2-2016

A Next-generation Marker Genotyping Platform (AmpSeq) in Heterozygous Crops: A Case Study for Marker-assisted Selection in Grapevine

SHanshan Yang

Jonathan Fresnedo-Ramírez

Minghui Wang

Linda Cote

Peter Schweitzer

See next page for additional authors

Authors
SHanshan Yang, Jonathan Fresnedo-Ramírez, Minghui Wang, Linda Cote, Peter Schweitzer, Paola Barba, Elizabeth M. Takacs, Matthew Clark, James Luby, David C. Manns, Gavin Sacks, Anna Katharine Mansfield, Jason Londo, Anne Fennell Dr, David Gadoury, Bruce Reisch, Lance Cadle-Davidson, and Qi Sun

OPEN

Citation: Horticulture Research (2016) **3**, 16002; doi:10.1038/hortres.2016.2

www.nature.com/hortres

ARTICLE

A next-generation marker genotyping platform (AmpSeq) in heterozygous crops: a case study for marker-assisted selection in grapevine

Shanshan Yang[1], Jonathan Fresnedo-Ramírez[2], Minghui Wang[2], Linda Cote[3], Peter Schweitzer[3], Paola Barba[4], Elizabeth M Takacs[1], Matthew Clark[5], James Luby[5], David C Manns[6], Gavin Sacks[7], Anna Katharine Mansfield[6], Jason Londo[8], Anne Fennell[9], David Gadoury[10], Bruce Reisch[1], Lance Cadle-Davidson[8] and Qi Sun[2]

Marker-assisted selection (MAS) is often employed in crop breeding programs to accelerate and enhance cultivar development, via selection during the juvenile phase and parental selection prior to crossing. Next-generation sequencing and its derivative technologies have been used for genome-wide molecular marker discovery. To bridge the gap between marker development and MAS implementation, this study developed a novel practical strategy with a semi-automated pipeline that incorporates trait-associated single nucleotide polymorphism marker discovery, low-cost genotyping through amplicon sequencing (AmpSeq) and decision making. The results document the development of a MAS package derived from genotyping-by-sequencing using three traits (flower sex, disease resistance and acylated anthocyanins) in grapevine breeding. The vast majority of sequence reads (⩾99%) were from the targeted regions. Across 380 individuals and up to 31 amplicons sequenced in each lane of MiSeq data, most amplicons (83 to 87%) had <10% missing data, and read depth had a median of 220–244×. Several strengths of the AmpSeq platform that make this approach of broad interest in diverse crop species include accuracy, flexibility, speed, high-throughput, low-cost and easily automated analysis.

Horticulture Research (2016) **3**, 16002; doi:10.1038/hortres.2016.2; Published online: 17 February 2016

INTRODUCTION

Marker-assisted selection (MAS) is now commonly employed in perennial crop breeding programs to pursue the acceleration of cultivar development.[1–3] In particular, MAS has been shown to provide advantages for selection during the juvenile phase;[4,5] for pyramiding disease resistance genes;[6,7] and for replacing expensive, time-consuming or technically difficult traits.[8,9] Simply inherited traits with Mendelian or near-Mendelian segregation patterns are major targets for MAS. Examples of MAS have been reported for seedlessness and flower sex in grape, and disease resistance in apple, grape and tomato breeding.[1,10,11] Markers have also been applied to quantitatively inherited traits, especially those with major quantitative trait loci (QTL) effect, including fruit acidity in peach,[12] fruit size in tomato,[13] peach and cherry,[14] grain yield in rice[15] and drought tolerance in chickpea.[9]

The development of molecular markers requires the detection of association between target traits and genotypes. Two approaches are often used to detect such associations: (a) QTL analysis with structured families, and (b) genome-wide association study, which takes advantage of linkage disequilibrium (LD) in diverse germplasm to capture the linkage between markers and causal genes.[16,17] However, for highly heterozygous and diverse crops, such as grape, genome-wide association study has limitations.[18–20] LD decays rapidly in species of *Vitis*, in which the square of the correlation coefficient (r^2) declines to 0.1 within 2.7 cM, making genome-wide association study an unsuitable method for genetic mapping with current genotyping platforms, such as single nucleotide polymorphism (SNP) microarrays.[21,22] QTL analysis in mapping families has been a more effective method for genetic mapping in species with diverse backgrounds.[21]

Multiple genotyping platforms and molecular marker types have been utilized for MAS in highly heterozygous crops, including simple sequence repeats (SSR) and single-locus or multi-locus SNP assays. SSRs are particularly well-suited for MAS because of their multi-allelic nature and high transferability among distinct species or genera,[23–25] which enables the analysis of complex crossing involving progenitors with multiple interspecific hybridizations. However, SSR as a genotyping platform has its own disadvantages including low-throughput, labor-intensive and time-consuming.[26] Low density of SSR markers could cause loss of linkage between markers and causal genes, or lack of segregation in certain families.[27,28] SNP microarrays emerged as an alternative high-throughput genotyping platform.[29] Commercially available high density oligonucleotide arrays allow parallel genotyping for thousands of individuals or markers.[30] However, SNP microarrays are closed platforms suffering from ascertainment bias,[31] resulting in poor flexibility and poor transferability across diverse germplasm.[32] In addition, the cost of microarray pre-design is still a major obstacle in adopting SNP arrays in horticultural breeding programs.[33,34]

[1]Horticulture Section, School of Integrative Plant Science, Cornell University, Geneva, NY 14456, USA; [2]Bioinformatics Facility, Cornell University, Ithaca, NY 14853, USA; [3]Institute of Biotechnology, Cornell University, Ithaca, NY 14853, USA; [4]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA; [5]Department of Horticultural Science, University of Minnesota, St Paul, MN 55108, USA; [6]Department of Food Science, Cornell University, Geneva, NY 14456, USA; [7]Department of Food Science, Cornell University, Ithaca, NY 14853, USA; [8]USDA-ARS Grape Genetics Research Unit, Geneva, NY 14456, USA; [9]Plant Science Department, South Dakota State University, Brookings, SD 57007, USA and [10]Plant Pathology and Plant-Microbe Biology Section, School of Integrative Plant Science, Cornell University, Geneva, NY 14456, USA.
Correspondence: L Cadle-Davidson (lance.cadledavidson@ars.usda.gov) or Q Sun (qisun@cornell.edu)

Received: 7 December 2015; Revised: 6 January 2016; Accepted: 8 January 2016

To fill the gaps between the above technical and economic considerations for breeding and actual implementation of MAS,[35] next-generation sequencing (NGS) technology offers a potential opportunity for unbiased genotyping with high-throughput and low per-sample cost. Genotyping-by-sequencing (GBS), with its simultaneous marker discovery and genotyping approach, delivers many benefits including availability of flanking DNA sequence information, high-sample throughput and scalability (multiplexing), and high resolution.[36] It has been successfully applied in marker discovery for many self-pollinating crops as well as outcrossing species.[37] However, successful implementation of GBS for MAS has not been reported in any heterozygous crop breeding programs. Many reasons may have contributed to the lack of adoptions. From a technical perspective, missing data, genotyping errors and heterozygote under-calling are common in GBS results due to uneven sequencing depth across sites and high level of sample multiplexing.[38–41] Rapid LD decay, large-scale genome structure variation coupled with lack of haplotype information makes it impractical to do genotype imputation in heterozygous perennial species with diverse backgrounds. From a practical perspective, the long turn-around time from sample collection to data analysis makes it difficult to fit into most breeding timeframes. Moreover, computational challenges in data processing further hinder breeders' interest in implementing GBS. In summary, GBS is currently impractical for MAS in heterozygous crops.

This study presents a novel and efficient strategy for molecular marker development and practical implementation in MAS, based on amplicon sequencing (AmpSeq). The semi-automated pipeline incorporates a machine learning model for primer design and uses Illumina's Nextera dual-barcoding and sequencing platforms for genotyping (Illumina, San Diego, CA, USA). After detecting a SNP from GBS, the strategy starts from the design of primers using the GBS sequence tags. The converted amplicon markers can then be used for genotyping through NGS. The design involves multiplexing of both samples and markers. As a case study, we document the use of AmpSeq in grapevine breeding programs for three traits including flower sex, powdery mildew (PM) resistance and acylated anthocyanins. We chose these three traits to represent a Mendelian trait and two QTL with differing effects on phenotypic variance: (a) a single gene for flower sex with three alleles[42]; (b) a QTL with moderately high R^2 (acylation of anthocyanins, initially reported here); and (c) a QTL with relatively low R^2 (*Ren2* locus for PM (*Erisyphe necator*) resistance from *Vitis cinerea* (Engelm. ex A. Gray) Engelm. ex Millard accession B9 (*V. cinerea* B9)[43,44]). All three loci are located on different chromosomes, and we were able to test the AmpSeq approach for flower sex (male versus female versus hermaphrodite) across four different families where the male flower allele descends from *V. cinerea*, the hermaphrodite flower allele descends from *V. vinifera* L. and the female allele descends from an unidentified North American *Vitis* species. Two of the three traits chosen for analysis would take 2–4 years to analyze phenotypically due to the time it takes for a seedling to produce flowers and fruit. We also report here the development of a pipeline package with tools for AmpSeq marker design and decision support.

MATERIALS AND METHODS

Plant materials

Four families were chosen for this study, all representing interspecific hybridization of diploid ($2n = 38$) *Vitis* species: *V. vinifera* 'Chardonnay' × *V. cinerea* B9; 'Horizon' (complex hybrid of *V. vinifera*, *V. labrusca* L., *V. rupestris* Scheele and *V. aestivalis* Michx.) × *V. cinerea* B9; 'Horizon' × Illinois 547-1 (*V. rupestris* B38 × *V. cinerea* B9) and MN1246 × MN1264 (both, complex hybrids with a genomic background including at least *V. vinifera, V. riparia* Michx., *V. rupestris, V. labrusca, V. cinerea* and *V. aestivalis*). The first three populations were grown in research vineyards operated by Cornell University, Geneva, New York. The latter population was grown in research vineyards of the University of Minnesota, St Paul, located at Excelsior, Minnesota. All four populations segregated for flower sex (male, female and hermaphrodite) a trait controlled by a single locus with three alleles (reviewed in the study by Hyma *et al.*[42]) In addition, our ongoing research indicated that two populations segregated for acylation of anthocyanins, and the three populations descending from *V. cinerea* B9 segregated for *Ren2* PM resistance.[43,44]

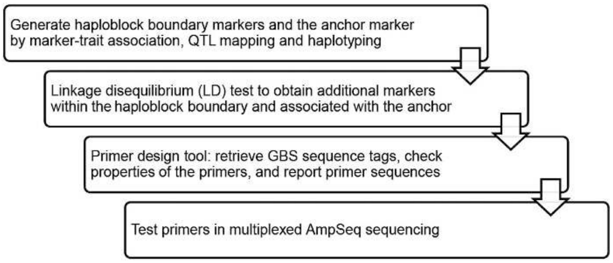
Generate haploblock boundary markers and the anchor marker by marker-trait association, QTL mapping and haplotyping
Linkage disequilibrium (LD) test to obtain additional markers within the haploblock boundary and associated with the anchor
Primer design tool: retrieve GBS sequence tags, check properties of the primers, and report primer sequences
Test primers in multiplexed AmpSeq sequencing

Figure 1. Workflow of the AmpSeq strategy.

AmpSeq marker development pipeline

The AmpSeq marker development procedure consists of four steps illustrated in Figure 1. First, GBS marker-trait associations were evaluated in TASSEL 4.3.13 (ref. 45). Genetic maps were constructed by the HetMappS strategy, and QTL were mapped in R/qtl ver. 1.37–11 as described by Hyma *et al.*[42] Marker phase and effect were estimated by a custom R script (Supplementary File 1) using the intermediate files 'phased' and 'LGmap' from the HetMappS pipeline.[42] For the second step, the range of the haploblock across the QTL region was defined by haploblock start and end markers with similar absolute value of marker effects. The anchor marker selected for each QTL was based on high association with trait within the 1.8-LOD (logarithm base 10 of odds) support interval of the QTL. A Perl script included in the package 'calculated_LD_distribution.pl' took in a text file 'test_marker' as input, which includes physical positions of the three markers: haploblock start, haploblock end and the anchor SNP. This script retrieved additional SNP markers from the GBS pipeline (pre-filtered by maximum of 5% missing data) that were within the haploblock range and verified to be in LD with the anchor marker.

After the SNP markers were defined, a Perl script 'parse_bam.pl' was used to identify genomic regions suitable for designing PCR primers. The template for the primer sequences was based on the actual sequences from GBS data, rather than the reference genome sequence itself. This was to ensure that there was no mismatch between the primer sequences and their targeted alleles. The 'parse_bam.pl' script required three inputs: (1) the reference genome FASTA file; (2) a BAM file, converted from the SAM file generated by TASSEL GBS pipeline,[45] which contained GBS reads aligned to the reference genome; and (3) a cutoff for the LD *P*-value as calculated from the previous step. By default, the script required that at least 10 bp of the left and right flanking sequences are covered by the GBS read, and ⩾90% alignment matching rate between the two alleles of up to 20 bp of each flanking sequences. Finally the script 'primer3.pl' was used to call the primer designing software Primer3 (ref. 46) and captured output properties from Primer3. The default primer sequence length from the pipeline was 22 bp. With a single SNP in the middle, the targeted amplicon size is 45 bp. From the output of the primer design pipeline followed by manual filtering, including removal of overlapping amplicons and primers with extreme annealing temperatures (T_m) outside of 47–79 °C, a total of 54 amplicons across the three loci were retained for testing: 19 for flower sex, 12 for PM resistance and 23 for acylated anthocyanins. The output file of the 54 amplicons is shown in Supplementary File 2. The documentation about the usage of the pipeline and the scripts are provided in Supplementary File 3.

Genotyping (PCR, library preparation and de-multiplex)

Two experiments were conducted to test three traits: flower sex, PM resistance due to the *Ren2* locus and acylated-anthocyanin concentration. In Experiment 1, 19 primers for flower sex and 12 primers for *Ren2* were pooled in one Illumina MiSeq lane (Illumina, San Diego, CA, USA) testing four 96-well plates (380 individuals and 4 blanks), each containing the parents and a subsample of progeny from each grapevine breeding family. Experiment 1 consisted of four breeding families: 'Horizon' × *V. cinerea* B9, 'Horizon' × Illinois 547-1, and 'Chardonnay' × *V. cinerea* B9 and MN1246 × MN1264. The other 23 primers for acylated-anthocyanin concentration were pooled in Experiment 2, testing two 96-well plates of 'Horizon' × Illinois 547-1 and two 96-well plates of *V. rupestris* B38 × 'Horizon' (380 individuals and 4 blanks).

For each vine, a single small leaf (< 1-cm diameter) was harvested and placed in one tube of a Costar 96-well cluster tube collection plate (Corning, Corning NY, USA), and DNA was isolated as described previously.[42] Briefly, each 96-well plate received up to 91 unique samples plus two sets of duplicated individuals and a blank well to serve as quality controls. Frozen samples were ground using stainless-steel beads in a Geno/Grinder 2000 (OPS Diagnostics LLC, Lebanon NJ, USA) and processed using DNeasy 96-well DNA extraction kits (Qiagen, Valencia CA, USA). The Qiagen AP1 lysis buffer was amended with PVP-40 (2% w/v) prior to heating of the buffer, to improve DNA quality and quantity. Due to historically consistent yields of 25–50 ng μl^{-1} within and between plates, DNA was used following twofold dilution without quantification.

AmpSeq uses two rounds of PCR: the 1st PCR to amplify a multiplex of markers and incorporate linker sequences, and the 2nd PCR to use the linker sequences to add a unique pair of indices, or barcodes, to each sample. The linker 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3′ was added to the 5′ end of each forward primer to accommodate S5xx barcode adapters, and the linker 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG -3′ was added to the 5′ end of each reverse primer to accommodate N7xx barcode adapters described in Supplementary File 4. For results presented here, samples were processed without previously testing the primers as a multiplex, and both rounds of PCR were conducted on 384 wells per batch, processed as four 96-well plates, which were pooled after the 2nd PCR for Illumina sequencing.

For the 1st PCR, up to 31 primer pairs (62 total, 10 pmol each) were combined with Qiagen Multiplex PCR Plus Mix (Qiagen, Valencia CA, USA), following the manufacturer's protocols calculated for 10 μl reaction volumes including 2 μl of DNA template. Thermocycling conditions for the 1st PCR were: 95 °C for 5 min; followed by 35 cycles of 95 °C for 30 s, 62 °C for 90 s and 72 °C for 30 s; followed by a final extension of 10 min at 68 °C. After the 1st PCR, 50 μl H_2O was added to each well prior to serving as a template for the 2nd PCR.

In the 2nd PCR, 384 unique combinations of 16 row and 24 column indices were incorporated in 10 μl reactions consisting of Platinum Taq (2 U per reaction, Invitrogen, Grand Island NY, USA), Platinum Taq buffer (1 × final), $MgCl_2$ (1.5 mM final), dNTPs (0.2 mM final each), primers (10 pmol each) and 2 μl of each amplicon pool from the 1st PCR reaction. Thermocycling conditions for the 2nd PCR were: 95 °C for 5 min; followed by eight cycles of 95 °C for 30 s, 53 °C for 30 s and 72 °C for 30 s; followed by a final extension of 10 min at 68 °C. Equal volumes of all 384 wells were pooled into a single tube, then purified by Ampure (Beckman Coulter, Indianapolis IN, USA), following manufacturer's instructions.

The amplicon pool was sequenced on an Illumina MiSeq instrument with single-end reads of 51 bp with 8 bp, dual index reads according to the manufacturer's instructions. Reads were de-multiplexed using the Illumina bcl2fastq pipeline software, allowing a single mismatch in the index reads.

Phenotyping and data analysis

The experimental design for phenotyping is described in Supplementary File 5 (refs 42, 47–49).

SNPs were called by a custom perl script 'run_gatk2.pl' (Supplementary File 6) from the de-multiplexed fastq file. Read counts were calculated in 500 bp sliding windows across the 12X.2 version of the PN40024 reference genome[50] using BEDTools[51] v2.22.1 to check the amplification specificity and efficiency of the AmqSeq markers. VCFtools[52] v0.1.12a was applied to calculate missing rate per site (--missing-site) and per individual (--missing-indv), and mean depth per site (--site-mean-depth) and per individual (--depth). VCF file generated from 'run_gatk2.pl' was input to TASSEL 4.3.13 for association detection. Genotyping and phenotyping data of the two parents were used for phasing.

To detect family-based marker specificity, the squared correlation coefficients between AmpSeq markers were calculated per family for each trait by VCFtools--geno-r2 function, following a dendrogram construction in R[53] v3.1.2. *P*-values of association between genotype and phenotype were calculated by a one-way analysis of variance co-segregation analysis General Linear Model (GLM) on a per-family basis and after pooling families with the same segregation pattern. The minor allele frequencies of the AmpSeq markers were also reported on a per-family and pooled basis.

Output from the primer design pipeline was used to classify primers as efficient and non-efficient based on their amplification and genotyping performance, and 4 statistical models (logistic regression, support vector machine, decision tree and random forest) were tested using the following cross-validation method for assessing 18 numerical parameters of each primer (num_tags, pvalue, phase, reject_code, a1, a1_count, a1_lseq_pct, a1_rseq_pct, a2, a2_count, a2_lseq_pct, a2_rseq_pct, ltemp, rtemp, a1_lseq_len, a1_rseq_len, a2_lseq_len, a2_rseq_len). In the cross-validation algorithm, the efficient and non-efficient categories are randomly partitioned into two groups, the training group and the testing group, and the testing group is used for testing the performance of four statistical models trained with the training group. The prediction accuracy of the cross-validation classification model was assessed using Receiver Operating Characteristic (ROC) curve analysis.[54–57] The ROC curve is a commonly used diagnostic to assess the prediction power of different classification methods, by means of a graphical plot of the true positive rate (sensitivity) against the false positive rate (1—specificity) for each classification method that predicts a dichotomous outcome. The plot is showed at several thresholds, which are used to designate whether the prediction of a given method as positive. The area under this curve (AUC) is one of the most important performance metrics that can be applied for selecting the most adequate classification method because it represents the accuracy of the prediction. In practice, AUC values range from 0.50 to 1.00, and a higher AUC value indicates better prediction accuracy for the classification models: values below 0.60 should be considered poor, 0.60–0.74 are moderate, 0.75–0.89 are good and ⩾0.90 are very good to excellent.[57] A value of exactly 0.50 would indicate that the model is useless for testing, while a value of exactly 1.00 would indicate the model is a perfect test. The models were evaluated using packages caret, e1071, rpart and random forest implemented in R v3.1.2.

RESULTS

Converting GBS-derived SNPs to amplicon sequencing (AmpSeq) markers

The procedure for the marker development and amplicon sequencing primer design pipeline is illustrated in Figure 1. Three traits were chosen for study: flower sex, *Ren2* PM resistance[43,44] and acylated anthocyanins (that is, anthocyanins esterified to hydroxycinnamic acids). The percentage of phenotypic variance explained by each QTL (R^2 value) confirmed these represent a qualitative trait (93%), a quantitative trait with a moderate QTL (13%), and a quantitative trait with a major QTL (54%), respectively (Table 1). For each locus, two flanking markers indicating the haploblock boundary, and one anchor marker having high association with the phenotype from the curated genetic maps are listed in Table 1. The primer design pipeline and curation resulted in 54 AmpSeq markers physically located between the two flanking markers and genetically linked to the anchor markers, for further testing (Supplementary File 2).

In Experiment 1, 12,677,206 reads were generated for the target QTL region on chromosome 2 (chr2) associated with flower sex, and 4,537,611 reads were generated for the target QTL region on chr14 associated with *Ren2* PM resistance. These 17,214,817 reads at the two target loci comprise 99.99% of the total reads, while only 2,237 reads aligned to off-target regions. Similar results were obtained in Experiment 2, with 14,927,137 reads (98.82%) aligning to the target QTL region on chr 3 associated with acylated anthocyanins.

The sequencing depth per individual and missing rate were evaluated for each amplicon (Tables 2–4). For both experiments, a majority of amplicons (80% in Experiment 1 and 83% in Experiment 2) had > 50-fold coverage per individual, in a relatively narrow range from 50–284 ×. Most amplicons (87% in Experiment 1 and 83% in Experiment 2) had < 10% missing data. Thus, in Experiment 1, the 380 individuals had 220 × median coverage per marker and 168 × mean coverage per marker. Similarly in Experiment 2 for 23 acylated-anthocyanin AmpSeq

Table 1. QTL parameters for the development of AmpSeq markers

Trait	*Population*	*Haploblock start*	*Haploblock end*	*Anchor marker*	*QTL* R^2[a]	*1.8 LOD interval (left border)*	*1.8 LOD interval (right border)*	*Highest LOD*
Flower sex (M/H)	'Chardonnay' × *V. cinerea* B9 (CC)	S2_4168128	S2_5507608	S2_5186869	99.5%	S2_4168128	S2_5186889	42.5
Flower sex (M/H)	'Horizon' × *V. cinerea* B9 (HC)	S2_4178835	S2_5333625	S2_5186889	92.8%	S2_5186894	S2_5528872	39.8
Flower sex (M/H)	'Horizon' × Illinois 547-1 (HI)	S2_4377285	S2_5333462	S2_5068641	97.0%	S2_4704546	S2_5068764	96.1
Flower sex (H/f)	MN1246 × MN1264 (MN)	S2_4142601	S2_6967072	S2_5181254	82.1%	S2_4142601	S2_5715366	17.0
Powdery mildew in controlled environment	'Horizon' × *V. cinerea* B9 (HC)	S14_25628594	S14_28890859	S14_27742679	12.7%	S14_26788064	S14_29638581	4.81
Acylated Anthocyanins	'Horizon' × Illinois 547-1 (HI)	S3_14851567	S3_18229426	S3_17225376	54.0%	S3_8793603	S3_18757460	23.9

Abbreviations: QTL, quantitative trait loci. [a] R^2 (coefficient of determination): percentage of phenotypic variance explained by the QTL.

markers, 380 individuals had genotyping data with 244× median coverage per marker and 188× mean coverage per marker. One amplicon per experiment failed because no polymorphism was identified during the SNP calling procedure. The amplification specificity, stable sequencing depth and consistent missing rate in the two experiments indicated the marker development strategy and amplicon primer design pipeline were effective for all three loci, with no obvious bias during amplification, pooling or sequencing.

Evaluation and validation of AmpSeq markers in grapevine breeding families

Flower sex. To evaluate the 19 AmpSeq markers for flower sex, genotypic data were obtained for four breeding families in Experiment 1. This trait is reportedly controlled by a single locus with three alleles, with male (M) dominant over hermaphrodite (H), which is dominant over female (f).[42] Three families segregated 1:1 for M/H—'Horizon' × Illinois 547-1, 'Horizon' × *V. cinerea* B9 and 'Chardonnay' × *V. cinerea* B9. One family segregated 1:1 for H/f—MN1246 × MN1264. Parents of the four families represent three phenotypes: 'Horizon' (HH), 'Chardonnay' (HH) and MN1264 (Hf) produce hermaphrodite flowers; Illinois 547-1 (Mf) and *V. cinerea* B9 (Mf) produce male flowers; and MN1246 (ff) produces female flowers.[42] Amplicon data were merged with flower sex field ratings for a one-way analysis of variance co-segregation analysis. For the pooled association test using families with the same segregation pattern, only 1 amplicon had a non-significant association, and 16 out of 18 amplicons had a $-\log_{10}$(*P*-value) > 13, suggesting a high association between AmpSeq markers and the recorded phenotypes (Figures 2a and b and Table 2). The high R^2 value of each marker was consistent with the high R^2 value of the flower sex QTL (93%, Table 1). Interpretation and direct application of AmpSeq genotypes for MAS required phase information, details of which are documented in Supplementary File 7.

To shed light on marker transferability among breeding populations, genotype-phenotype association and minor allele frequency were reported by family in Table 3 and compared using a dendrogram (Figures 3a–d). AmpSeq markers with similar *P*-values and R^2 patterns clustered together. Some AmpSeq markers in certain families have an R^2 value equal to 1, suggesting the potential of 100% predictive accuracy in explaining 100% of the phenotypic variance.

PM resistance. To evaluate the 12 AmpSeq markers for *Ren2* PM resistance, genotypic data from Experiment 1 were merged with phenotypes from two families, each having a different phenotyping approach: Set 1 used 78 progeny of 'Horizon' × *V. cinerea* B9 scored using the transformed mean of total hyphal transects *in vitro*; Set 2 used visual ratings (on a 1–5 scale) of natural infection in the vineyard for 91 progeny of 'Horizon' × Illinois 547-1.

For Set 1, eight markers significantly predicted resistance [$-\log_{10}$(*P*-value) > 2], and the marginal *P*-values and small R^2 values (Figure 2c, Table 3) reflected the moderate QTL ($R^2 = 13\%$) discovered by GBS. For Set 2, 11 of 12 markers were significant ($-\log_{10}$(*P*-value) > 2) (Figure 2d) with an average of 5.3, and all 12 markers explained more variance for vineyard ratings than for *in vitro* hyphal transects. The results confirmed that AmpSeq markers for *Ren2* PM resistance worked in a separate but related breeding family, even when evaluated by a different phenotyping method. A majority of the amplicons (9/12) can be used to track resistant alleles in phase, meaning that the minor allele is associated with PM resistance. AmpSeq markers with similar *P*-values and R^2 patterns clustered together (Figures 3e and f).

Acylated anthocyanins. To evaluate the 22 AmpSeq markers for acylated-anthocyanin content, association tests were executed on 182 individuals including: 82 progeny from 'Horizon' × Illinois 547-1, a subset of the original family for marker development, and 100 progeny from *V. rupestris* B38 × 'Horizon' to evaluate marker transferability. In a pooled analysis, 1 amplicon had a non-significant association, and 15 of 22 amplicons had a $-\log_{10}$(*P*-value) > 10 (Figure 2e, Table 4). Individually, these 15 AmpSeq markers explained an average of 33% of the phenotypic variance, comparable to the QTL contributing 54% of the phenotypic variance. As with the flower sex dendrogram, AmpSeq markers with similar *P*-values and R^2 patterns clustered together (Figures 3g and h).

Criteria of efficient AmpSeq markers

To reduce costs for primer synthesis and testing, we tested models to determine *post hoc* the relative importance of AmpSeq marker parameters to predict the most effective markers. Four statistical models (logistic regression, support vector machine, decision tree and random forest) were tested using quantified parameters of each primer to classify the primer output from the primer design pipeline into efficient and non-efficient categories. In all cases, regardless of the training set size, the random forest model (with over 90% predictive accuracy) outperformed the other three models (Figure 4). Thus, we developed a decision support tool based on the random forest model (Supplementary File 8), which can be used to facilitate AmpSeq marker selection following the primer design pipeline. The most important parameters in this model included the *P*-value of LD with the anchor marker, the primer annealing temperatures, the rejection code and the length

Table 2. AmpSeq marker-trait associations for flower sex

Chr	Position	Family of AmpSeq design	Mean depth	CV	Missing individuals	Pooled (258)[a]			'Horizon' × Illinois 547-1 (87)			'Horizon' × *V. cinerea* B9 (82)			'Chardonnay' × *V. cinerea* B9 (89)			MN1246 × MN1264 (47)		
						Flower sex (M/H)			Flower sex (M/H)			Flower sex (M/H)			Flower sex (M/H)			Flower sex (H/f)		
						P-value	marker R^2	Minor allele frequency	P-value	marker R^2	Minor allele frequency	P-value	marker R^2	Minor allele frequency	P-value	marker R^2	Minor allele frequency	P-value	marker R^2	Minor allele frequency
chr2	4245542	CC	248	1	0	1.64E − 98	0.8238	0.27	1.92E − 33	0.8204	0.31	2.32E − 27	0.7714	0.24	2.99E − 40	0.8696	0.24	NaN	0	0.08
chr2	4422069	CC	13	6	4	1.65E − 13	0.1938	0.08	1.18E − 03	0.1197	0.04	3.10E − 05	0.1961	0.08	2.34E − 06	0.2295	0.11	NaN	0	0
chr2	4428879	HI	109	42	1	7.18E − 16	0.2256	0.10	4.57E − 57	0.9517	0.30	NaN	0	0	NaN	0	0	NaN	0	0.08
chr2	4448335	HI	6	3	14	3.35E − 06	0.0967	0.06	1.33E − 06	0.2782	0.17	NaN	0	0	3.59E − 01	0.0099	0.01	3.57E − 01	0.0207	0.15
chr2	4457222	HI	248	2	0	1.66E − 16	0.2336	0.10	2.96E − 58	0.9531	0.29	NaN	0	0	NaN	0	0	4.54E − 07	0.4852	0.48
chr2	4560461	HI	9	11	16	5.07E − 146	0.9363	0.29	4.53E − 49	0.9486	0.25	1.19E − 32	0.8555	0.31	0.00E+00	1	0.29	1.29E − 01	0.0505	0.01
chr2	4726521	HC	248	2	0	6.34E − 17	0.2392	0.10	0.00E+00	1	0.29	NaN	0	0	NaN	0	0	5.23E − 16	0.7711	0.25
chr2	4726692	HI	248	2	0	3.43E − 173	0.9540	0.23	0.00E+00	1	0.20	7.68E − 36	0.8595	0.24	0.00E+00	1	0.26	NaN	0	0
chr2	4798009	HI	96	49	1	8.30E − 17	0.2385	0.10	0.00E+00	1	0.29	NaN	0	0	NaN	0	0	3.55E − 01	0.0191	0.21
chr2	5005636	HI	248	1	0	5.52E − 16	0.2264	0.10	4.58E − 58	0.9526	0.30	NaN	0	0	NaN	0	0	5.23E − 16	0.7711	0.25
chr2	5064267	MN	248	1	0	5.23E − 16	0.7711	0.25	NaN[b]	0	0.49	NaN	0	0.48	3.23E − 01	0.0112	0.24	5.23E − 16	0.7711	0.25
chr2	5068641	HI	249	1	0	6.34E − 17	0.2392	0.39	0.00E+00	1	0.20	NaN	0	0.48	NaN	0	0.50	NaN	0	0
chr2	5181254	MN	132	51	8	6.74E − 14	0.2170	0.19	9.09E − 41	0.9002	0.47	5.44E − 01	0.0159	0.06	NaN	0	0	NaN	0	0.04
chr2	5207140	HI	7	5	25	3.47E − 106	0.8712	0.26	1.76E − 39	0.8985	0.21	2.59E − 28	0.8414	0.29	1.54E − 37	0.8671	0.27	NaN	0	0
chr2	5219708	HI	248	1	0	1.66E − 16	0.2336	0.39	2.96E − 58	0.9531	0.20	NaN	0	0.48	NaN	0	0.50	4.54E − 07	0.4852	0.47
chr2	5219969	CC	248	1	0	3.38E − 145	0.9239	0.23	2.96E − 58	0.9531	0.20	7.68E − 36	0.8595	0.24	9.85E − 61	0.9559	0.26	NaN	0	0
chr2	5220041	HI	248	1	0	1.66E − 16	0.2336	0.10	2.96E − 58	0.9531	0.29	NaN	0	0	NaN	0	0	4.64E − 04	0.2945	0.47
chr2	5233638	HI	181	68	107	1.20E − 01	0.0257	0.06	5.79E − 03	0.1554	0.02	6.22E − 01	0.0373	0.18	2.34E − 01	0.0196	0.03	NaN	0	0

Abbreviation: CC, 'Chardonnay' × V. cinerea B9; CV, coefficient of variation of read depth across individuals; HC, 'Horizon' × V. cinerea B9; HI, 'Horizon' × Illinois 547-1; MN, MN1246 × MN1264; QTL, quantitative trait loci. [a]Pooled includes all three families segregating M/H. Number of progeny analyzed is presented in parentheses. [b]NaN means no association detected.

of the sequence flanking each side of the SNP (Supplementary File 8).

DISCUSSION

This study had two main goals: (1) to develop a cost-effective and robust genotyping platform for MAS in heterozygous crops, and (2) to generate marker sets for MAS in grapevine breeding. A semi-automated primer design pipeline was developed to convert GBS tags to AmpSeq markers. The primer design procedure was facilitated by a decision support tool to predict which primers would perform appropriately in terms of predictability based on a pre-trained random forest model. For grapevine breeding, AmpSeq markers were developed for three economically important and representative traits with high breeding value: flower sex, PM resistance and acylated-anthocyanin concentration. The strategy was effective for all three traits given the diverse background and high heterozygosity of grapevine. In the process, 54 markers were tested on a total of 760 individuals in 6 breeding families involving 7 *Vitis* species across 2 breeding programs. The results indicated that the majority of AmpSeq markers have potential for accurate trait predictions, and a MAS package can be implemented for interspecific hybrid families.

Technical considerations of the AmpSeq strategy

Two key decision points for the usage of the AmpSeq primer design pipeline were the selection of an anchor marker and the definition of haploblock boundaries of each QTL. The optimal parameters depended on the nature of the trait and mapping family, and here we describe what worked for these three traits. First, each anchor marker was selected from our final GBS linkage map, after the HetMappS pipeline and genetic map curation[42] removed SNPs that had segregation bias, low genotyping rate, putative sequencing error, redundancy or aligned to repetitive genome regions. Thus, the anchor marker was selected from high-quality, mapped SNPs. Second, the anchor marker had a high *P*-value for marker-trait association. Third, the anchor marker was within the 1.8 LOD interval of the QTL to ensure tight genetic linkage. The anchor marker was used to retrieve SNPs in LD from an un-filtered VCF file within the haploblock defined by the two boundary markers, to increase marker density under the QTL region and provide more alternatives for AmpSeq marker design. The range of the haploblock can be defined by the marker effect as suggested in 'Materials and methods', or by other haplotyping software specific for heterozygous species, like iXora[58] and SHAPEIT2 (ref. 59).

Two custom options can be set for the amplicon primer design pipeline: one is *P*-value threshold of the LD test, and the other one is the primer size, which was specified as 22 in this study to accommodate the 50 bp sequencing length and data analysis pipeline. To optimize the *P*-value of the LD test, all the parameters output from the pipeline were explored by four models (logistic regression, support vector machine, decision tree and random forest), in which the random forest model was the most useful to predict high-quality markers. In the random forest model, four primer parameters primarily contributed to the decision: the *P*-value of LD with the anchor marker, the rejection code and the annealing temperatures (T_m) of the forward and reverse primers. Supported by the split value of the decision tree model (Supplementary File 9) and general guidelines for PCR based on experimental experience and thermocycling conditions used here, the recommended criteria for efficient amplicon markers includes: (1) the default *P*-value of LD with the anchor marker should be 1e −25; (2) T_m of primers should be between 52 and 70 °C to obtain high depth and avoid missing data; (3) the rejection code should be 0, representing exactly two alleles identified in GBS tags; (4) extremely high read depths of GBS tags, which may imply

Table 3. AmpSeq marker-trait associations for *Ren2* powdery mildew resistance

Chr	*Position*	*Family of AmpSeq design*	*Mean depth*	*CV*	*Missing individuals*	*'Horizon' × V. cinerea B9 (78)*[a]			*'Horizon' × Illinois 547-1 (91)*		
						Powdery mildew in controlled environment			*Powdery mildew field rating*		
						P-*value*	*marker R^2*	*Minor allele frequency*	P-*value*	*marker R^2*	*Minor allele frequency*
chr14	26205721	HC	65	38	2	1.56E − 02	0.1050	0.20	1.56E − 05	0.1940	0.27
chr14	26954543	HC	207	40	0	9.23E − 03	0.1174	0.27	2.24E − 03	0.1309	0.47
chr14	27052970	HC	1	0	380	NaN[b]	0	0.00	NaN	0	0.00
chr14	27118778	HC	52	21	1	3.72E − 03	0.1054	0.18	1.13E − 07	0.2776	0.26
chr14	27263681	HC	232	5	2	2.42E − 03	0.1148	0.19	1.45E − 08	0.3101	0.26
chr14	27742679	HC	244	2	1	1.34E − 03	0.1274	0.18	2.14E − 08	0.3041	0.27
chr14	28408093	HC	248	2	0	6.12E − 04	0.1440	0.19	3.82E − 06	0.2165	0.27
chr14	28415565	HC	248	2	0	6.12E − 04	0.1440	0.19	9.53E − 06	0.2007	0.26
chr14	28448817	HC	51	48	42	3.78E − 03	0.1399	0.34	1.05E − 02	0.1205	0.37
chr14	28686983	HC	248	2	0	8.90E − 03	0.0866	0.20	8.29E − 06	0.2031	0.27
chr14	28791319	HC	50	38	2	1.32E − 02	0.1089	0.25	6.44E − 05	0.2010	0.34
chr14	28824733	HC	101	49	2	1.32E − 02	0.0782	0.20	4.53E − 05	0.1749	0.28

Abbreviations: CV, coefficient of variation of read depth across individuals; HC, 'Horizon' × V. cinerea B9; QTL, quantitative trait loci. [a]Number of progeny analyzed is presented in parentheses. [b]NaN means no association detected.

presence of a repetitive genome region, should be filtered; and (5) overlapping amplicons, which would result in cross-amplification, should be excluded. The pooling strategy employed in the study (Experiment 1: 384 individuals with 31 amplicon markers for 2 traits; Experiment 2: 384 individuals with 23 amplicon markers for 1 trait) resulted in read depths greater than 50 ×, low missing rate and equal amplification efficiency, suggesting that multiplexing could be increased for more traits, markers or individuals.

In terms of predictive power of converted markers, the R^2 value of the QTL plays a major role. For a Mendelian trait such as flower sex, in which one major locus can explain 93% of phenotypic variance, the association between genotypes and phenotypes is highly significant. Some AmpSeq markers approached 100% predictive accuracy, as they can explain 100% of phenotypic variance in a given family. In contrast, for the major QTLs contributing $< 50\%$ phenotypic variation, the predictive power is lower, which is consistent with theoretical framework.[60] However, the predictive power is still attractive in breeding, when phenotypic screening is expensive and challenging. The notion was supported by several studies showing selection performed combining MAS with phenotypic evaluation is more efficient than selection based on phenotyping alone, especially when the family is large and trait heritability is low.[8,61–65] Minor allele frequencies were calculated of each AmpSeq marker for each segregating family (Table 2). Segregation distortion, defined as observed minor allele frequency deviating from expected minor allele frequency of 0.25, was observed for some AmpSeq markers in all analyses. Similar distortion has been reported for SSR markers in grape,[43] for GBS markers in rice[15] and for sequence-tagged microsatellite site in chickpea.[66] Although simulations claim that the presence of marker segregation distortion has little effect on linkage map construction and QTL analysis,[67,68] distortion has resulted in less effective markers and spurious conclusions in practice.[43,66] Here, AmpSeq markers with high *P*-values typically had extreme segregation distortion, indicating the segregation pattern of AmpSeq markers should be taken into consideration when choosing efficient markers.

Advantage of AmpSeq genotyping platform

The most significant advantage of AmpSeq for MAS is the ability to harness the high-resolution GBS or other NGS techniques during marker development, which provides resiliency against rapid LD decay in species with high heterozygosity and diversity,[69] while nearly eliminating issues of missing data and heterozygote under-calling common to GBS. An average density of 55, 218 and 133 kb per AmpSeq marker was obtained for flower sex, PM resistance and acylated anthocyanins, respectively. In grapevine, SSR and indel markers reported to predict flower sex[27] have failed in progenies resulting from the cross between complex North American hybrids (BI Reisch and JJ Luby, personal communication), which may be due to the loss of linkage between the causal gene and markers. The situation has been similar for MAS of PM resistance. The closest pairs of *Ren2* SSR markers were 550-kb apart in the 12X.2 reference genome.[70–72] Four amplicon markers were generated by the AmpSeq strategy within the 550-kb region, targeting a lower probability of recombination between causal gene and markers, and improved ability to detect recombinations near the locus.

AmpSeq gains efficiency via the NGS capacity to multiplex more markers per trait, more traits and more individuals. Our approach results in the genotyping of a haplotype block in one test for MAS decision making, which showed higher accuracy in cattle breeding[73,74] and potential application in heterozygous crops such as cocoa[75] and cotton,[76] as well as for flowering time, cluster width and berry size in grapevine.[77] While biallelic SNPs were traditionally criticized by limited polymorphisms per marker,[78] the pooling of multiple SNP markers provides flexibility that overcomes this limitation.[79] Further, multiple traits can be genotyped simultaneously for pyramiding, maximizing the application of MAS in contrast to phenotype screening. For example, anthocyanin pigments acylated to hydroxycinnamic acids are of interest to wine and juice producers because their absorbance behavior is not highly pH dependent, as compared with non-esterified anthocyanins in grapes which exist primarily in colorless forms over a juice pH range of 3–4 (ref. 80). Our results indicate that a hermaphrodite vine with increased resistance to PM and higher concentrations of berry acylated anthocyanins could be selected in a single AmpSeq run. While our current barcoding system is for 380 individuals (accommodating 4 blank negative controls), additional barcodes could be used for increased sample throughput.[39]

Table 4. AmpSeq marker-trait associations for acylated anthocyanins

Chr	*Position*	*Family of AmpSeq design*	*Mean depth*	*CV*	*Missing individuals*	*Pooled (182)*[a]			*'Horizon'* × *Illinois 547-1 (82)*			*V. rupestris B38* × *'Horizon' (100)*		
						Acylated anthocyanins			*Acylated anthocyanins*			*Acylated anthocyanins*		
						P*-value*	*marker* R^2	*Minor allele frequency*	P*-value*	*marker* R^2	*Minor allele frequency*	P*-value*	*marker* R^2	*Minor allele frequency*
chr3	14851567	HI	250	0	0	1.07E – 16	0.3370	0.33	7.36E – 06	0.2586	0.42	4.90E – 13	0.4147	0.25
chr3	15171738	HI	247	2	0	2.05E – 17	0.3491	0.28	7.31E – 06	0.2587	0.30	5.33E – 13	0.4137	0.25
chr3	15828002	HI	250	0	127	1.76E – 10	0.3102	0.47	8.82E – 03	0.1724	0.44	8.35E – 09	0.4214	0.50
chr3	16007996	HI	245	3	0	1.26E – 16	0.3357	0.26	5.35E – 06	0.2646	0.26	4.10E – 12	0.3892	0.26
chr3	16008017	HI	244	3	0	1.26E – 16	0.3357	0.26	5.35E – 06	0.2646	0.26	4.10E – 12	0.3892	0.26
chr3	16162132	HI	250	0	67	2.70E – 03	0.0763	0.31	2.73E – 04	0.2502	0.43	3.78E – 02	0.0710	0.21
chr3	16163517	HI	249	1	0	3.64E – 17	0.3265	0.24	1.15E – 06	0.2574	0.22	4.10E – 12	0.3892	0.26
chr3	16246446	HI	241	5	0	5.22E – 03	0.0425	0.05	6.13E – 02	0.0431	0.06	5.29E – 02	0.0377	0.05
chr3	16375440	HI	249	1	0	3.64E – 17	0.3265	0.24	1.15E – 06	0.2574	0.22	4.10E – 12	0.3892	0.26
chr3	16402674	HI	113	39	1	1.61E – 06	0.1385	0.42	6.88E – 06	0.2599	0.35	1.35E – 04	0.1678	0.50
chr3	16402706	HI	14	6	2	2.15E – 15	0.3143	0.33	7.62E – 06	0.2580	0.29	1.59E – 10	0.3720	0.37
chr3	16473116	HI	249	1	0	1.67E – 16	0.3336	0.43	2.06E – 05	0.2391	0.36	1.86E – 11	0.3992	0.49
chr3	16516083	HI	4	4	200	1.80E – 05	0.2417	0.35	8.50E – 04	0.2507	0.46	NaN	0	0.01
chr3	16520617	HI	249	1	0	8.04E – 17	0.3390	0.24	1.15E – 06	0.2574	0.23	4.78E – 12	0.4158	0.26
chr3	16523313	HI	73	53	3	1.85E – 16	0.3329	0.30	1.15E – 06	0.2574	0.22	6.12E – 11	0.3842	0.38
chr3	16774466	HI	249	1	0	1.06E – 17	0.3356	0.26	1.26E – 06	0.2557	0.29	5.02E – 13	0.4145	0.24
chr3	16873018	HI	248	2	0	NaN[b]	0	0.00	NaN	0	0.01	NaN	0	0.00
chr3	16873164	HI	88	46	0	1.06E – 17	0.3356	0.26	1.26E – 06	0.2557	0.28	5.02E – 13	0.4145	0.24
chr3	17225240	HI	20	27	2	9.37E – 14	0.2861	0.22	6.59E – 05	0.2163	0.18	2.53E – 10	0.3393	0.27
chr3	17225412	HI	149	0	31	8.45E – 03	0.0413	0.06	2.17E – 02	0.0720	0.06	1.06E – 01	0.0281	0.06
chr3	17225585	HI	210	14	0	2.28E – 17	0.3483	0.49	6.66E – 07	0.3024	0.45	1.96E – 11	0.3985	0.48
chr3	17774559	HI	62	151	124	1.97E – 02	0.0649	0.31	1.08E – 02	0.1319	0.22	4.35E – 01	0.0328	0.44

Abbreviations: CV, coefficient of variation of read depth across individuals; HI, 'Horizon' × Illinois 547-1; QTL, quantitative trait loci. [a]Pooled includes all two families. Number of progeny analyzed is presented in parentheses. [b]NaN means no association detected.

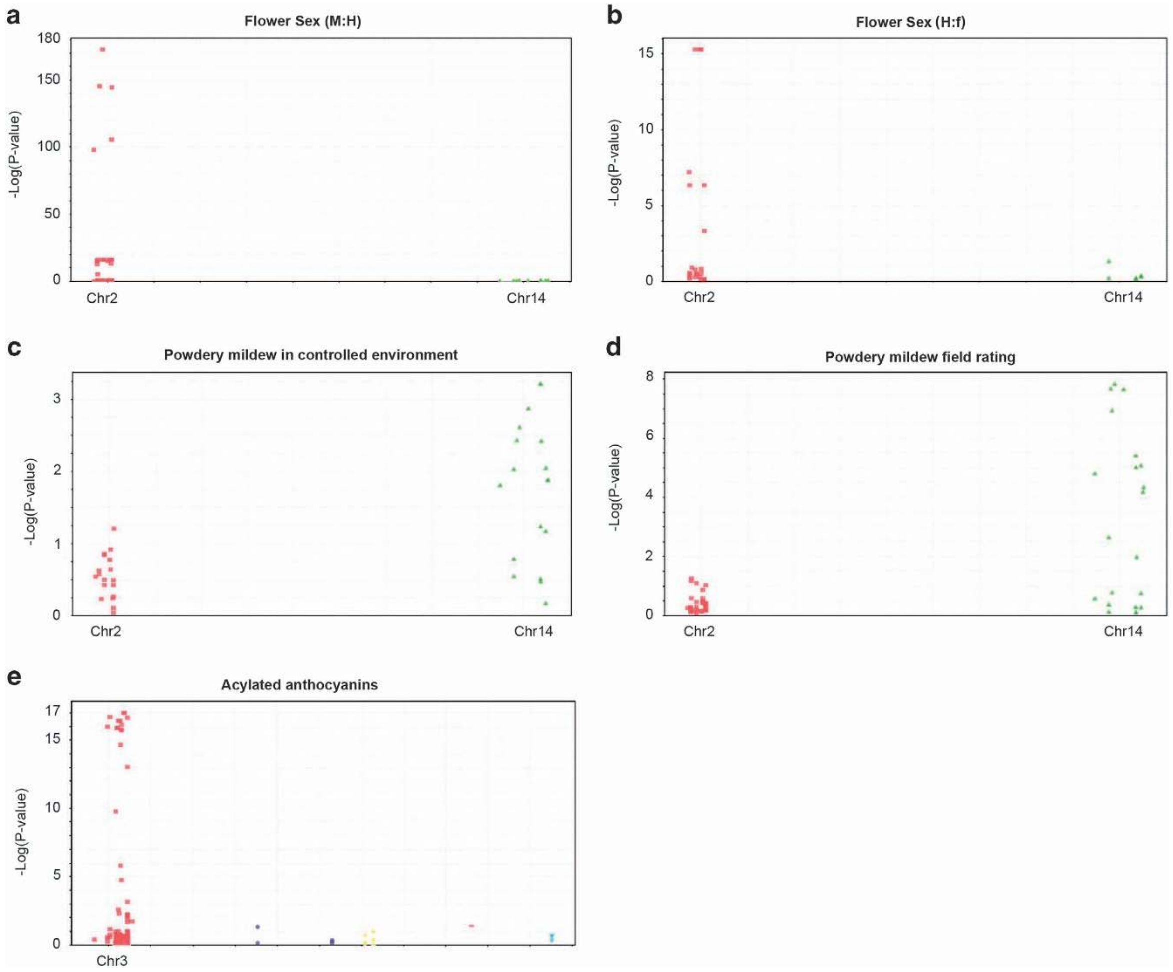

Figure 2. Manhattan plots of pooled association tests between AmpSeq markers and three traits. (**a**) Flower sex trait with male/hermaphrodite (M/H) segregation ('Horizon' × Illinois 547-1, 'Horizon' × *V. cinerea* B9 and 'Chardonnay' × *V. cinerea* B9); (**b**) Flower sex trait with hermaphrodite/female (H/f) segregation (MN1246 × MN1264); (**c**) Powdery mildew resistance assessed in a controlled environment (transformed mean of total hyphal transects *in vitro*) ('Horizon' × *V. cinerea* B9); (**d**) Powdery mildew resistance assessed visually in the field (1–5 scale based on natural infection) ('Horizon' × Illinois 547-1); (**e**) Acylated-anthocyanin content ('Horizon' × Illinois 547-1 and *V. rupestris* B38 × 'Horizon').

As mentioned above, marker transferability is another big concern in a breeding program with diverse germplasm spanning more than one species,[30] and SNP microarrays for grapevine and apple are both reported to suffer from this constraint.[1,81,82] Loss of polymorphism of SSR markers has also been reported when used in unrelated germplasm, as is the case for the marker *gwm261* for the locus *Reduced height 8* (*Rht8*) in wheat[83] or *Ren4* in grapevine.[84] This can result from sampling bias in the original study, since a distinct allele may not be amplified for the marker. The current work indicates that the amplicon markers are transferable in related families in grapevine. Among AmpSeq markers available in the current MAS package, dendrogram construction by genotype clustering could guide the choice of useful markers when transferring to new families without phenotypic data. In some, but not all cases, the markers most significantly associated with the trait clustered together (Figure 3), and in all cases markers not associated with the trait had poor correlations with the other markers. Still, because the amplicons are likely not genotyping the causal allele, re-training markers pools through validating, correcting and supplementing the current MAS package with markers developed in other breeding families by GBS or other genotyping platforms, following the same AmpSeq strategy, is recommended.

Finally, AmpSeq also provides advantages in terms of cost, time and ease of application. Actual cost to obtain AmpSeq data is comparable to single-locus SSRs and cheaper than most other multi-locus platforms. With the simple and straightforward PCR and library preparation protocol, the turn-around time from marker development through testing can be reduced to 1 month, which in our experience is quicker than other SNP platforms and SSRs. In addition, the data analysis pipeline can be automated to output results in a spreadsheet format, which may be attractive to some plant breeders, eliminating bioinformatic challenges typical of GBS data and removing the need for graphic interpretation typical of other genotyping platforms.

A case of MAS implementation in grape breeding

The three traits selected for testing are economically important and representative targets for North American hybrid winegrape cultivar development: flower sex, PM resistance and acylated anthocyanins. Hybridization with wild species takes place to introgress positive adaptive traits including disease resistance, pest resistance and cold hardiness,[20,85–87] but fruit quality and yield from the cultivated species *V. vinifera* are also critical. The families used for testing are five winegrape breeding families from

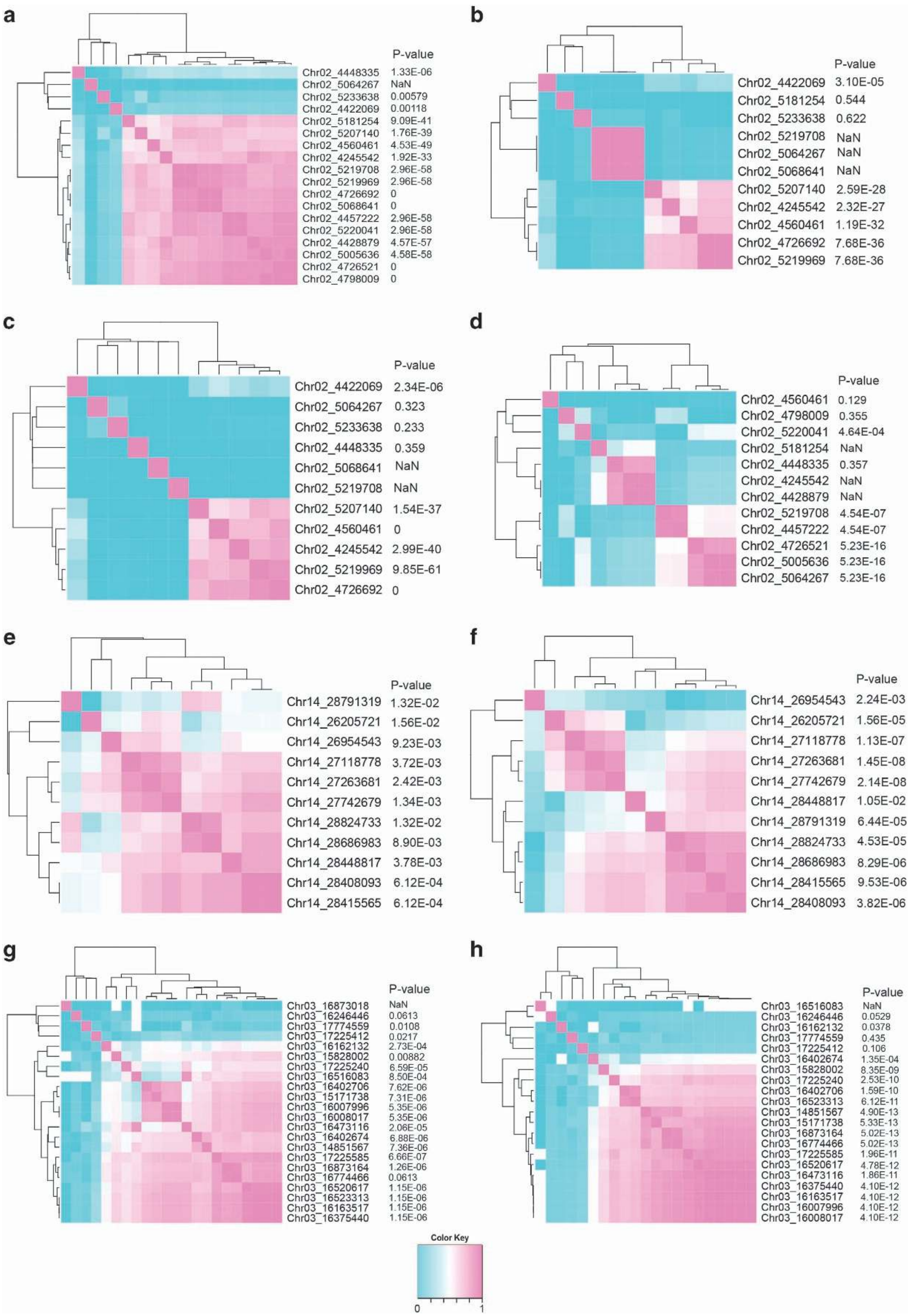

Figure 3. Marker transferability explored by dendrogram analysis with *P*-values of marker-trait association. (**a**) Flower sex for 'Horizon' × Illinois 547-1; (**b**) Flower sex for 'Horizon' × *V. cinerea* B9; (**c**) Flower sex for 'Chardonnay' × *V. cinerea* B9; (**d**) Flower sex for MN1246 × MN1264; (**e**) Powdery mildew resistance assessed in controlled environment for 'Horizon' × *V. cinerea* B9; (**f**) Powdery mildew resistance assessed visually in the field for 'Horizon' × Illinois 547-1; (**g**) Acylated-anthocyanin content for 'Horizon' × Illinois 547-1; (**h**) Acylated-anthocyanin content for *V. rupestris* B38 × 'Horizon'. Color key below indicates that shades of red color represent high correlation, while shades of blue color represent low correlation.

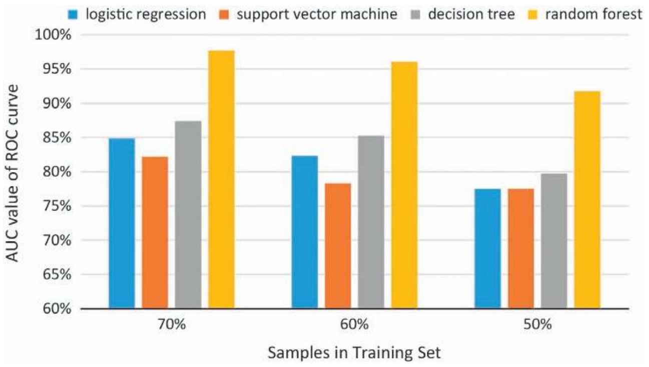

Figure 4. Evaluation of four machine-learning models for performance prediction of AmpSeq markers.

Cornell University and the University of Minnesota, with diverse backgrounds including *V. vinifera*, *V. aestivalis*, *V. cinerea*, *V. labrusca*, *V. riparia* and *V. rupestris*. The New York and Minnesota breeding programs are connected by the historical use of 'Seyval blanc', which is a complex interspecific hybrid of *V. vinifera* (55% by pedigree) with wild species, and is a relatively cold hardy white wine cultivar, resistant to disease and the phylloxera aphid *Daktulosphaira vitifoliae* (Fitch, 1855).[85,88,89] The current results demonstrate that the amplicon markers developed for all three traits are portable within the New York breeding program. Further, the flower sex amplicons are transferable between the New York and Minnesota programs. Since this study, we have implemented AmpSeq in breeding programs, and for each trait we have selected a subset of amplicons that appear to be robust and transferable. As an example, for 84 progeny of a *V. riparia* 37 × Seyval blanc F_2 family[90] from the South Dakota State University breeding program, seven AmpSeq markers exceeded a 10^{-8} significance threshold for predicting female flower sex (data not shown)—an improvement over the results presented here, indicating the initial success of the MAS package implementation in an additional population.

Customization of the AmpSeq strategy for other crops

In summary, the AmpSeq platform described here should provide several considerable advantages for breeders of other crop species due to its reliability, flexibility, high-throughput, cost-effectiveness, ease-of-automation and speed. This approach is among the first practical examples showing how SNP-based markers can be applied in a high-throughput screening, which may increase the application of GBS tags and leverage the massive power of NGS. This practicality may help fill the gap between genomic discoveries and breeding applications. With some customization, we propose that the AmpSeq strategy can be widely used by other crops.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We would like to thank Michelle Schaub, Hema Kasinathan, Anna Nowogrodzki, Paige Appleton and Jackie Lillis for their technical support phenotyping PM resistance; Steve Luce and Peter Hemstad for helping to maintain the mapping families used in this study; Mani Awale for phenotyping flower sex at the South Dakota State University breeding program; Steve Bogdanowicz for helpful discussions when designing the wet lab techniques and barcoding strategy; and the Cornell University Biotechnology Resource Center for providing the necessary infrastructure and personnel to generate the raw genetic data analyzed and presented here. The US Department of Agriculture National Institute of Food and Agriculture Specialty Crop Research Initiative provided funding for this project (award no. 2011-51181-30635).

REFERENCES

1 Myles S. Improving fruit and wine: what does genomics have to offer? *Trends Genet* 2013; **29**: 190–196.

2 Ru S, Main D, Evans K, Peace C. Current applications, challenges, and perspectives of marker-assisted seedling selection in Rosaceae tree fruit breeding. *Tree Genet Genomes* 2015; **11**: 1–12.

3 van Nocker S, Gardiner SE. Breeding better cultivars, faster: applications of new technologies for the rapid deployment of superior horticultural tree crops. *Hortic Res* 2014; **1**: 14022.

4 Bus VM, Esmenjaud D, Buck E, Laurens FApplication of genetic markers in rosaceous cropsIn:Folta K, Gardiner S (eds). *Genetics and Genomics of Rosaceae* vol. 6. Springer International Publishing: New York, NY, USA. 2009 pp 563–599.

5 Karaagac E, Vargas A, de Andrés M, Carreño I, Ibáñez J, Carreño J *et al.* Marker-assisted selection for seedlessness in table grape breeding. *Tree Genet Genomes* 2012; **8**: 1003–1015.

6 Eibach R, Zyprian E, Welter L, Topfer R. The use of molecular markers for pyramiding resistance genes in grapevine breeding. *Vitis* 2007; **46**: 120–124.

7 Zini E, Raffeiner M, Di Gaspero G, Eibach R, Grando MS, Letschka T. Applying a defined set of molecular markers to improve selections of resistant grapevine accessions. *Acta Hortic* 2015; **1082**: 73–78.

8 Collard BCY, Mackill DJ. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos Trans R Soc Lond B Biol Sci* 2008; **363**: 557–572.

9 Varshney RK, Thudi M, Nayak SN, Gaur PM, Kashiwagi J, Krishnamurthy L *et al.* Genetic dissection of drought tolerance in chickpea (*Cicer arietinum* L.). *Theor Appl Genet* 2014; **127**: 445–462.

10 Reisch B, Owens C, Cousins PGrapeIn:Badenes ML, Byrne DH(eds). *Fruit Breeding* vol. 8. Springer International Publishing: New York, NY. USA. 2012, 225–262.

11 Shirasawa K, Hirakawa H. DNA marker applications to molecular genetics and genomics in tomato. *Breeding Sci* 2013; **63**: 21–30.

12 Boudehri K, Bendahmane A, Cardinet G, Troadec C, Moing A, Dirlewanger E. Phenotypic and fine genetic characterization of the *D* locus controlling fruit acidity in peach. *BMC Plant Biol* 2009; **9**: 59.

13 Alpert KB, Tanksley SD. High-resolution mapping and isolation of a yeast artificial chromosome contig containing *fw*2.2: A major fruit weight quantitative trait locus in tomato. *Proc Natl Acad Sci USA* 1996; **93**: 15503–15507.

14 De Franceschi P, Stegmeir T, Cabrera A, van der Knaap E, Rosyara UR, Sebolt AM *et al.* Cell number regulator genes in *Prunus* provide candidate genes for the control of fruit size in sweet and sour cherry. *Mol Breed* 2013; **32**: 311–326.

15 Spindel J, Wright M, Chen C, Cobb J, Gage J, Harrington S *et al.* Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theor Appl Genet* 2013; **126**: 2699–2716.

16 Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 2013; **9**: 29.

17 Huang XH, Han B. Natural variations and genome-wide association studies in crop plants. *Annu Rev Plant Biol* 2014; **65**: 531–551.

18 Fodor A, Segura V, Denis M, Neuenschwander S, Fournier-Level A, Chatelet P *et al.* Genome-wide prediction methods in highly diverse and heterozygous species: Proof-of-concept through simulation in grapevine. *PLoS One* 2014; **9**: e110436.

19 Khan MA, Korban SS. Association mapping in forest trees and fruit crops. *J Exp Bot* 2012; **63**: 4045–4060.

20 Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK *et al.* Genetic structure and domestication history of the grape. *Proc Natl Acad Sci USA* 2011; **108**: 3530–3535.

21 Myles S, Chia J-M, Hurwitz B, Simon C, Zhong GY, Buckler E *et al.* Rapid genomic characterization of the genus*Vitis*. *PLoS One* 2010; **5**: e8219.

22 Cattonaro F, Testolin R, Scalabrin S, Morgante M, Gaspero GDGenetic diversity in the grapevine germplasmIn:Tuberosa R, Graner A, Frison E(eds). *Genomics of Plant Genetic Resources*. Springer International Publishing: New York, NY, USA. 2014 pp 683–704.

23 Arnold C, Rossetto M, McNally J, Henry RJ. The application of SSRs characterized for grape (*Vitis vinifera*) to conservation studies in Vitaceae. *Am J Bot* 2002; **89**: 22–28.

24 Dirlewanger E, Denoyes-Rothan B, Yamamoto T, Chagné DGenomics tools across *Rosaceae* speciesIn:Folta K, Gardiner S(eds). *Genetics and Genomics of Rosaceae* vol. 6. Springer International Publishing: New York, NY, USA. 2009 pp 539–561.

25 Mnejja M, Garcia-Mas J, Audergon JM, Arus P. *Prunus* microsatellite marker transferability across rosaceous crops. *Tree Genet Genomes* 2010; **6**: 689–700.

26 Deschamps S, Llaca V, May GD. Genotyping-by-sequencing in plants. *Biology* 2012; **1**: 460–483.

27 Fechter I, Hausmann L, Daum M, Sorensen TR, Viehover P, Weisshaar B *et al.* Candidate genes within a 143 kb region of the flower sex locus in *Vitis*. *Mol Genet Genomics* 2012; **287**: 247–259.

28 Picq S, Santoni S, Lacombe T, Latreille M, Weber A, Ardisson M *et al.* A small XY chromosomal region explains sex determination in wild dioecious *V. vinifera* and the reversal to hermaphroditism in domesticated grapevines. *BMC Plant Biol* 2014; **14**: 229.

29 Grattapaglia D, Silva-Junior OB, Kirst M, de Lima BM, Faria DA, Pappas GJ Jr. High-throughput SNP genotyping in the highly heterozygous genome of *Eucalyptus*: assay success, polymorphism and transferability across species. *BMC Plant Biol* 2011; **11**: 65.

30 Yang HA, Li CD, Lam HM, Clements J, Yan GJ, Zhao SC. Sequencing consolidates molecular markers with plant breeding practice. *Theor Appl Genet* 2015; **128**: 779–795.

31 Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR. Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theor Appl Genet* 2010; **120**: 1525–1534.

32 Thomson MJ. High-throughput SNP genotyping to accelerate crop improvement. *Plant Breed Biotech* 2014; **2**: 195–212.

33 Miller AJ, Matasci N, Schwaninger H, Aradhya MK, Prins B, Zhong GY *et al. Vitis* phylogenomics: hybridization intensities from a SNP array outperform genotype calls. *PLoS One* 2013; **8**: e78680.

34 Myles S, Mahanil S, Harriman J, Gardner KM, Franklin JL, Reisch BI *et al.* Genetic mapping in grapevine using SNP microarray intensity values. *Mol Breed* 2015; **35**: 88–100.

35 Luby JJ, Shaw DV. Does marker-assisted selection make dollars and sense in a fruit breeding program? *Hort Sci* 2001; **36**: 872–879.

36 Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES *et al.* A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. *PLoS One* 2011; **6**: e19379.

37 He JF, Zhao XQ, Laroche A, Lu ZX, Liu HK, Li ZQ. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front Plant Sci* 2014; **5**: 484.

38 Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G *et al.* Marker density and read depth for genotyping populations using Genotyping-by-Sequencing. *Genetics* 2013; **193**: 1073–1081.

39 Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q *et al.* TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* 2014; **9**: e90346.

40 Swarts K, Li HH, Navarro JAR, An D, Romay MC, Hearne S *et al.* Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. *Plant Genome* 2014; **7**(3): 1–12.

41 Bianco L, Cestaro A, Sargent DJ, Banchi E, Derdak S, Di Guardo M *et al.* Development and validation of a 20 K Single Nucleotide Polymorphism (SNP) whole genome genotyping array for apple (*Malus × domestica* Borkh). *PLoS One* 2014; **9**: e110377.

42 Hyma KE, Barba P, Wang M, Londo JP, Acharya CB, Mitchell SE *et al.* Heterozygous mapping strategy (HetMappS) for high resolution Genotyping-By-Sequencing markers: a case study in grapevine. *PLoS One* 2015; **10**: e0134880.

43 Dalbò MA, Ye GN, Weeden NF, Wilcox WF, Reisch BI. Marker-assisted selection for powdery mildew resistance in grapes. *J Am Soc Hortic Sci* 2001; **126**: 83–89.

44 Reisch BI, Mahanil S, Consolie N, Luce RS, Wallace PG, Cadle-Davidson L. Examination of marker-assited selection for powdery mildew resistance. *Acta Hortic* 2014; **1046**: 151–155.

45 Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 2007; **23**: 2633–2635.

46 Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* 2007; **23**: 1289–1291.

47 Cadle-Davidson L, Chicoine DR, Consolie NH. Variation within and among *Vitis* spp. for foliar resistance to the powdery mildew pathogen *Erysiphe necator*. *Plant Dis* 2011; **95**: 202–211.

48 Feechan A, Kocsis M, Riaz S, Zhang W, Gadoury DM, Walker MA *et al.* Strategies for *RUN1* deployment using *RUN2* and *REN2* to manage grapevine powdery mildew informed by studies of race specificity. *Phytopathology* 2015; **105**: 1104–1113.

49 Manns DC, Mansfield AK. A core-shell column approach to a comprehensive high-performance liquid chromatography phenolic analysis of *Vitis vinifera* L. and interspecific hybrid grape juices, wines, and other matrices following either solid phase extraction or direct injection. *J Chromatogr A* 2012; **1251**: 111–121.

50 Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007; **449**: 463–467.

51 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**: 841–842.

52 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA *et al.* The variant call format and VCFtools. *Bioinformatics* 2011; **27**: 2156–2158.

53 *R core team: A language and environment for statistical computing ver v3.1.2*. R Foundation for Statistical Computing: Vienna, Austria. 2014.

54 Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging* 1989; **29**: 307–335.

55 Grzybowski M, Younger JG. Statistical methodology: III. Receiver operating characteristic (ROC) curves. *Acad Emerg Med* 1997; **4**: 818–826.

56 Pepe MS. Receiver operating characteristic methodology. *J Am Stat Assoc* 2000; **95**: 308–311.

57 Lawal BProbability and probability distributionsIn:. *Applied Statistical Methods in Agriculture, Health and Life Sciences*. Springer International Publishing: New York, NY, USA. 2014 pp 59–113.

58 Utro F, Haiminen N, Livingstone D, Cornejo OE, Royaert S, Schnell RJ *et al.* iXora: exact haplotype inferencing and trait association. *BMC Genet* 2013; **14**: 48–63.

59 Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 2013; **10**: 5–6.

60 Knapp SJ. Marker-assisted selection as a strategy for increasing the probability of selecting superior genotypes. *Crop Sci* 1998; **38**: 1164–1174.

61 Bernardo R. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Sci* 2008; **48**: 1649–1664.

62 Guimarães EP. Food and Agriculture Organization of the United NationsIn:. *Marker-Assisted Selection: Current Status and Future Perspectives in Crops, Livestock, Forestry and Fish*. Food and Agriculture Organization of the United Nations: Rome, Italy. 2007.

63 Hospital F, Moreau L, Lacoudre F, Charcosset A, Gallais A. More on the efficiency of marker-assisted selection. *Theor Appl Genet* 1997; **95**: 1181–1189.

64 Iftekharuddaula KM, Salam MA, Newaz MA, Ahmed HU, Collard BCY, Septiningsih EM *et al.* Comparison of phenotypic versus marker-assisted background selection for the *SUB1* QTL during backcrossing in rice. *Breeding Sci* 2012; **62**: 216–222.

65 Xu Y. *Molecular Plant Breeding*. CABI Publishing: Wallingford, CT, USA. 2010.

66 Castro P, Rubio J, Madrid E, Fernandez-Romero MD, Millan T, Gil J. Efficiency of marker-assisted selection for ascochyta blight in chickpea. *J Agr Sci* 2015; **153**: 56–67.

67 Hackett CA, Broadfoot LB. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity* 2003; **90**: 33–38.

68 Zhang L, Wang S, Li H, Deng Q, Zheng A, Li S *et al.* Effects of missing marker and segregation distortion on QTL mapping in F_2 populations. *Theor Appl Genet* 2010; **121**: 1071–1082.

69 Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doeblay J *et al.* Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* 2001; **98**: 11479–11484.

70 Di Gaspero G, Cipriani G, Adam-Blondon AF, Testolin R. Linkage maps of grapevine displaying the chromosomal locations of 420 microsatellite markers and 82 markers for R-gene candidates. *Theor Appl Genet* 2007; **114**: 1249–1263.

71 Di Gaspero G, Cipriani G, Marrazzo MT, Andreetta D, Castro MJP, Peterlunger E *et al.* Isolation of (AC)n-microsatellites in *Vitis vinifera* L. and analysis of genetic background in grapevines under Marker-assisted selection. *Mol Breed* 2005; **15**: 11–20.

72 Merdinoglu D, Butterlin G, Bevilacqua L, Chiquet V, Adam-Blondon AF, Decroocq S. Development and characterization of a large set of microsatellite markers in grapevine (*Vitis vinifera* L.) suitable for multiplex PCR. *Mol Breed* 2005; **15**: 349–366.

73 Calus MPL, Meuwissen THE, Windig JJ, Knol EF, Schrooten C, Vereijken ALJ *et al.* Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. *Genet Sel Evol* 2009; **41**: 11.

74 Hayes BJ, Chamberlain AJ, McPartlan H, Macleod I, Sethuraman L, Goddard ME. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genet Res* 2007; **89**: 215–220.

75 Cervantes-Martinez C, Brown JS. A haplotype-based method for QTL mapping of F_1 populations in outbred plant species. *Crop Sci* 2004; **44**: 1572–1583.

76 Feng XH, Keim D, Wanjugi H, Coulibaly I, Fu Y, Schwarz J *et al.* Development of molecular markers for genetic male sterility in *Gossypium hirsutum*. *Mol Breed* 2015; **35**: 141.

77 Fernandez L, Le Cunff L, Tello J, Lacombe T, Boursiquot JM, Fournier-Level A *et al.* Haplotype diversity of *VvTFL1A* gene and association with cluster traits in grapevine (*V. vinifera*). *BMC Plant Biol* 2014; **14**: 209.

78 Rafalski A. Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin Plant Biol* 2002; **5**: 94–100.

79 Jones ES, Sullivan H, Bhattramakki D, Smith JSC. A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (*Zea mays* L.). *Theor Appl Genet* 2007; **115**: 361–371.

80 Lambert SG, Asenstorfer RE, Williamson NM, Iland PG, Jones GP. Copigmentation between malvidin-3-glucoside and some wine constituents and its importance to colour expression in red wine. *Food Chem* 2011; **125**: 106–115.

81 Chagne D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C *et al.* Genome-wide SNP detection, validation, and development of an 8 K SNP array for apple. *PLoS One* 2012; **7**: e31745.

82 Vezzulli S, Micheletti D, Riaz S, Pindo M, Viola R, This P *et al.* A SNP transferability survey within the genus *Vitis*. *BMC Plant Biol* 2008; **8**: 128.

83 Gasperini D, Greenland A, Hedden P, Dreos R, Harwood W, Griffiths S. Genetic and physiological analysis of *Rht8* in bread wheat: an alternative source of semi-dwarfism with a reduced sensitivity to brassinosteroids. *J Exp Bot* 2012; **63**: 4419–4436.

84 Mahanil S, Ramming D, Cadle-Davidson M, Owens C, Garris A, Myles S *et al.* Development of marker sets useful in the early selection of *Ren4* powdery mildew resistance and seedlessness for table and raisin grape breeding. *Theor Appl Genet* 2012; **124**: 23–33.

85 Fennell A. Genomics and functional genomics of winter low temperature tolerance in temperate fruit crops. *Crit Rev Plant Sci* 2014; **33**: 125–140.

86 Fennell A, Mathiason K, Luby J. Genetic segregation for indicators of photoperiod control of dormancy induction in *Vitis* species. *Acta Hortic* 2005; **689**: 533–539.

87 Poland JA, Rife TW. Genotyping-by-Sequencing for plant breeding and genetics. *Plant Genome* 2012; **5**: 92–102.

88 Mathiason K, He D, Grimplet J, Venkateswari J, Galbraith DW, Or E *et al.* Transcript profiling in *Vitis riparia* during chilling requirement fulfillment reveals coordination of gene expression patterns with optimized bud break. *Funct Integr Genomics* 2009; **9**: 81–96.

89 Sreekantan L, Mathiason K, Grimplet J, Schlauch K, Dickerson JA, Fennell AY. Differential floral development and gene expression in grapevines during long and short photoperiods suggests a role for floral genes in dormancy transitioning. *Plant Mol Biol* 2010; **73**: 191–205.

90 Garris A, Clark L, Owens C, McKay S, Luby J, Mathiason K *et al.* Mapping of photoperiod-induced growth cessation in the wild grape *Vitis riparia*. *J Am Soc Hortic Sci* 2009; **134**: 261–272.

Supplemental Information for this article can be found on the *Horticulture Research* website (http://www.nature.com/hortres).