GENOMIC ANCESTRY ESTIMATION IN INTERSPECIFIC GRAPE HYBRIDS

by

Jason Sawler

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
March 2014

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The genus *Vitis* (the grapevine) is a group of highly diverse, diploid woody perennial vines consisting of approximately 60 species from across the northern hemisphere. To gain insights into the use of wild *Vitis* species during the past century of interspecific grape breeding and to provide a foundation for marker-assisted breeding programmes, we present a principal components analysis based ancestry estimation method to calculate admixture proportions of hybrid grapes in the United States Department of Agriculture grape germplasm collection using genome-wide polymorphism data. We find that grape breeders have backcrossed to both the domesticated *V. vinifera* and wild *Vitis* species and that reasonably accurate genome-wide ancestry estimation can be performed on interspecific *Vitis* hybrids using a panel of fewer than 50 ancestry informative genetic markers.

LIST OF ABBREVIATIONS AND SYMBOLS USED

AIM	ancestry informative marker
AOC	Appellation d'Origine Contrôlée
CI	confidence interval
DNA	deoxyribonucleic acid
DOC	Denominazione di Origine Controllata
F_{ST}	fixation index
GRIN	Germplasm Resources Information Network
LM	linear model
MAB	marker assisted backcrossing
MAF	minor allele frequency
MAS	marker assisted selection
PC	principal component
PCA	principal components analysis
QmP	Qualitätswein mit Prädikat
QTL	quantitative trait locus
SNP	single-nucleotide polymorphism
USDA	United States Department of Agriculture
VQA	Vintners Quality Alliance
V. vinifera	*Vitis vinifera*

ACKNOWLEDGEMENTS

First and foremost I would like to thank my supervisor, Dr. Sean Myles. I am incredibly fortunate to have such an enthusiastic, insightful and committed mentor. The opportunities and experiences I have had as a member of his research group, as well as his contributions to my academic development are beyond measure.

I would like to thank all my lab mates. Laura Butler's support of my work and assistance in navigating the administrative side of graduate studies (and being located off-campus) has been essential to finishing this thesis before the next ice age. Dr. Kyle Gardner's humor and willingness to give impromptu lessons on population and quantitative genetics whenever I have questions has been an invaluable contribution to my time in graduate school.

I would like to acknowledge my co-authors on the manuscript that formed the basis of this thesis, as well as the United States Department of Agriculture for providing the data and resources used for this project.

My family and friends – I give my thanks for all of their support and motivation. I appreciate their bearing with my lengthy rambling about wild and domesticated grapes whenever I've been asked about my work for the past two years.

Finally, I would like to thank my high-school biology teacher Jennifer Osmond. Without her passion and talent for teaching I would have never pursued genetics.

CHAPTER 1: INTRODUCTION

Reproduced from: Sawler J, Reisch B, Aradhya MK, Prins B, Zhong G-Y, et al. (2013) Genomics Assisted Ancestry Deconvolution in Grape. PLoS ONE 8(11): e80791. doi:10.1371/journal.pone.0080791 [1]

The genus *Vitis* (the grapevine) is a group of highly diverse, diploid woody perennial vines consisting of approximately 60 species from across the northern hemisphere [2]. According to the archaeological record, cultivation of the domesticated grapevine, *Vitis vinifera*, began 6000-8000 years ago in the Near East [3]. Today, the grape is the world's most valuable horticultural crop with ~8 million hectares planted, most of which is processed into wine (http://faostat.fao.org/). Grapes from the domesticated species, *V. vinifera*, account for more than 95% of the grapes grown worldwide [2] and the world's vineyards are dominated by a small number of closely related *V. vinifera* cultivars that have often been vegetatively propagated for centuries [4]. Because they are perpetually propagated, elite grape cultivars require increasingly intense chemical applications to combat evolving pathogen pressures. It is widely recognized that the exploitation of wild *Vitis* species' resistance to disease is crucial to the continued success and expansion of the grape and wine industries, and that the grape is well-poised to benefit from the use of marker-assisted breeding for this purpose [2,5,6].

In plant breeding, marker-assisted backcrossing can be used to incorporate traits into elite cultivars while minimizing the transfer of undesirable alleles from the donor genome [7]. This process involves both foreground and background selection. Foreground selection refers to the screening and selection of offspring based on the presence or absence of a specific allele that is associated with a trait of interest. In contrast, background selection is the selection of offspring on the basis of genomic ancestry estimates. A breeder may wish to introgress a specific trait from a wild species into an elite cultivar, while minimizing the genomic contribution from the wild species unrelated to that trait [5,6]. Recombinant selection (through backcrossing) aims to reduce the size of the chromosomal segment carrying the desired locus. Wild species often possess genes that negatively affect crop performance, making it advantageous to remove any additional background contribution from these wild species to the genomes of the resulting progeny [7]. While backcrossing in many crops is performed by crossing offspring back to one of the parents, "pseudo-backcrossing" is the method used to perform backcrosses in grapes. Pseudo-backcrossing involves crossing hybrid offspring back to a cultivated *V. vinifera* cultivar that is not one of the parents from the original cross. This form of backcrossing is performed because grapes suffer from severe inbreeding depression and thus crosses between closely related cultivars must be avoided [6]. Background selection relies on accurate estimation

of the percentages of the donor and recurrent parental genomes present in the resulting progeny.

To gain insights into the use of wild *Vitis* species during the past century of interspecific grape breeding and to provide a foundation for background selection in marker-assisted breeding programmes, we present a principal components analysis (PCA) based ancestry estimation method to calculate admixture proportions of hybrid grapes in the US Department of Agriculture (USDA) grape germplasm collection using genome-wide polymorphism data from the Vitis9kSNP microarray [4]. We find that grape breeders have backcrossed to both *V. vinifera* and wild *Vitis* species and that reasonably accurate genome-wide ancestry estimation can be performed on interspecific *Vitis* hybrids using a panel of fewer than 50 ancestry informative markers (AIMs). Our method of ancestry deconvolution provides a first step towards selection at the seed or seedling stage for desirable admixture profiles, which will facilitate marker-assisted breeding that aims to introgress traits from wild *Vitis* species while retaining the desirable characteristics of elite *V. vinifera* cultivars.

CHAPTER 2: LITERATURE REVIEW

2.1 Genetic Characteristics

The genus *Vitis* is divided in to two subgenera, *Vitis* (2n = 38 chromosomes) containing all but four taxa and *Muscadinia* (2n = 40 chromosomes) [8]. The grapevine genome size is approximately 475 Mb [9] and has been completely sequenced from Pinot Noir using Sanger shotgun and pyrosequencing [10,11]. Wild grapes are highly heterozygous [12] and interspecific breeding usually produces viable hybrid offspring [8,10]. Variation in recombination frequencies between species and within specific genome regions has been suggested in this genus [13]. One of several types of genetic markers commonly used to study the genus *Vitis* are single nucleotide polymorphisms (SNPs) [14]. Others include microsatellites, amplified fragment length polymorphisms (AFLPs), restriction fragment length polymorphisms (RFLPs), and sequence characterized amplified region (SCAR) markers [15,16]. A SNP is simply a single base pair position in a genome for which there is nucleotide variation (alternative alleles) between individuals [17]. While the exact distinction of SNPs from rare variants depends on the context of usage, a generally accepted definition is that a SNP's least abundant allele is found at a frequency of 1% or higher [17]. Approximately 2 million SNPs were identified across the grape

genome after its initial sequencing [10]. The development of the Vitis9K SNP microarray [14] in 2009 allowed for high-throughput genotyping of *Vitis* accessions in the USDA germplasm repository for cultivar identification and population level analyses [4].

Several marker-trait associations have been determined in grapes. Polymorphisms and transposable element insertions in the gene *VvmybA1*, which is responsible for the transcriptional regulation of anthocyanin biosynthesis, have been associated with variation of grape skin color [18]. Large-effect quantitative trait loci (QTL) have been identified for seedlessness, berry weight and leaf morphology via genetic mapping [19,20]. The loci *Rpv3*, *Rpv8*, *Rpv10* and *Rpv12* for downy mildew (*Plasmopara viticola*) resistance and SCAR markers linked to powdery mildew (*Uncinula necator*) resistance have been identified [21-24]. A single "gain of function" SNP at the *VvDXS* locus is responsible for the muscat flavor in grapevine, which is a trait of interest in winemaking [25]. Genetic control of plant sex in *Vitis*, described under domestication history below, has also been characterized.

2.2 Domestication History

Vitis is comprised of three geographical and evolutionary distinct groups: North American, Eurasian and Asiatic [26]. The common grapevine *Vitis vinifera*

was domesticated 6000-8000 years ago in the Near East from its wild progenitor *Vitis sylvestris* [4]. This aggregate of wild and feral forms of *Vitis* is widely distributed over southern Europe and Western Asia, and was previously thought to be an independent species from *V. vinifera* [26,27]. The wild progenitor and domesticated grapes are now botanically named *V. vinifera* subsp. *sylvestris* and *V. vinifera* subsp. *vinifera* respectively [28]. The frequency of SNPs in grapevines has been estimated at one SNP every 78 bp between species and every 119 bp in *V. vinifera* [29]. Genetic analyses suggests that the domestication of grape took place over relatively few generations, as there is limited reduction in haplotype diversity relative to wild types. This diversity has been maintained by clonal propagation of varieties with desirable traits [4]. One trait that was selected for during domestication was hermaphroditism (for the purpose of self-fertilization), which occurs rarely in natural populations [30]. Wild grape species are almost entirely dioecious, with sex being determined by 3 alleles of a single gene (Su^{+}, Su^{F}, Su^{m}) [31]. Female plants have a $Su^{m}Su^{m}$ homozygous recessive genotype at this locus, whereas males ($Su^{F}Su^{m}$) have a dominant Su^{F} allele that suppresses the formation of a pistil. The Su^{+} mutation causes pistil and anther formation in each flower, allowing self-fertilization in cultivars with the $Su^{+}Su^{+}$ and $Su^{+}Su^{m}$ genotypes. The dominance hierarchy for these three alleles is $Su^{F} > Su^{+} > Su^{m}$ [26-28,31].

Fruit quality has been greatly improved through domestication [18]. Grape clusters are large and elongated, and berries are larger and have higher sugar content for better fermentation relative to the wild forms [26]. Domesticated grapes have fewer and morphologically different seeds compared to wild species, which has been a key trait in archeological studies of *Vitis* [32].

2.3 Cultivation and Usage

Grapes are cultivated on every continent except Antarctica [33] as a result of their ability to grow in a wide range of environmental conditions [12]. They are a valuable commercial fruit crop with approximately 8 million hectares grown worldwide [4]. The main commercial product of cultivated grapes is wine, however they are also grown for fresh and dried fruit as well as the production of unfermented juice and concentrate [2]. A study of grape accessions in the USDA germplasm collection revealed that 74.8% of the 583 unique *V. vinifera* cultivars genotyped were related to at least one other cultivar by a first-degree relationship [4].

2.4 Breeding and Marker Assisted Selection (MAS)

Since domestication, nearly all cultivated grapevines have been clonally (or vegetatively) propagated from other cultivated vines. Propagation can be

achieved by either planting a cutting (scion) directly, or grafting it on to existing rootstock [27]. This produces genetically identical offspring, which allows for consistency in production and the maintenance of traditional cultivars such as Chardonnay and Pinot Noir. These varieties have been grown for wine production since the middle ages [32]. Frequent mutations in grapevine mean that new cultivars with morphological and agronomical differences can arise through vegetative propagation when genetic changes accumulate over many generations [32]. While this process can generate a small amount of diversity in cultivated grapevines, it does not compare to sexual reproduction (either through natural or artificial selection) in terms of its ability to drive evolution in plants [34]. *Vinifera* grapes are highly susceptible to many diseases, and a lack of genetic recombination in grapes is partly responsible for intense pathogen pressure and increasing chemical input by grape growers [4]. *Vinifera* grapes account for more than 95% of the worldwide market [2]. The breeding of new cultivars, and the introgression of traits from wild *Vitis* species that are unavailable within the *V. vinifera* gene pool is crucial to the continued success and expansion of the grape industry [4,6].

For most of their domesticated history grapes were relatively pest and disease resistant [35]. With the advent of large-scale globalization in the 19th century, grapevines in Europe began to face pressure from pathogens originating

from North America to which they had little or no resistance [36]. Powdery and downy mildew became more prevalent during this period, however the greatest example is the *Phylloxera* crisis which led to near extinction of the grapevine in Europe during the late 1800s [35]. This event greatly reduced the genetic diversity of *V. vinifera* in Europe [32]. *Phylloxera* is a louse that infects the roots of grapevines, eventually causing death to the entire plant. Various North American wild species are resistant to *Phylloxera* [37], and some of the first interspecific hybrids were bred using these species for the purpose of saving the European wine industry from this insect. Eventually these new hybrids were replaced by traditional varieties grafted on to rootstocks bred for *Phylloxera* resistance [12]. These were the first intentional interspecific *Vitis* hybrids, and marked the beginning of efforts to introgress genetically controlled traits from the diverse gene pool of wild species in to commercially grown grapes [35].

Marker Assisted Selection (MAS) is a modern tool used by breeders to screen for offspring with certain genotypes associated with desired traits. Tomato, maize, wheat, rice, barley and soybean are major crops for which MAS is routinely used to improve cultivars [38]. In table grapes, MAS has been used to select for seedlessness in progeny by detecting the presence of a 198-bp allele at the VMC7F2 marker [39]. There are many advantages of MAS over traditional plant breeding approaches. It can be performed on seedlings to reduce the time

and associated growing cost for genotyping progeny [40]. In certain crops with large seeds such as maize, protocols have been developed for genotyping endosperm prior to germination [41]. This permits high-throughput selection for advantageous genotypes without the need to grow plants and collect leaf tissue for DNA extraction. In traditional breeding programs the time between successive generations can be a costly and limiting factor. This is particularly true in long-lived perennial crops like grape for which fruit cannot be evaluated for 3-7 years after planting [5]. MAS can also be used to assess a plant's resistance to pathogens independent of environmental conditions, if markers are available for the resistance trait of interest [40]. This is important, as one of the main goals of many grape breeding programs is the improvement of fungal resistance [2]. Marker Assisted Selection does have limitations, such as the reliance on previously characterized marker-trait associations and the initial start-up costs for a genotyping program. Recombination may occur between the marker and desired gene and result in the retention of progeny that do not actually carry the trait of interest. Markers that have been developed for one population may not necessarily be transferrable to other populations [5,40].

MAS can be incorporated in to plant breeding programs in many different ways. Analyses typically used in population genetics can be applied to assess breeding material by identifying cultivars, measuring diversity and heterosis, and

identifying genomic regions under selection [7]. Marker-assisted pyramiding, or gene pyramiding, is the process of breeding multiple different alleles (from two or more sources) into a genome for the same purpose (e.g. resistance to a specific pathogen). The offspring populations are genotyped and screened for individuals carrying all desired markers [40]. Pathogens may be able to overcome one or more resistance genes; therefore pyramiding can provide broad-spectrum resistance when multiple QTLs have been identified [7]. The native Asian grape species *Vitis amurensis* possesses the *Rpv10* and *Rpv12* loci for downy mildew resistance, and has been used to introgress this resistance into *V. vinifera* cultivars in combination with the *Rpv3* locus originating from North American wild species [21,22,42].

Marker-assisted backcrossing (MAB) is used to incorporate traits into successful varieties while minimizing the probability of including undesirable alleles also present in the donor genome [7]. Foreground selection refers to the screening and selection of offspring based on the presence or absence of a specific allele. In contrast, background selection is the selection of offspring on the basis of genomic ancestry estimates [5]. A breeder may wish to introgress a specific trait from a wild species into an elite cultivar, while minimizing the genomic contribution unrelated to that trait [6]. Recombinant selection (through

backcrossing) aims to reduce the size of the chromosomal segment carrying the desired locus. Wild species or varieties often possess genes that negatively affect crop performance, making it advantageous to remove any additional background contribution to the genome of progeny [7]. This process is also known as "whole genome selection". Inbreeding depression in Vinifera grapes prohibits selfing, which is used to create inbred lines for other crops [43].

Background selection in MAS relies on accurate estimation of the percentages of the donor and recurrent parental genome in progeny. The number of ancestry informative markers (AIMs) required for accurate estimation is influenced by the level of genetic divergence between the two source populations [44].

2.5 Ancestry Estimation

An admixed population is defined as "a population formed recently from the mixing of two or more groups whose ancestors had long been separated" [45]. In human genetics this generally refers to structured populations such as African Americans in which relatively recent intercontinental gene flow makes it possible to calculate the proportion of African and European ancestry in an individual's genome [46]. Genome-wide studies of admixture have mainly focused on

identifying population structure in humans due to the low cost of genotyping individuals and the availability of polymorphism data from projects like HAPMAP [46,47]. Software developed for ancestry estimation has been used less commonly to study the degree of admixture in plants [48]. Interspecific grape hybrids are an admixed population and can potentially be characterized in terms of the wild and *V. vinifera* content in their genomes [49]. The following software packages are available for estimating genetic ancestry [50]: STRUCTURE, *frappe*, ADMIXTURE, EIGENSTRAT/smartpca, ipPCA/EigenDev, GEMTools, PLINK, LAMP, SABER, HAPMIX and ANCESTRYMAP.

Principal Components Analysis (PCA) is a multivariate statistical method that is widely used to quantify patterns of population structure in high-density genotype data [51]. PCA is described as follows by Reich, Price and Patterson [52]: "PCA is a statistical method for exploring and making sense of datasets with a large number of measurements (which can be thought of as dimensions) by reducing the dimensions to the few principal components (PCs) that explain the main patterns. Thus, the first PC is the mathematical combination of measurements that accounts for the largest amount of variability in the data." By performing PCA on genotype data using smartpca (part of the EIGENSTRAT package), it is possible to see individuals segregate in PC space based on species,

population, and geographical or evolutionary distance [53,54]. For situations of two-way admixture, PCA can be used to estimate the degree of contribution from each ancestral population [55]. For each principal component (or eigenvector) EIGENSTRAT calculates a weighting (or loading) for each marker used in the analysis. This weighting is a measure of how much a genetic marker contributes to a sample's position on a given axis of variation [47]. In this method, the weighting of a marker for PC1 can also be considered a measure of ancestry informativeness.

Genetic markers are not equal in their ability to infer ancestry. An ancestry informative marker (AIM) is most useful when it is fixed between the two ancestral populations. In other words, an allele should be present at a frequency of 1.0 in one ancestral population, and not present in the other [44]. For this reason, F_{ST} is often used to choose AIMs for admixture analysis [56]. Fisher Information Content (FIC), Shannon Information Content (SIC), Informativeness for Assignment Measure (I_n) and the Absolute Allele Frequency Differences (delta, δ) are also measures that can be used to evaluate the ancestry informativeness of genetic markers [44].

CHAPTER 3: MATERIALS AND METHODS

Reproduced from: Sawler J, Reisch B, Aradhya MK, Prins B, Zhong G-Y, et al. (2013) Genomics Assisted Ancestry Deconvolution in Grape. PLoS ONE 8(11): e80791. doi:10.1371/journal.pone.0080791 [1]

3.1 Sample Collection and Genotype Calling

Leaf tissue was collected from the USDA grape germplasm collections in Davis, California and Geneva, New York. Permission for tissue collection was obtained from the local USDA authority. DNA was extracted using commercial extraction kits. Genotype data were generated from the custom Illumina Vitis9KSNP array, which assays 8,898 single nucleotide polymorphisms (SNPs). After quality filters (GenTrain Score ≥ 0.3 and GenCall ≥ 0.2) 6114 SNPs in 1817 *Vitis* samples remained for analysis [4].

3.2 Data Curation

Samples with >10% missing data were removed, and SNPs with >10% missing data and minor allele frequency (MAF) <0.10 were removed using PLINK [57]. After these filters, 1599 samples and 2959 SNPs remained. PCA was performed on this data set using SMARTPCA [58] and 60 samples were removed due to mislabeling. For example, some samples labeled as *V. vinifera* clustered with wild species and some samples labeled as hybrids clustered with wild or *V. vinifera* (Figure S1). DNA sample mix-up is an unlikely explanation for these errors because sample processing was done primarily with robotics and no

genotype discordance for 145 pairwise comparisons between replicate samples placed randomly across sample plates was observed [4]. Thus, the cases of mislabeling are likely due to curation error. Our ancestry estimates of putatively mislabeled individuals are currently being verified by direct observation in the vineyard and the USDA Germplasm Resources Information Network (GRIN) online database [59] will be updated accordingly.

Our PCA plot of the full data set revealed a clear separation of North American wild species from *V. vinifera* along the first principal component (PC1; Figure S1). Eurasian wild species fell between these two groups. Although they are occasionally used in grape breeding, we excluded Eurasian wild species and hybrids with known Eurasian wild ancestry from the remaining analysis because the number of samples was low and their position in PC space complicates ancestry estimation. The present study thus focuses on hybrids with ancestry from North American wild *Vitis* species (hereafter referred to simply as wild *Vitis*) and *V. vinifera*.

3.3 Admixture Analyses

Principal components were computed using 333 wild *Vitis* samples and a random sample of 333 *V. vinifera* samples. Equal sample sizes (N = 333) for ancestral populations were selected as this has been shown to be a crucial factor in accurately inferring genetic relatedness based on PCA [55]. After establishing

the PC axes based on these ancestral populations, all 1599 samples were subsequently projected on to these axes, and individuals between -0.02 and 0.02 on PC1 (with the exception of Eurasian wild samples and hybrids with known Eurasian wild ancestry) were considered hybrids for the remainder of the analysis (N = 127 hybrids; Figure 1A). These conservative thresholds were chosen because the projection space between them included all samples labeled as "hybrid" in the USDA Germplasm Resources Information Network database [59].

Our method of calculating ancestry coefficients employs the approach described in [47] and [46], where admixture proportions are equal to the coordinate distance in PC space between the admixed individual (hybrid) and the two ancestral populations (*V. vinifera* and wild *Vitis*). For each purported hybrid grape, the genome-wide proportion of *V. vinifera* is estimated as P = b/(a+b), where b and a are the chord distances from the wild *Vitis* and *V. vinifera* centroids, respectively, for the given hybrid along PC1 (Figure 1B). We also estimated ancestry proportions using the model-based software STRUCTURE [60]. STRUCTURE was run with a burn-in period of 20,000 iterations followed by 100,000 iterations using the admixture model where each sample draws some fraction of its genome from each of the K populations where K = 2. As with previous work [51], our PCA-based ancestry estimates are highly similar to those generated from STRUCTURE (R^2 = 0.998; Figure S2).

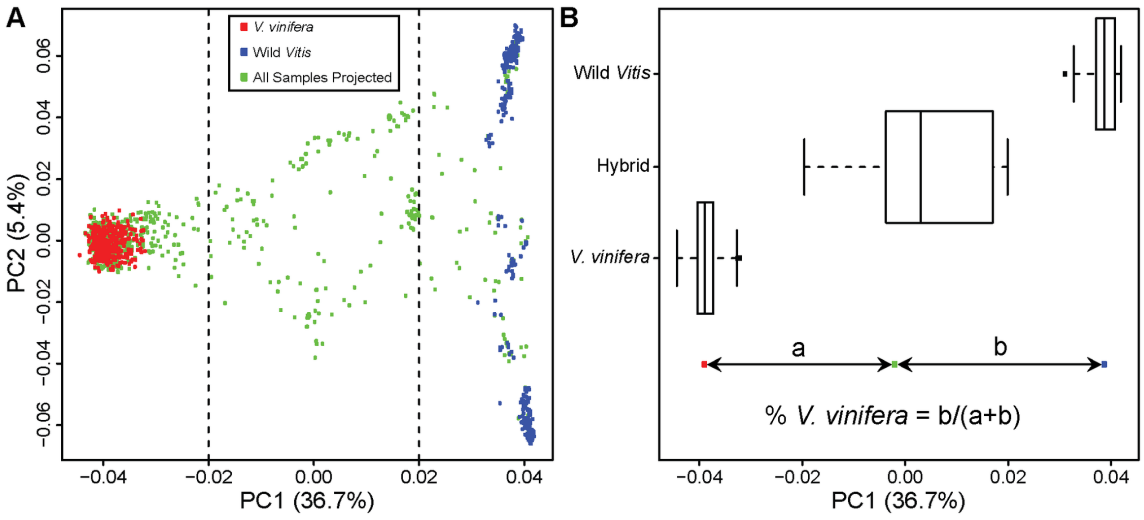

Figure 1. PCA based ancestry estimation. (A) PC axis 1 (PC1) and PC2 were calculated using 2959 SNPs from 333 *V. vinifera* and 333 wild *Vitis* samples. The proportion of the variance explained by each PC is shown in parentheses along each axis. Subsequently, 1599 samples, including various *Vitis* species and hybrids, were projected onto these axes (green dots). Samples lying between the dotted vertical lines were considered hybrids for the remaining analyses. (B) Boxplots show the range of PC1 values for the two ancestral populations (*V. vinifera* and wild *Vitis*) and the hybrids identified in (A). Boxes denote upper and lower quartiles and whiskers extend to 2.7 SD. Below the boxplots, an illustration of how ancestry proportions are calculated is provided (see Methods for details).
doi:10.1371/journal.pone.0080791.g001

3.4 Measures of Ancestry Informativeness

We ranked 2959 SNPs according to four measures of ancestry informativeness: PC1 Weight, PC1 Positive Weight, F_{ST} and a linear model described below. We evaluated the ability of reduced marker sets (1-100 SNPs) based on these measures to predict ancestry relative to the full set of 2959 SNPs. The "classification accuracy" of a set of AIMs is the R^2 value generated from a Pearson correlation between the hybrid ancestry estimates based on the reduced

marker set and the estimates based on the full set of 2959 SNPs. PC1 Weight is the absolute value of the PC1 loading given to each SNP by SMARTPCA, whereas PC1 Positive Weight excludes any values given a negative loading by the software. F_{ST}, a measure of allele frequency difference between the ancestral populations, was calculated according to [61] for each SNP using allele frequencies output by PLINK. For the Linear Model (LM) measure, linear regression was performed in R [62] using genome-wide *V. vinifera* content (estimated from all 2959 markers) as a response variable and genotypes for a given SNP across wild *Vitis*, *V. vinifera* and hybrid samples as an explanatory variable (0 = homozygous reference allele; 1 = heterozygous; 2 = homozygous non-reference allele). SNPs were ranked according to their R^2 from the linear model as a measure of ancestry informativeness.

3.5 Simulations of Admixture

To evaluate the accuracy of our PCA-based ancestry estimation method, *in silico* crosses between *V. vinifera* and wild species were simulated in R. Simulated F1 offspring were generated by randomly sampling one of the 333 *V. vinifera* and one of the 333 wild *Vitis* as parents. Parental genotypes were combined to produce offspring genotypes by sampling one allele at random from each parent at each SNP. Linkage disequilibrium between SNPs was ignored. This procedure was repeated 10,000 times to generate 10,000 F1 offspring. To

generate simulated F2 populations this process was repeated, using the F1 individuals as one ancestral population and either wild or *V.vinifera* accessions as the other to simulate backcrossing (n = 10000 for F1 backcrossed to wild, and n = 10000 for F1 backcrossed to *V. vinifera*).

CHAPTER 4: RESULTS AND DISCUSSION

Reproduced from: Sawler J, Reisch B, Aradhya MK, Prins B, Zhong G-Y, et al. (2013) Genomics Assisted Ancestry Deconvolution in Grape. PLoS ONE 8(11): e80791. doi:10.1371/journal.pone.0080791 [1]

4.1 PCA-Based Ancestry Estimation

PCA is a useful tool for revealing patterns of population structure and relatedness among samples for which genome-wide SNP data are available [54,63,64]. A genotyping microarray for the grape, the Vitis9KSNP array, was recently developed with probes designed for SNPs segregating within the domesticated species, *V. vinifera*, and a small number of probes designed to assay variation among *Vitis* species [14]. When PCA is applied to Vitis9KSNP array data from a diverse collection of *V. vinifera* cultivars, wild *Vitis* species and hybrid cultivars, PC1 clearly separates wild *Vitis* species from *V. vinifera* while hybrid cultivars lie between these two groups (Figure S1). This observation motivated us to apply methods developed previously [e.g. 46,47] to use a hybrid cultivar's projected position along PC1 to estimate the proportion of its ancestry derived from *V. vinifera* and wild *Vitis* species (Figure 1). Our PCA-based method provides highly similar ancestry estimates to those generated from the model-based approach in STRUCTURE (Figure S2).

For the present study, we considered a sample a "hybrid" if its projected position along PC1 was between -0.02 and 0.02. After removing obvious errors

(see Methods), all samples labeled as "hybrid" in the USDA Germplasm Resources Information Network (GRIN) online database [59] fall within this range, but many samples labeled as either wild *Vitis* species or *V. vinifera* fall within this range as well (Figure S1). Our ancestry estimates are therefore being used to improve the accuracy of the ancestry assignments associated with each accession in the USDA grape germplasm collection. We acknowledge, however, that samples outside of our defined "hybrid" range may in fact represent hybrid samples that resulted from extensive backcrossing to either *V. vinifera* or wild *Vitis* species. Further studies will be required to verify the ancestry of such hybrid samples and distinguish them unequivocally from the ancestral groups.

4.2 Verification of Ancestry Estimation Method

To verify the accuracy of our PCA-based ancestry estimation method, we simulated F1 (*V. vinifera* x wild *Vitis*) and F2 hybrids (F1 simulated hybrids backcrossed to *V. vinifera* or wild *Vitis*) using real genotype calls from the ancestral populations. The PCA plot of the simulated progeny and ancestral populations is shown in Figure 2A. The mean estimated genome-wide proportion of *V. vinifera* in the simulated F1 hybrids was 0.499, 95% CI [0.460, 0.537]. We expect the proportion of *V. vinifera* in these individuals to be 0.5, with the remainder of the genome (0.5) being contributed from the wild *Vitis* population. For offspring of the simulated F1 x *V. vinifera* cross we estimate the mean

genome-wide proportion of *V. vinifera* at 0.743, 95% CI [0.699, 0.782], with an expected value of 0.75. For offspring of the simulated F1 x wild *Vitis* cross, we estimate the mean genome-wide proportion of *V. vinifera* at 0.257, 95% CI [0.222, 0.294], with an expected value of 0.25. Distributions of the estimated *V. vinifera* genomic content of the three simulated crosses using PCA-based ancestry estimation are shown in Figure 2B. These results demonstrate that our PCA-based method provides reasonably accurate ancestry estimates for hybrid grape cultivars generated from a highly diverse collection of grape germplasm.

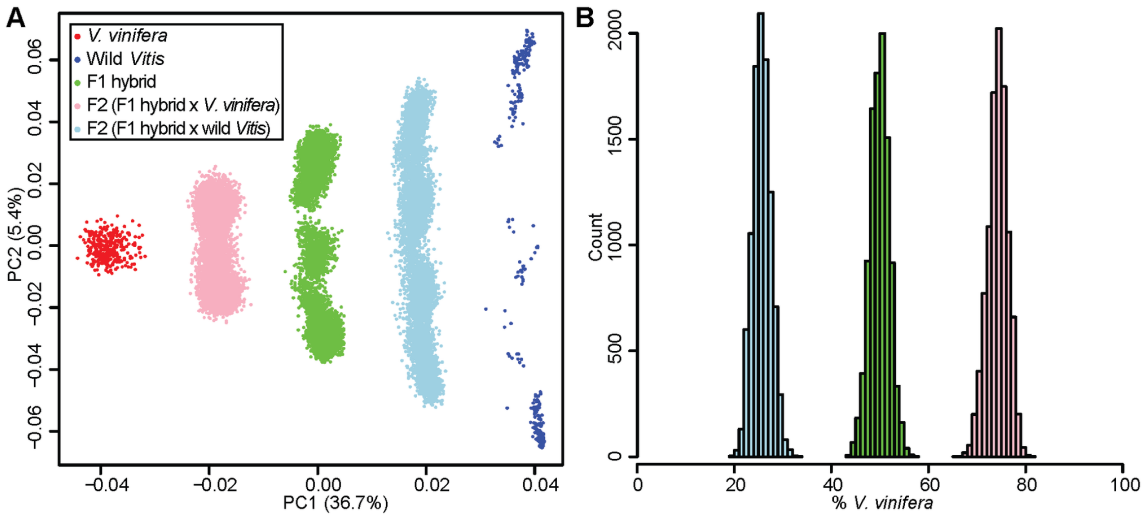

Figure 2. Verification of PCA-based ancestry estimates through simulation. (A) 10,000 F1 hybrids (green) were generated by simulating *V. vinifera* x wild *Vitis* crosses. Using the simulated genotype data, these hybrid samples were then projected onto the PC axes defined by the 333 *V. vinifera* (red) and 333 wild *Vitis* samples (blue) and the proportion of each F1 hybrid's ancestry derived from each ancestral population was estimated using our PCA-based approach. The same method was applied to F2 populations derived from backcrossing F1 hybrids to *V. vinifera* (pink) and backcrossing F1 hybrids to wild *Vitis* (light blue). (B) The distribution of *V. vinifera* ancestry proportions for the F1 and F2 populations. doi:10.1371/journal.pone.0080791.g002

4.3 Grape Ancestry Estimation

Genome-wide *V. vinifera* content was estimated for the 127 samples identified as being hybrids with *V. vinifera* and wild *Vitis* ancestry (Figure 3; Table S1). The mean proportion of *V. vinifera* content for these hybrids is 0.474 (Min: 0.2506, Max: 0.7508). The range of observed admixture estimates in hybrid grapes suggests that backcrossing to both wild *Vitis* and *V. vinifera* has occurred in the past century of interspecific grape breeding. The relatively large number of samples with approximately 50% *V. vinifera* contribution to their genome suggests the existence of many first-generation interspecific hybrids in the USDA grape collection. An F1 hybrid included in this analysis, Baco Noir, is a cross between Folle Blanche (*V. vinifera*) and an accession of *Vitis riparia* (North American wild *Vitis* species). Based on its pedigree, we expect ancestry proportions of 50% wild and 50% *V. vinifera*, and our method using the full 2959 SNPs provides estimates of 49% and 51%, respectively. The cultivar Alicante Ganzin is the result of a cross between Alicante Bouschet (*V. vinifera*) and Ganzin No. 4. Ganzin No. 4 is an F1 hybrid between *V. rupestris* (North American wild *Vitis*) and Aramon Noir (*V. vinifera*). This pseudo-backcross pedigree suggests the genome-wide *V. vinifera* content for Alicante Ganzin will be 75% on average. Our method provides an estimate of 75.1% *V. vinifera* for this sample (Table S1).

According to the distribution in Figure 3, it appears that, unlike breeders of many other crops, grape breeders have not explicitly aimed to introgress specific genetic loci from wild *Vitis* species by repeatedly backcrossing to the domesticated species, *V. vinifera*. In fact, the large number of cultivars with low % *V. vinifera* ancestry suggests that backcrossing to wild *Vitis* may have been more frequent than backcrossing to *V. vinifera*. However, hybrids with *V. vinifera* content outside a particular range are not included in this analysis due to thresholds established in PC space for hybrid classification (see Methods). If breeders have historically aimed to minimize wild *Vitis* content in hybrid grapes by backcrossing extensively to *V. vinifera*, it is possible that commercially successful hybrid cultivars fall outside our established thresholds and thus may be underrepresented here. In addition, the sample of hybrids from the USDA collection may not be representative of interspecific grape breeding in general. Thus, ancestry estimation of a large sample of hybrid grape cultivars from breeding programmes worldwide is currently underway to verify this claim.

When PCA is applied to the genotypes generated from the Vitis9KSNP array, there is a clear distinction between *V. vinifera* and wild North American *Vitis* species along the first principal component (Figure 1A, Figure S1). The genotype data are also sufficient to enable the various wild species to be distinguished from each other. For example, wild *Vitis* samples clearly cluster by species along PC2

(Figure S1). This suggests that our present method could be extended to identify the precise wild *Vitis* species that has contributed to a hybrid's ancestry: a hybrid's position on PC2 is likely an indicator of the wild *Vitis* species that has contributed to its ancestry. However, many hybrid grape cultivars have complex pedigrees with genetic contributions from multiple wild *Vitis* species. For example, the hybrid Brianna derives its ancestry from seven different wild *Vitis* species and *V. vinifera* [65].

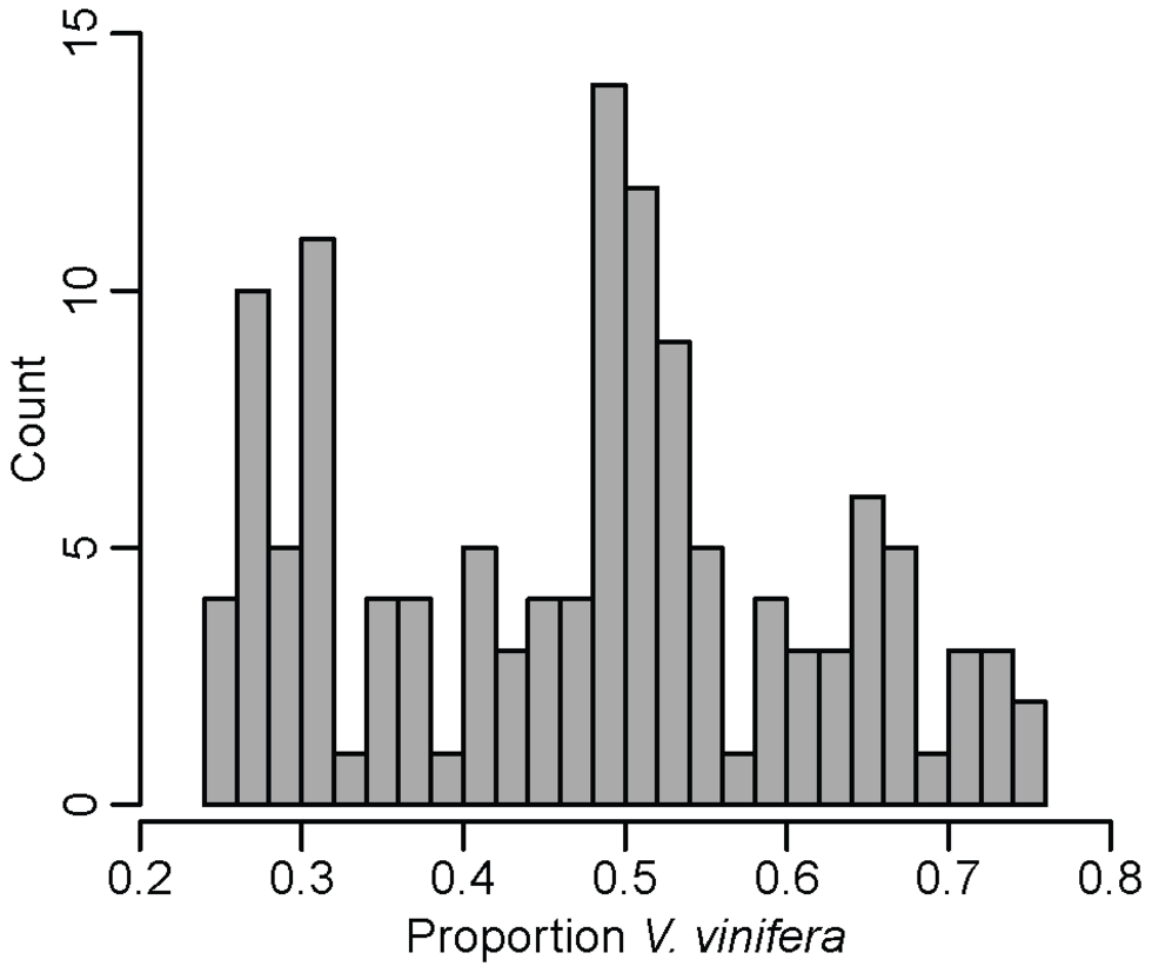

Figure 3. Estimated *V. vinifera* content in grape hybrids. The distribution of *V. vinifera* ancestry proportions in 127 hybrids from the USDA germplasm repository. Estimates are based on the full set of 2959 SNPs. A table of cultivar names, information and proportion *V. vinifera* ancestry is provided in Table S1. doi:10.1371/journal.pone.0080791.g003

Extensions of the present method beyond a simple 2-way admixture model and higher density genotype data will be required to generate accurate estimates of the genetic contributions of each individual wild *Vitis* species in complex hybrids. A high-density set of SNPs for this purpose could be generated using a genotyping by sequencing (GBS) approach [e.g. 66].

4.4 Selection of Ancestry Informative Markers (AIMs)

To enable ancestry estimation in grapes not included in this study, we investigated several methods for identifying a small number of SNPs, or ancestry informative markers (AIMs), that most effectively capture the ancestry information contained within the full set of 2959 SNPs. In admixed populations, an ideal AIM should have alleles that are fixed between the two ancestral populations and thus have an $F_{ST} = 1.0$ [44]. In addition, PCA generates weights for each SNP indicating the degree to which a SNP contributes to each PC. SNPs with extreme PC1 weights differentiate *V. vinifera* from wild *Vitis* along PC1 and are thus also good candidate AIMs [47]. We find that Fst values and PC1 weights are highly correlated ($R^2 = 0.979$; Figure S3) and that both metrics

are useful for the selection of AIMs (Figure 4A). This relationship between Fst and the first principal component has been previously described in [55].

We reasoned that the effectiveness of an AIM should not only depend on its frequency difference between the ancestral populations, but also on the extent to which the segregation pattern of its alleles in the hybrid population correlates with the ancestry of the hybrids. Thus, we developed a linear model of informativeness (LM; see Methods) and found that it outperformed both F_{st} and PC1 weights when fewer than 20 SNPs are used, but failed to improve when additional SNPs were added (Figure 4A).

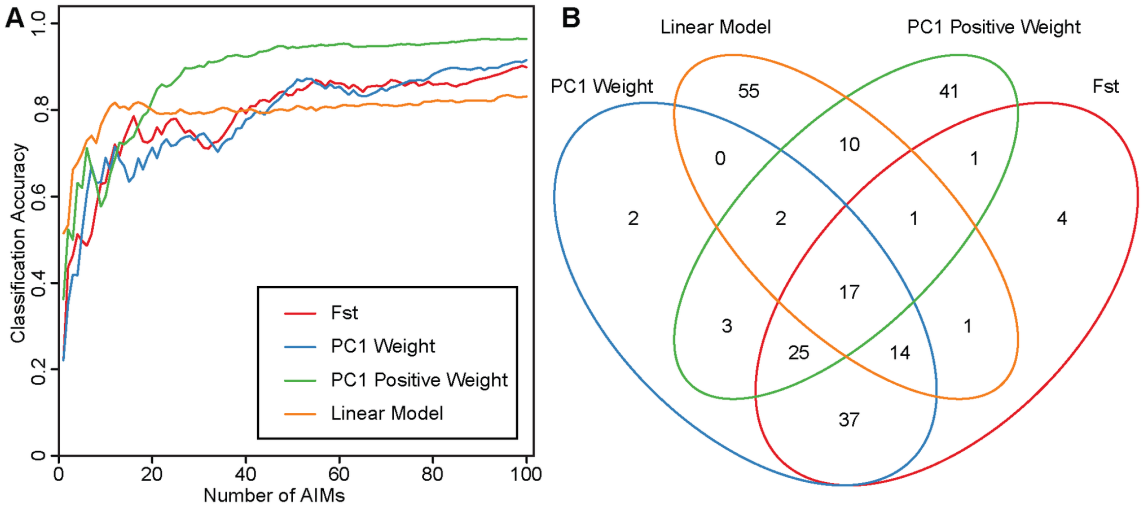

Figure 4. Comparison of measures of informativeness for identifying AIMs. (A) Each measure used to rank AIMs is shown in the legend. For each measure, the proportion of *V. vinifera* ancestry across the 127 hybrids was estimated using 1-100 SNPs ranked according to that measure and the result was compared to the proportion of *V. vinifera* ancestry estimated from the full set of 2959 SNPs. The classification accuracy (Y axis) is the squared Pearson correlation coefficient (R2) between the estimate derived from the reduced set of AIMs and the estimate from the full set of SNPs. (B) A Venn diagram showing

the overlap of the top 100 AIMs identified using each of the measures.
doi:10.1371/journal.pone.0080791.g004

The PC1 weight of a SNP can be interpreted as a scaled regression coefficient that takes on negative values when PC1 values are negatively correlated with genotypes and positive values when this correlation is positive. The sign of the weight depends on how the genotypes are encoded in the input file. For the present study, genotypes were encoded as follows: 0 = homozygous reference allele; 1 = heterozygous; 2 = homozygous alternative allele. Because the grape reference genome is *V. vinifera* [11], reference alleles are more frequent in the *V. vinifera* samples which occupy the lower values along PC1 (Figure 1). This results in a PC1 weight distribution that is highly skewed towards positive values (Figure S4). Thus, SNPs with extreme negative PC1 weights, where *V. vinifera* are homozygous for the alternative allele and wild *Vitis* species are homozygous for the reference allele, are rare. We reasoned that most of the useful ancestry information would therefore be contained within the positive PC1 weights. We therefore not only tested the ancestry informativeness of markers based on the absolute value of the PC1 weights as is normally done [47,58,67], but also ranked SNPs by their positive PC1 weights only. We found that ignoring the negative weights and only considering the positive weights significantly reduced the number of AIMs required to accurately infer ancestry

(Figure 4A). This observation should serve as a cautionary note to future uses of PC1 weights for the purposes of AIM identification.

The physical coordinates of the AIMs identified in the present study can be found in Table S2. Each of the four measures we used to rank SNPs by ancestry informativeness resulted in a different set of AIMs. The overlap in the top 100 AIMs identified by each measure is shown as a Venn diagram in Figure 4B. Within the four sets of 100 markers, 55 and 41 SNPs were unique to LM and PC1 Positive Weight, respectively. The PC1 Weight and F_{ST} panels had 93 SNPs in common. Thus, the selection of AIMs on the basis of PC1 weight and F_{ST} produce highly similar marker panels, however additional informative markers are overlooked if other measures are not taken into consideration. Although each measure is useful in identifying a set of AIMs, there is clearly a need for a method that can conclusively identify the optimal set of AIMs that maximizes ancestry informativeness.

CHAPTER 5: SUPPORTING INFORMATION

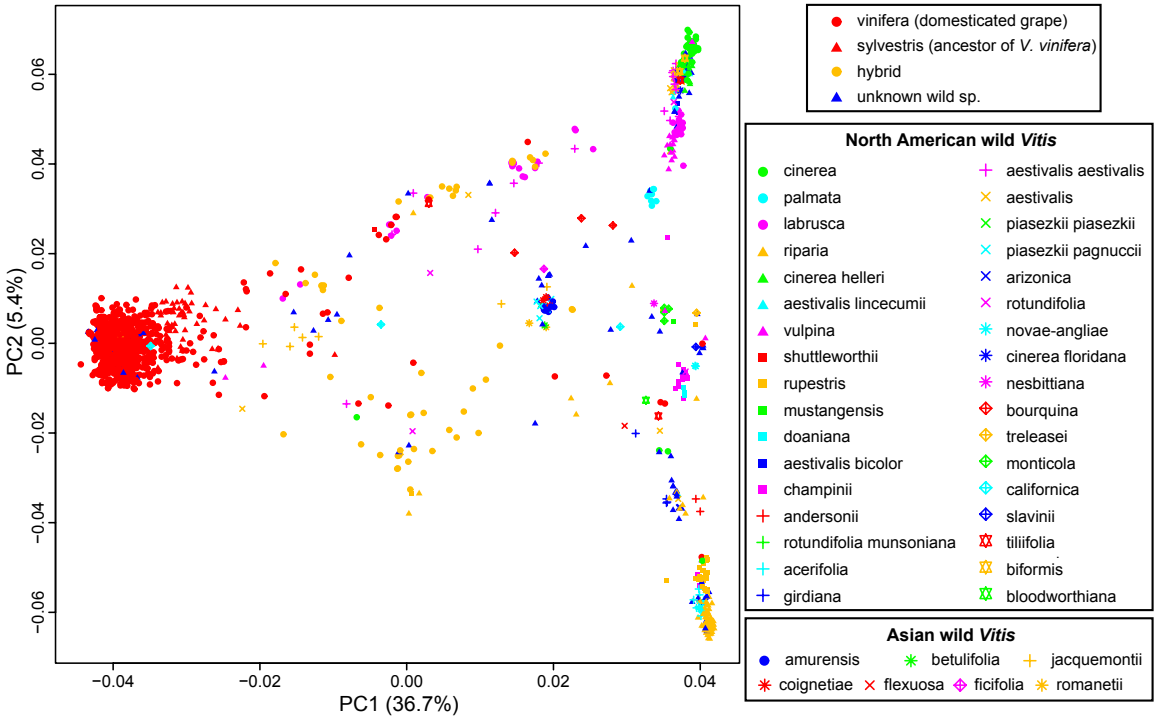

Figure S1. PCA of 1599 samples from USDA grape germplasm collection. (A) PC axis 1 (PC1) and PC2 were calculated using 2959 SNPs from 333 *V. vinifera* and 333 wild *Vitis* samples. The proportion of the variance explained by each PC is shown in parentheses along each axis. Subsequently, 1599 samples, including various *Vitis* species and hybrids, were projected onto these axes. This is the same plot as Figure 1 in the main manuscript, but each sample is labeled with the species identifier associated with that sample. Species identifiers were obtained from the Germplasm Resources Information Network (GRIN) database managed by the USDA. It is evident that many samples are mislabeled. For example, some samples labeled as *V. vinifera* clearly cluster far to the right of PC1 with the wild species. In cases where there was an obvious error and it interfered with downstream analyses, the samples were removed from analysis (N = 60). Eurasian wild *Vitis* samples and hybrids with known ancestry from Eurasian wild species were removed from the analysis. See Materials and Methods on how we defined "hybrid" for the present study.

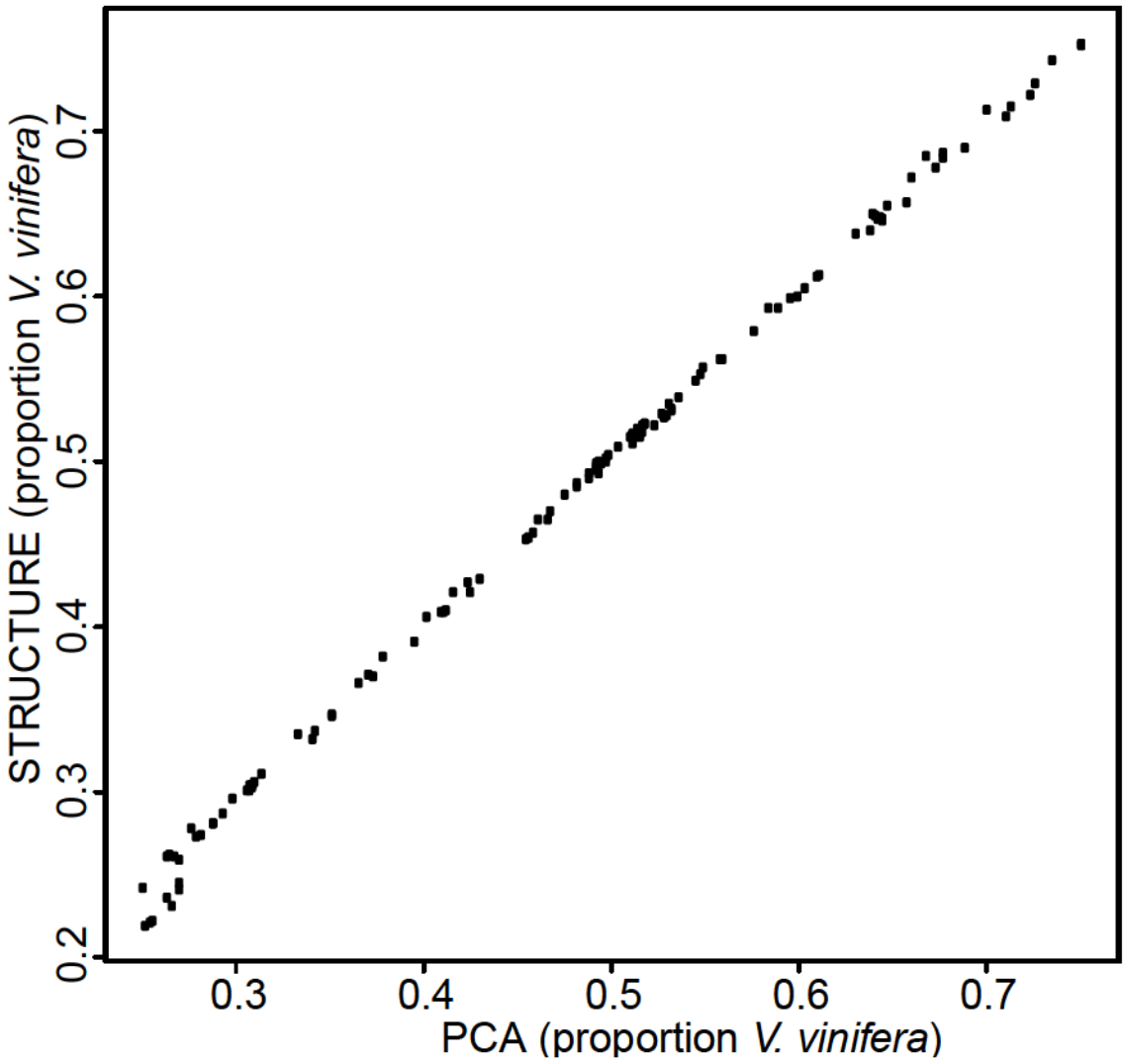

Figure S2. A comparison of ancestry estimates derived from our PCA-based method and the model-based method STRUCTURE. The proportion *V. vinifera* ancestry estimated using the PCA-based method and the program STRUCTURE are shown on the X and Y axes, respectively, for the 127 hybrid samples analyzed in the present study.

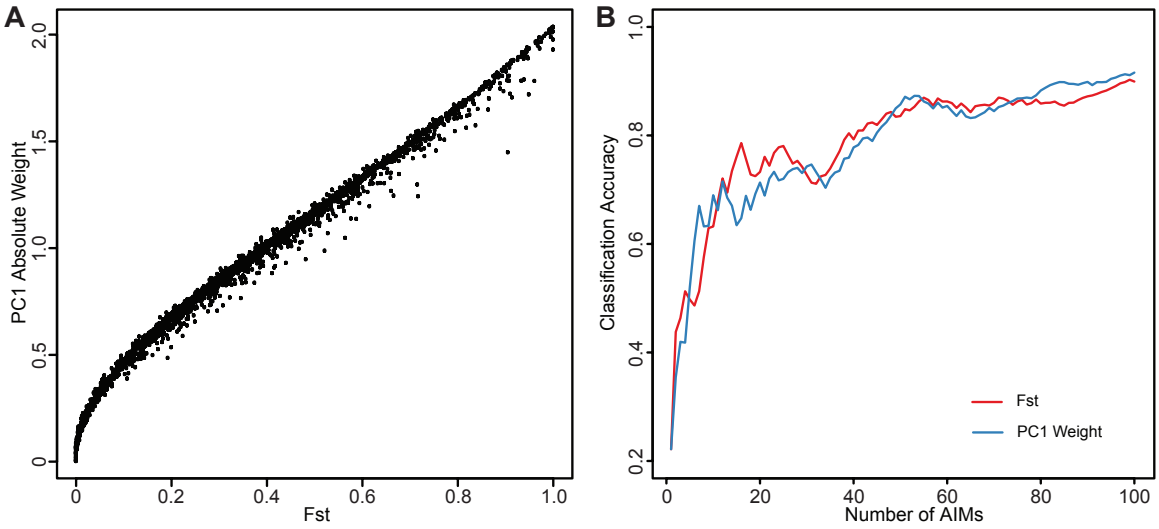

Figure S3. (A) FST and the absolute value of PC1 weights are highly correlated. (B) The classification accuracy of AIMs ranked by FST and PC1 absolute weight are highly similar.

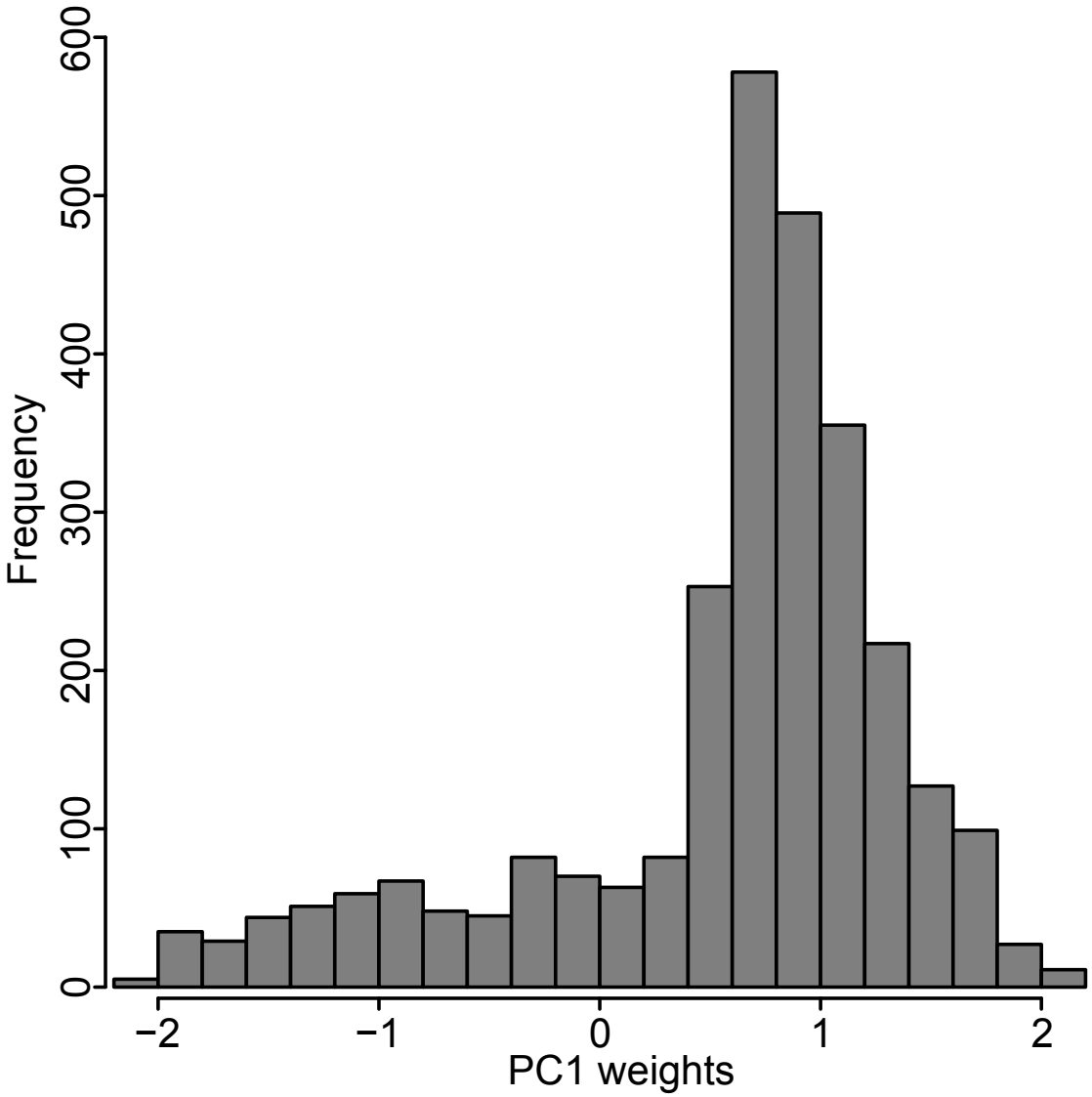

Figure S4. The distribution of PC1 weights from running SMARTPCA on 333 *V. vinifera* and 333 wild *Vitis* samples. The distribution is skewed towards positive values. doi:10.1371/journal.pone.0080791.s004

Table S1. A list of the 127 accessions from the USDA grape germplasm collection considered "hybrid" in the current study based on their positions along PC1 and associated cultivar name for these accessions from the USDA Germplasm Resources Information Network (GRIN) database.
doi:10.1371/journal.pone.0080791.s005

Cultivar Name	USDA ID	Proportion vinifera
(V56-33)	DVIT_1848	0.49
035-23	DVIT_1810	0.71
333_Em_4x_(foex_4x)	DVIT_1676	0.50
Agwam	DVIT_2630	0.53
Alicante_Ganzin	DVIT_3155	0.75
AN5-74	DVIT_2591	0.70
b_41-47_M31	DVIT_1875	0.60
Bertille-seyve_5563	DVIT_169	0.33
Big_Brown_F560	DVIT_2580	0.35
Black_Damascus	DVIT_355	0.29
Burgaw	DVIT_2628	0.50
Chatauqua	DVIT_31	0.31
Choultu_White	DVIT_2909	0.64
CN1-90	DVIT_2572	0.61
Concord_Sport	DVIT_40	0.31
Corbeau	DVIT_696	0.55
Criolla_Negra	DVIT_1091	0.52
Cynthiana	DVIT_43	0.28
Everglade__F272	DVIT_2581	0.34
Hayes	DVIT_66	0.30
Hicks	DVIT_69	0.31
Honey_Dew	DVIT_71	0.29
Hybrid_Pardes	DVIT_1990	0.52
i_8-13_M31	DVIT_1947	0.28
I-167-034	DVIT_2554	0.27
I-167-048	DVIT_2559	0.26
Jackson_Sel._number3	DVIT_2916	0.46
Joannes-seyve_23-416	DVIT_197	0.48
Khawngi	DVIT_2919	0.53
King	DVIT_81	0.31
Kuhlmann_187-1	DVIT_213	0.51

Kuhlmann_188-2	DVIT_214	0.55
Kuhlmann_191-1	DVIT_215	0.52
Lady	DVIT_87	0.52
Louisiana	DVIT_94	0.31
Marengo_(P._205)	DVIT_455	0.53
Mathiasz_Y_Ne_number40	DVIT_457	0.49
Mcpike	DVIT_102	0.31
Mureto	DVIT_854	0.69
NA	PI_597295.02	0.25
Himrod	PI_588095.02	0.64
Bertille Seyve 2862	PI_597226.02	0.42
Bertille Seyve 2667	PI_588244.02	0.42
Bertille Seyve 3408	PI_597146.02	0.48
Couderc 299-35	PI_588241.02	0.71
NA	PI_588247.02	0.40
Alexander	PI_594349.02	0.53
Shimek	PI_588675.02	0.26
NA	PI_588185.02	0.75
Ontario	PI_588074.04	0.41
Vignoles (Ravat 51)	PI_181481.02	0.49
Lucile	PI_588283.02	0.28
Wheeler	PI_597172.02	0.31
NA	PI_588436.02	0.51
Melody	PI_597229.02	0.60
Pinard (Kuhlmann 191-1)	PI_215419.02	0.52
Marechal Joffre (Kuhlmann 187-1)	PI_588254.02	0.51
Seyve Villard 5-276	PI_588309.02	0.38
Sugar Plum	PI_597228.02	0.46
Seyve-Villard 5-267	PI_597244.02	0.46
Glenfeld	PI_597203.02	0.31
Golden Muscat	PI_588111.04	0.66
Diamond	PI_588120.02	0.41
Fredonia	PI_597098.04	0.26
Diamond 4X	GVIT_1591.02	0.41
Bertille Seyve 2758	PI_279505.02	0.37
Concord	PI_588077.15	0.31
Suffolk Red	PI_597242.04	0.64

NA	PI_597298.02	0.66
NA	PI_597293.02	0.25
Suffolk Red	PI_597242.06	0.64
Chambourcin (Johannes Seyve 26-205)	PI_588075.02	0.56
Vidal Blanc (Vidal 256)	PI_200684.02	0.63
Schuyler	PI_588099.09	0.72
NA	PI_597294.02	0.27
NA	PI_237621.02	0.49
Leon Millot (Kuhlmann 194-2)	PI_588112.02	0.51
Niagara Seedless	PI_588151.02	0.46
Concord	PI_588077.17	0.31
Niagara	PI_588106.04	0.45
NA	PI_597292.02	0.26
Albany Surprise 4X	GVIT_1587.02	0.53
Cayuga White (GW 3)	PI_588079.02	0.54
NA	DVIT_2180	0.27
Athens	PI_588158.02	0.43
Alden	PI_588102.02	0.67
Chancellor (Seibel 7053)	PI_588072.02	0.37
Aramon Rupestris Ganzin 1	PI_588092.02	0.49
Winchell	PI_588130.02	0.43
Baco 37-16	PI_188588.03	0.58
Gladwin 113	PI_588169.02	0.73
Catawba	PI_588070.02	0.51
Marechal Foch (Kuhlmann 188-2)	PI_588107.02	0.50
Baco Noir (Baco 1)	PI_594334.02	0.49
Concord Seedless	PI_588101.02	0.25
NA	DVIT_2099	0.60
Vergennes	PI_588128.02	0.52
number2	DVIT_1361	0.55
number580	DVIT_1411	0.48
Olmo_(035-64)	DVIT_1816	0.67
Olmo_(U67-64)	DVIT_1704	0.53
Olmo_(U68-2)	DVIT_1705	0.49
Olmo_(U68-53)	DVIT_1711	0.27
Olmo_(U69-11)	DVIT_1716	0.47
Olmo_(U69-30)	DVIT_1718	0.31

Olmo_(U69-37)	DVIT_1717	0.37
Olmo_(U69-50)	DVIT_1719	0.56
Olmo_H24-36_Or_37	DVIT_1349	0.50
Ozark	DVIT_119	0.40
Perbos_205	DVIT_2144	0.58
Pocklington	DVIT_124	0.68
Pulliat	DVIT_128	0.49
Reflex_(RF5)	DVIT_2704	0.68
Rofar_Vidor	DVIT_2258	0.52
Rx_41-1	DVIT_1731	0.47
Siewiernyi	DVIT_2686	0.64
Skiathopoulo	DVIT_957	0.74
Suffolk_Red	DVIT_1315	0.64
Suffolk_Red	DVIT_1128	0.64
Thelma	DVIT_168	0.53
Triumphant	DVIT_2720	0.34
Unknown_vinifera_Cultivar	DVIT_2564	0.59
Van_Buren	DVIT_1129	0.27
Venus_(seedless)	DVIT_1130	0.61
Vignoles	DVIT_2741	0.49
Woodsprite	DVIT_2578	0.35
Wyoming	DVIT_162	0.29

Table S2. A list of the AIMs identified in the present study. The top 100 AIMs are listed for each of the four measures used in the present study. The AIMs are ranked according to the measure listed at the top of each column. The name of each SNP contains the physical coordinates of the SNP according to the 8x Pinot Noir reference genome, where the chromosome name is separated by the physical position by a colon. Chromosome numbers outside of the range of 1-19 refer to the unanchored contigs found in the 8x Pinot Noir reference genome.
doi:10.1371/journal.pone.0080791.s006

Rank	FST	PC1 Absolute Weight	PC1 Positive Weight	Linear Model
1	2:141291	2:141291	15:7577453	4:1783623
2	5:21669686	6:3792284	1:13794046	1:2619751
3	6:3792284	7:4934118	1:2622531	3858:755890
4	7:4934118	14:19163590	7:13094346	4:2322806
5	12:9066350	15:7577453	13:10692443	15:6401787
6	14:19163590	16:6890855	18:11239810	1:366545
7	17:5140114	1:13794046	12:9066350	4:3763620
8	15:7577453	3860:2941213	7:2151751	4:2322779
9	16:6890855	1:2622531	13:10692442	7:13094346
10	7:2151751	7:13094346	18:11788433	12:1363731
11	7:13094346	13:10692443	1:7802183	11:2312261
12	1:13794046	18:11239810	8:18515027	5:20822461
13	13:10692443	12:9066350	12:1363731	5:21669686
14	18:11239810	7:2151751	5:133260	10:643669
15	3858:755890	13:10692442	19:1245768	13:1778602
16	8:18515027	18:11788433	7:1898848	17:9185022
17	3860:2941213	1:7802183	6:664571	13:2653439
18	13:10692442	3:5354193	17:5140114	4:5205328
19	1:2622531	3858:755890	17:3740979	4:5205336
20	18:11788433	8:18515027	4:2322806	8:19931676
21	1:7802183	1:14526073	11:3842444	7:13984078
22	3:5354193	12:1363731	2:5546719	8:19164005
23	12:1363731	5:21669686	5:20111846	19:48390
24	14:561355	3864:1766072	5:21728007	5:17037585
25	19:1245768	14:561355	14:11321597	15:6181731
26	3:3772919	5:133260	14:717049	9:678759
27	1:14526073	19:1245768	4:13333712	5:4661133

28	14:561356	14:561356	1:7667975	8:18515027
29	4:19278660	4:19278657	6:5350487	15:7577453
30	3864:1766072	7:1898848	17:13881	1:6577130
31	4:19278657	17:7429913	9:10425035	1:4402635
32	17:7429913	4:19278660	7:7180291	9:3394641
33	7:1898848	8:10679565	17:4478453	12:6506582
34	5:133260	6:664571	2:1059092	12:2314340
35	6:16862053	3:3772919	17:9087066	12:2314313
36	6:664571	17:5140114	4:19226186	5:20806305
37	17:13992	17:3740979	4:3763620	5:21728007
38	5:22798649	4:2322806	3850:2913814	1:2619760
39	5:21728007	18:1226553	6:16672304	1:6609058
40	8:10679565	11:3842444	6:16862053	19:2460096
41	4:2322806	5:3864908	8:13700787	7:1898848
42	17:3740979	5:3864833	3:1434899	5:21002705
43	7:4627189	2:5546719	13:2653439	9:15546926
44	18:1226553	5:20111846	7:10693037	1:960128
45	6:16672304	7:4627189	13:1551073	5:3864833
46	3861:2904660	4:1783623	15:6181731	6:214423
47	2:5546719	5:21728007	14:18790751	6:5350487
48	5:3864908	14:11321597	12:18461120	6:664571
49	5:20111846	5:22798649	7:1667778	5:3864908
50	5:3864833	3861:2904660	3:1409233	7:3124287
51	11:3842444	14:717049	17:6018575	16:6890855
52	11:2170516	1:14526353	5:4829662	11:954150
53	4:1783623	6:20768367	9:2617627	6:804458
54	14:11321597	6:7391824	18:152927	18:12623855
55	14:717049	19:13205401	18:1280003	6:2797680
56	6:20768367	16:108498	1:366545	19:1245768
57	19:13205401	6:20768355	5:17037585	18:7801150
58	1:14526353	3:1013821	12:17973652	4:7094602
59	17:6149926	17:6645289	7:3124287	5:3404641
60	6:7391824	17:7157963	19:6602925	9:4807390
61	16:108498	6:3675021	2:17384063	13:5989356
62	17:6645289	4:1228865	14:9871232	12:2314262
63	3:1013821	4:13333712	17:1373655	19:305125
64	6:20768355	5:6157321	9:6489555	15:5865310

65	6:15916489	1:7667975	19:305125	9:6489555
66	4:13333712	9:4934873	3:3436524	6:16672304
67	17:7157963	6:5350487	8:8693377	1:4154869
68	2:2336810	17:13881	14:12442991	13:6989040
69	3862:1396759	7:7180291	5:20822461	11:954243
70	10:643669	9:10425035	3:1521643	13:10692443
71	18:1226162	3859:3075643	5:6678558	5:18002243
72	4:11618949	6:6045443	2:2057306	16:7228842
73	1:4154869	2:2391182	6:804458	14:717049
74	6:3675021	17:4478453	7:13901320	1:15368583
75	3859:3075643	4:2322779	2:2057255	17:8811102
76	6:5350487	2:1059092	1:8974759	5:4466326
77	5:6157321	3862:1396759	13:15098620	9:3252966
78	17:13881	17:9087066	11:7424642	1:7802183
79	7:7180291	18:1226162	14:1571132	4:1185523
80	4:1228865	15:5082017	8:18221627	18:596586
81	9:4934873	4:19226186	8:18221623	13:359456
82	15:6401787	5:20806305	2:3655671	1:2622531
83	17:9087066	3850:2913814	4:15175943	6:2794950
84	9:10425035	4:3763620	16:7157916	18:148579
85	1:7667975	6:16672304	8:15537913	13:184462
86	1:366545	1:4154869	11:630969	19:6602925
87	2:1059092	11:2170516	6:214423	15:499416
88	13:1551073	2:2336810	2:4412982	13:10692442
89	3850:2913814	10:643669	5:6685568	12:6009737
90	17:4478453	6:16862053	4:1185523	3864:1766072
91	6:6045443	5:12161644	7:7179779	1:1709404
92	2:2391182	14:4173981	5:299722	7:13518424
93	4:2322779	8:13700787	17:6230861	18:9540738
94	4:3763620	3:1434899	18:3490918	3:1013821
95	4:19226186	13:2653439	5:5423758	6:18163408
96	5:20806305	7:10693037	2:5854759	13:7156955
97	14:4173981	13:1551073	8:15853918	4:1228865
98	7:10693037	15:6181731	18:8762900	6:4228995
99	15:5082017	18:3611276	6:18163408	11:4774143
100	7:1667778	14:18790751	3:5564543	18:1226162

CHAPTER 6: CONCLUSIONS

Reproduced from: Sawler J, Reisch B, Aradhya MK, Prins B, Zhong G-Y, et al. (2013) Genomics Assisted Ancestry Deconvolution in Grape. PLoS ONE 8(11): e80791. doi:10.1371/journal.pone.0080791 [1]

Over the past century, grape breeders have generated interspecific hybrid grapes by crossing cultivars of the cultivated *V. vinifera* species with numerous wild *Vitis* species. Our PCA-based ancestry estimates of 127 hybrid cultivars indicate that F1 hybrids (*V. vinifera* x wild *Vitis*) are common and that backcrossing to wild *Vitis* was equally or even more frequent than backcrossing to *V. vinifera.* However, estimates from a more reliable and representative sample of hybrids are required to verify this claim. Our method provides a framework for enabling marker-assisted breeding of seedling populations based on ancestry estimates, but the application of such background selection in bi-parental populations will require higher marker densities than those provided by the Vitis9KSNP array.

We identify sets of AIMs and demonstrate that genotypes from only ~50 SNPs are sufficient to accurately estimate the proportion of ancestry a hybrid grape derives from *V. vinifera* and wild *Vitis* species. Not only can the AIMs identified here be employed to curate germplasm collections, but they can also be used for forensic purposes. Regulatory and appellation systems around the world like the AOC (France), DOC (Italy), QmP (Germany) and VQA (Canada) exist

to verify and guarantee the authenticity of the origin of their wines. Often, these systems only approve the use of cultivars with 100% *V. vinifera* ancestry, yet the ancestry inferences they employ are often based on questionable morphological analyses, error-prone breeding records or pure conjecture. Although it is widely recognized by the scientific community that the restriction of cultivar use by these organizations poses a serious threat to the future of the wine and grape industry [5], the set of AIMs and the method presented here provide a robust forensic tool that can be used to definitively verify the ancestry criteria these regulatory agencies attempt to apply.

REFERENCES

1. Sawler J, Reisch B, Aradhya MK, Prins B, Zhong G-Y, et al. (2013) Genomics Assisted Ancestry Deconvolution in Grape. PLoS ONE 8: e80791.

2. Reisch B, Owens C, Cousins P (2012) Grape. In: Badenes ML, Byrne DH, editors. Fruit Breeding: Springer US. pp. 225-262.

3. McGovern PE (2003) Ancient Wine: The Search for the Origins of Viniculture. Princeton, NJ: Princeton University Press.

4. Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, et al. (2011) Genetic structure and domestication history of the grape. Proceedings of the National Academy of Sciences 108: 3530-3535.

5. Myles S (2013) Improving fruit and wine: what does genomics have to offer? Trends in Genetics 29: 190-196.

6. Di Gaspero G, Cattonaro F (2010) Application of genomics to grapevine improvement. Australian Journal of Grape and Wine Research 16: 122-130.

7. Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences 363: 557-572.

8. Aradhya M, Wang Y, Walker MA, Prins BH, Koehmstedt AM, et al. (2013) Genetic diversity, structure, and patterns of differentiation in the genus *Vitis*. Plant Systematics and Evolution 299: 317-330.

9. Lodhi MA, Reisch BI (1995) Nuclear DNA content of *Vitis* species, cultivars, and other genera of the *Vitaceae*. Theoretical and Applied Genetics 90: 11-16.

10. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, et al. (2007) A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. PLoS ONE 2: e1326.

11. 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.

12. Alleweldt G, Possingham JV (1988) Progress in grapevine breeding. Theoretical and Applied Genetics 75: 669-673.

13. Lowe K, Riaz S, Walker MA (2009) Variation in recombination rates across *Vitis* species. Tree Genetics & Genomes 5: 71-80.

14. Myles S, Chia J-M, Hurwitz B, Simon C, Zhong GY, et al. (2010) Rapid Genomic Characterization of the Genus *Vitis*. PLoS ONE 5: e8219.

15. Bowers JE, Meredith CP (1996) Genetic Similarities among Wine Grape Cultivars Revealed by Restriction Fragment-length Polymorphism (RFLP) Analysis. Journal of the American Society for Horticultural Science 121: 620-624.

16. Aradhya M, Pitcher A, Prins B, Dangls G, Stover E (2007) Genetic structure, differentiation, and phylogeny of the genus *Vitis*: implications for genetic conservation. Acta Horticulture Proceedings.

17. Brookes AJ (1999) The essence of SNPs. Gene 234: 177-186.

18. This P, Lacombe T, Cadle-Davidson M, Owens C (2007) Wine grape (*Vitis vinifera L.*) color associates with allelic variation in the domestication gene VvmybA1. Theoretical and Applied Genetics 114: 723-730.

19. Welter LJ, Göktürk-Baydar N, Akkurt M, Maul E, Eibach R, et al. (2007) Genetic mapping and localization of quantitative trait loci affecting fungal disease resistance and leaf morphology in grapevine (*Vitis vinifera L*). Molecular Breeding 20: 359-374.

20. Doligez A, Bouquet A, Danglot Y, Lahogue F, Riaz S, et al. (2002) Genetic mapping of grapevine (*Vitis vinifera L.*) applied to the detection of QTLs for seedlessness and berry weight. Theoretical and Applied Genetics 105: 780-795.

21. Venuti S, Copetti D, Foria S, Falginella L, Hoffmann S, et al. (2013) Historical Introgression of the Downy Mildew Resistance Gene Rpv12 from the Asian Species *Vitis amurensis* into Grapevine Varieties. PLoS ONE 8: e61228.

22. Schwander F, Eibach R, Fechter I, Hausmann L, Zyprian E, et al. (2012) Rpv10: a new locus from the Asian *Vitis* gene pool for pyramiding downy mildew resistance loci in grapevine. Theoretical and Applied Genetics 124: 163-176.

23. Akkurt M, Welter L, Maul E, Töpfer R, Zyprian E (2007) Development of SCAR markers linked to powdery mildew (*Uncinula necator*) resistance in grapevine (*Vitis vinifera L.* and *Vitis* sp.). Molecular Breeding 19: 103-111.

24. Blasi P, Blanc S, Wiedemann-Merdinoglu S, Prado E, Rühl E, et al. (2011) Construction of a reference linkage map of *Vitis amurensis* and genetic mapping of Rpv8, a locus conferring resistance to grapevine downy mildew. Theoretical and Applied Genetics 123: 43-53.

25. Emanuelli F, Battilana J, Costantini L, Le Cunff L, Boursiquot J-M, et al. (2010) A candidate gene association study on muscat flavor in grapevine (*Vitis vinifera L.*). BMC Plant Biology 10: 241.

26. Olmo H (1996) The Origin and Domestication of the *Vinifera* Grape. In: Patrick McGovern SF, Solomon Katz, editor. Origins and Ancient History of Wine. pp. 31-43.

27. Zohary D, Hopf M (2000) Grape vine: *Vitis vinifera*. Domestication of Plants in the Old World: The Origin and Spread of Cultivated Plants in West Asia, Europe, and the Nile Valley. 3 ed: Oxford University Press. pp. 151-159.

28. Zohary D (1997) The Domestication of the Grapevine *Vitis Vinifera L.* in the Near East. The Origins and Ancient History of Wine. pp. 23-30.

29. Salmaso M, Faes G, Segala C, Stefanini M, Salakhutdinov I, et al. (2004) Genome diversity and gene haplotypes in the grapevine (*Vitis vinifera L.*), as revealed by single nucleotide polymorphisms. Molecular Breeding 14: 385-395.

30. Renfrew JM (2003) Archaeology and origins of wine production. Wine: A Scientific Exploration. pp. 56-59.

31. Olmo H, Smartt J, Simmonds N (1995) Grapes. Evolution of Crop Plants. 3rd ed. pp. 495-490.

32. This P, Lacombe T, Thomas MR (2006) Historical origins and genetic diversity of wine grapes. Trends in Genetics 22: 511-519.

33. Michael Mullins AB, Larry Williams (2007) The growing of grapes. Biology of the Grapevine. pp. 10.

34. Zohary D (2004) Unconscious selection and the evolution of domesticated Plants. Economic Botany 58: 5-10.

35. Joseph Smartt NS (1995) Evolution of Crop Plants: Wiley-Blackwell. 496 p.

36. Santiago JL, Boso S, Gago P, Alonso-Villaverde V, Martínez MC (2008) A contribution to the maintenance of grapevine diversity: The rescue of Tinta Castañal (*Vitis vinifera L.*), a variety on the edge of extinction. Scientia Horticulturae 116: 199-204.

37. Grzegorczyk W, Walker MA (1998) Evaluating Resistance to Grape *Phylloxera* in *Vitis* Species with an in vitro Dual Culture Assay. American Journal of Enology and Viticulture 49: 17-22.

38. Boopathi NM (2013) Success Stories in MAS. Genetic Mapping and Marker Assisted Selection: Springer India. pp. 187-192.

39. Karaagac E, Vargas A, Andrés M, Carreño I, Ibáñez J, et al. (2012) Marker assisted selection for seedlessness in table grape breeding. Tree Genetics & Genomes 8: 1003-1015.

40. Boopathi NM (2013) Marker-Assisted Selection. Genetic Mapping and Marker Assisted Selection: Springer India. pp. 173-186.

41. Gao S, Martinez C, Skinner D, Krivanek A, Crouch J, et al. (2008) Development of a seed DNA-based genotyping system for marker-assisted selection in maize. Molecular Breeding 22: 477-494.

42. Di Gaspero G, Copetti D, Coleman C, Castellarin S, Eibach R, et al. (2012) Selective sweep at the Rpv3 locus during grapevine breeding for downy mildew resistance. Theoretical and Applied Genetics 124: 277-286.

43. Scorza R, Cordts JM, Gray DJ, Gonsalves D, Emershad RL, et al. (1996) Producing Transgenic `Thompson Seedless' Grape (*Vitis vinifera L.*) Plants. Journal of the American Society for Horticultural Science 121: 616-619.

44. Ding L, Wiener H, Abebe T, Altaye M, Go RC, et al. (2011) Comparison of measures of marker informativeness for ancestry and admixture mapping. BMC Genomics 12: 622.

45. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, et al. (2010) Genome-wide association studies in diverse populations. Nat Rev Genet 11: 356-366.

46. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, et al. (2009) Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proceedings of the National Academy of Sciences.

47. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, et al. (2007) PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations. PLoS Genet 3: e160.

48. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, et al. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proceedings of the National Academy of Sciences 98: 11479-11484.

49. Cipriani G, Spadotto A, Jurman I, Di Gaspero G, Crespan M, et al. (2010) The SSR-based molecular profile of 1005 grapevine (*Vitis vinifera L.*) accessions uncovers new synonymy and parentages, and reveals a large admixture amongst varieties of different geographic origin. Theoretical and Applied Genetics 121: 1569-1585.

50. Liu Y, Nyunoya T, Leng S, Belinsky SA, Tesfaigzi Y, et al. (2013) Softwares and methods for estimating genetic ancestry in human populations. Human genomics 7: 1-7.

51. Ma J, Amos CI (2012) Principal Components Analysis of Population Admixture. PLoS One 7: e40115.

52. Reich D, Price AL, Patterson N (2008) Principal component analysis of genetic data. Nat Genet 40: 491-492.

53. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

54. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98-101.

55. McVean G (2009) A Genealogical Interpretation of Principal Components Analysis. PLoS Genet 5: e1000686.

56. Anamarija F, Birgit G, Urs S, Ino C, Johann S (2011) How to Use Fewer Markers in Admixture Studies. Agriculturae Conspectus Scientificus; Vol 76, No 3 (2011).

57. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81: 559-575.

58. Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genet 2: e190.

59. USDA-ARS (2013) National Genetic Resources Program. Germplasm Resources Information Network - (GRIN) http://www.ars-grin.gov/cgi-bin/npgs/html/index.pl.

60. Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155: 945-959.

61. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.

62. R Development Core Team (2008) R: A language and environment for statistical computing (http://www.R-project.org). Vienna, Austria.

63. The Bovine HapMap C, Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, et al. (2009) Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 324: 528-532.

64. Boyko AR, Boyko RH, Boyko CM, Parker HG, Castelhano M, et al. (2009) Complex population structure in African village dogs and its implications for inferring dog domestication history. Proc Natl Acad Sci U S A.

65. Robinson J, Harding J, Vouillamoz JF (2012) Wine Grapes: A complete guide to 1368 vine varieties, including their origins and flavours. New York: HarperCollins.

66. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6: e19379.

67. Paschou P, Drineas P, Lewis J, Nievergelt CM, Nickerson DA, et al. (2008) Tracing Sub-Structure in the European American Population with PCA-Informative Markers. PLoS Genet 4: e1000114.

APPENDIX A: COPYRIGHT PERMISSION

Sawler J, Reisch B, Aradhya MK, Prins B, Zhong G-Y, et al. (2013) Genomics Assisted Ancestry Deconvolution in Grape. PLoS ONE 8(11): e80791. doi:10.1371/journal.pone.0080791

APPENDIX B: STUDENT CONTRIBUTION TO MANUSCRIPT

Sawler J, Reisch B, Aradhya MK, Prins B, Zhong G-Y, et al. (2013) Genomics Assisted Ancestry Deconvolution in Grape. PLoS ONE 8(11): e80791. doi:10.1371/journal.pone.0080791

Conceived and designed the experiments: **Jason Sawler**, Sean Myles.

Performed the experiments: **Jason Sawler**, Sean Myles.

Analyzed the data: **Jason Sawler**, Sean Myles.

Contributed reagents/materials/analysis tools: Bruce Reisch, Mallikarjuna K. Aradhya, Bernard Prins, Gan-Yuan Zhong, Heidi Schwaninger, Charles Simon, Edward Buckler.

Wrote the manuscript: **Jason Sawler**, Sean Myles.