipeline can be easily adapted for other taxa for ecological and evolutionary studies, QTL mapping, GWAS, and molecular breeding. The costs for rhAmpSeq highly depend upon the number of markers and samples, as primer and reagent prices are subject to scale, and sequencing costs are reduced by greater sample multiplexing. Given existing DNA and perfect efficiency, a large project could generate sequencing reads for 2000 rhAmpSeq markers for as little as \$4 per sample plus rhAmpSeq reagent costs. With the recent availability of inexpensive linked-read sequencing, this cost-effective strategy will be particularly useful for crops (or other eukaryotes) with poor genomic resources where it will now be possible to develop core genomes, and especially for genera or families (e.g., Poaceae or Rosaceae) that would benefit from a universal set of highly transferable markers.

Methods

DNA processing and genome assembly. For de novo sequenced vines, rapidly expanding leaves about two-centimeter (cm) long were collected for six samples from vineyards in Geneva, New York (Supplementary Table 1). High-molecular-weight (HMW) genomic DNA (gDNA) was isolated using a CTAB protocol modified from Japelaghi et al. and Haley et al.^{47,48}. The genomic DNA was quantified with Quant-iT kits using a Qubit fluorometer (Thermo Fisher Scientific), quality checked with a Thermo Nanodrop™ 2000 Spectrophotometer (Thermo Fisher Scientific), and sized via Pulsed Field Gel Electrophoresis. The bulk of the gDNA smear was required to be >60 kb before further processing and was typically centered on 100 kb. For seven of the genomes listed in Supplementary Table 1, HMW DNA was shipped to 10X Genomics (10X Genomics Inc., Pleasanton, CA, USA) for preparation and sequencing of libraries following their standard protocols (10X Genomics; Pleasanton, CA). Each 10X Genomics library was sequenced to between 32- and 66-fold coverage on an Illumina NovaSeq sequencer to generate linked-reads with a mean read length of 139.5 bp after trimming. The whole genome sequencing (WGS) library preparation and sequencing was performed at 10X Genomics. The linked-read data were assembled using Supernova v.2.0.1 assembler⁴⁹ with default settings. The weighted mean molecule size was estimated by the Supernova software as 63.18 kb and mean read coverage as ~68 fold. The BUSCO score was calculated based on lineage-specific sets of Eudicotyledons odb10 with genome model (BUSCO: version 3.0.2, AUGUSTUS: Version 3.3).

Syntenic core-genome construction. Repetitive regions in both the reference genome and the assemblies were masked using a kmer frequency-based approach with BBduk, part of the BBTools package (version 35.50)⁵⁰. Sequences with a kmer ($K = 31$) frequency larger than two were replaced with Ns. The masked assembly was aligned to the reference genome, PN40024 (version 12X.v2)²⁶, using Minimap2 with parameter presets tuned for cross-species alignment, denoted as “asm10” in the manual⁵¹. By the definition in minimap2, each alignment has a global alignment score larger than 400 (defined by $-z$) and with >90% identity with the reference (defined by $-x$ asm10). The results were transformed to a bed-like format for chaining and identifying one-to-one matches of chains with at least three alignments derived from minimap2 using `quota_alignment` (<https://github.com/tanghaibao/quota-alignment>)⁵². The total length of each collinear core-genome alignment that is larger than 10 kb were kept in the downstream analysis. In this study, we defined the core genome as the chromosomal regions of reference PN40024 that had collinear, syntenic alignment with all other genome assemblies. The core genome coverage, gene density, transposable element (TE) density, and other genome features were calculated with window size of 1 Mb using BEDTools (v2.27.1). The correlation between each pair of genome features was calculated using Spearman's rank correlation coefficient in R software (Version 3.5.0).

Genus-wide variant calling. A total of 47 *Vitis* accessions with 3- to 93-fold paired-end Illumina sequencing data were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). If one accession had more than 20-fold read coverage, we randomly down-sampled to 20-fold read coverage for variant calling to avoid bias in variant calling from high-depth samples and to save processing time. We also sequenced seven accessions with 8- to 160-fold paired-end sequencing using the Illumina HiSeq 2500 platform. Reads were processed and variants were called based on PN40024 (version 12X.v2)²⁶ using the Sentieon DNA Software Package (version, Golden Helix)⁵³ with default settings. This Sentieon package is a speed-up software that rebuilt the Genome Analysis Toolkit HaplotypeCaller and returns the same result as GATK 3.3. Principal component analyses (PCA) was conducted using R/Bioconductor Package SNPrelate⁵⁴. To avoid the strong influence of clustered SNPs in the PCA analysis, the SNPs were filtered using LD-based pruning algorithm implemented in SNPrelate with LD threshold 0.2. We arbitrarily selected 20 *V. vinifera* samples and 20 non-*vinifera* samples.

Marker design pipeline. The VARIANT CALLING FORmat (VCF) file generated from the genus-wide variants calling was loaded into R software. For each aligned region in the core-genome, the length, diversity and missing rate was calculated. The regions that were shorter than 200 bp, with diversity larger than 7% or smaller than 2%, or with average missing rate larger than 50% were removed. These steps were conducted in R using the bioconductor (version 3.8) package VariantAnnotation⁵⁵. The candidate regions were then picked to ensure one marker per 200 kb. If no qualified candidate region could be found in a 1 Mbp window, we included the regions that had highest coverage in the core genome construction. To achieve a better representation of gene rich regions with high recombination rates, we included more candidate regions with high gene density. A total of 2500 candidate regions were sent to Integrated DNA Technologies, Inc. (IDT, Coralville, IA, USA) for primer design and pooling compatibility test. Primers could be designed for 99.6% of the regions, and 98.1% of them were pooling compatible in a single PCR amplification reaction. A total of 2000 rhAmpSeq primer pair assays were synthesized by IDT.

rhAmpSeq sequencing and genotyping. The DNA of the MN family was extracted by Intertek AgriTech (Sweden) using an automated magnetic bead

pipeline with beadex kit provided by LGC (Teddington, United Kingdom). The DNA of other families was extracted using QIAGEN DNeasy 96 Plant Kits manually. We modified the protocol to include 3% w/v PVP40 to the lysis buffer prior to extractions to remove PCR inhibitors. rhAmpSeq amplification enrichment using the 2000 marker panel was conducted following manufacturer's protocol. Briefly, the first PCR used 14 cycles with annealing temperature at 61 °C for each sample. The PCR products were diluted 1:20 and indexed with IDT indexing primers using 24 cycles with an annealing temperature at 60 °C. The indexed PCR products were pooled, cleaned with Agencourt AMPure beads, quantified, and sequenced on an Illumina (Illumina, San Diego, CA, USA) MiSeq (2 × 150 bp) or NextSeq (2 × 150 bp) sequencer. rhAmpSeq sequencing data for all the four genetic mapping families were initially analyzed using the Perl script `analyze_amplicon.pl` (https://github.com/avinashkarn/analyze_amplicon/blob/master/analyze_amplicon.pl), and later re-analyzed with an upgraded pipeline optimized for rhAmpSeq data analysis (https://bitbucket.org/cornell_bioinformatics/amplicon). The pipelines de-multiplex the sequencing reads based on PCR primer sequence, obtain haplotype variants for each marker across all vines in the four families, and generate a sample to haplotype allele matrix. Both PCR errors and sequencing errors produce false haplotype alleles. To correct these errors, haplotype variants caused by genotyping errors were collapsed by filtering out nucleotide sites within homo-polymer repeats or deviating from expected Mendelian segregation ratios of bi-parental families. In rare cases where a haplotype marker had no sites segregating at the expected ratio, the site with the biggest minor allele frequency was used to represent the haplotype allele. Monomorphic markers and markers with greater than seventy five percent missing data in 'hapgeno' file were manually removed from the further analysis. Finally, using a custom Perl script, `haplotype_to_VCF.pl` (https://github.com/avinashkarn/analyze_amplicon/blob/master/haplotype_to_VCF.pl), the four most frequent haplotype alleles for each marker (within a family) in the `hapgeno` file were converted to a VCF file, where each haplotype allele of a marker was converted to a pseudo A, C, G or T allele, for further marker validation analyses that are discussed hereafter.

Imputation and filtering. The raw converted VCF files for each grapevine family were imported in TASSEL (Trait Analysis by association, Evolution and Linkage) 5.2.51 software⁵⁶ and the genotypes were imputed using the LD-kNNi imputation plugin also known as LinkImpute (v1.1.4)⁵⁷ using the default parameters (High LD Sites = 30, Number of nearest neighbors = 10, and Max distance between site to find LD = 10,000,000). Post-imputation, vines with >90% missing data were removed from the analysis.

Multidimensional scaling for quality control analysis. Post-imputed and filtered markers were used to calculate a genome-wide pairwise identity-by-state (IBS) distance matrix for each family in TASSEL software using 1-IBS followed by Multidimensional scaling (MDS) analysis⁵⁶. The first three principal coordinates in the multidimensional scaling of each family were graphically depicted using the R statistical software. The MDS analysis predicted the hidden population structure by separating four clusters of progeny vines relative to their corresponding parents and grandparents, and indicated vines that were self-pollinated, outcrosses, or mislabeled, which were removed from linkage mapping and GWAS analysis.

Construction of genetic maps. Genetic maps were constructed in Lep-MAP3 v.0.2⁵⁸ (LM3) using the VCF file of post-imputed and filtered markers as well as pedigree information for each family. The following LM3 modules and steps were used to construct the genetic maps: (1) *ParentCall2* module of LM3 was used to call parental genotypes; (2) the resulting output was filtered by using *Filtering2* module (parameter *dataTolerance* = 1.00E-3 for F₁ and mBC₁ families and 1.00E-10 for the F₂ family), and the markers were filtered out based on a two-sided χ^2 (chi-squared) test (testing if the allele ratio is significantly deviated from the expected mendelian ratio, at the above tolerance thresholds) or monomorphism; (3) *SeparateChromosomes2* module was used to identify linkage groups using logarithm of odds (LOD) score limit ranging from none to 20 for the individual family (Table 1); (4) Finally, *OrderMarkers2* module was used to compute the parental genetic distances (sex specific for the F₁ and mBC₁ families and sex averaged for the F₂ family) of the markers in the linkage groups using 20 iterations per group. Correlation plots of genetic and physical distances of individual markers per chromosome in each family were plotted to evaluate the consistency of the maps, genome organization and structural variation.

Further, a consensus genetic map was constructed in LM3, where 150 progeny vines from each family were randomly chosen, and their genotypes were merged into a single VCF file in TASSEL using 'Merge Genotype Table' plugin. The merged VCF file and pedigree information containing all four families were used in LM3 as an input to construct the sex averaged consensus genetic map as described above.

Genome-wide association study of flower sex. The association between a well-characterized trait (flower sex) and genotypes was used to further evaluate rhAmpSeq marker transferability. Specifically, a genome-wide association study (GWAS) was conducted to map the flower sex locus in HC (F₁) and RS (F₂). The male allele (M) is dominant to hermaphroditic (H), which is dominant to female

(f), that is, M > H > f. HC represents a cross of homozygous hermaphroditic flowers (HH) emasculated and pollinated with pollen from a male (Mf) vine, and should segregate 1 male (MH): 1 hermaphrodite (Hf). RS represents self-pollination of heterozygous hermaphroditic flowers (Hf) and should segregate three hermaphrodite (HH/Hf/Hf): one female (ff).

GWAS was conducted in TASSEL on post-imputed and filtered markers in the two families with their respective flower sex phenotypic values and phenotypes using a mixed linear model (MLM). As covariates, the first two principal components of MDS analysis were used to correct for population structure (P), and the kinship matrix (K), the proportion of alleles shared between each pair of vines was used to correct for familial relatedness as covariates. The Eq. (1) for MLM (P + K) model was:

$$y = X\alpha + P\beta + Ku + \epsilon \quad (1)$$

where, y is a vector of a phenotypic data, α is the fixed effects related to the marker, β is a vector of the fixed effects related to the population structure, u is a vector of the random effects related to the relatedness among the vines, and ϵ is a vector of the residual effects. X is the genotypes of the marker, P is the matrix of principle components, K is the centered identity by state (IBS) kinship matrix.

Phenotypic variability explained by each significant marker was estimated by R^2 values generated in MLM statistics output from TASSEL software⁵⁶. Further, the Bonferroni-corrected threshold was determined for each association analysis using $1/N$ ($\alpha = 0.05$), where N is the number of markers tested, and quantile-quantile (QQ) plots were utilized to examine model fitness for the flower sex trait in each family.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this Article is available as a Supplementary Information file. The datasets generated and analyzed during the current study are available from the corresponding author upon request. All the raw sequencing reads that support the findings of this study and its supplementary information file have been deposited in the in the National Center for Biotechnology Information Sequence Read Archive (SRA) and are accessible through BioProject ID PRJNA281110 [<https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA281110>]. All the information of SRA, including project number, total base pairs, and name of accession are list in Supplementary Table 2. The source data underlying Fig. 5 are provided as a Source Data file.

Code availability

All custom scripts are available at Github (https://github.com/avinashkarn/analyze_amplicon) or at Bitbucket (https://bitbucket.org/cornell_bioinformatics/amplicon).

Received: 1 November 2019; Accepted: 19 December 2019;

Published online: 21 January 2020

References

- Botstein, D., White, R., Skolnick, M. & Davis, R. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
- Tautz, D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res.* **17**, 6463–6471 (1989).
- Davey, J. W. et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510 (2011).
- Miller, M., Dunham, J., Amores, A., Cresko, W. & Johnson, E. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* **17**, 240–248 (2007).
- Elshire, R. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379 (2011).
- Yang, S. et al. A next-generation marker genotyping platform (AmpSeq) in heterozygous crops: a case study for marker-assisted selection in grapevine. *Hortic. Res.* **3**, 16002 (2016).
- Grattapaglia, D. et al. High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. *BMC Plant Biol.* **11**, 65 (2011).
- Wan, Y. et al. A phylogenetic analysis of the grape genus (*Vitis* L.) reveals broad reticulation and concurrent diversification during neogene and quaternary climate change. *BMC Evol. Biol.* **13**, 141 (2013).
- McGovern, P. E. *Ancient wine: the search for the origins of viticulture* (Princeton University Press, 2003).

10. Myles, S. et al. Genetic structure and domestication history of the grape. *Proc. Natl Acad. Sci. USA* **108**, 3530–3535 (2011).
11. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
12. Zhang, J. et al. Cold-resistance evaluation in 25 wild grape species. *Vitis* **51**, 153–160 (2012).
13. Fresnedo-Ramírez, J. et al. An integrative AmpSeq platform for highly multiplexed marker-assisted pyramiding of grapevine powdery mildew resistance loci. *Mol. Breed.* **37**, 145 (2017).
14. Riaz, S., Tenscher, A. C., Ramming, D. W. & Walker, M. A. Using a limited mapping strategy to identify major QTLs for resistance to grapevine powdery mildew (*Erysiphe necator*) and their use in marker-assisted breeding. *Theor. Appl. Genet. Theor. Angew. Genet.* **122**, 1059–1073 (2011).
15. Miller, J. M., Kijas, J. W., Heaton, M. P., McEwan, J. C. & Coltman, D. W. Consistent divergence times and allele sharing measured from cross-species application of SNP chips developed for three domestic species. *Mol. Ecol. Resour.* **12**, 1145–1150 (2012).
16. Bauer, E. et al. Intraspecific variation of recombination rate in maize. *Genome Biol.* **14**, R103 (2013).
17. Vezzulli, S. et al. A SNP transferability survey within the genus *Vitis*. *BMC Plant Biol.* **8**, 128 (2008).
18. Michelizzi, V. N. et al. A global view of 54,001 single nucleotide polymorphisms (SNPs) on the Illumina BovineSNP50 BeadChip and their transferability to water buffalo. *Int. J. Biol. Sci.* **7**, 18–27 (2010).
19. Wu, J. J. et al. Investigation of transferability of BovineSNP50 BeadChip from cattle to water buffalo for genome wide association study. *Mol. Biol. Rep.* **40**, 743–750 (2013).
20. Lu, F. et al. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun.* **6**, 6914 (2015).
21. Zhou, Y. et al. The population genetics of structural variants in grapevine domestication. *Nat. Plants* **5**, 965–979 (2019).
22. Veeckman, E., Ruttink, T. & Vandepoele, K. Are we there yet? Reliably estimating the completeness of plant genome sequences. *Plant Cell* **28**, 1759–1768 (2016).
23. Dobosy, J. et al. RNase H-dependent PCR (rhPCR): improved specificity and single nucleotide polymorphism detection using blocked cleavable primers. *BMC Biotechnol.* **11**, 80 (2011).
24. Patel, S. et al. Comparison of three assembly strategies for a heterozygous seedless grapevine genome assembly. *BMC Genomics* **19**, 57 (2018).
25. Minio, A., Lin, J., Gaut, B. S. & Cantu, D. How single molecule real-time sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. *Front. Plant Sci.* **8**, 826 (2017).
26. Canaguier, A. et al. A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). *Genomics Data* **14**, 56–62 (2017).
27. Rodgers-Melnick, E. et al. Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl Acad. Sci. USA* **112**, 3823–3828 (2015).
28. Hyma, K. E. et al. Heterozygous mapping strategy (HetMappS) for high resolution genotyping-by-sequencing markers: a case study in grapevine. *PLoS ONE* **10**, e0134880 (2015).
29. Teh, S. L. et al. Genetic dissection of powdery mildew resistance in interspecific half-sib grapevine families using SNP-based maps. *Mol. Breed.* **37**, 1 (2016).
30. Yang, S. et al. Next generation mapping of enological traits in an F2 interspecific grapevine hybrid family. *PLoS ONE* **11**, e0149560 (2016).
31. Ramming, D. W. et al. Identification of race-specific resistance in North American *Vitis* spp. limiting *Erysiphe necator* hyphal growth. *Phytopathology* **102**, 83–93 (2011).
32. Zhou, Y., Massonnet, M., Sanjak, J. S., Cantu, D. & Gaut, B. S. Evolutionary genomics of grape (*Vitis vinifera* spp. *vinifera*) domestication. *Proc. Natl Acad. Sci. USA* **114**, 11715–11720 (2017).
33. Fechter, I. et al. Candidate genes within a 143 kb region of the flower sex locus in *Vitis*. *Mol. Genet. Genomics* **287**, 247–259 (2012).
34. Battilana, J. et al. Linkage mapping and molecular diversity at the flower sex locus in wild and cultivated grapevine reveal a prominent SSR haplotype in hermaphrodite plants. *Mol. Biotechnol.* **54**, 1031–1037 (2013).
35. Conner, P. et al. Development and characterization of molecular markers associated with female plants in muscadine grape. *J. Am. Soc. Hortic. Sci.* **142**, 143–150 (2017).
36. Lewter, J. et al. High-density linkage maps and loci for berry color and flower sex in muscadine grape (*Vitis rotundifolia*). *Theor. Appl. Genet.* **132**, 1571–1585 (2019).
37. Chagné, D. et al. Cross-species transferability and mapping of genomic and cDNA SSRs in pines. *Theor. Appl. Genet.* **109**, 1204–1214 (2004).
38. Brondani, R., Williams, E., Brondani, C. & Grattapaglia, D. A microsatellite-based consensus linkage map for species of *Eucalyptus* and a novel set of 230 microsatellite markers for the genus. *BMC Plant Biol.* **6**, 20 (2006).
39. Diaz, A. et al. A consensus linkage map for molecular markers and quantitative trait loci associated with economically important traits in melon (*Cucumis melo* L.). *BMC Plant Biol.* **11**, 111 (2011).
40. Singh, B. K. et al. Transferability of Brassica-derived microsatellites to related genera and their implications for phylogenetic analysis. *Natl Acad. Sci. Lett.* **35**, 37–44 (2012).
41. Bernardes, C. et al. Transferability of Psidium microsatellite loci in Myrtaeae (*Myrtaceae*): a phylogenetic signal. *Euphytica* **214**, 150 (2018).
42. Kuleung, C., Baenziger, P. & Dweikat, I. Transferability of SSR markers among wheat, rye, and triticale. *Theor. Appl. Genet.* **108**, 1147–1150 (2004).
43. Pan, L. et al. EST-SSR marker characterization based on RNA-sequencing of *Lolium multiflorum* and cross transferability to related species. *Mol. Breed.* **38**, 80 (2018).
44. Barbará, T. et al. Cross-species transfer of nuclear microsatellite markers: potential and limitations. *Mol. Ecol.* **16**, 3759–3767 (2007).
45. Chagné, D. et al. Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS ONE* **7**, e31745 (2012).
46. Design targeted amplicon sequencing panels rhAmpSeq IDT. <https://www.idtdna.com/pages/products/next-generation-sequencing/amplicon-sequencing/custom-rhampseq-panels> (2019).
47. Japelaghi, R. H., Haddad, R. & Garoosi, G.-A. Rapid and efficient isolation of high quality nucleic acids from plant tissues rich in polyphenols and polysaccharides. *Mol. Biotechnol.* **49**, 129–137 (2011).
48. Healey, A., Furtado, A., Cooper, T. & Henry, R. J. Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods* **10**, 21 (2014).
49. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).
50. Bushnell, B. BBMap: a fast, accurate, splice-aware aligner. <https://sourceforge.net/projects/bbmap/> (2014).
51. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
52. Tang, H. et al. Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinforma.* **12**, 102 (2011).
53. Kendig, K. I. et al. Sentieon DNaseq variant calling workflow demonstrates strong computational performance and accuracy. *Front. Genet.* **10**, 736 (2019).
54. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
55. Obenchain, V. et al. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **30**, 2076–2078 (2014).
56. Bradbury, P. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
57. Money, D. et al. LinkImpute: fast and accurate genotype imputation for nonmodel organisms. *G3: Genes, Genomes, Genet.* **5**, 2383–2390 (2015).
58. Rastas, P. Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics* **33**, 3726–3732 (2017).

Acknowledgements

We would like to thank Linda Cote, Peter Schweitzer, and the Cornell University Biotechnology Resource Center for providing the necessary infrastructure, personnel, and expertise to amplify and sequence the DNA samples. Mike Colizzi and Steve Luce helped maintain vineyard plantings and collect tissue samples at Cornell University. The US Department of Agriculture (USDA)-National Institute of Food and Agriculture (NIFA) Specialty Crop Research Initiative provided funding for this project (award No. 2017-51181-26829). Additional funding supported this work, including USDA-Agricultural Research Service CRIS projects 8062-21000-044-00D and 8060-21220-007-00-D; the National Science Foundation project NPGI PRFB1523793; and the USDA-NIFA Hatch project SD00R668-18. Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. IDT and 10X Genomics co-authors did not have influence over which data was or was not included in the manuscript.

Author contributions

L.C., Q.S. and J.P.L. initiated the idea to design universal markers applicable across the *Vitis* genus. X.X. implemented techniques for high molecular weight DNA isolation. D.W., M.S.C., S.W. and D.C. constructed seven 10X de novo assemblies. S.P. and A.F. sequenced and assembled *V. riparia* 37, and optimized the re-assembly of Sultania. C.Z. setup the pipeline and constructed the core genome. A.K. and C.Z. conducted the *Vitis*-wide diversity study. C.Z. picked candidate regions for marker design. Y.S., A.N. and Y.B. designed the primers targeting the candidate regions. Q.S. and A.K. developed the rhAmpSeq analysis pipeline. A.K. constructed the genetic maps and performed GWAS.

A.K., C.Z., Q.S. and L.C. summarized marker validation in four families. B.R., A.F., M.C. and C.A.L. provided the validation populations. C.Z. and A.K. drafted the paper with L.C., J.G., Q.S. and all co-authors contributed to paper review and revisions.

Competing interests

Authors affiliated with Integrated DNA Technologies, Inc. (IDT, A.N., Y.S., and Y.B.) or 10X Genomics (M.S.C., D.C., S.W) are employees receiving compensation in the form of salary from their respective company. Those companies offer equipment and/or reagents for sale similar to those described in the manuscript. All other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-019-14280-1>.

Correspondence and requests for materials should be addressed to L.C.-D.

Peer review information *Nature Communications* thanks Korbinian Schneeberger, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <http://creativecommons.org/licenses/by/4.0/>.

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2020