Mol Breeding (2015) 35:88
DOI 10.1007/s11032-015-0288-3

Genetic mapping in grapevine using SNP microarray intensity values

Sean Myles · Siraprapa Mahanil · James Harriman · Kyle M. Gardner · Jeffrey L. Franklin · Bruce I. Reisch · David W. Ramming · Christopher L. Owens · Lin Li · Edward S. Buckler · Lance Cadle-Davidson

Received: 14 October 2014 / Accepted: 16 February 2015 / Published online: 25 February 2015

Abstract Genotyping microarrays are widely used for genetic mapping, but in high-diversity organisms, the quality of SNP calls can be diminished by genetic variation near the assayed nucleotide. To address this limitation in grapevine, we developed a simple heuristic that uses hybridization intensity to genetically map phenotypes without the need to distinguish between polymorphic states. We applied this approach to the mapping of three previously mapped traits, each controlled by single major effect loci—color, flower sex, and powdery mildew resistance—and confirmed that intensity values outperform SNP calls in all cases. Further, because per sample cost is a major limitation to the adoption of genotyping microarrays in applied genetic research and plant breeding, we tested how many samples were required to map a Mendelian trait in an F1 grape population and found that we could identify the correct genomic region with as few as 12 samples. For high-diversity species for which genotyping arrays are available or under development, our findings suggest a powerful and cost-effective approach to identify large-effect QTL when faced with poor SNP quality.

Keywords *Vitis* · Grapevine · SNP discovery · Genotyping microarray

Introduction

There is currently strong demand for genome-wide genotype data from various organisms. The demand is

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. USDA is an equal opportunity provider and employer.

Electronic supplementary material The online version of this article (doi:10.1007/s11032-015-0288-3) contains supplementary material, which is available to authorized users.

S. Myles (✉) · K. M. Gardner
Department of Plant and Animal Sciences, Faculty of Agriculture, Dalhousie University, Truro, NS B2N 5E3, Canada
e-mail: sean.myles@dal.ca

S. Mahanil · J. Harriman · C. L. Owens · L. Cadle-Davidson (✉)
USDA-Agricultural Research Service (ARS) Grape Genetics Research Unit, 630 W. North St., Geneva, NY, USA
e-mail: lance.cadledavidson@ars.usda.gov

J. Harriman · E. S. Buckler
Institute for Genomic Diversity, USDA-ARS, Cornell University, Ithaca, NY, USA

J. L. Franklin
Agriculture and Agri-Food Canada, Kentville, NS, Canada

B. I. Reisch
Department of Horticulture, New York State Agricultural Experiment Station, Cornell University, Geneva, NY, USA

especially high in species of agricultural importance as genome-wide genotype data are required to improve the efficiency of breeding (Morrell et al. 2012; Heffner et al. 2009; Hamblin et al. 2011). To date, genotyping microarrays that assay thousands or hundreds of thousands of single nucleotide polymorphisms (SNPs) have been the technology of choice for acquiring genome-wide genotype data. These arrays have arguably had their largest impact by enabling genome-wide association (GWA) studies in humans (Stranger et al. 2011). Following their successful use in human genomics, microarrays have been developed for important agricultural species including cow (Matukumalli et al. 2009), pig (Ramos et al. 2009), sheep (Kijas et al. 2009), rice (Zhao et al. 2011), corn (Ganal et al. 2011), grape (Myles et al. 2010), peach (Verde et al. 2012), and apple (Chagne et al. 2012).

While SNP microarrays have proven to be effective in many organisms, there is increasing evidence that they are unsuitable for assaying genetic variation in organisms with high levels of genetic diversity. The main reason for this is that microarray probes are designed based on a reference genome sequence, and genotype calls from samples that are highly divergent from the reference are often unreliable because of reduced probe-sequence hybridization. This has been demonstrated in maize, where a sample's SNP calling failure rate is correlated with its genetic distance to the maize reference genome (Ganal et al. 2011). Moreover, mismatches between a probe and a sample's DNA sequence have been shown to reduce SNP quality scores using a grapevine SNP array (Myles et al. 2010). The reliability of SNP calls is further complicated by variation in genome content: The presence of insertion/deletion polymorphisms (indels), copy number variants (CNVs), and presence–absence variants (PAVs) interfere with hybridization signals and are known to be ubiquitous in several high-diversity species including maize (Morgante et al. 2005; Springer et al. 2009) and soybean (Kim et al. 2010). Finally, SNP calling algorithms often rely on clustering approaches that assume the presence of three distinct SNP genotypes, two homozygous genotypes, and a heterozygous genotype, roughly in Hardy–Weinberg equilibrium (Teo et al. 2007; Rabbee and Speed 2006). This assumption is often violated in high-diversity plants, where population structure results in genotype frequencies that deviate significantly from Hardy–Weinberg equilibrium, for example when alternative alleles are fixed between species or subspecies and heterozygotes are absent. The assumption is violated even in full-sibling populations when the probe sequence no longer assays the bi-allelic SNP it was designed to detect, for reasons including sequence divergence, CNVs, and PAVs.

Here we focus on the Vitis9KSNP array, a SNP array designed to genotype diverse grapevine germplasm (Myles et al. 2010). The domesticated grapevine, *Vitis vinifera*, is as diverse as maize with a SNP every 60 bp (Lijavetzky et al. 2007). Grapevines are also highly heterozygous: The highly divergent alternative haplotypes within a single cultivar make genome assembly challenging (Velasco et al. 2007). However, the Vitis9KSNP array was designed to genotype not only *V. vinifera* cultivars, but to assay variation across the entire genus *Vitis*, which includes species that have likely been geographically isolated from *V. vinifera* for tens of millions of years. Thus, wild North American grapevines and hybrids generated from crosses between *V. vinifera* and wild species are expected to have highly divergent sequences compared with the Pinot noir inbred 'PN40024' reference genome sequence on which the Vitis9KSNP array probes are based (The French-Italian Public Consortium for Grapevine Genome Characterization 2007). Our focus on a highly diverse crop such as the grapevine provides a useful template for extending genetic mapping and marker-assisted breeding to other highly diverse agricultural species (Reisch et al. 2012).

The present study investigates the utility of using probe-sequence hybridization scores rather than SNP calls from a SNP genotyping microarray in genetic mapping experiments of Mendelian phenotypes involving highly diverse grapevine germplasm. Our approach is motivated by previous work that has used oligonucleotide probes to call genotypes without directly querying alternative alleles at a locus. For example, comparative genomic hybridization (CGH) has been used to identify polymorphisms between maize inbred lines (Fu et al. 2010; Springer et al. 2009); association mapping has been performed in

D. W. Ramming
USDA-ARS San Joaquin Valley Agricultural Sciences
Center, 9611 South Riverbend Avenue, Parlier, CA, USA

L. Li
Department of Biostatistics, Harvard University, Boston,
MA, USA

Arabidopsis by hybridizing DNA to expression arrays and calling single feature polymorphisms (SFPs) (Kim et al. 2006); AFLP marker band intensities have been used for association mapping in tetraploid potato (D'hoop et al. 2008); and CNVs and loss-of-heterozygosity (LOH) regions have been detected using SNP microarray data in humans (Peiffer et al. 2006). These approaches mostly do not distinguish among different types of polymorphisms as the observed differences in hybridization signal can be caused by SNPs, indels, CNVs, or PAVs. A wide variety of statistical methods are employed to distinguish between polymorphic states and thereby assign each sample a "genotype" based on the hybridization signal from a single probe sequence or by integrating across adjacent probes. Here we demonstrate that genotype calls from a grapevine genotyping microarray are often unreliable, and we introduce a simple heuristic that uses hybridization intensity to genetically map phenotypes without the need to distinguish between polymorphic states. Our method not only circumvents the need to call genotypes, but it is model free, simple to compute, and more powerful than relying on genotype calls. In addition, we demonstrate that as few as 12 offspring from an F_1 cross-segregating for a Mendelian phenotype are required to localize the underlying causal genomic region using our approach.

Materials and methods

Phenotyping

Biparental grapevine populations were selected for this study based on qualitative marker-trait associations previously mapped (Table 1). Grape berry skin color was unambiguously scored as either blue or white. Flower sex was scored as either hermaphroditic or staminate. Powdery mildew resistance was evaluated following natural infection in a California vineyard and on inoculated detached leaves in New York, the results of which had full agreement (Ramming et al. 2011). Vines with any sporulation were considered susceptible.

Vitis9KSNP genotyping microarray

Oligonucleotide probes on the Vitis9KSNP genotyping microarray each consist of a 50 bp query sequence

Table 1 Description of grapevine linkage mapping populations analyzed in the present study

Phenotype	Parent 1 [ancestry] (phenotype)	Parent 2 [ancestry] (phenotype)	Offspring		No. of tested SNPs	Locus Name	Locus position*	References
			N (phenotype)	*N* (phenotype)				
Color	St. Pepin [*vinifera*, *riparia*, *aestevalis*, *labrusca*] (white)	Cabernet Franc [*vinifera*] (blue)	7 (white)	9 (blue)	4036	VvMybA	Chr 2: 12.4–12.5 Mb	Fournier-Level et al. (2009)
Flower sex	Horizon [*vinifera*, *rupestris*, *aestivalis*, *labrusca*] (hermaphroditic)	Ill. 547-1 [*rupestris*, *cinerea*] (staminate)	6 (hermaphroditic)	6 (staminate)	3374	Sex	Chr 2: 3.7–5.0 Mb	Lowe and Walker (2006), Marguerit et al. (2009), Dalbo et al. (2000)
Powdery mildew resistance	C87-41 [*romanetii*, *vinifera*] (resistant)	B70-57 [*vinifera*] (susceptible)	12 (resistant)	4 (susceptible)	3630	*Ren4*	Chr 18: 15.2–16.5 Mb	Mahanil et al. (2012)

Each of these F_1 double pseudo-testcross populations segregates for the dominant, qualitative phenotype listed in parentheses, as detailed in previous references. The diverse ancestries of the parents are provided as *Vitis* species names in square brackets

* Locus position is according to the 8× PN40024 genome sequence

that is complementary to the 8× PN40024 reference genome (Myles et al. 2010). Each probe hybridizes with one target locus, and either an allele-specific primer extension (ASPE) step or a single base extension (SBE) reaction is performed to assay the two alleles at each SNP. Each of the 8898 SNPs on the Vitis9KSNP array is queried by an average of 30 identical probes. An intensity value for allele A and allele B of each SNP is generated by averaging the intensities from the probes querying that SNP. The A and B intensity information is then converted to normalized polar coordinates (Gunderson et al. 2005, 2006). We refer to these normalized intensity values as *X* and *Y* (Fig. 2).

Genotype data

Genotype calls from the Vitis9KSNP array were made using Illumina's probabilistic model in which probabilities are assigned for membership into three genotype clusters: AA, AB, or BB, where A and B are the two alleles at a SNP. The probabilistic assignment to these clusters is based on the normalized intensity values that we refer to as *X* and *Y*. The degree to which samples cluster into the three distinct genotype classes is summarized as a clustering quality score (GenTrain score), which is assigned to each SNP. For each SNP, a genotype quality score (GenCall score) is assigned to each sample based on how well the sample clusters within one of the defined genotypic classes. We used BeadStudio to generate genotype calls from the Vitis9KSNP genotyping microarray, which assays 8898 SNPs across the grapevine genome (Myles et al. 2010). We did not implement a SNP quality score filter as we did not want true positive associations to be filtered out at the SNP calling stage. To control the number of statistical tests performed within each association analysis, we restricted the association analysis within each mapping population to SNPs with a minor allele frequency >0.1 and with <20 % missing data within each population in order to ensure that we tested only segregating SNPs. The final number of SNPs considered within each population is shown in Table 1.

Genetic mapping of each phenotype was performed in PLINK using the –assoc command, which performs a Chi-squared test assessing the magnitude of the difference in allele frequency between cases and controls. The resulting *P* values indicate the strength of the genotype–phenotype association. The "inflation factor" of each association test (λ) was calculated as the median of the Chi-squared test statistics divided by 0.455, which is the expected median Chi-square value under the null hypothesis of no association for a one degree of freedom test.

Intensity data

We tested several summary statistics of the intensity values for their power to map phenotypes in the grapevine F1 population segregating for powdery mildew resistance. We found that all of our tested statistics performed better than genotype calls, regardless of whether they captured allele-specific information (e.g., *X* or *Y* only) or whether they summarized intensity information from both alleles (e.g., ln(*X*/(*X* + *Y*)), ln(*X*/*Y*); Figure S1). This was the case despite our observation that the statistics we tested were often only weakly correlated with one another (Figure S2). We present results from using ln(*X*/*Y*) as this summary statistic had the lowest false-positive inflation factor (λ; Figure S3), produced the largest number of significant SNPs at the powdery mildew resistance (*Ren4*) locus, and produced the most significant *P* value of all statistics tested (Table S1).

The ln(*X*/*Y*) statistic was calculated for each sample at each of the 8898 SNPs, or probe sets, on the Vitis9KSNP array. To test for an association with the phenotype, we performed a Student's *t* test between the ln(*X*/*Y*) values from samples with one phenotype (e.g., white skin) and the ln(*X*/*Y*) values of samples with the other phenotype (e.g., blue skin). The resulting *P* values indicate the strength of the genotype–phenotype association. Because the *t* distribution is approximately normal with a mean of 0 and a standard deviation of 1, the "inflation factor" of each association test (λ) was calculated as the median of the squared *t* statistic divided by 0.455.

Sample size analysis

The ability to genetically map a Mendelian phenotype with varying sample sizes was evaluated in the F1 population segregating for color (see Table 1). We resampled individuals such that every possible combination of cases (white skin) and controls (blue skin) were compared at varying sample sizes. For example, for case/control sample size = 2, association analyses using intensity values were performed for every possible combination of two white-skinned offspring versus two blue-skinned offspring. Thus, for each case/control sample size, the number of analyses varied: For the case/control sample sizes in parentheses, the numbers of analyses performed were 756 (2), 2940 (3), 4410 (4), 2646 (5), 588 (6), and 36 (7), respectively. For each case/control sample size, we calculated the proportion of analyses that correctly mapped grape skin color to chromosome 2. The mapping was considered "correct" if the most significant *P* value was found at the correct location on chromosome 2 and was below the genome-wide significance threshold. The analysis was performed in *R*, and the *R* script for this analysis is available here: www.cultivatingdiversity.org/software.

Results

The Vitis9KSNP array is designed to assay 8898 SNPs across the grapevine genome, with each SNP queried by a probe set made up of several identical probes distributed across the array (Myles et al. 2010; see "Materials and methods" section). Using this array, we genotyped three F_1 double pseudo-testcross populations (i.e., parents and offspring are highly heterozygous), which segregate 1:1 for dominant, Mendelian phenotypes including berry color, flower sex, and powdery mildew resistance. Table 1 provides a description of these three populations.

In analyzing an F_1 double pseudo-testcross using a genotyping microarray, only SNPs that are heterozygous in one parent and homozygous in the other parent are informative. For each mapping population, if we restrict our analysis to only SNPs with this parental segregation pattern, exclude offsprings' genotypes that are inconsistent with Mendelian inheritance, implement liberal quality thresholds (GenTrain and GenCall scores >0.25), and remove SNPs with >20 % missing data, only 15 % or fewer SNPs remained for analysis in the populations segregating for color (1377 SNPs), flower sex (939 SNPs), and powdery mildew resistance (1320 SNPs). Thus, using these filters, over 85 % of the data collected from the Vitis9KSNP array are ignored.

To include as many useful genotypes as possible, we implemented a more liberal filter without SNP quality thresholds. SNPs with a minor allele frequency >0.1 and with <20 % missing data within each population were included. The inclusion of low-quality genotypes could arguably result in false-positive associations. However, the genomic positions of the causal loci in each of the three populations are known, and our goal is to assess the power of genetic mapping with genotype calls compared with intensity values. Thus, a liberal threshold is most appropriate in order to include as many genotype calls in the analysis as possible. The implementation of our liberal filter results in 4036 SNPs, 3374 SNPs, and 3630 SNPs in the populations segregating for color, flower sex, and powdery mildew resistance, respectively (Table 1).

Our motivation for using intensity values from a SNP array to map phenotypes in the grapevine stems from the poor quality of the genotypes generated from the Vitis9KSNP array. Figure 1 demonstrates that both the quality of the genotype clusters (i.e., GenTrain scores) and the quality of the genotype calls (i.e., GenCall scores) are far poorer in the present data set from the Vitis9KSNP array compared with data from the human 650 K genotyping array. Although both of the arrays are Infinium BeadChips from Illumina and thus make use of identical technologies, the performance of the grapevine array is clearly inferior to the human array. This is not surprising since the BeadChips were first designed to query human genetic variation, which can be an order of magnitude lower than variation in some high-diversity plant species. Additional metrics also indicate poor quality from the Vitis9KSNP array. For example, the percentage of genotypes that are inconsistent with Mendelian inheritance is 14, 18, and 12 % in the populations segregating for color, flower sex, and powdery mildew resistance, respectively. Moreover, between 33 and 44 % of SNPs would be excluded from the present analyses if we implemented the recommended quality thresholds from Illumina by excluding SNPs with GenCall scores <0.25 (Fan et al. 2003).

We hypothesize that the reason for the poor quality of the genotype data from the Vitis9KSNP array is due to the high levels of genetic diversity in the grapevine. Polymorphism within the probe sequence and variation in genome content (i.e., indels, CNVs and PAVs) produce hybridization signals that cannot be easily interpreted as bi-allelic genotype calls. To determine whether hybridization signals could be used for genetic mapping without the need to call genotypes,

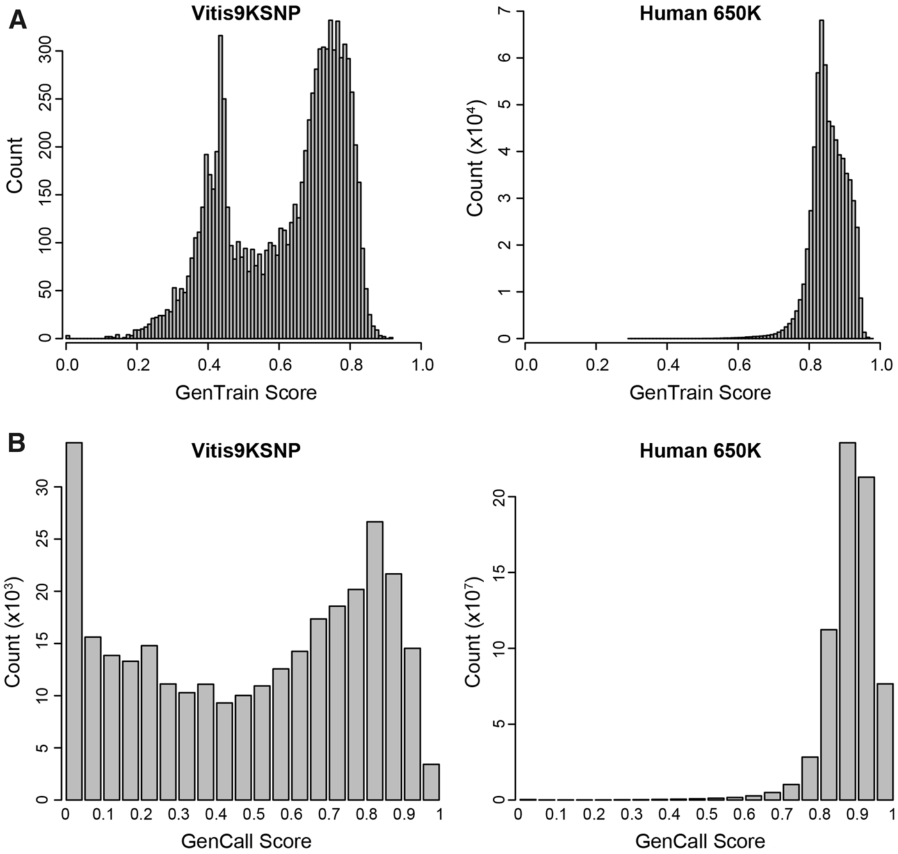

Fig. 1 Genotype calls from a human genotyping microarray are more reliable than genotype calls from the grapevine genotyping microarray. The quality scores for the Vitis9KSNP array (*left*) were obtained from the grapevine samples analyzed in the present study. The quality scores for the Human 650 K SNP array (*right*) were obtained from a set of 1043 humans from 51 different human populations (Li et al. 2008). Both the SNP quality scores (GenTrain scores) and the genotype quality scores (GenCall scores) vary from 0 (low) to 1 (high). **a** The distribution of GenTrain scores, which summarize the overall quality of a SNP based on its tendency to cluster into distinct genotypic classes. **b** The distribution of GenCall scores, which are quality scores assigned to each genotype. For each SNP, GenCall scores are assigned to each sample based on how well the sample clusters within one of the defined genotypic classes

we investigated several summary statistics of the fluorescence intensity values generated from the Vitis9KSNP array. All of the tested summary statistics outperformed the use of genotype calls. We found that summarizing each of the 8898 assayed loci using ln(X/Y), where X and Y are the normalized fluorescence intensity values from allele A and B at a locus, was a suitable summary statistic based on several metrics (see Methods and Figure S1, S2, Table S1). Figure 2 demonstrates the utility of using the ln(X/Y) summary of the fluorescence intensity values when dealing with hybridization signals that result in low-quality genotype calls.

To assess the utility of our ln(X/Y) statistic for genetic mapping compared with the use of genotype calls, we mapped three simply inherited grapevine phenotypes using both approaches. We found that using ln(X/Y) to summarize fluorescence intensity values provided a dramatic increase in the ability to genetically map these phenotypes in all three mapping populations, relative to using genotype calls (Fig. 3; Figure S3). Upon closer investigation of the signals surrounding the known causal loci in each of the mapping populations, we found that the association signals generated from using intensity values were near or overlap with the known causal loci in all three mapping populations (Fig. 4). Finally, we assessed how few samples were required in order to map a Mendelian phenotype in the grapevine using intensity values from the Vitis9KSNP array. We found that

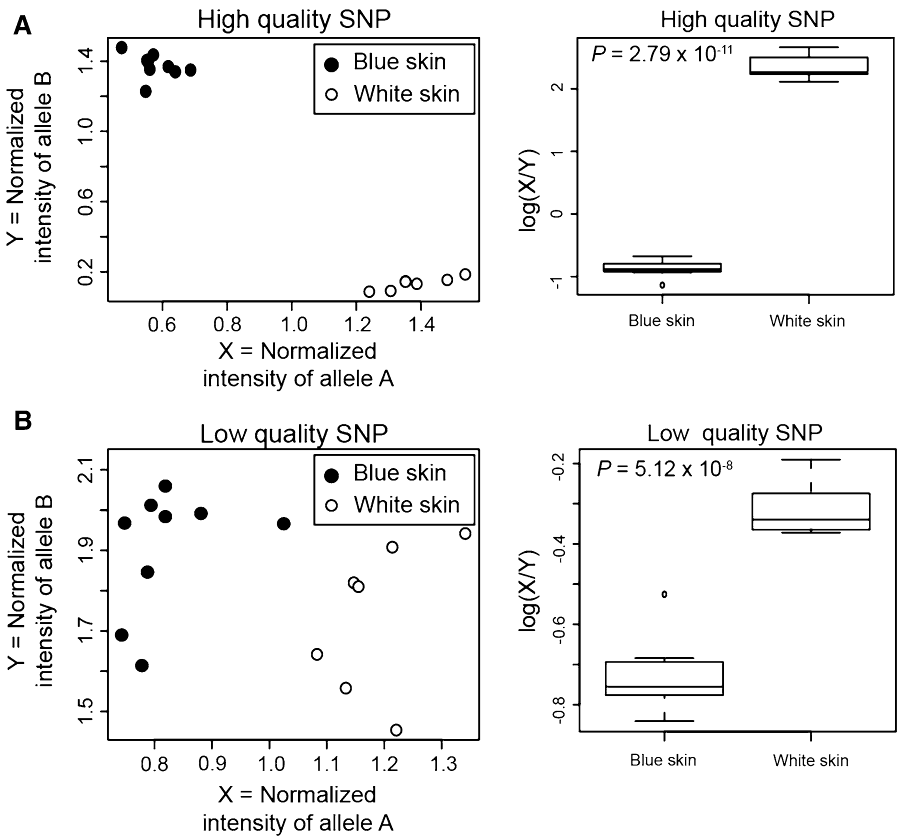

Fig. 2 Fluorescence intensity values enable trait mapping even with low-quality SNPs. Reliable SNP calls are only generated when fluorescence intensities corresponding to alleles A and B are distinct. **a** A high-quality SNP (GenTrain score = 0.77, GenCall scores > 0.8) and **b** A low-quality SNP (GenTrain score = 0.31, GenCall scores < 0.3) from the Vitis9KSNP array. For a high-quality SNP (**a**) and a low-quality SNP (**b**) that fall within the genomic region associated with grape skin color, the genotype cluster plots are shown to the *left*. Each offspring from an F_1 grapevine population segregating for skin color is represented as a *dot,* and each offspring's position in the cluster plot reflects its normalized fluorescence intensity from allele A (x-axis) and allele B (y-axis). The *dots* in the cluster plot of **b** do not cluster clearly into genotypic classes, which results in low SNP (GenTrain) and genotype (GenCall) quality scores. The SNP represented in **a** is strongly associated with skin color when using genotype calls ($P = 8.74 \times 10^{-4}$) but does not fall below the Bonferroni-corrected *P* value threshold ($P = 1.24 \times 10^{-5}$). The SNP represented in **b** does not associate significantly with skin color when using genotype calls ($P = 0.765$) and would likely be excluded from standard analyses because of low quality. *Boxplots* on the *right* of each panel show the result of an association test between grape skin color and intensity values (ln(X/Y)). Intensity values for both the high-quality SNP (**a**) and the low-quality SNP (**b**) produce strongly significant associations with grape skin color. Thus, while probes on genotyping microarrays may produce low-quality genotype calls, their intensity values nevertheless contain useful information for genetic mapping

using only 6 cases and 6 controls sufficed to genetically map grape color (Figure S4).

Discussion

High levels of genetic diversity in many species make assaying genetic polymorphism problematic. While genotyping microarrays work well for species with reduced diversity (e.g., humans, mice), they present challenges when used in high-diversity species, which includes many agriculturally important plants such as grapevines (Fig. 1). High polymorphism levels can interfere with hybridization signals. For example, we previously found that the genotype quality scores from the Vitis9KSNP array decrease significantly in the presence of mismatches between the probe and the sequence (Myles et al. 2010). Genotyping microarrays

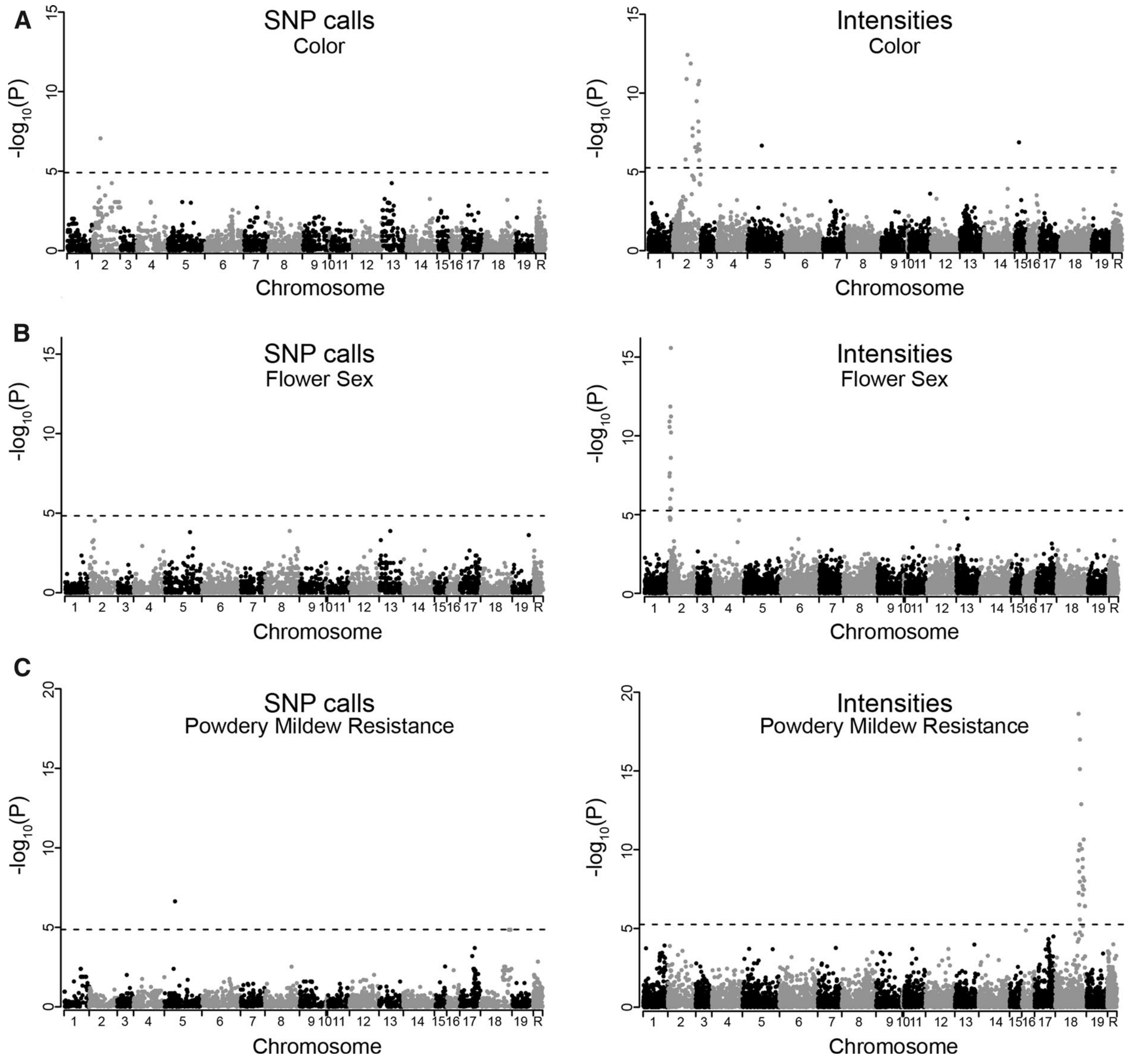

Fig. 3 Fluorescence intensity values from the Vitis9KSNP microarray increase the power to genetically map grapevine phenotypes, relative to using SNP calls. Each panel shows two Manhattan plots with genome-wide $-\log_{10}(P)$ values from a case/control association analysis using SNP calls (*left*) and intensity values (*right*). The analysis was performed in three F_1 populations segregating for different qualitative phenotypes: skin color (**a**), flower sex (**b**), and powdery mildew resistance (**c**). The genome-wide Bonferroni-corrected significance thresholds are indicated by the *dotted horizontal lines*. Chromosome "*R*" refers to SNPs that are unanchored to the genome assembly. For each phenotype, the association signal is stronger using the intensity values than the SNP calls

are often developed for the purpose of genetic mapping, yet in high-diversity species such as the grapevine, we find that a significant proportion of the genotypes called from the array are too low in quality to be useful. Here we show that genotype calling can be circumvented altogether and that an array's fluorescence intensity values can be more useful than genotype calls for mapping simply inherited phenotypes.

There are several possible ways in which high levels of genetic diversity can interfere with clear genotype clustering and calling. For example, the 50 bp probes that lie adjacent to SNPs on the Vitis9KSNP array were designed based on the

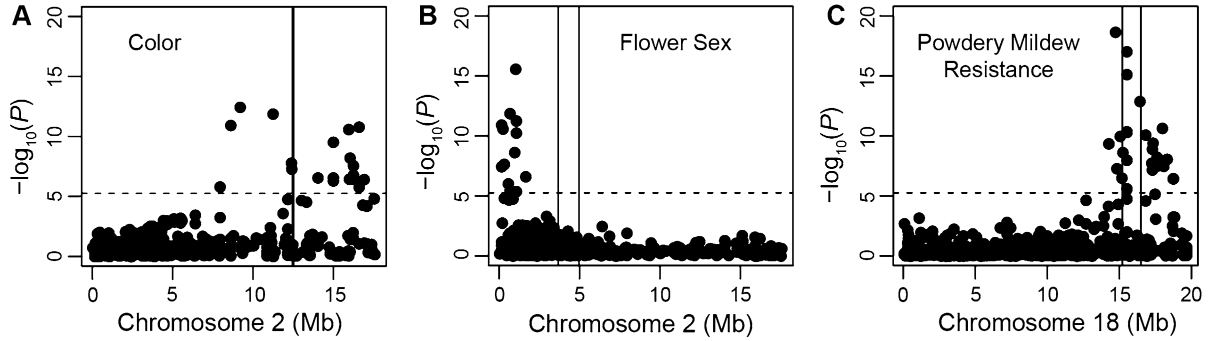

Fig. 4 Association scores resulting from the use of Vitis9KSNP microarray intensity values across the chromosomes harboring known causal loci for three grapevine phenotypes. Intensity values from the Vitis9KSNP microarray were used in F_1 populations to map three phenotypes: skin color (**a**), flower sex (**b**), and powdery mildew resistance (**c**). The previously determined mapping intervals for each phenotype are indicated by *solid vertical lines*. The known genomic interval for the skin color locus is so small that it appears as a *single solid vertical line* in **a**. The Bonferroni-corrected genome-wide significance thresholds are indicated by the *dotted horizontal lines*

PN40024 reference genome. When assaying a cultivar that has a mismatch in the sequence queried by the 50 bp probe, we can expect reduced hybridization and thus some interference with the clustering of that cultivar in genotype calling space. Alternatively, CNVs and PAVs can increase or decrease hybridization signals in particular cultivars, either in a locus or allele-specific manner. While these types of polymorphism lead to low-quality genotype calls, they may nonetheless be useful for genetic mapping. In order to make use of these other forms of variation for mapping purposes, we developed a simple mapping method that makes use of the fluorescence intensity values without assigning each sample a particular “genotype.”

We tested several summary statistics of the fluorescence intensities, some of which assayed variation in an allele-specific manner, and some of which only assayed fluorescence intensity at the locus in general. Both types of statistics outperformed genotype calling in their ability to map Mendelian phenotypes to known loci, and many of the summaries were correlated with one another. It is unclear from the present analysis what the optimal summary statistic is and what factors should be considered in choosing a particular summary statistic. For example, a combination of an allele-specific and a locus-specific summary may outperform our chosen summary, depending on the application.

When using genotype calls to map, we did not discard SNPs based on quality scores but rather implemented liberal minor allele frequency and missing data thresholds. Despite this, there were SNPs discarded from analysis whose fluorescence intensity values were significantly associated with the trait. For example, for the 25 probe sets whose intensity values were significantly associated with powdery mildew resistance, 5 resulted in genotype calls that did not pass our liberal filters. Moreover, in cases where SNPs did pass our filter, highly significant loci based on the intensity values were found to be nonsignificant based on genotype calls. From Fig. 3, it is clear that the intensity values are capturing more information and provide more power for mapping phenotypes and this is likely due to the fact that intensity values capture information about genetic polymorphism that is not captured by genotype calling algorithms. However, the small sample sizes in the present study may have resulted in poorer than usual genotype calls and reduced power to detect associations using the Chi-squared test implemented here. Thus, larger mapping populations will be required to definitively test whether intensity values are more useful than genotype calls.

While the genetic mapping results using intensity values clearly delineate single genomic regions associated with the three Mendelian phenotypes of interest, overlap with the known causal locus was not always complete (Fig. 4). However, the most significant values are not always expected to be found at the causal locus. For example, in a recent GWA study using more than 500 rice landraces, there were several occasions in which known causative markers showed weaker signals than nearby markers (Huang et al. 2010). Similar results have been observed in GWA studies conducted in *Arabidopsis* (Atwell et al. 2010). Thus, our observation of the most significant markers not overlapping with known causal loci is not unusual (Fig. 4). However, we find a strong association signal for flower sex that is several Mb from the interval identified in previous studies (Lowe and Walker 2006; Marguerit et al. 2009; Dalbo et al. 2000; Fechter et al. 2012; Picq et al. 2014; Riaz et al. 2006). Most likely, this inaccurate localization is due to a technical artifact of the microarray, since when this trait is mapped using genotyping-by-sequencing across the full population of 'Horizon' x Ill. 547-1, the sex locus localized between 4.75 and 5.12 Mb on chromosome 2 of the PN400024 version 12X.0 (Barba, Cadle-Davidson, and Reisch, personal communication). Unexpected hybridization of nontargeted paralogous regions may also explain this discrepancy.

For much applied genetic research and plant breeding, per sample cost is a major limitation to the adoption of genotyping microarrays. Here we demonstrate that qualitative Mendelian traits can be mapped in an F1 double pseudo-testcross population by collecting sufficiently dense microarray data from as few as 12 samples (Figure S4). We identify an average of 16 trait-associated markers in populations of 6 cases and 6 controls segregating for a Mendelian trait. To identify simple bi-allelic SNPs of high quality for marker-assisted breeding, larger populations of samples segregating for the trait could be genotyped for SNPs in and around the trait-associated markers using a cheaper, lower-plex technology for interval mapping (Mahanil et al. 2012). In this manner, the cost of identifying and verifying markers useful for marker-assisted breeding may be reduced. We anticipate that our approach can be extended to non-Mendelian traits, i.e., traits controlled by a larger number of loci with smaller effects, but further investigation is required to determine the power of our approach for mapping such traits.

It is becoming increasingly obvious that genotyping microarrays are not suitable for many purposes in high-diversity species. We find that 33–44 % of genotype calls from the Vitis9KSNP array should be discarded according to the most liberal criteria recommended by Illumina (Fig. 1). Such a high rate of missing data results in a substantial increase in the per genotype cost of useful data. This scenario is not unique to the grapevine. Despite extensive measures to control the quality of the SNPs on the recently developed SNP genotyping array for the apple, 28 % of the SNPs on the array exhibited poor-quality genotype clustering or were monomorphic and were thus excluded (Chagne et al. 2012). Efforts to develop arrays are currently in progress for a number of high-diversity species despite the poor quality of the genotype data generated from the grapevine and apple arrays. For high-diversity species already committed to the development or application of genotyping arrays, our findings provide a powerful approach to identifying marker-trait associations for Mendelian traits when faced with poor SNP quality. However, it is evident that high-throughput genotyping methods that make use of next-generation DNA sequencing technologies [e.g., genotyping-by-sequencing (Elshire et al. 2011)] are now the preferred method in place of genotyping microarrays for most genome-wide genotyping applications in high-diversity species.

Acknowledgments We thank Jeremiah Degenhardt and Carlos D. Bustamante for useful discussion. This research was undertaken thanks to funding from the USDA-Agricultural Research Service, the National Science and Engineering Research Council (NSERC) of Canada and the Canada Research Chairs program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethical standards The research presented here did not involve the use of human participants or animal.

Conflict of interest The authors declare no conflict of interests.

References

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627–631

Chagne D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, Vanderzande S, Hellens RP, Kumar S, Cestaro A, Velasco R, Main D, Rees JD, Iezzoni A, Mockler T, Wilhelm L, Van de Weg E, Gardiner SE, Bassil N, Peace C (2012) Genome-wide SNP detection, validation, and development of an 8 K SNP array for apple. PLoS ONE 7:e31745. doi:10.1371/journal.pone.0031745

D'hoop BB, Paulo MJ, Mank RA, van Eck HJ, van Eeuwijk FA (2008) Association mapping of quality traits in potato (*Solanum tuberosum L.*). Euphytica 161:47–60

Dalbo MA, Ye GN, Weeden NF, Steinkellner H, Sefc KM, Reisch BI (2000) A gene controlling sex in grapevines placed on a molecular marker-based genetic map. Genome 43:333–340

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379

Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, McBride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelpnagel J, Chee MS (2003) Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol 68:69–78

Fechter I, Hausmann L, Daum M, Rosleff Sörensen T, Viehöver P, Weisshaar B, Töpfer R (2012) Candidate genes within a 143 kb region of the flower sex locus in vitis. Mol Genetics Genomics 287:247–259. doi:10.1007/s00438-012-0674-z

Fournier-Level A, Le Cunff L, Gomez C, Doligez A, Ageorges A, Roux C, Bertrand Y, Souquet J-M, Cheynier V, This P (2009) Quantitative genetic bases of anthocyanin variation in grape (*Vitis vinifera L.* ssp. *sativa*) berry: a quantitative trait locus to quantitative trait nucleotide integrated study. Genetics 183:1127–1139. doi:10.1534/genetics.109.103929

Fu Y, Springer NM, Ying K, Yeh C-T, Iniguez AL, Richmond T, Wu W, Barbazuk B, Nettleton D, Jeddeloh J, Schnable PS (2010) High-resolution genotyping via whole genome hybridizations to microarrays containing long oligonucleotide probes. PLoS ONE 5:e14178

Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, Clarke JD, Graner E-M, Hansen M, Joets J, Le Paslier M-C, McMullen MD, Montalent P, Rose M, Schön C-C, Sun Q, Walter H, Martin OC, Falque M (2011) A large maize (Zea mays L.) SNP Genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS ONE 6:e28334. doi:10.1371/journal.pone.0028334

Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS (2005) A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 37:549

Gunderson KL, Steemers FJ, Ren H, Ng P, Zhou L, Tsan C, Chang W, Bullis D, Musmacker J, King C, Lebruska LL, Barker D, Oliphant A, Kuhn KM, Shen R (2006) Whole-genome genotyping. Methods Enzymol 410:359–376

Hamblin MT, Buckler ES, Jannink J-L (2011) Population genetics of genomics-based crop improvement methods. Trend Genetics 27:98

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci 49:1–12. doi:10.2135/cropsci2008.08.0512

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967

Kijas JW, Townley D, Dalrymple BP, Heaton MP, Maddox JF, McGrath A, Wilson P, Ingersoll RG, McCulloch R, McWilliam S, Tang D, McEwan J, Cockett N, Oddy VH, Nicholas FW, Raadsma H, for the International Sheep Genomics C (2009) A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS ONE 4:e4668

Kim S, Zhao K, Jiang R, Molitor J, Borevitz JO, Nordborg M, Marjoram P (2006) Association mapping with single-feature polymorphisms. Genetics 173:1125–1133. doi:10.1534/genetics.105.052720

Kim MY, Lee S, Van K, Kim T-H, Jeong S-C, Choi I-Y, Kim D-S, Lee Y-S, Park D, Ma J, Kim W-Y, Kim B-C, Park S, Lee K-A, Kim DH, Kim KH, Shin JH, Jang YE, Kim KD, Liu WX, Chaisan T, Kang YJ, Lee Y-H, Kim K-H, Moon J-K, Schmutz J, Jackson SA, Bhak J, Lee S-H (2010) Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. Proc Natl Acad Sci. doi:10.1073/pnas.1009526107

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104. doi:10.1126/science.1153717

Lijavetzky D, Cabezas JA, Ibanez A, Rodriguez V, Martinez-Zapater JM (2007) High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. BMC Genom 8:424

Lowe KM, Walker MA (2006) Genetic linkage map of the interspecific grape rootstock cross Ramsey (Vitis champinii) × Riparia Gloire (Vitis riparia). Theor Appl Genet 112:1582–1592

Mahanil S, Ramming DW, Cadle-Davidson MM, Owens C, Garris AJ, Myles S, Cadle-Davidson L (2012) Development of marker sets useful in the early selection of Ren4 powdery mildew penetration resistance and seedlessness for table and raisin grape breeding. Theor Appl Genet 124:23

Marguerit E, Boury C, Manicki A, Donnart M, Butterlin G, Nemorin A, Wiedemann-Merdinoglu S, Merdinoglu D, Ollat N, Decroocq S (2009) Genetic dissection of sex determinism, inflorescence morphology and downy mildew resistance in grapevine. Theor Appl Genet 118:1261–1278

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TPL, Sonstegard TS, Van Tassell CP (2009) Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4:e5350

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37:997–1002

Morrell PL, Buckler ES, Ross-Ibarra J (2012) Crop genomics: advances and applications. Nat Rev Genet 13:85–96

Myles S, Chia J-M, Hurwitz B, Simon C, Zhong GY, Buckler E, Ware D (2010) Rapid genomic characterization of the genus vitis. PLoS ONE 5:e8219

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL (2006) High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. Genome Res 16:1136–1148. doi:10.1101/gr.5402306

Picq S, Santoni S, Lacombe T, Latreille M, Weber A, Ardisson M, Ivorra S, Maghradze D, Arroyo-Garcia R, Chatelet P, This P, Terral J, Bacilieri R (2014) A small XY chromosomal region explains sex determination in wild dioecious V-vinifera and the reversal to hermaphroditism in domesticated grapevines. BMC Plant Biol 14:229

Rabbee N, Speed TP (2006) A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22:7–12. doi:10.1093/bioinformatics/bti741

Ramming DW, Gabler F, Smilanick J, Cadle-Davidson MM, Barba P, Mahanil S, Cadle-Davidson L (2011) A single dominant locus Ren4 confers non-race-specific penetration resistance to grapevine powdery mildew. Phytopathology. doi:10.1094/PHYTO-09-10-0237

Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P, Hansen MS, Hedegaard J, Hu Z-L, Kerstens HH, Law AS, Megens H-J, Milan D, Nonneman DJ, Rohrer GA, Rothschild MF, Smith TPL, Schnabel RD, Van Tassell CP, Taylor JF, Wiedmann RT, Schook LB, Groenen MAM (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS ONE 4:e6524

Reisch BI, Owens CL, Cousins PS (2012) Grapes. In: Badenes ML, Byrne DH (eds) Handbook of plant breeding: fruit breeding, vol 8. Springer, USA, pp 225–262. doi:10.1007/978-1-4419-0763-9_7

Riaz S, Krivanek AF, Xu K, Walker MA (2006) Refined mapping of the Pierce's disease resistance locus, PdR1, and sex on an extended genetic map of *Vitis rupestris* × *V. arizonica*. Theor Appl Genet 113:1317–1329. doi:10.1007/s00122-006-0385-0

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D, Schnable PS (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5:e1000734

Stranger BE, Stahl EA, Raj T (2011) Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187:367–383

Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, Kwiatkowski DP, Clark TG (2007) A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics 23:2741–2746

The French-Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Dematte L, Mraz A, Battilana J, Stormo K, Costa F, Tao Q, Si-Ammour A, Harkins T, Lackey A, Perbost C, Taillon B, Stella A, Solovyev V, Fawcett JA, Sterck L, Vandepoele K, Grando SM, Toppo S, Moser C, Lanchbury J, Bogden R, Skolnick M, Sgaramella V, Bhatnagar SK, Fontana P, Gutin A, Van de Peer Y, Salamini F, Viola R (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE 2:e1326

Verde I, Bassil N, Scalabrin S, Gilmore B, Lawley CT, Gasic K, Micheletti D, Rosyara UR, Cattonaro F, Vendramin E, Main D, Aramini V, Blas AL, Mockler TC, Bryant DW, Wilhelm L, Troggio M, Sosinski B, Aranzana MJ, Arús P, Iezzoni A, Morgante M, Peace C (2012) Development and evaluation of a 9 K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. PLoS ONE 7:e35668. doi:10.1371/journal.pone.0035668

Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat Commun 2:467