Neuroscience Area

PhD course in Molecular Biology

Analysis of genetic determinants of gene expression variation in grapevine (*Vitis Vinifera*)

Candidate:

Giovanni Gabelli

Advisor:

Prof. Michele Morgante

Co-advisors:

Prof. Fabio Marroni

Academic Year 2020-2021

CONTENTS

SUMMARY

While the effect of genetic variants on phenotypes has been widely investigated and led to several groundbreaking findings in the field of human, animal, and plant genetics, our understanding of the effects of such variants on gene expression is still relatively limited.

In this study, we present a whole genome analysis of regulatory variation in grapevine, analyzing gene expression of three tissues, namely leaves, hard, and soft berries in 98 cultivars of *Vitis vinifera*, selected to represent the diversity of the spp. *sativa*.

We performed an expression Quantitative Trait Loci (eQTL) analysis, finding that genomic variants explain the variation of the expression of thousands of genes in each tissue (eGenes). Both *cis*- and *trans*-eQTL were mapped, highlighting the effect of different variants, acting on near genes or distally, through diffusible factors. We then characterized the eGenes, finding that they show higher variability and lower selective constraints on their coding regions than non-eGenes. Subsequently, we performed an allele-specific expression (ASE) analysis on the same samples to identify differences in expression between the two alleles of each polymorphic gene. This corresponds to indirectly testing the *cis*-acting effect of genomic variants on the regulation of gene expression. Our results showed that *cis*-acting variants have largely tissue-dependent effects and that a single gene can show differences in *cis*-regulation between tissues. Moreover, we found that the genes in berries tissues showed overall higher levels of ASE than in leaves. eQTL mapping and ASE analysis represent two independent experiments, both studying the same problem from different perspectives, and they produced overall consistent results. In fact, according to our findings, the eGenes have a higher probability to show ASE. Both eGenes and genes with noticeable ASE patterns, show fewer selection constraints.

Finally, we inferred for every gene its allelic variants at the DNA sequence level in the available cultivars of *V. vinifera* spp. *sativa*, and combined this information with the ASE analysis to elaborate a model capable of measuring the net contribution of every allele to the expression of the gene. This enabled the simultaneous comparison of all the possible allelic variants of a gene and a deeper comprehension of their expression patterns within the *V. vinifera* group, highlighting the very frequent occurrence of multiple *cis*-regulatory alleles with very different effects on gene expression.

This project gave us a first perspective of how genomic variants influence gene expression in *V. vinifera* and created a database of information, gathering for every gene a set of information about the *trans*- and *cis*-acting variants that influence its expression as well as the number, sequences, and level of expression of its alleles.

1. INTRODUCTION

1.1. The role of non-coding DNA

Since the discovery that the majority of eukaryotic genomes is made of non-coding DNA (Ohno, 1972), scientists struggled to define its role and evolutionary origin. In the beginning, it was thought to be useless and famously named “junk DNA”. Most of the researchers investigating genome function focused on the study of the genes, but soon some studies aimed to investigate the function of non-coding DNA were undertaken (Zuckerandl, 1992; Nowak, 1994).

Non-coding DNA plays a major role in gene expression regulation and understanding how this happens is pivotal to comprehending the structure and organization of the genomes. Great progress has been made in classifying regulatory sequences and understanding their functions, especially in the human genome, with the contribution of projects such as the Encyclopedia of DNA Elements (The ENCODE Project Consortium, 2020) and the Genotype-Tissue Expression Project (GTEx consortium, 2020). Regulatory elements can be divided into two categories, based on their action mechanism: *cis*-regulatory DNA elements and sequences that encode for diffusible factors carrying out their regulatory effect in *trans*. Regulatory elements acting in *cis* can be sequences adjacent to the target genes (promoter elements, typically found from 1000 bp upstream to 200 bp downstream the TSS of the gene), or DNA elements located far from the gene, upstream or downstream on the chromosome (enhancer, silencer, insulator elements). Compared to a promoter, it is difficult to identify the location of an enhancer and its target gene, since the two elements could be distant up to several hundred kilobases from each other and other elements and genes could be positioned between the two (Elkon and Agami, 2016). Moreover, multiple *cis*- and *trans*-non-coding elements can affect the expression of a gene simultaneously, and their activity can be synchronous or be carried out through transcription factor proteins, which presence or concentration is in turn regulated by other coding and non-coding sequences affecting the expression of the corresponding genes. In conclusion, locating a regulatory element or identifying its effect could be a harsh challenge.

1.2. Genome-Wide Association Studies and eQTL analysis

GWAS are experiments that aim to find associations between genetic variants and a phenotype. To do so the genomes of a large number of individuals are scanned in order to find variants with a correlation with a given phenotype. This type of study was first introduced in human genetic epidemiology, which enabled the study of complex diseases, controlled by multiple genetic loci. This approach brought a better understanding of various conditions in different medical fields including major depressive disorder, anorexia nervosa, cancers, and coronary artery disease (Tam *et al*., 2019).

If the phenotype of interest is the expression of all the genes of the transcriptome (or a subset of interest), the genomic variants associated with gene expression will be variants potentially acting as regulatory elements of said gene. This kind of analysis is called eQTLs (expression Quantitative Trait Loci) mapping. An expression Quantitative Trait Locus is defined as a genomic variant that can explain the variation of expression of a gene, across a population (Nica and Dermitzakis, 2013).

Ultimately, to discover an eQTL we look for a statistical association between the genotype of a variant and the expression levels of a gene in a population of samples. In figure 1.1a (from Elkon and Agami, 2017), we see a graphical representation of the phenomenon: the boxplot depicts the expression levels in a population of individuals for a given gene, divided into three classes depending on the genotype of the variant, in this case, a SNP located in an enhancer region. The two homozygous conditions, reference and alternative, are correlated with respectively lower and higher values, while the heterozygous samples show intermediate levels of expression of the gene. So, it is possible to test the statistical significance of such a correlation, in this case using an additive model, where each allele of the SNP is associated with a given level of expression and the two can be added to determine the genotype's value of the expected gene expression. In an eQTL mapping study, DNA sequence variants genotyped in a population will be tested against the expression values of all the genes, looking for significant correlations. The genes and the variant involved in an eQTL are commonly referred in the literature as eGenes and eVariants. eVariants can also be divided into eSNPs, eInsertions, and eDeletions according to the sequence variant that is associated with the phenotype. The distance and chromosomal location between eGene and eVariant is generally taken as a proxy to assume the mechanism of control of the expression, thus dividing the eQTLs into *cis*- or *trans*-eQTLs (Gilad *et al*., 2009).

1.2.1. The Linkage Disequilibrium

Association studies (thus including eQTL mapping) rely on the presence of Linkage Disequilibrium. Linkage Disequilibrium (LD) is the non-random association of alleles at two or more loci (Single and Thomson, 2016). This means that, in a given population, if two variants are in LD, we will observe them together in the individuals more frequently than expected. In the context of an eQTL mapping, this could lead to the situation depicted in figure 1.1b, taken from the study of Astle and Balding (2009): two variants are in a condition of strong LD. One, unobserved in the experiment, is the cause of a variation in a genotype (in our case the expression level in a gene). Due to the LD, the second variant can establish a correlation with the gene expression level, resulting in the mapping of an eQTL. The value of LD that links different variants is specific for a given population and depends on several factors. Generally speaking, the more diverse a population is and the more distant in time are the ancestors that link the individuals of the population, the less LD we will register among variants, due to the accumulation of historical recombination events. This means that two variants located close to each other will be in LD with high probability and, in an eQTL mapping study, it's common to see clusters of variants establish eQTLs with the same gene. These correlations are probably the record of a unique biological event, where one single causal variant causes numerous indirect associations.

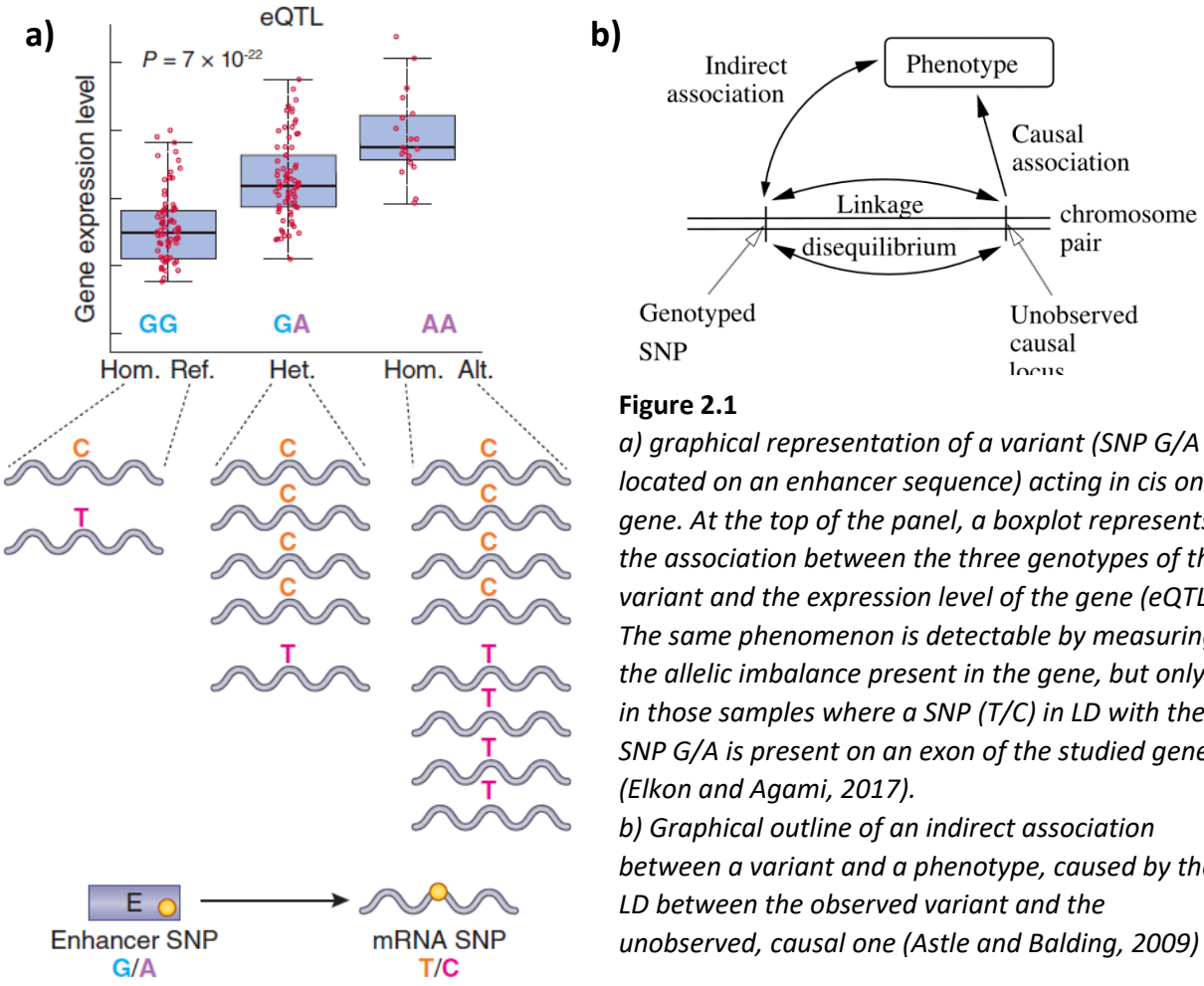

Figure 2.1

a) graphical representation of a variant (SNP G/A located on an enhancer sequence) acting in cis on a gene. At the top of the panel, a boxplot represents the association between the three genotypes of the variant and the expression level of the gene (eQTL). The same phenomenon is detectable by measuring the allelic imbalance present in the gene, but only in those samples where a SNP (T/C) in LD with the SNP G/A is present on an exon of the studied gene (Elkon and Agami, 2017).

b) Graphical outline of an indirect association between a variant and a phenotype, caused by the LD between the observed variant and the unobserved, causal one (Astle and Balding, 2009)

1.2.2. Bias in eQTLs studies due to population characteristics

A key factor to consider when designing a GWAS is the structure and characteristics of the given population, which could cause confounding effects and produce significant biases in the results, mainly in the form of false positive associations. The main concerns are relative to population structure or stratification, cryptic relatedness, and rare variants (Price *et al*., 2010).

Population Stratification is the situation where in a population there are groups of individuals with large-scale systematic differences in ancestry (Astle and Balding, 2009). The consequence is that there are groups with differences in the frequency of some alleles compared with the rest of the population. When population stratification occurs, we should theoretically be able to distinguish one or more sub-populations by observing their genotypes (Hellwege *et al*., 2017), but the situation is not always so well-defined. When the population is not constituted by distinguishable sub-populations (island model) but has a more complex composition, the term Population Structure is preferred (Astle and Balding, 2009). When population structure is present, several variants across the genome may be informative about an individual's subpopulation of origin, and thus be (wrongly) associated with any phenotype that varies across subpopulations. Another occurrence that can happen in a population is cryptic relatedness, which means the undetected presence of close relatives in a population of otherwise unrelated individuals (*ibidem*). If this occurs, we will have some similar genomes among the rest and this will pose as a confounder in the study, causing once again spurious associations if not adjusted.

Finally, rare variants can introduce biases in a GWAS. Several studies have focused on rare variants, especially in human genetic epidemiology (Bush and Moore, 2012). That is because rare variants represent the majority of the annotated variants and possibly explain a substantial share of the heritability of phenotypes (Tennessen *et al.*, 2012 analyzing common, complex diseases) and genes (Hernandez *et al.*, 2019). Moreover, they have a frequently larger effect than common variants, as can be explained by the effect of purifying selection (Eyre-Walker, 2010). However, in an eQTL mapping, since the number of individuals that carry a rare variant is low and the number of tested associations is equal to the number of tested genes, there is a high risk of false positives, indistinguishable from the real ones. To prevent such cases, in eQTL studies a filter on minimal allele frequency in the population is applied, or alternatively a threshold on the minimal number of occurrences of a certain genotype. Such filters are set according to statistical power given by the size and the diversity of the population used for the study (Huang *et al*. 2018).

1.2.3. eQTLs effect size estimation

The expression of a gene is determined by both genetic as well as environmental factors. When we perform an eQTL study, we aim to minimize the second component by the choice of an appropriate experimental

design and search for a possible explanation for the latter, called heritability. The definition of broad-sense heritability, valid for gene expression as well as any other measurable phenotype, is "the proportion of phenotypic variance that is due to all genetic effects" (Holland *et al*., 2002). When we call an eQTL, we are trying to explain part of this variability based on the genotypes observed for a genomic variant. The magnitude of the influence of the eVariant on the eGene is called effect size of the eQTL (Mohammadi *et al*., 2017). An eQTL with a small effect size represents a variant with a small impact on the expression of a gene, therefore requiring a greater statistical power to be correctly identified. This case can occur for different reasons: a gene with small heritability, a gene influenced by several variants, each accountable for a small portion of his variance and so with small effect sizes, an eVariant linked to the causal variant by a low LD, since the effect size of the eQTL is reduced by a low correlation between the genotypes of the two in the population.

1.3. Allele-specific expression analysis

Another way to explore the role of *cis*-acting variants in the regulation of the genes is the Allelic Specific Expression (ASE) analysis. This kind of study aims to assess the difference in expression between the two alleles of the same gene in a heterozygous individual. Microarrays (Schena *et al*., 1995) or quantitative RT-PCR (Yan *et al*., 2002) were used in early experiments to assess these metrics in a few genes, but NGS technologies gave the possibility to measure ASE of all the genes of a species in a population of individuals through RNA-seq (Pastinen, 2010). A graphical example of ASE in a gene and the requirement to measure it is in figure 1.1. As we can see, the variation of the SNP of genotype G/A located in the enhancer is associated with variations in the level of gene expression, due to a boost of the transcription of the G genotype, which causes a boost of transcription of the gene in the same haplotype. If we identify one or more SNP in the transcribed sequence of the gene, we can measure the impact of the variation on the gene, as seen in the quantification of reads carrying the genotype C of the SNP C/G, located in the exon of that gene. In other words, the ASE of a gene doesn't give us any information on the causal *cis*-variant, but measures its effect with precision.

1.3.1. Haplotype reconstruction

The first step in the measurement of the ASE of a gene is the phasing of the SNPs present on the transcribed region, obtaining the two haplotypes of the gene. This goal can be achieved with different methods, based on laboratory or computational approaches.

The most used method consists of a family-based approach and is performed by sequencing the parents and other relatives of the studied individual and then inferring the haplotypes using Mendel's laws. This method,

although straightforward and capable to provide good-quality results, can leave some loci un-resolved and requires the genomes of the relatives, not always available (Choi *et al*., 2017).

Laboratory-based methods provide sequencing protocol to obtain phased assembly for the samples. For example, 10X Genomics is a system that manages to barcode sequences obtained by a single molecule of ss-DNA, enabling a strand-aware assembly. An example of a project that used this approach can be found in Zheng *et al*. (2016). In the last few years, single molecule sequencing gave the possibility of obtaining haplotype resolved assemblies or, when mapping reads to a genome, very long stretches of haplotypes reconstructed by a single DNA molecule. This is the case, for example, of the PacBio and Oxford Nanopore technology. An advantage that laboratory-based methods have is that they don't need any accessory information about the genotypes of the family or the population to which the studied individual belongs. They share this characteristic with computational reads-based approaches, such as HapCUT2 (Edge *et al*., 2016). This software uses genome sequencing output to reconstruct haplotypes, overlapping sequencing fragments with two or more polymorphic sites. This method doesn't require a dedicated sequencing step, and therefore is more approachable, but, even if its performance is related to the quality and the length of the input reads, the length of the phased blocks achievable is limited compared to the other methods (Choi *et al*., 2017).

If the genome-wide genotypes of a population of individuals are available, population-based approaches offer a cheap and accurate alternative. With this approach, various computational processes based on hidden Markov models, are used to infer the most probable haplotypes present in the population, and then assign them to each individual. Examples of these software are SHAPEIT (Delanau *et al*., 2019), Eagle2 (Loh *et al*., 2016), and Beagle (Browning *et al*., 2021). In some cases, accessory information can be provided to the algorithm, in order to boost its velocity and accuracy. For example, sequencing reads, known population haplotypes, and parental genotypes can be fed to SHAPEIT, improving its performance (Choi *et al*., 2017).

1.3.2. Haplotype-aware reads alignment

Once obtained the two haplotypes of each individual under study, the RNA-seq reads will be aligned on the sequences and all the reads aligned on a heterozygous site will be considered informative for allele-specific expression analysis. The number of aligned reads on the two different haplotypes will give an indication of the allelic expression imbalance present between the two alleles. In order to prevent possible errors, besides the necessary procedures to perform a quality RNA-seq alignment (Conesa *et al*., 2016), it is necessary to account for a systematic bias known as "reference bias" (Degner *et al*., 2009). This systematic error can happen when one of the two alleles is more similar to the reference used for the alignment. In this case, the software can align reads to this allele with greater efficiency, inflating its expression level. To avoid this, Pandey et al. (2013) proposed a two-part solution implemented in their software ALLIM: first with the

information of the phased SNPs two reference sequences are built, one for each haplotype, and both are used for the alignment process. Moreover, a simulation tool estimates for every gene the probable reference bias produced, and this information is used to correct the final output.

1.4. Comparing the results of *cis*-acting variant detection obtained with eQTL mapping and ASE analyses

eQTL mapping and ASE analysis are two different, independent methods to study the effect of variants on the *cis*-regulation of the genes. eQTL mapping is a population-oriented analysis, where each gene is studied by comparing all the individuals, aiming to understand if it can be considered an eGene or not. Therefore, the resulting information for the single individual is not complete: we can obtain in a single individual the genotype of the eVariant and therefore infer the gene regulation, but the cases where a lot of eVariants are correlated with the same eGene are the majority, and the regulation of gene expression is probably the result of a multi-factor action determined by numerous SNPs, each one concurring at a single eQTL and explaining only a fraction of the variance of the expression level. In this scenario, it is difficult to single out the regulation of the eGene in a single individual. On the other hand, ASE analysis considers one sample at a time, assessing if the gene shows an allelic imbalance in expression in that specific genetic makeup. In this case, we don't have any information about the hypothetical causal variant or variants, their type, number, or location, but we register their effect. We can do so only in the genes that can be analyzed, which means the genes with heterozygous variants in their transcribed sequence, and that have enough aligned reads covering their polymorphic sites, in order to perform the analysis with statistical relevance. In this scenario, we are not able to study the gene ASE in all individuals of the population. Different factors can affect the sensibility of the two analyses. eQTLs mapping requires a certain size of population and information about its structure to have enough statistical power to correctly identify eQTLs. Even with correction procedures, the risk of unknown confounders causing spurious associations cannot be eliminated, and environmental effects on gene expression introduce undesired variation in the data. Other limitations in terms of sensibility have to be taken into account: eQTL with small effect sizes are difficult to identify. This is an important factor when it comes to assessing the regulation of expression of eGenes influenced by many different variants. Moreover, as recalled in paragraph 1.2.2., we must exclude from the analysis the rare variants. This is a big loss of power in the analysis, as this category of variants is a major driver of genome diversity, and they are frequently associated with bigger effect sizes than average. Another confounding effect can be the presence of multiple *cis*-acting eVariants, with different effect sizes, if they are not adequately tagged by eSNPs, when the GWAS analysis is performed using individual SNP genotypes and not haplotypes.

ASE analysis isn't affected by the size or structure of the analyzed population, and a big advantage of this method is that isn't affected by the environment: factors independent from the genome affect equally both

alleles of the gene. Since the output of the analysis is the ratio between the expression of the two alleles, factors that affect both in equal measure are less confounding.

In conclusion, the use of both ASE analysis and eQTLs mapping on the same subject can offer different views on the same phenomenon, independent validation of the results, and reciprocally cover their blind spots. For this reason, the two approaches are often used together (for example in studies such as GTEx consortium, 2017; Cheng *et al.*, 2021; Khansefid *et al*., 2018; Hasin-Brumshtein *et al*., 2014). At the same time, the literature underlines the difficulty to integrate the results of the two methods, since the set of observable cases in one analysis only partially overlaps the input set of the other. Moreover, the different focus of the two (analysis of one gene in a whole population in eQTL mapping and analysis of all the heterozygous genes in a single individual) forces us to adopt some simplifications in order to be able to compare the two.

1.5. The study of regulatory variants in *Vitis vinifera*

1.5.1. eQTLs studies in plants

Understanding how genetic diversity in a population shapes phenotypic traits was the first goal of GWAS and QTL studies. We have previously cited the contributions of this methodology applied to human populations regarding the epidemiology field. The application of association studies in plants was initially focused on the identification of genes related to phenotypes of commercial interest in crops, often using highly homozygous Recombinant Inbred Lines (RILs). For example, regulation of gene expression in association with the response of *Brassica rapa* to various soil phosphorus concentrations (Hammond *et al*., 2011), or with dry weight in *Oryza sativa* shoots (Wang *et al*., 2010). However, eQTL studies can help us to understand the mechanism that regulates heritable expression traits and the evolutionary forces that shape the process (Cubillos *et al*., 2012). Several studies focused on the non-coding regions of plants genomes, characterizing the role of transposable elements (Morgante, 2006; Catlin and Josephs, 2022) and structural variants (Marroni *et al*., 2014), and their contribution to the phenotype determination is assessed. Association studies could help to better define the general mechanism that underlies these regulations as proven by the work in other species of interest as *Populus trichocarpa* (Mähler *et al*., 2017) and *Zingiber officinalis* (Cheng *et al.*, 2021).

1.5.2 *Vitis vinifera*

Grapevine as a crop has a long history: its domestication dates more than 6000 years ago in Southwestern Asia (Ramos-Madrigal *et al*., 2019) and since then it dispersed across Europe, northern Africa, and Western Asia following human trades and migration. The domesticated varieties shared the habitat with their wild relatives, and that could have favored the adoption of vegetative propagation as a way to preserve valuable phenotypes arising from spontaneous crosses, in fact, we know that such a practice is at least 900 years old, since the discovery of grape seeds from 1100 AD with a genome matching the one of Savagnin Blanc, a cultivar used nowadays (*ibidem*).

Both domestication and vegetative propagation are two factors that could represent an evolutionary bottleneck, lowering the diversity of germplasm of *V. vinifera* spp. s*ativa*. However, Magris *et al*. (2021) proved that the genomic diversity among cultivated grapevine is similar to the one registered among their wild relatives, an effect probably caused by gene flow between the two groups.

Vitis vinifera, given its cultural and economic relevance, is an extensively studied species. It was the fourth plant to be sequenced, the second woody one after *Populus trichocarpa,* and the first among the perennial crops. An obstacle to its sequencing was that *Vitis vinifera* is a highly heterozygous species, so The French-Italian Public Consortium for Grapevine Genome Characterization, that produced its first genome draft (2007), selected and sequenced a highly homozygous clone, named PN40024. The size of the genome of *Vitis vinifera* resulted in approximately 500 Mb, divided into 19 chromosomes. A total of 31 922 genes were predicted by Vitulo *et al*. (2014). Subsequently de-novo sequencing of other *Vitis vinifera* spp. *sativa* accessions was performed, such as a heterozygous clone of the Pinot cultivar (Velasco *et al*., 2007) and a clone of the Nebbiolo cultivar (Gambino *et al*., 2017). In recent years a number of specimens of both *V. vinifera* and its wild relatives were sequenced (Minio and Cantu, 2022). Thanks to the technological and computational advances (Chin *et al*., 2016; Minio *et al*., 2022), many of the reference now available are phased diploid genome, meaning that both haplotypes are resolved (Minio *et al*., 2017; Minio *et al*., 2019).

Given its relevance, many of its phenotypes of interest were the subject of QTL studies or GWAS, for example, the levels of metabolites in the ripening of the berries (Reshef *et al.*, 2022), resistance to diseases (Fu *et al.*, 2020), or abiotic stresses (Trenti *et al*, 2021; Wang *et al*., 2021), but they don't provide a general picture of the expression regulation of the genes through the genome. Similarly, Martinez-Garcìa *et al*. (2022) mapped the *cis*-eQTL to validate a small number of genes differentially expressed in different ripening stages of berries among 10 cultivated grapevines. Finally, Magris et al. (2019) studied the gene expression regulation among 10 *Vitis vinifera* cultivars comparing genes in equal or different *cis*-regulation conditions.

2. OBJECTIVES

The present work aims to perform a whole genome analysis of the effect of genetic variants on gene expression in Vitis vinifera.

Several studies can be found in the literature concerning the role of regulatory sequences in cultivated grapevine (paragraph 1.5.2), but a comprehensive description of how genomic variants affect gene expression has never been published. Our study aims to fill this gap using two independent analyses: an eQTL mapping and an ASE analysis. The interest in this subject is due not only to the economic relevance of *Vitis vinifera* but also to its unique genomic characteristics. This species has a relatively small genome with high genomic diversity among the cultivars, despite its long domestication history and being reproduced through vegetative propagation. Moreover, a study of allele-specific expression of its genes is of particular interest, given its well-known high level of heterozygosity. This second step of our study is rarely applied to crops, given that these projects often involve inbred lines. Given these particularities, this project can achieve two goals: describe a comprehensive picture of *Vitis vinifera* transcriptome regulation and provide a useful set of data for every gene, ready to be browsed in case of need for information about the regulation of genes of interest.

We selected 98 cultivars representative of the variability present in the population of *Vitis vinifera* spp. *sativa* from which we obtained RNA-seq data from three tissues: leaves, and berries at two different stages of development: hard berries (target developmental stage: 5.2 °Brix) and soft berries (target developmental stage: 6.4 °Brix). Using a database of variants consisting of known genotypes of SNPs and large indels obtained from whole genome resequencing data, we mapped the variants correlated with variations in expression levels and characterized the genes targeted by this effect.

Allele-specific expression analysis, performed on the same samples, provided new information on the same subject from a different point of view. We measured ASE levels for the different tissues and cultivars, assessing noticeable differences between the samples and the tissues.

Lastly, we performed an investigation gathering all the allelic variants present for every gene in the selected population, assessing for each its net contribution to the expression of the gene. This contributed to giving us an understanding of the existing diversity for *cis*-regulatory alleles across the population.

3. MATERIALS AND METHODS

3.1. SNP and SV genotype data

We used a grapevine variants database previously produced by our research group (Magris *et al*., 2021), consisting of 10 393 171 SNPs, 52 427 large insertions, and 22 312 large deletions. Each variant, whether it is a SNP or a SV, is biallelic, so every cultivar can be homozygous with both alleles equal to the reference genome, heterozygous, homozygous with both alleles alternative to the reference, or not known. The variants were filtered independently for every RNASeq dataset according to the following rules:

- A variant must be polymorphic
- Every genotype, if present in that variant, must be recorded in at least five cultivars
- The genotype of a variant must be known in at least half of the cultivars

The numbers of variants selected for the analyses are listed in table 3.1.
The d_N/d_S values for *Vitis* genes were taken from the same study (*ibidem*)

	SNPs	**deletions**	**insertions**
Leaves	3 095 950	8 837	18 731
Hard Berries	2 287 437	6 171	12 960
Soft Berries	2 251 894	6 081	12 697

Table 3.1: *Number of input variants in eQTL analyses*

3.2. Sampling and sequencing for RNASeq data production

Berries and leaves samples were collected, used for library preparation, and sequenced by our research group (Magris *et al*. 2021).
The varieties (replicated twice) were forced to root in potted soil. A single shoot per cutting was raised until the stage of 10–12 leaves in a common garden experiment. At that stage, the fourth distal (fully expanded) leaf was sampled from each replicate and variety at the same time and frozen immediately for RNA extraction. Berries were sampled at the same developmental stage on different dates, from two replicated field plots. From each plot, two batches of asynchronous berries were collected over the same bunches, one composed of hard berries (target developmental stage: 5.2 °Brix), the other composed of soft berries (target developmental stage: 6.4 °Brix), both sorted by firmness to the touch. The accuracy of berry sorting was validated by subsampling from each batch random subsets of berries for destructive measurements, e.g.

soluble solids concentration. Each accession was sampled at a single time-point, corresponding to the exact day when hard and berries coexisted on the same bunches.

RNA was extracted using the Spectrum Plant Total RNA Kit (Sigma-Aldrich, Saint Louis, MO). Approximately 500 ng of RNA was used for library construction with the TruSeq Stranded mRNA Kit (Illumina, San Diego, CA) for leaf RNA and with the Universal Plus mRNA-Seq Library Preparation Kit (Tecan Genomics, Redwood City, CA) for berry RNA. Paired-end reads were obtained from Illumina HiSeq2000 and HiSeq2500 sequencers

The raw reads were filtered with ERNE-filter v.1.4.6 (Del Fabbro *et al*., 2013), to filter chloroplast reads and remove reads of low quality and shorter than 50 pb. The alignment of the reads to the reference genome of *Vitis vinifera* was performed with the software STAR v2.5 (Dobin *et al*., 2013) with the subsequent setting of parameters: --outMultimapperOrder Random, --outSAMmultNmax 10, --outWigStrand Stranded, --twopassMode Basic. The total count of filtered and aligned reads is in Supplementary material, table 1.

A quality assessment of the samples was performed, considering the percentage of read duplicates in the alignment and the outliers obtained with a Principal Component Analysis of the aligned reads. This led to the exclusion of the subsequent samples in Soft Berries: Nebbiolo replicates (rep) 1 and 2, Cabernet Sauvignon rep1, Sultanina rep2, Falanghina rep1, Raboso Piave rep 1 and 2, and V294 rep 1 and 2. In Hard Berries Greco di Tufo rep 1 and 2, Raboso Piave rep 1 and 2, Schioppettino rep1, 411 rep1, and Nebbiolo rep1 were discarded. In Leaves samples none of the replicates showed high levels of duplicates.

An analysis of the SNPs in RNA samples was executed, in order to check that every sample was assigned to the correct cultivar. The 300 most expressed genes for every tissue were selected and a SNP calling for the reads mapped on those genes was performed. The tool used was GATK v.3.3 HaplotypeCaller (Poplin *et al*., 2018), with the following parameters: heterozygosity = 0.01, maximum number of alleles per position = 6. For each sample we compared the resulting SNPs genotypes with the corresponding data from genomic sequences, verifying that the transcriptomic sequences belong to the correct cultivar.

3.3 eQTL analysis

3.3.1. Expression data filtering and normalization

The number of reads mapping on each gene was computed using STAR. In samples with replicates available, the reads count of the replicates was summed across genes. The genes were then filtered to exclude the less expressed genes. In every tissue, only the genes that registered at least 10 reads in more than 10% of the cultivars (9 CVs in leaves and 6 CVs in both berries soft and berries hard) were kept. Moreover, identified all

the genes that were potentially pseudogenes, transposable element products or misannotated. This quality control was performed as described below:

- The sequences of the primary transcript of the V2.1 annotation (Vitulo *et al*., 2014) were used as a query for a blastx analysis (Altschul *et al*., 1990) against a database with all Viridiplantae (NCBI: txid33090) non-redundant proteins. From this database, we excluded the *Vitis vinifera* (NCBI: txid29760) proteins. A total of 246 transcript sequences did not align against any other sequence. Those genes were discarded from all the subsequent analyses of the project.
- A list of probable transposable elements to discard from the analysis was compiled using the program repeatmasker (Smit and Hubley, 2008) with the default setting, adding a library of identified TE in Vitis, previously elaborated by our group, with the function "-lib".

A total of 22 546 genes were retained for the analysis in leaves, 24 179 genes in hard berries, and 23 865 genes in soft berries.

Read counts were normalized using the method median of ratios implemented in the R package *DESeq2* version 1.26.0 (Love *et al*., 2014). The normalization was performed independently for the three tissues, setting at one the number of different conditions across the samples.

3.3.2. eQTL mapping

The eQTL analysis was performed using the R function Matrix_eQTL_main of the *MatrixEQTL* package (Shabalin *et al*. 2012). The three tissues were analyzed independently, using as input the genotype and expression data previously described. The effect of the genotypes on the gene expression was assumed to be additive linear.
After the first run of the model, we analyzed the distribution of the resulting p-values, and a significant deviation from the normal distribution was observed. This was expected, caused by false association due to population structure and hidden confounders. A measure of this effect is the inflation factor λ, which is the ratio of the median of the observed distribution of the test statistic to the expected median. This index was calculated with the R function P_lambda of the package *QCEWAS* (Van der Most *et al*., 2017).

In order to correct for population structure and hidden confounders (cf. paragraph 1.2.2.), we adopted the principal component (PC) adjustment method (Astle and Balding, 2009). To carry out this method principal components of both the genotypes and expression data were computed. matrixEQTL was then run including in the linear model as covariates the first components of the two analyses. For every tissue the model was executed using as covariates from 0 to 22 PCs of the genotypes and from 0 to 10 PCs of the expression data, testing all the combinations between the two. For these analyses we used a reduced

dataset as input, randomly selecting 1000 genes and 40000 variants, computing for every analysis the λ factor using a distribution of 8 x 10^8 p-values.

Finally, we selected the smallest number of PCs necessary to achieve a λ factor near 1. For the final eQTL analysis 20 genotypes PCs and 2 expression PCs for leaves (estimated λ factor= 1.0306), 4 genotypes PCs and 0 expression PCs for hard berries (estimated λ factor = 1.005196), and 3 genotypes PCs and 0 expression PCs for soft berries (estimated λ factor = 1.029024) were used as covariates in the model.

The analysis was performed by separating the output between local eQTL and distant eQTL depending on the relative distance between the TSS of the gene and the position of the variant. Gene and variant closer to each other than 100 kb were assigned to local eQTL, while the ones separated by more than 100 kb or belonging to different chromosomes were considered distant eQTLs. This threshold wassup decided by observing the decrease of eQTLs p-values according to the distance of eVariant to the TSS of eGenes (supplementary material, figure S1) and is the same to the one adopted in a similar study in *Populus trichocarpa* (Mähler *et al*., 2017). The positions of the genes were obtained from the V2.1 annotation (Vitulo *et al*., 2014), while the coordinates of SNPs and SVs were present in the source files. Since deletions could consist of more than one base, the middle point between start and end was used, while for insertions we assume as their position the mean point of insertion.

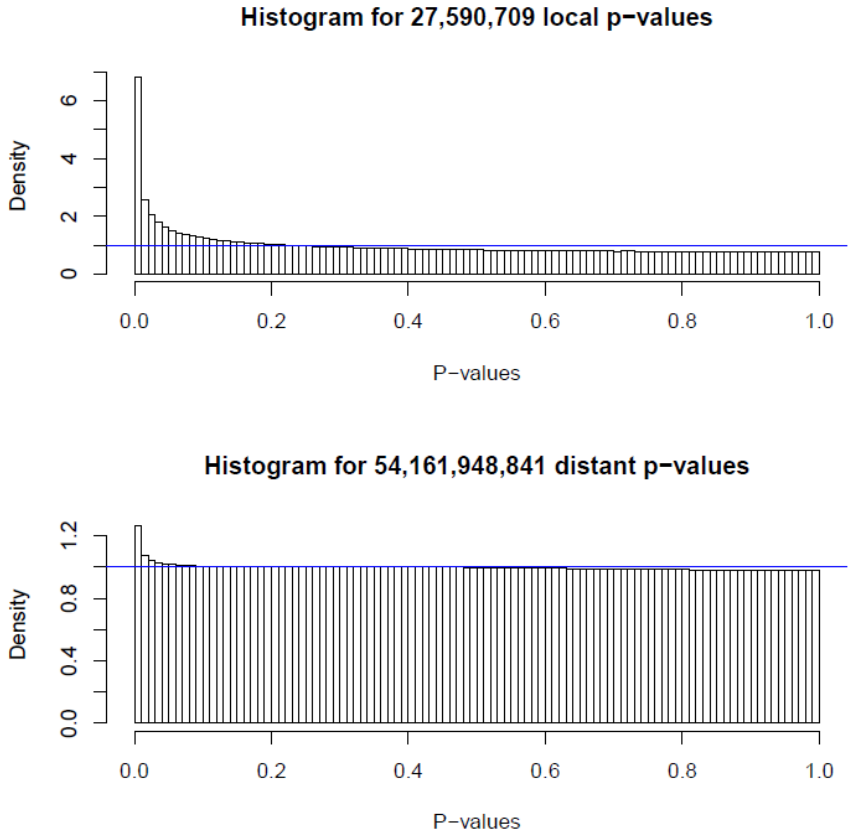

Figure 3.1: *Histograms with p-values distribution of cis-eQTL (above) and* trans*-eQTLs (below) after the correction with the principal component adjustment method*

3.3.3. FDR correction

We assessed the false discovery rate (FDR) of the p-values using two different methods. For local-eQTLs we adopted a hierarchical procedure described by Huang *et al*. (2018). Since this method cannot be used in the case of distant-eQTLs we implemented a permutation-based procedure as in Peters *et al*. (2016). The hierarchical method of FDR control is structured in three steps:

1. Taking into account one gene at a time, the p-values of its eQTL are adjusted for multiple testing (locally-adjusted p-values)
2. A table is built with the eQTLs with the smallest p-value for every gene and their locally-adjusted p-value are once again adjusted for multiple testing (globally-adjusted p-values). The highest locally-adjusted p-value corresponding to a globally-adjusted p-value smaller than 0.05 is set as the threshold value
3. For every gene, the eQTLs with locally-adjusted p-values equal to or smaller than the threshold are considered significant

Taking the leaves matrix-eQTL output as testing data, we performed nine times the described procedure, testing three different methods of p-value adjustment in each of the two adjustment steps: Bonferroni's, Benjamini-Hochberg's (Benjamini and Hochberg, 1995) and Benjamini-Yekuteli's (Benjamini and Yekuteli, 2001) FDR correction methods. Finally, we selected the more stringent correction approach, formed by Benjamini-Yekuteli's method for the local adjustment and Bonferroni's method for the global adjustment.

The adjustment of the distant-eQTL p-values was performed with the use of permutations. It was carried out one gene at a time, as follows:

1. The expression values in all the cultivars for that gene were randomly shuffled 10 000 times and all the vectors were put together in a matrix of 10 000 lines.
2. This matrix was used as an expression input for a matrix-eQTL analysis performed with the same other inputs and parameters as in the analysis previously described
3. For every eQTL's p-value belonging to the analyzed gene an FDR threshold was computed following the subsequent formula, given an eQTL "i":

$$FDR_i = \frac{R_i - 1}{r_i \times N}$$

Where R_i is the position of eQTL i p-value in a ranked vector from lowest to highest of all the p-values of eQTLs generated by the permutations, r_i is the position of eQTL i p-value in a ranked vector from lowest to highest of all the p-value of the eQTLs belonging to the analyzed gene, N is the number of permutations

4. All the eQTLs with an FDR ≤ 0.01 were considered significant.

Consistently with the current literature, we called eGenes the genes with one or more significant eQTLs, and we called eSNPs, eInsertion, and eDeletion the variants involved in a significant eQTL.

The estimate of eQTLs effect size was performed by taking the proportion of explained variance (r^2) as a measure of the effect size (Mohammadi et al., 2017), computed by the formula:

$$r^2 = \frac{t^2}{t^2 + df}$$

With t indicating the t-statistic value of every eQTL (obtained by matrixEQTL output) and df are the degrees of freedom in the linear model (i.e. the number of samples considered for that eQTL minus one)

3.3.4. Functional annotation and Gene Ontology category enrichment

A functional annotation of all *Vitis vinifera* genes was obtained from an analysis previously carried out by our group with the program blast2go (Götz *et al*., 2008). The category enrichment analyses performed on the list of eGenes were performed with the function *runTest of* R package topGO (Alexa and Rahnenfuhrer, 2022), using the parameters algorithm = "classic" and statistic = "fisher". We considered as enriched the categories with fisher's test p-value < 0.05 and a number of total genes belonging to that category higher than 10.

3.4. Allele-Specific Expression analysis

3.4.1. Haplotype phasing

SHAPEIT2 (Delaneau *et al*. 2013) was used to infer the haplotypes of the samples. Haplotype inference was performed using a previously obtained set of SNPs (see paragraph above). The performance of population-based software for haplotypes reconstruction depends also on the size of the population in input: the bigger the population, the more accurate are the output haplotypes. So, we used the genomic data of all the cultivar of *Vitis vinifera* spp. *sativa* in the disponibility of our group, consisting in a pool of 144 samples.
We filtered these SNPs discarding the ones without a known genotype in less than 50% of the cultivars. We then selected only the SNPs located in the exonic regions, according to the V2.1 annotation (Vitulo *et al*., 2014).

We chose the read-aware option of SHAPEIT2; therefore, we performed the extractPIRs function of the program, using as inputs the alignment files of genomic sequence reads of the cultivars and the genotype

data in VCF format; both alignment and VCF data were previously obtained by our research group. In order to maximize the accuracy, we tested different combinations of parameters. For every combination, we ran SHAPEIT2 on two chromosome regions: chromosome 4 from 1Mb to 6Mb and chromosome 15 from 12.5Mb to 17.5Mb. The results were then compared to the known haplotypes of 5 cultivars, previously obtained with the sequencing of selfed progenies (Alice Fornasiero Ph.D. thesis). We counted the number of identical exons and identical genes between the SHAPEIT2-obtained haplotype and the reference haplotypes, obtaining a percentage representing the quality of the phasing. Referring to SHAPEIT2 manual, we tested the following values: --family: 10000 and 1000; --rho: 1e-2, 5e-3, 1e-3 and 1e-4; --burn: 7,10 and 15; --prune: 8 and 10; --main: 20 and 50; --states: 100, 150, 200 and 250.

During this set of tests, we noticed that the majority of errors were caused by missing data in the SNPs genotype, so we performed one analysis for each cultivar, providing SHAPEIT2 with the genotypes of all the cultivars as inputs, but excluding all the positions for which the SNPs was not known in the studied cultivar. After this evaluation on the test dataset, SHAPEIT2 was performed on the whole genome with the following settings: --family 1000, --rho 1e-3, --burn 7, --prune 8, --main 50, --states 250, --window 0.1. Aiming to make the analysis repeatable we used a single thread and set the seed for the random numbers' generator at 5. With these parameters, the comparison between the output of the analysis and the aforementioned data of 5 cultivars, indicated a percentage of identical genic haplotypes between 96 and 98% of all the genes in the analysis. This result is similar to the one presented by Choi *et al*. (2017) as the best achievable for a similar analysis. The quality of the phasing process was measured again for every chromosome of the 5 reference cultivars, to check that no regions of the genome registered a lower quality of the analysis (Table 3.2).

3.4.2. Allele-specific mapping of transcriptomic reads

Based on a previous analysis of the chromosomal regions where each cultivar is homozygous (Magris *et al*., 2021) SNPs eventually located in those regions were pruned from the SHAPEIT2 output, as they most probably represent genotyping errors.

We then integrated the information of the phased SNPs with the reference genome in FASTA format obtaining the sequences of the two haplotypes in all the cultivars. This operation was carried out with the functions readDNAStringSet, DNAString, replaceLetterAt, and writeXStringSet from the R package *Biostrings* (Pagès *et al*., 2022).

We then used the software ALLIM v1.1 (Pandey *et al*. 2013) to estimate the allele-specific expression of the genes in all the tissues and cultivars. For every sample, the inputs were the two FASTA files with the haplotypes, a GTF file with the reference annotation v.2.1, the calculated insert size, and the fastq files with the RNA-seq reads. The program assigned the transcriptomic reads to one of the two alleles, measuring the

allele-specific expression. We filtered all the genes that, among the different replicas of the sample, didn't reach a total of 50 reads informative about the ASE.

The significance of an allelic imbalance was assessed with the Stouffer method for meta-analysis (Stouffer *et al*., 1949)

	Cabernet Franc	Pinot Noir	Rkatsiteli	Sangiovese	Savagnin Blanc
chr1	0.96	0.98	0.98	0.98	0.98
chr2	0.96	0.98	0.98	0.97	0.98
chr3	0.95	0.97	0.97	0.95	0.96
chr4	0.96	0.99	0.98	0.97	0.97
chr5	0.97	0.98	0.97	0.97	0.98
chr6	0.98	0.98	0.97	0.98	0.98
chr7	0.98	0.98	0.97	0.98	0.98
chr8	0.99	0.99	0.98	0.98	0.97
chr9	0.94	0.95	0.95	0.95	0.92
chr10	0.95	0.98	0.96	0.96	0.94
chr11	0.97	0.96	0.97	0.98	0.97
chr12	0.96	0.98	0.97	0.97	0.96
chr13	0.97	0.98	0.95	0.97	0.96
chr14	0.97	0.98	0.96	0.96	0.96
chr15	0.97	0.99	0.96	0.95	0.98
chr16	0.96	0.97	0.98	0.97	0.95
chr17	0.95	0.97	0.98	0.99	0.97
chr18	0.97	0.99	0.97	0.97	0.97
chr19	0.96	0.96	0.96	0.95	0.95
chrUn	0.95	0.98	0.97	0.94	0.97

Table 3.2: *Percentage of genes in every chromosome where the SHAPEIT2 obtained haplotypes identical to the haplotypes used as reference.*

3.4.3 Allelic Imbalance assessment

For this analysis, a gene was considered homozygous when all the SNPs present in the same collection used for eQTL mapping (cf. Par. 3.1.), with a position between 1 kb upstream of the TSS of the gene and 1kb downstream of the end of the transcribed region were homozygous. The information in annotation v.2.1 was used to derive gene coordinates. If a gene was not considered homozygous but didn't have any SNPs in exonic regions, and therefore was impossible to analyze with ALLIM, it was assigned to the category "no exonic SNPs".

As stated before, the genes that were not expressed or with less than 50 informative reads for ALLIM analysis were filtered (category "low informative reads"), while for all the other genes their allelic imbalance (AI) was calculated with the subsequent formula

$$AI = \left| \log_2 \frac{n.\ reads\ hapA}{n.\ reads\ hapB} \right|$$

Where *hapA* and *hapB* are the two haplotypes of the gene. If only one of the two haplotypes was expressed, and therefore was no longer possible to compute the logarithm, we substituted the zero with a value equal to 1/100 the number of reads of the other haplotypes, obtaining an AI of 6.64. All the AI values were then evaluated for significance as described before.

A functional annotation and GO enrichment analysis was performed on the genes with significant AI in more than 80% of the cultivars analyzed or homozygous and the genes with only one expressed allele in at least one cultivar. The method used is as described in paragraph 3.3.4.

3.5. Haplotype reconstruction and analysis of the allelic variant population

3.5.1. Reconstruction of haplotypes present in the population

We used the 144 couples of genic haplotypes produced with SHAPEIT2 (cf. paragraph. 3.4.1) as starting data to reconstruct the allele population of every gene. We chose to adopt a parsimonious procedure, aiming to explain the variability of the population with as few as possible alleles. We did it taking one gene at a time, with a procedure structured as follows:

1. We searched among the 288 haplotypes the ones without missing data. For every one of them, if not already present in the "library" of existing alleles of that gene, we added to it
2. Every haplotype with missing data was compared, one at a time, with the library of existing alleles excluding the missing position. If only one match is found, the haplotype is assigned to that allele. If there is more than one match, the haplotype is marked as ambiguous with the names of all the

"candidate" alleles. If no match is found the considered sequence is put in a second collection of incomplete alleles.

3. In this second collection the first two incomplete alleles are compared. If the two are compatible (i.e., they are identical once the missing positions of the two are excluded) they are merged in one allele, filling some of the missing positions they didn't share. If not compatible, they are listed as two separate alleles. Then the third haplotype is taken into account and compared with the ones listed before him and so on. If the considered sequence matches with more than one allele, it isn't taken into account for the rest of the analysis, due to the high number of missing data.
   This comparison between all the alleles with missing data is performed a second time because some alleles are progressively filled with information in this process, and a haplotype that wasn't assigned in the first step could find a suitable sequence in a second run of the algorithm.
4. The alleles obtained in step 2 and 3 are, together, our collection of alleles for the gene. We then assign to every haplotype of cultivars one allele, a combination of possible alleles (as in step 2 if the sequence could identify more than one allele), or a marker for missing data (as in step 3 if too many positions are unknown)

These operations were performed in R with the packages *dplyr* and *data.table.*

The final result is the list of all the alleles observed in the study sample for each gene.

3.5.2. Estimating the *cis*-regulatory values of alleles from ASE population data

For each gene, an assessment was made about the contribution of each haplotype to allelic imbalance. For every gene in a specific tissue all the samples with the following characteristics were considered:
a) they are heterozygous
b) the two alleles are both known
c) the two alleles are not unique (i.e. are present in other samples of the filtered collection)
d) the sample had more than 50 informative reads between the two haplotypes for the AI measurement (as in par. 3.4.2.)

For every allele present in this population, its relative contribution to allele-specific expression was computed as the ratio between the sum of the reads attributed to the haplotype and the sum of all the reads attributed to the alternative haplotypes. A comparison between two alleles was considered significant according to the same procedure adopted for assessing the significance of AI in the ALLIM analysis with the Stouffer method for meta-analysis (Stouffer *et al*., 1949). Here were considered as replicates all the samples for which the two considered alleles were the haplotypes.

Spearman's correlation was measured between the relative expression observed in the data and the estimated relative expression.

All the statistical operations were performed in R.

3.5.3. Identification of the genes with an abnormal distribution of low or high AI among the sample population

For every gene for which was possible to perform the analysis described in the previous paragraph all the alleles without a computed estimate of net contribution on general ASE were discarded, and the frequencies of the remaining alleles were calculated. An index, here called ψ, was calculated with this formula

$$\psi = \sum_{(i,j) \in A \times A} f_{(i,j)} \left| \varepsilon_{(i)} - \varepsilon_{(j)} \right|$$

where A refers to the set of alleles, i and j to two alleles, f to the frequency observed of the genotype, and ε to the estimate of net contribution on the general ASE of the allele.

The alleles present in the population for that gene were then randomly associated in genotypes, according to their frequencies, generating a new population of casual genotypes, for which the index ψ was calculated. The process was repeated 1000 times, generating a distribution of indexes ψ that constituted our null population. The index ψ of the observed gene population was then ranked among the null distribution and its genotypes distribution was considered significantly divergent from the expected if its position was in the top or bottom 5% of the values.

4. RESULTS AND DISCUSSION

4.1 RNA sequencing

RNA samples from 90 cultivars in leaves, 58 in hard berries, and 57 in soft berries were sequenced. Among the three tissues, we have at least one sample for 98 different cultivars, while 49 have data for all three tissues. One of the two replicas could be missing due to samples with technical problems or batch errors. The metrics of the sequencing results are listed in tables S1, S2, and S3 of supplementary materials.

4.2. eQTL analysis

4.2.1 eQTLs overview

The results of the eQTL analysis are summarized in table 4.1. The higher number of eQTLs identified in leaves is expected, as a result of the higher number of cultivars used for the analysis and therefore higher statistical power. In all the tissues the majority of eQTLs were identified between genes and variants located in the same chromosome, and a significant fraction of the variants were distant less than 100 Kbp from the TSS of the genes. In these cases, the eQTL capture the regulatory effect of a variant situated near the gene, and therefore we assign them to the "*cis*-eQTL" category. On the other hand, the effect of a variant on a gene located on a different chromosome is an effect mediated by a diffusible factor. We assigned these signals to the category of "*trans*-eQTL *inter* chromosomes". Another category consists of variants linked to a gene with the TSS located further than 100 Kbp, but on the same chromosome. These are variants that could be due to a *trans*-effect, but their very large number (7 to 8 times larger than the number of *trans*-effects detected in the other 18 chromosomes, when we would expect that *inter*-chromosome effects are 18 times more frequent than *intra*-chromosome ones if *trans*-acting effects were randomly distributed among the 19 grapevine chromosomes) and the observation that LD blocks in the analysed population often extends over more than 100 Kbp, makes us believe that these could be the effect of long LD relationships with *cis*-regulatory variants closer to the gene. We assign them to a third class, the "*trans*-eQTL *intra* chromosome". Lastly, a fourth group is made of all the eQTLs that have the eGene or the eVariant located in a scaffold not assembled in a chromosome in the genome assembly used for our analysis. In this case, we cannot define the distance between the gene and the variant, and therefore we cannot assign these eQTLs to one of the other three groups. The number of eQTLs in this last category is substantial, since the total size of the aforementioned scaffold is roughly 60 Mbp, more than 12% of the total size of the genome and is sufficient that one between the eGene or the eVariant is situated on this region to assign the eQTL to this class. Since the information on the distance between the two members of the eQTL is not available, we will not consider these eQTLs when assessing the relative abundance of the different groups.

categories	Leaves	Hard Berries	Soft Berries
eQTL number	487 293	266 415	364 959
cis-eQTL	104 205	39 570	46 796
trans-eQTL *intra* chromosome	252 340	146 512	216 235
trans-eQTL *inter* chromosomes	28 726	20 734	32 166
eQTL on non-assembled scaffold	102 022	59 599	69 762
eSNP	327 369	196 656	246 991
cis-eSNP	86 867	34 935	41 702
trans-eSNP *intra* chromosome	187 697	120 396	159 790
trans-eSNP inter chromosomes	26 926	17 275	28 774
eSNP of eQTL on non-assembled scaffold	57 499	40 331	45 068
eInsertion	1 397	787	1 046
cis-eInsertion	283	110	135
trans-eInsertion *intra* chromosome	856	501	711
trans-eInsertion *inter* chromosomes	153	69	151
eInsertion of eQTL on non-assembled scaffold	239	172	180
eDeletion	714	489	612
cis-eDeletion	192	68	81
trans-eDeletion *intra* chromosome	433	325	466
trans-eDeletion *inter* chromosomes	41	37	56
eDeletion of eQTL on non-assembled scaffold	136	82	84
eGenes	3 132	1 740	1 978
eGenes non on random	2 590	1 418	1 608
eGenes with only *cis*-eQTL	793	445	443
eGenes with only *trans*-eQTL *intra* chr	255	157	183
eGenes with only *trans*-eQTL *inter* chr	289	238	299
eGenes with only *cis* and *trans*-eQTL (*intra*)	597	255	294
eGenes with only *cis* and *trans*-eQTL (*inter*)	27	22	18
eGenes with only *trans*-eQTL	100	76	91
eGenes with eQTL of all categories	529	225	280
eGenes with eQTL on non-assembled scaffold	542	322	370

Table 4.1 – *Summary of eQTL mapping results for the three tissues*

	Leaves			Hard Berries			Soft Berries		
	number	r^2 mean	r^2 ds	number	r^2 mean	r^2 ds	number	r^2 mean	r^2 ds
All eQTLs best pv	3132	0.510	0.116	1740	0.615	0.093	1978	0.616	0.096
cis-eQTLs	1564	0.516	0.111	729	0.614	0.088	770	0.625	0.090
trans-eQTLs *intra* chromosome	671	0.491	0.118	388	0.609	0.103	462	0.616	0.103
trans-eQTLs *inter* chromosomes	316	0.493	0.115	269	0.612	0.094	338	0.590	0.094
eQTLs on non-assembled scaffold	581	0.524	0.122	354	0.624	0.092	408	0.620	0.097
eQTLs best pv with eSNPs	2957	0.509	0.116	1649	0.615	0.093	1898	0.616	0.096
cis-eQTLs	1487	0.516	0.111	704	0.614	0.087	745	0.625	0.090
trans-eQTLs *intra* chromosome	647	0.491	0.118	378	0.610	0.103	449	0.615	0.103
trans-eQTLs *inter* chromosomes	294	0.493	0.115	239	0.612	0.094	318	0.591	0.095
eQTLs on non-assembled scaffold	544	0.524	0.122	336	0.624	0.093	389	0.620	0.097
eQTLs best pv with eInsertions	17	0.517	0.086	11	0.590	0.096	16	0.604	0.103
cis-eQTLs	9	0.557	0.076	4	0.633	0.126	4	0.618	0.030
trans-eQTLs *intra* chromosome	5	0.492	0.086	4	0.559	0.093	4	0.703	0.172
trans-eQTLs *inter* chromosomes	0	--	--	2	0.567	0.113	7	0.562	0.067
eQTLs on non-assembled scaffold	3	0.439	0.065	2	0.588	0.057	1	0.562	--
eQTLs best pv with eDeletions	6	0.509	0.128	10	0.623	0.099	6	0.674	0.060
cis-eQTLs	4	0.497	0.101	7	0.632	0.118	2	0.691	0.095
trans-eQTLs *intra* chromosome	1	0.378	--	1	0.562	--	4	0.666	0.052
trans-eQTLs *inter* chromosomes	0	--	--	0	--	--	0	--	--
eQTLs on non-assembled scaffold	1	0.692	--	2	0.619	0.018	0	--	--

Table 4.2 – *count of eQTLs with lowest p-value for every eGene, and their mean and standard deviation of r^2, according to the type of eVariant and of distance between eGene and eVariant*

In the table, we see that the bigger group in all the tissues is the one of the *trans*-eQTLs *intra* chromosome (65% - 73%). This could be a consequence of the high linkage disequilibrium and high haplotype sharing observed in *Vitis* (Magris *et al*., 2021). Association between a variant and a gene is registered in our results as some *cis*-eQTL and other *trans*-eQTL on the same chromosome, due to variants in linkage disequilibrium. This phenomenon can also be observed in figure 4.1: the bisector of the three scatterplots represents *cis*-eQTLs and some *trans*-eQTLs *intra* chromosome, identifiable as segments slightly off from the bisector. If we compare the number of variants involved in eQTLs with the overall number of variants selected for the analysis, we find that neither the tissue nor the type of variants has a big effect on this ratio (8.6% - 11.0% for eSNPs, 6.1% - 8.2%, for eInsertions, 7.9% - 10.1% for eDeletions).
Several studies identified eQTL hotspots, *i.e.* regions containing variants influencing a large number of genes through the genome, (Tian. *et al*. 2016; Qu *et al*., 2018; Velez-Irizarry *et al*., 2019). In our analysis, we did not identify any eQTL hotspot.

eQTLs analyses are based on statistical correlation and, without other types of tests, we cannot define the relationship as causal. However, we can assume that, among the eQTLs established in an eGene, the one

with the lowest p-value is the best candidate to be the causal variant or to be its best proxy. If we consider this data set, we find the *cis*-eQTLs grow from representing 16% - 27% of the eQTLs to 49% - 61%. This could be an effect of the different FDR threshold setting methods used for *cis*-eQTLs, more effective in pruning the low p-value eQTLs, but an overall higher significance of *cis*-eQTLs compared to *trans*-eQTLs (Zhang *et al*., 2020; Mahler *et al*., 2017) is commonly reported. As a measure of the effect size of the eQTLs, we calculated the proportion of explained variance (r^2) taking into consideration, for every eGenes, the eQTLs with the lower p-value. In table 3 we reported the mean and standard deviation of r^2 of the various groups of eQTLs, divided according to the causal variant. Consistently to the literature, the *cis*-eQTLs have higher effect sizes (Zhang *et al*., 2020; Mähler *et al*., 2017; Wang *et al*., 2010), while the higher standard deviation in the berries' samples could be a consequence of the smaller sample size in the two analyses in comparison to the leaves one.

If we select the eSNPs, among these candidate causal variants, we can see if there is any enrichment for the genomic context of the SNP (figure 4.2) compared with the total of SNPs used as input in the analysis. SNPs situated in the exonic regions of genes or their UTRs are significantly enriched (p-value of a fisher test < 0.01 with the exception of the exonic SNPs in Soft Berries), whereas the SNPs situated in intergenic regions are significantly less than expected. We do not register other significant effects on SNPs belonging to other regions (splicing, intronic, and upstream and downstream regions).

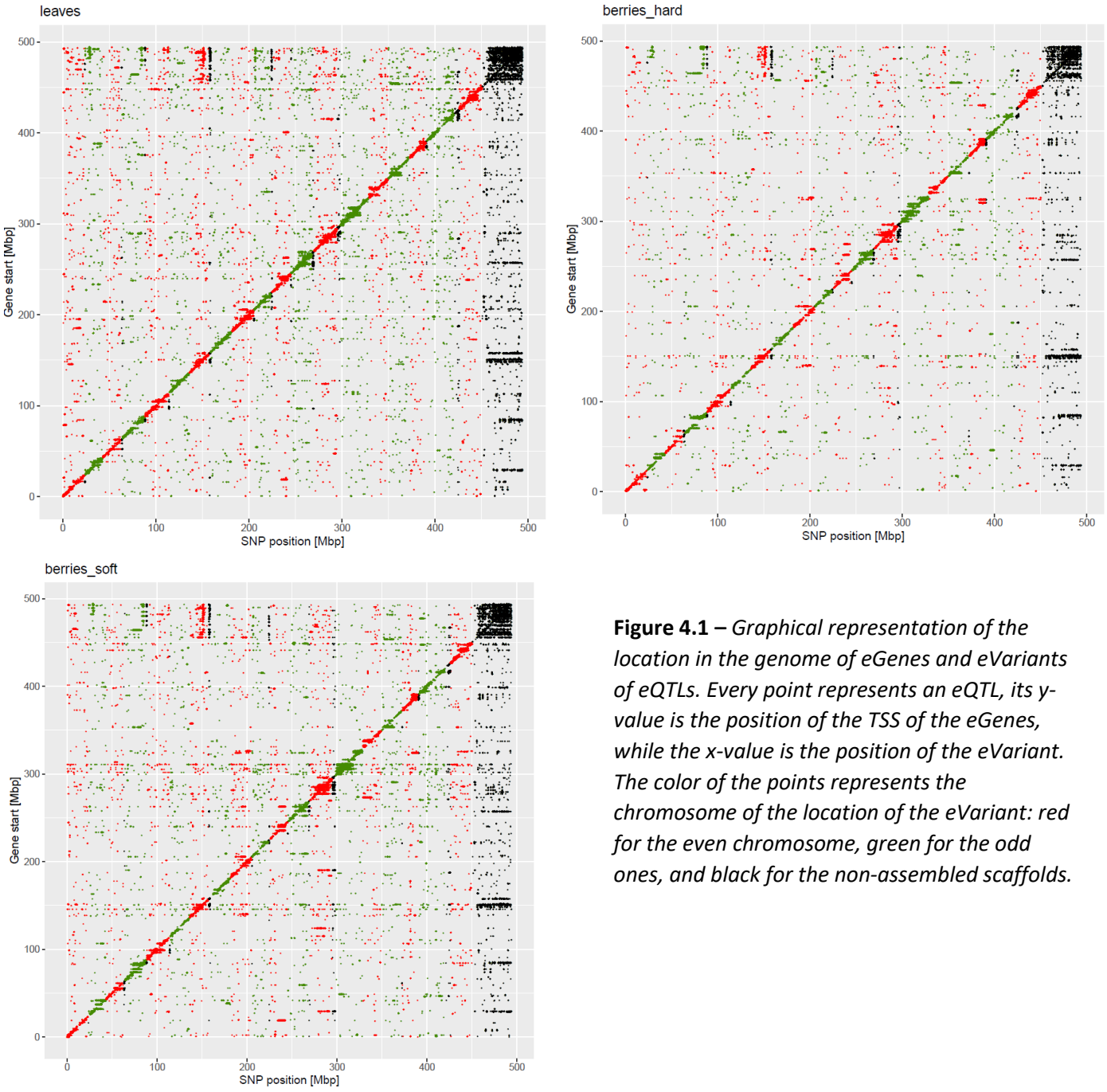

Figure 4.1 – *Graphical representation of the location in the genome of eGenes and eVariants of eQTLs. Every point represents an eQTL, its y-value is the position of the TSS of the eGenes, while the x-value is the position of the eVariant. The color of the points represents the chromosome of the location of the eVariant: red for the even chromosome, green for the odd ones, and black for the non-assembled scaffolds.*

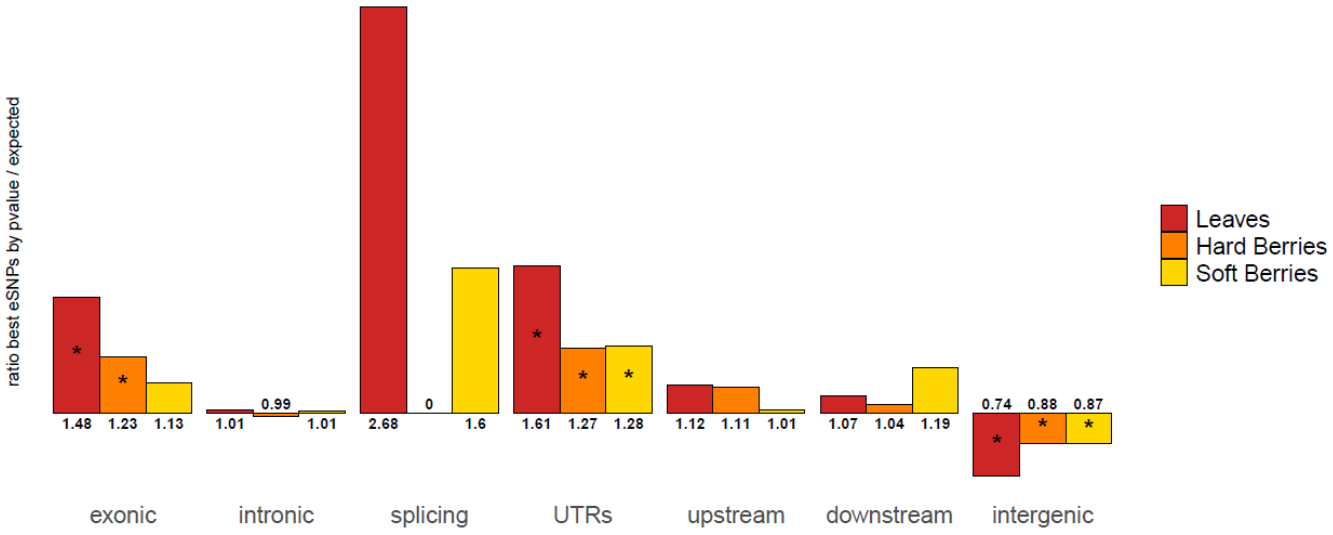

Figure 4.2 – *Graphical representation of the genomic distribution of the lowest p-value scoring eSNP per eGene compared with the distribution of all the SNP used as input in the analysis. Height of the bars and value at their bottom represent the ratio between best-scoring eSNPs number observed and expected, divided for gene regions and tissues. The asterisks on the bars mark the comparisons with a fisher test p-value < 0.01*

4.2.2 eGenes overview

Analysing the category of eQTLs established by the eGenes and their relative abundance, we must exclude all the eGenes situated in non-assembled scaffolds and eGenes correlated only with eVariants situated in said scaffolds (17% - 19% of the total). Most eGenes show *cis*- and *trans*- eQTL (41% - 48%). However, the biggest share of eGenes is the one with eQTLs only in *cis* (28% - 31%), followed by the eGenes with eQTLs in all the three categories (16% - 20%). Overall, the majority of eGenes have at least one cis-eQTL (64% - 75%). In this study, we call this group *cis*-eGenes, while the eGenes with only *trans*-eQTLs will be indicated as *trans*-eGenes and the genes that don't have any significant eQTLs as non-eGenes. If we compare the coefficient of variation of expression between the three groups (figure 4.3) we find that the *cis*-eGenes have higher levels than the non-eGenes (p-value of Wilcoxon test < 0.01), while the trans eGenes have an intermediate level of variation of expression, being significantly higher than non-eGenes and significantly lower than *cis*-eGenes in all the three tissues. These data are consistent with the assumption that the genes whose expression can be affected by variants are genes relatively dispensable in the plant, that can undergo a change in expression without compromising the fitness of the individual. It is reported that, in an expression network analysis, we will find few core genes among the eGenes, while a bigger number of peripheral and dispensable genes will be affected by genomic variants (Mähler *et al*., 2017). We can further confirm this observation with an

analysis of the d_N/d_S ratio of the genes, a measure of strength and mode of selection acting on genes (Jeffares *et al*., 2015). If eGenes preferentially consist of genes on which selection constraints are relaxed, we expect that they will have a higher ratio between the number of nonsynonymous substitutions per non-synonymous site (d_N), to the number of synonymous substitutions per synonymous site (d_S). This trend is confirmed: *cis*-eGenes, *trans*-eGenes, and non-eGenes establish three significantly different groups with the smallest values in the latter. A strong correlation between genes' variance in expression levels and d_N/d_S values was found in a population of *Vitis vinifera* by Magris *et al*. (2019), a result that goes in the same direction as the one presented here.

A comparison between the eGenes found in the three tissues can give us a measure of the tissue-specificity of this regulatory mechanism (figure 4.4). Unsurprisingly the tissues that share the bigger number of eGenes are the two berries (58% of the hard berries eGenes and 51% of the soft berries ones), but more than half of these are also eGenes in leaves, suggesting the existence of a group of gene consistently controlled by variants. Understandably the majority of leaves eGenes are unique of that tissue (67%). This is probably an effect of more than one factor: the higher sensitivity of the analysis compared to the berries ones, the differences in the set of cultivars used and, above all, the overall bigger diversity of the tissue compared to the difference between two stages of the same tissue.
The same patterns can be observed in the distribution of *cis*-eGenes.

To characterize the biological process affected by the regulation of genes by eQTL, we performed a functional annotation of the eGenes and highlighted the Gene Ontology category enriched in those lists (Supplementary material, tables S4, S5, and S6).
In leaves, the majority of the enriched categories are involved in response to external stimuli both biotic (21) or abiotic (5). Other notable categories are the production and regulation of primary metabolites (12), and nucleic acid replication and maintenance (5). In hard berries, two groups emerge: as in leaves several enriched categories can be connected to response to biotic (13) and abiotic stimuli (2); while the second group is made of categories related to the regulation of the cell cycle, duplication, and maintenance of nucleic acids (20). The soft berries list includes fewer categories related to response to biotic (3) and abiotic stimuli (3); on the other hand, there are 19 categories relevant to the regulation of the cell cycle, duplication, and maintenance of nucleic acids, 7 to production and transport of primary metabolite and 8 related to the transport of RNA molecules from the nucleus.

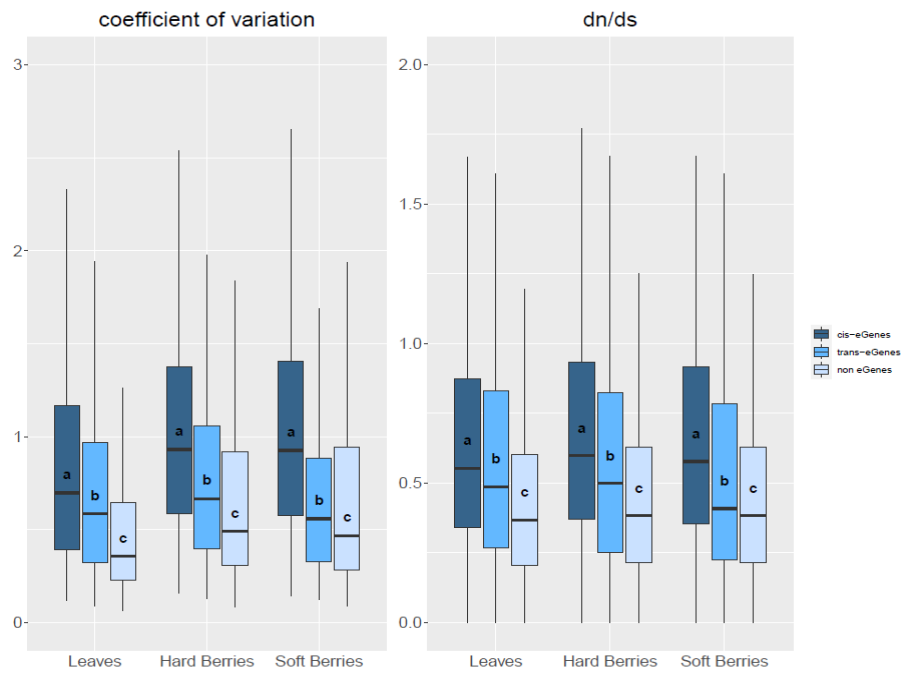

Figure 4.3 – *boxplots of the distribution of the coefficient of variation of expression and d_N/d_S values of Vitis genes, divided in tissues and categories of the eQTL analysis: eGenes with at least one cis-eQTL (cis-eGenes), eGenes with only trans-eQTLs (trans-eGenes) and eGenes without any eQTL (non eGenes). Different letters between bars indicate a significant difference between distributions calculated with a Wilcoxon test (p-value < 0.01)*

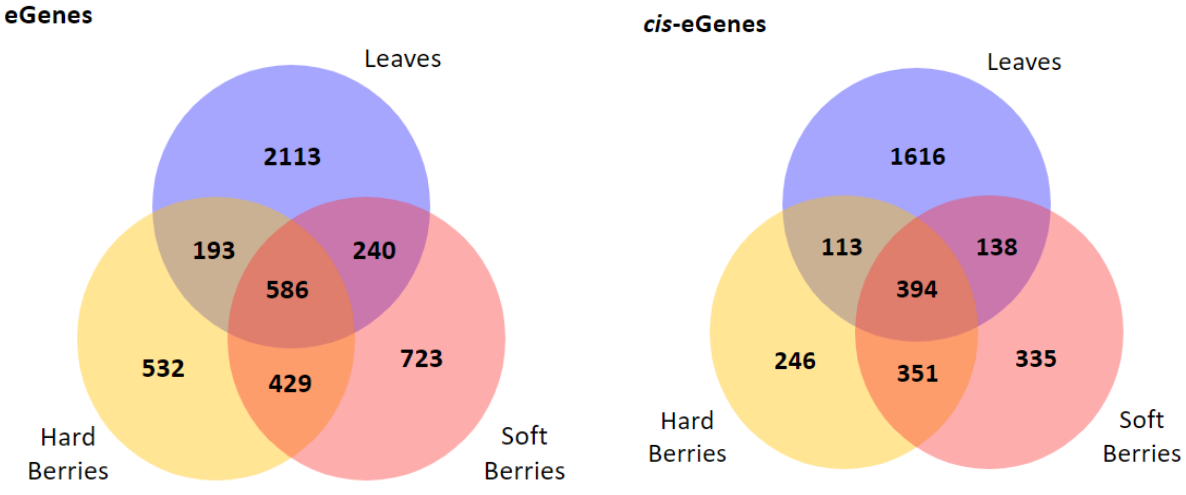

Figure 4.4 – *Venn diagrams with the numbers of eGenes (left) and cis-eGenes (right) for every tissue or shared between them*

4.3. ASE analysis

4.3.1 Overview and classification of genes

We used the reconstructed haplotypes to perform an ASE analysis on all the genes with the program ALLIM (Pandey *et al.*, 2015). In figure 4.5 we present a classification of all the genes in the different cultivars after this analysis. We classified a gene as homozygous in a cultivar when no heterozygous SNPs are identified in the transcribed portion or in a region between 1kb upstream of its TSS or 1kb downstream of its end. The percentage of homozygous genes in the cultivars varies from a minimum of 7.8% (Schioppettino) to a maximum of 25.8% (Henab Turki), the mean value across the 98 cultivars is 14.8%. If a gene presents a heterozygous SNP in the flanking regions, but not on its coding sequence, even if the SNP is causing allelic imbalance, we are not able to detect it, since the transcripts from the two alleles are identical. These genes are classified in the category "no exonic SNPs" (minimum Verdicchio Bianco and Savagnin Blanc: 12.0%, maximum Falanghina: 24.0%, overall mean 14.4%). All the genes suitable for ASE analysis underwent further quality control, as described in paragraph 3.4.3. Among the analysed genes, some do not present a significant allelic imbalance (p-value > 0.05), while we divided the others into three levels based on the measure of allelic imbalance (AI), calculated as described in paragraph 3.4.3.

The number of genes that display a statistically significant allelic imbalance can vary substantially across the cultivars in the same tissue. In the leaves samples, the minimum is 13.9% of the genes (Carignan), while the maximum is 30.4% (Nebbiolo), with a mean of 28.8%. The levels are higher in the berries' samples: the range is between 24.3% and 44.2% (respectively Aglianico and Heunisch Weiss), with a mean of 30.8% for hard berries, while in the case of soft berries the minimum is 18.5% (Falanghina), the maximum 43.8% (Heunisch Weiss again) and the mean 29.8%.

Taking into consideration only the genes analysed with ALLIM (figure 4.6) once again we see the difference between tissues. Nearly half of the genes didn't show any allelic imbalance in leaves (higher: Lambrusco Grasparossa 54.1%, lower Heunisch Weiss 35.6%, mean 44.4%), while in berries the proportion is lower and similar between the two: the mean across the cultivars is in both of 31.0%, with ranges only slightly differ: in hard berries from 39.5% (Ribolla Gialla) to 17.5% (Heunisch Weiss), in soft berries from 18.8 (Heunisch Weiss) to 39.9 (Falanghina). Some cultivars stand out for their high percentages of genes with high levels of allelic imbalance. Schioppettino shows a high share of genes with AI>2: 11.9% in leaves (mean 5.4%), 18.4% in hard berries (mean 7.7%), and 16.8% in soft berries (mean 7.6%), other cultivars have an overall high percentage of genes with AI >1 if compared with the mean level, (for example Kölner Blau and Heunisch Weiss in berries) but there is no other outlier as clear as Schioppettino.

4.3.2. Regulatory heterozygosity

With this data, we can estimate for every sample the level of what we called regulatory heterozygosity, an assessment of how much *cis*-regulatory diversity is present within each individual. The computation includes all expressed genes in each tissue, those where ASE could be estimated by ALLIM as well as those that were homozygous in each individual (where an ASE level of 0 was considered). In figure 4.7 we plotted the distribution of AI levels in the 49 cultivars in which we have samples for all three tissues. We include the genes analysed by ALLIM that pass the quality controls previously described. As evident in the boxplot, levels of ASE observed in the samples from the berries are higher than the ones from leaves in all the cultivars but one, Pignoletto. All the differences were significant according to Wilcoxon's test. In 19 cultivars the Wilcoxon test underlined a significant difference (p-value < 0.01) between the two berries samples, notably the hard berries' ASE levels are significantly higher in Cesanese d'Affile, Chasselas Blanc, Falanghina, Garganega, Bombino Bianco, Rkatsiteli, Sangiovese, Schioppettino, Sultanina, and Verdicchio Bianco. On the other hand, soft berries show higher ASE in Barbera, Cabernet Franc, Cabernet Sauvignon, Lambrusco di Sorbara, Merlot Noir, Montepulciano, Picolit, Riesling Weiss, Savagnin Blanc, Tribidrag, and Vernaccia S.G.
The overall higher levels of AI detected in berries compared to leaves are consistent with a precedent study of our group (unpublished data). Other works underlined the difference in ASE levels between different tissues (Cheng *et al*., 2021).

4.3.3. Comparison of ASE among tissues

We then assessed if the *cis*-regulation of gene expression detected with ASE was tissue-specific. For every cultivar, if present in more than one tissue dataset, we performed a Fisher test for every gene, comparing the number of reads in the two haplotypes and the two tissues. Being the null hypothesis of the test "the two samples have the same distribution of reads in the two haplotypes", in presence of p-values < 0.05 the genes have different regulations of the allele-specific expression. In table 4.3 we show, for every cultivar, the percentage of genes for which Fisher's test p-value is lower than an FDR threshold calculated with Benjamin-Hochberg's method (Benjamini and Hochberg, 1995), i.e. those showing tissue-specific levels of ASE. In the first column, the results concern a test executed with all three tissues, indicating the percentage of genes that undergo different haplotypic regulation, while the other three relate to comparisons of two tissues at a time, underlying genes with different behaviours between the two. As we can see in all the cultivars more than half of the genes show a difference in the regulation of allelic-specific expression (mean 63.7%). Unsurprisingly the comparison between the two berries tissues is the one with the lowest share of differentially regulated genes in most of the cultivars (mean: 35.0, mean of comparisons between leaves and berries: 45.1 for both hard and soft). Some cultivars make exceptions: the two berries are most different in

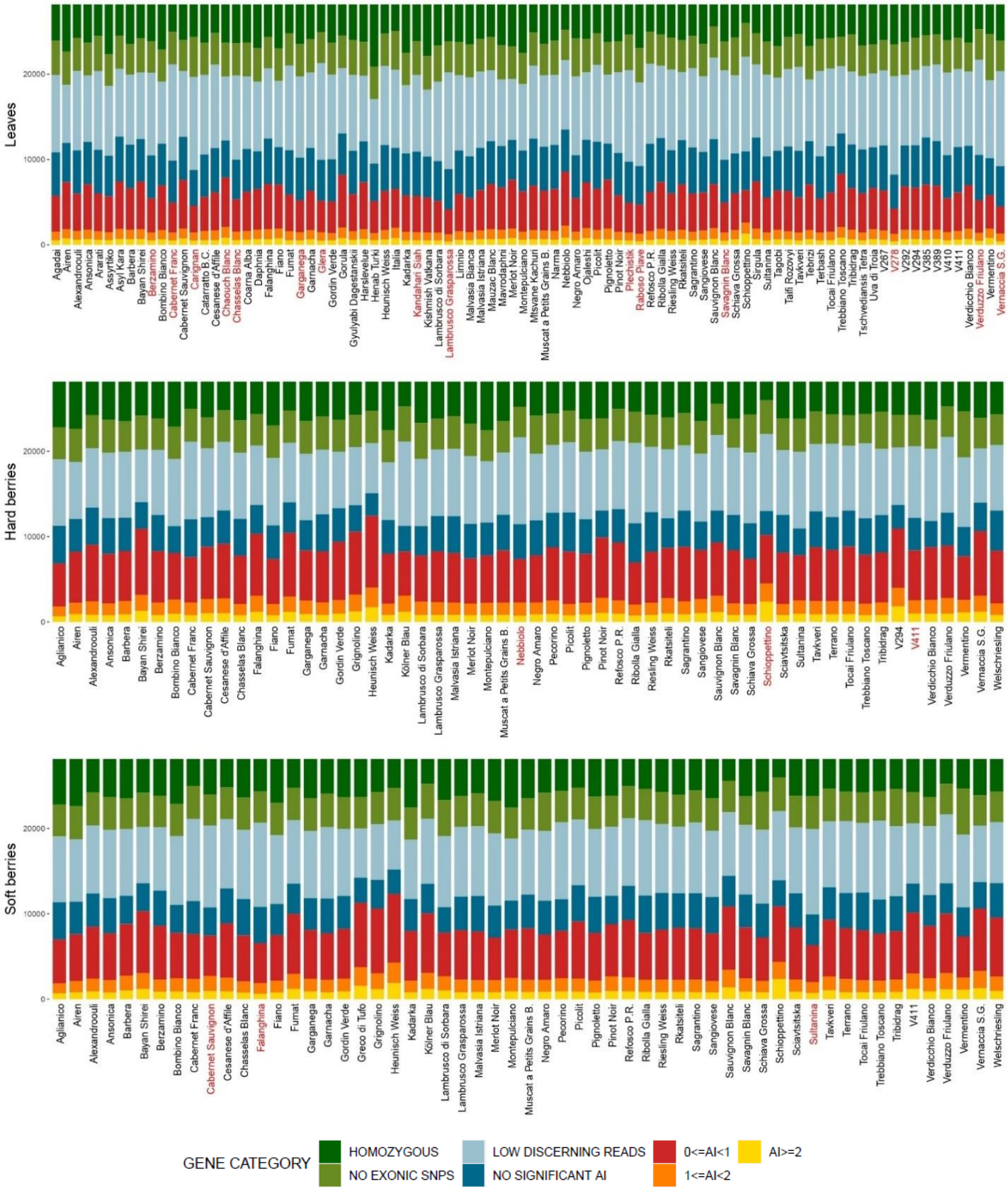

Figure 4.5 – *Number of genes in each cultivar and tissue belonging to the following categories: "homozygous" (without heterozygous SNPs from 1kb upstream the TSS of the gene to 1kb downstream the end of the transcribed region), "no exonic SNPs" (with heterozygous SNPs but not in exonic regions and therefore not analyzed by ALLIM), "low informative reads" (genes with less than 50 reads on exonic SNPs), "no significant AI", "0 ≤ AI < 1", "1 ≤ AI < 2", "AI ≥ 2. Red labels point out cultivars with only one replicate.*

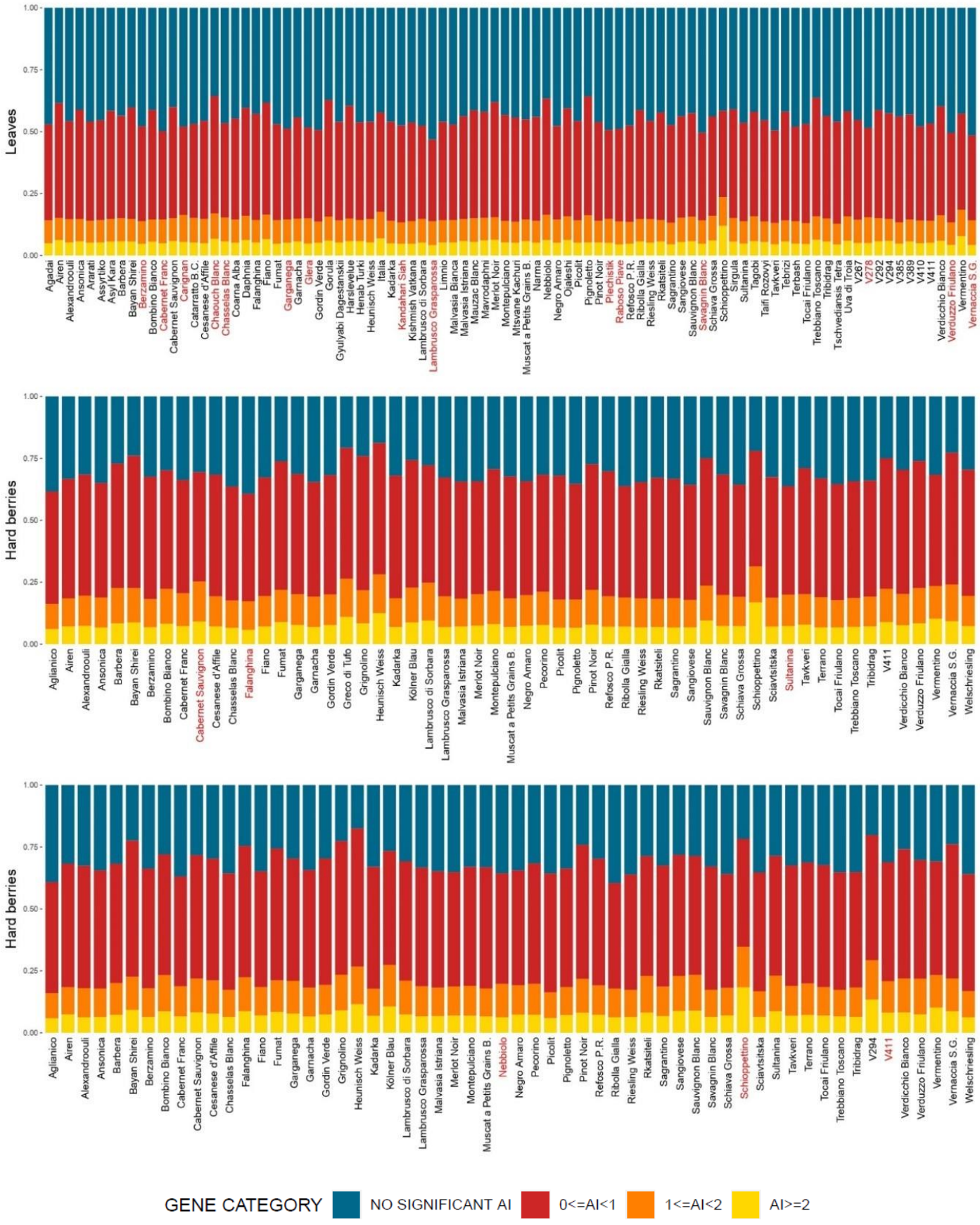

Figure 4.6 – *Percentage of genes belonging to different category in ALLIM analysis output. Red labels point out cultivars with only one replicate.*

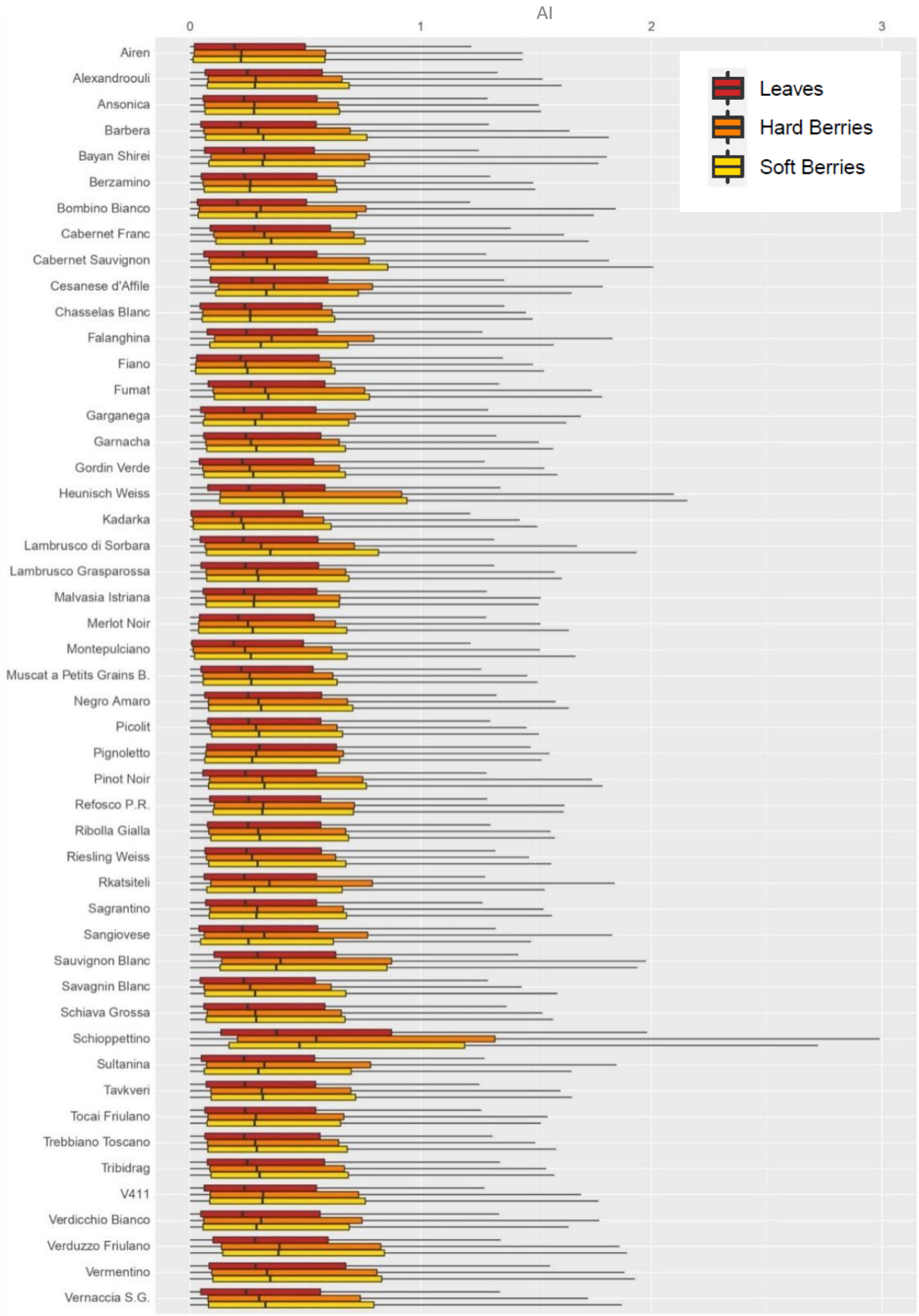

Figure 4.7 – *Distribution of AI levels of genes in cultivars with at least one replicate for every tissue. Both genes with significant AI calculated by ALLIM as well as homozygous genes expressed in the sample (to which a value of AI = 0 is attributed) are used in the graph.*

	three way test	Leaves vs Hard Berries	Leaves vs Soft Berries	Hard vs Soft Berries
Airen	61.3	48.7	46.6	24.3
Alexandroouli	64.4	43.8	43.6	39.8
Ansonica	61.8	45.8	45.4	31.0
Barbera	70.7	47.4	52.0	44.1
Bayan Shirei	75.8	55.0	53.3	50.9
Berzamino	56.1	41.1	41.1	23.3
Bombino Bianco	72.5	53.7	51.9	42.1
Cabernet Franc	60.4	38.3	39.8	35.2
Cabernet Sauvignon	76.3	53.7	55.8	49.2
Cesanese d'Affile	68.6	48.5	45.4	42.1
Chasselas Blanc	52.2	39.0	38.4	21.7
Falanghina	68.6	54.1	42.2	36.1
Fiano	63.4	46.5	48.9	31.4
Fumat	68.4	46.0	46.7	43.2
Garganega	62.8	43.0	41.1	38.1
Garnacha	57.2	42.2	43.2	24.4
Gordin Verde	57.3	40.9	41.4	30.5
Heunisch Weiss	82.7	55.8	54.2	65.1
Kadarka	57.4	42.1	44.1	23.6
Lambrusco di Sorbara	71.5	45.0	48.9	47.7
Lambrusco Grasparossa	59.6	36.9	38.4	35.8
Malvasia Istriana	55.3	41.7	41.8	21.1
Merlot Noir	64.6	48.1	49.4	28.7
Montepulciano	66.7	46.2	50.6	37.7
Muscat a Petits Grains B.	56.8	42.0	44.1	25.2
Negro Amaro	54.3	38.3	40.2	25.1
Picolit	55.7	40.9	43.1	21.4
Pignoletto	65.0	50.9	50.1	26.4
Pinot Noir	73.7	48.1	48.7	52.3
Refosco P.R.	59.6	42.6	43.4	29.4
Ribolla Gialla	55.5	41.5	44.9	17.7
Riesling Weiss	56.3	39.8	42.2	25.1
Rkatsiteli	69.5	51.5	46.4	42.5
Sagrantino	57.2	42.8	41.5	26.0
Sangiovese	66.3	49.0	40.7	41.4
Sauvignon Blanc	73.0	50.2	49.1	50.6
Savagnin Blanc	58.8	38.0	41.3	34.7
Schiava Grossa	54.3	40.1	40.0	22.3
Schioppettino	72.3	48.9	45.1	50.7
Sultanina	64.1	48.6	41.4	34.0
Tavkveri	62.6	41.3	42.5	37.3
Tocai Friulano	56.1	41.9	40.5	24.4
Trebbiano Toscano	61.6	47.5	47.8	25.0
Tribidrag	57.4	41.3	42.9	26.2
V411	70.6	46.4	48.7	44.6
Verdicchio Bianco	69.4	51.4	49.9	38.7
Verduzzo Friulano	68.8	41.8	43.8	47.3
Vermentino	56.5	40.6	42.0	25.6
Vernaccia S.G.	71.3	42.4	44.1	54.0

Table 4.3 - *Genes with allele-specific expression different between tissues. Column "three way test": percentage of genes showing significant deviations from expectation when comparing all the three tissues with a fisher's test. Columns "Leaves vs Hard Berries", "Leaves vs Soft Berries" and "Hard vs Soft Berries": percentage of genes showing significant deviations from expectation when comparing the two titular tissues.*

Heunisch Weiss, Pinot Noir, Sauvignon Blanc, Schioppettino, Verduzzo Friulano, and Vernaccia S.G. These results indicate the presence of a tissue-dependent regulation of ASE. This is consistent with the diversity detected between tissues in the eGene determination. Cheng et al (2021), although detecting an overall difference in AI levels between tissues, underline that individual genes have similar AI levels between tissues. However, in order to determine that, they divide the genes into categories according to their fold change values using as class limits 0, 2, 4 and 8, and then count the genes changing groups between tissues. This method is understandably less sensible than a Fisher test on the number of reads and doesn't register eventual cases of reversion of more and less expressed haplotypes in different tissues.

4.3.4. Genes with monoallelic expression

Allele-specific expression analysis highlighted a group of genes with a noticeable expression pattern: all the reads were attributable only to one of the alleles. We called these genes "monoallelic genes" (MAG) and counted their distribution among the cultivars with a sample for every tissue (fig 4.8). The median of the monoallelic genes is similar in the three tissues (35 in leaves, 31 in hard berries, 34 in soft berries). Unsurprisingly, samples with high median AI tend to have more monoallelic genes (for example Heunisch Weiss is the second among the cultivars in berries for both AI median values and monoallelic counts), but the two measures aren't always coupled as can be seen for example in Pignoletto and Sauvignon Blanc. A separate discussion must be made for the Schioppettino case. The number of monoallelic genes in this cultivar is particularly high in all the tissue and this could partly explain the high level of AI previously described in this cultivar, even if the low number of homozygous genes on this sample also concurs to raise the AI levels. We didn't find a biological explanation for this clear outlier, so we have to consider a possible technical one. Theoretically, if the sample of one tissue was mislabeled and belonged to another cultivar, we should see high numbers of monoallelic genes. That is because if we perform a haplotype-aware alignment between two different cultivars, for every gene for which the two share only one allele, the reads will be aligned only to that allele and none to the other, making it a monoallelic gene. However, a label switching of the tissue samples is highly improbable because the three belong to different biological samples and were collected independently. The genome sample may be mislabeled as Schioppettino, but in the early stage of this work the correspondence of cultivars between genome and transcriptome samples was checked (paragraph 4.1). Moreover, we executed the analysis with a mislabeled cultivar (an unknown cultivar assigned to Lambrusco Sorbara) and the number of monoallelic genes obtained was far higher than the one observed for Schioppettino (1492 in leaves). A possible explanation of the Schioppettino results is that in the tissue used for the genomic sequencing some contaminants from other cultivars were present. This could have caused the identification of a lower number of homozygous genes and a higher number of genes with

monoallelic expression. If the level of contaminants was low, this could have gone undetected in our analysis for cultivar correspondence.

We checked if the monoallelic genes also had a tissue-dependent distribution. To do so, for every tissue, we verify the ALLIM output in the other two tissues of its monoallelic genes. Even if it's not frequent that a gene displays a monoallelic expression in all three tissues (a mean of 10.3% in leaves, 27.1% in hard berries, and 11.2% in soft berries), the vast majority of MAGs are included in one of the following categories in other tissues: either have still a monoallelic expression, or an AI > 2, or have too few reads to be analyzed with ALLIM (Supplementary figures S2, S3, and S4). In conclusion, if in a tissue we observe a silenced allele, it's highly probable that in other tissues too we will observe a strong imbalance between the two alleles.

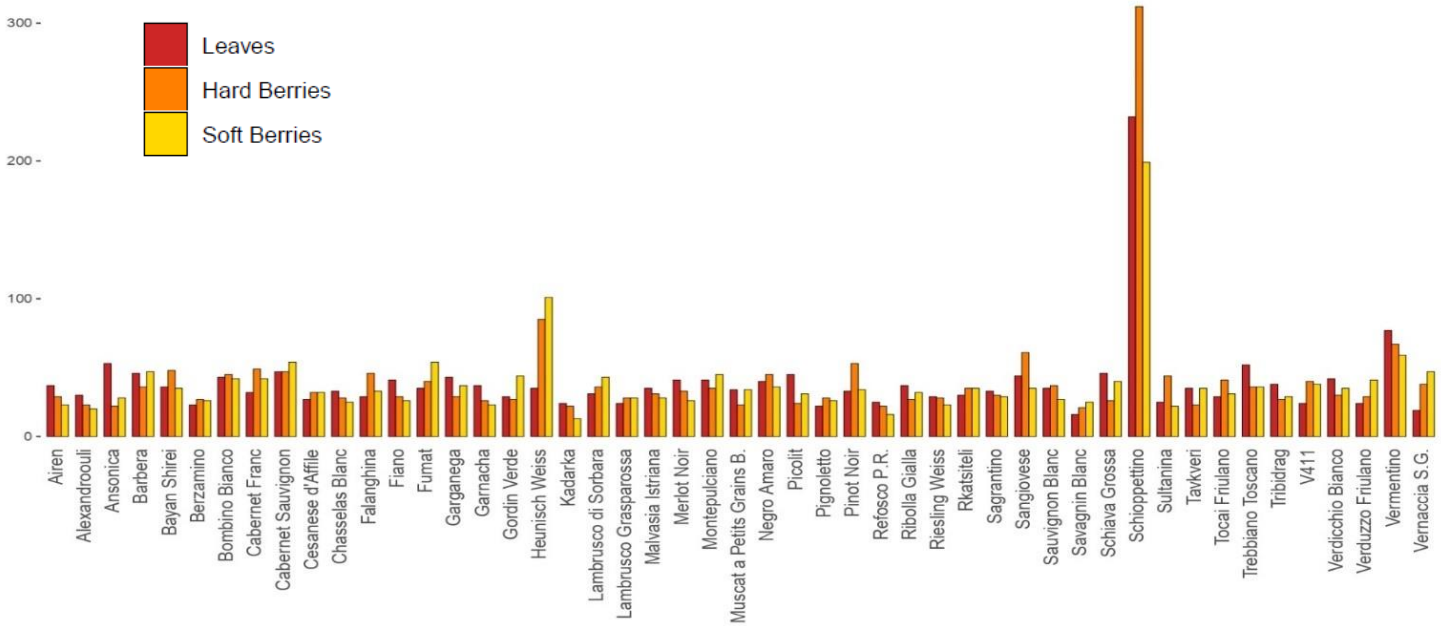

Figure 4.8 – *Number of monoallelic genes in cultivar with at least one sample for every tissue*

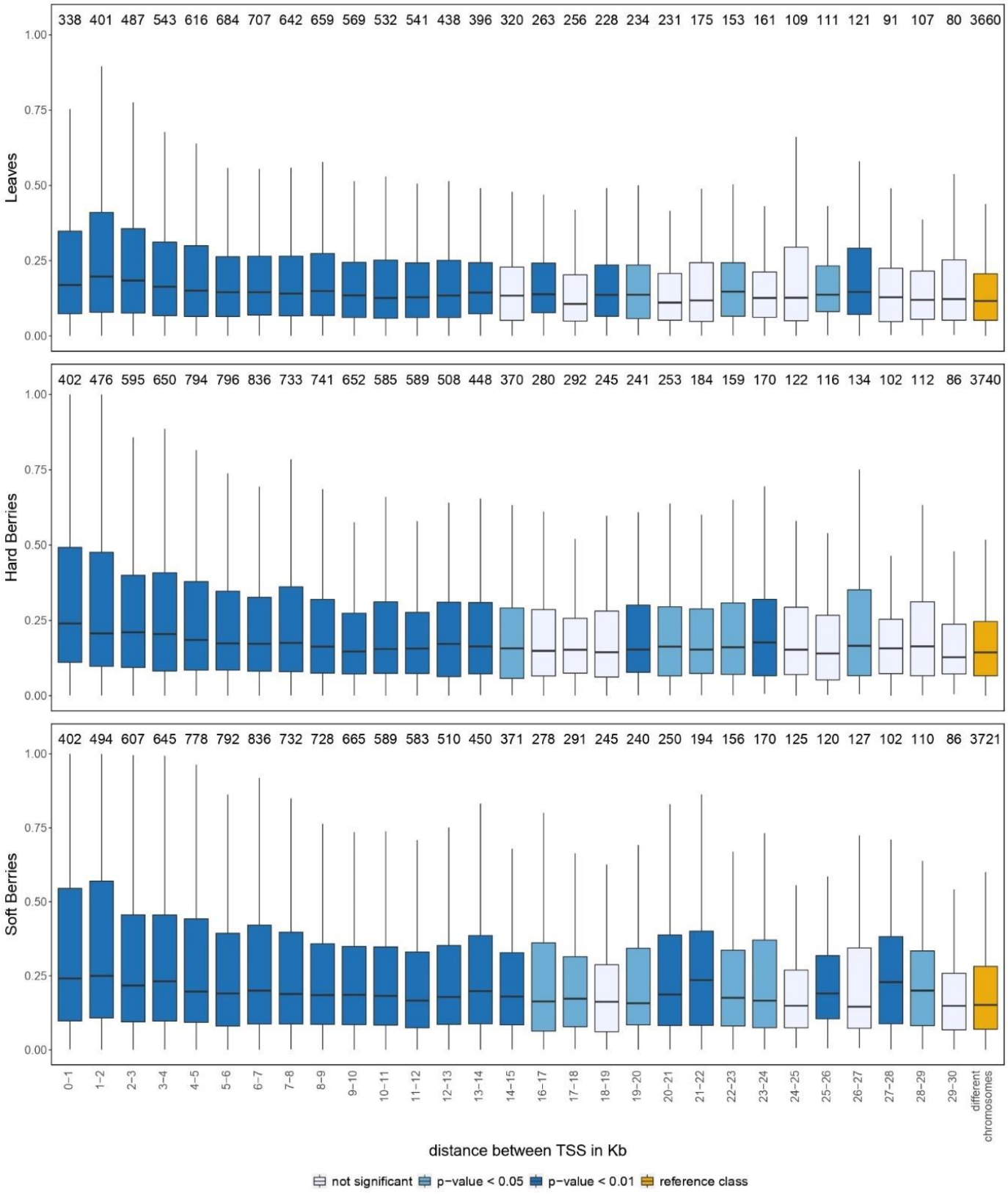

Figure 4.9 - *Correlation of allele-specific expression regulation between consecutive genes according to reciprocal distance. The boxes represent the distribution of the coefficient of correlation (in absolute value) between vectors constituted by AI values in different cultivars of two consecutive genes. On the x-axis the consecutive genes are divided in class depending on reciprocal distance, the labels above the boxes are the numerosity of the class of genes. The boxes are colored depending on the p-value of a Wilcoxon test between the distribution of correlation coefficients in that class and a similar distribution obtained from a population of a randomly selected set of gene pairs belonging to different chromosomes.*

4.3.5. Correlation of ASE values in contiguous genes

Since a different expression of two alleles has its roots in heterozygous *cis*-elements that regulate differentially the two haplotypes, we tried to assess if these elements may exert their action not on a single gene but rather on multiple physically close genes by testing if physically proximal genes have an allele-specific expression pattern more correlated compared to two randomly selected genes. Figure 4.9 summarizes the distribution of the correlation coefficient between AI values of a given pair of genes in each cultivar. Consecutive genes were divided into classes according to the distance of the two TSS and compared, with a Wilcoxon test, with a distribution of correlation values from randomly generated couples of genes belonging to different chromosomes. In all tissues, the statistical test gives a p-value lower than 0.01 to continuous genes with a mutual distance of up to 14 kbp. Moreover, the medians of the distribution of each class tend to decrease with each increment of the TSS distance. A notable exception is, in leaves and soft berries, the median of class 0-1kbp lower than the one belonging to class 1-2 kbp, however, this doesn't surprise us, as this class is probably enriched for genes transcribed on opposite directions, with TSS that are near each other, but the rest of sequences are diverging. Class of pairs of genes with distances from 14 to 27 kb in leaves and hard berries and to 29 kb in soft berries, sometimes are significantly more correlated than the reference class, but give us more uncertain results, due to a lowering of the median of the distribution of the class, but also due to the lower number of the population of genes available.

These results are consistent with the hypothesis that the allele-specific regulation of genes has a local effect, and that more than one gene could be affected by the same mechanism of haplotype regulation. This finding is consistent with the result of a similar analysis carried out in *Vitis vinifera*'s berries by Magris *et al*. (2019). In that work, the authors measured the correlation of expression levels of consecutive genes dividing them into groups according to the distance of their TSS and compared them to values relative to couples of genes randomly paired. Consecutive genes showed significantly higher correlation values up to a TSS distance of 27 kb.

4.3.6. Gene ontology categories enrichment in genes with strong evidence of ASE

Lastly, this data gives us the opportunity to investigate if some genes more than others are subject to *cis*-regulatory effects and, if so, which characteristics they have. We selected the genes that, among all the cultivars in which they have been tested with ALLIM plus the cultivars where the gene is homozygous, have 80% or more of the results with significant AI. On these genes, we performed a functional annotation and selected the gene ontology categories enriched according to a Fisher's test (Supplementary material, tables S7, S8 and S9). Even if the results vary among the tissues, we can aggregate the GO tags in macro-category as: reaction to biotic or abiotic stress or stimuli (14 in leaves and hard berries, 13 in soft berries),

transmembrane transport, especially of ions and water (11 in leaves, 7 in hard berries, and 3 in soft berries) and most importantly biosynthetic and metabolic process of primary metabolites. It is worthy of mention that the type of metabolite varies among tissues: in leaves (29 categories) are primarily carbohydrates, lipids, waxes, and cellulose; in hard berries (37 categories) we find some carbohydrates and lipids, but most of all amino acids and ribonucleotide related compounds; in soft berries (43 categories) in addition to the above mentioned ones there are also categories related to glycolysis and oxidative phosphorylation.

We performed a similar analysis for the monoallelic genes in each tissue (Supplementary material, tables S10, S11, S12). In leaves most of the enriched GO categories are related to a response to an external stimulus (24), but we find also 8 categories related to secondary metabolism, a process we didn't find enriched in the other cases studied here. In berries is more difficult to find common features between GO terms: the categories related to response to external stimuli are 9 in hard berries and 8 in soft, the ones related to primary metabolism 6 and 5, respectively, and the GO terms attributable to a secondary metabolism process are 4 in hard berries and 8 in soft. A similar analysis was performed in *Zingiber officinalis* by Cheng *et al*. (2021), and the pathways more represented were related to resistance (response to toxic substance, terpenoid biosynthetic process and alkaloid metabolic process)

4.3.7. Comparison between eQTL and ASE analysis results

Using the same criterion mentioned before (percentage of cultivars where a gene has significant AI) as a proxy of how much is a gene subject to allele-specific expression regulation, we can do a comparison between the ASE behavior of a gene and its status of *cis*-eGenes or not. As discussed in the introduction, the two measures have complementary blind spots and even if a *cis*-regulation of the expression of a gene can be effectively measured by both, there are many cases where only one of the two is able to detect and assess the phenomenon. This is well described in figure 4.10. The more are the cultivars in which a gene shows an AI, more likely it is that it has a *cis*-eQTL, even if the majority of the genes in all the categories are not identified as eGenes. In the graphic, we can also appreciate the better sensitivity of the eQTL leaves analysis. The proportion of non tested *cis*-eQTL is higher in the group of genes with a low percentage of "AI-cultivars". This is probably due to the fact that in these groups there is a bigger share of homozygous genes, likely to be discarded as non-informative or with heterozygous genotypes too rare for the eQTL analysis (and filtered out as in paragraph 3.4.3.). Moreover, a decrease in significant ASE cases is expected in genes with overall low expression, and therefore not tested in eQTL analysis: in these cases it will be more probable that moderate AI will be assessed as not significant, due to the small sample size of informative reads. The same phenomenon can explain the slightly higher share of *cis*-eGenes in the class of genes with AI in less than 20% of the tested cultivar. Still, if the genotypes heterozygous and alternative homozygous pass the filter, there will be few points for the linear model of the eQTL analysis, and in this situation can occur the risk of high p-

value due to low sample size.

The difficulty to cross the results of the two analyses was registered by other authors. For example, Khansefid *et al*. (2018) in their work on *Bos taurus* noticed that, although the eQTL mapping and the ASE analysis gave similar results in terms of the number and effect of *cis*-acting variants described, only half of the cases were identified by both the analyses.

A second approach could be to assess if genes with high ASE regulation have low selective constrain, sharing this feature with eGenes. As we can see in figure 4.11 there is a positive correlation between the d_N/d_S index of the *Vitis* genes and the mean AI measured across all samples. Similarly, genes with high d_N/d_S show significant AI in more cultivars than the others. The same correlation was observed by Cheng *et al*. (2021). This is an indirect observation, but we can conclude that both eGenes (and particularly *cis*-eGenes) and genes with frequent or high AI regulations are genes with fewer selective constraints.

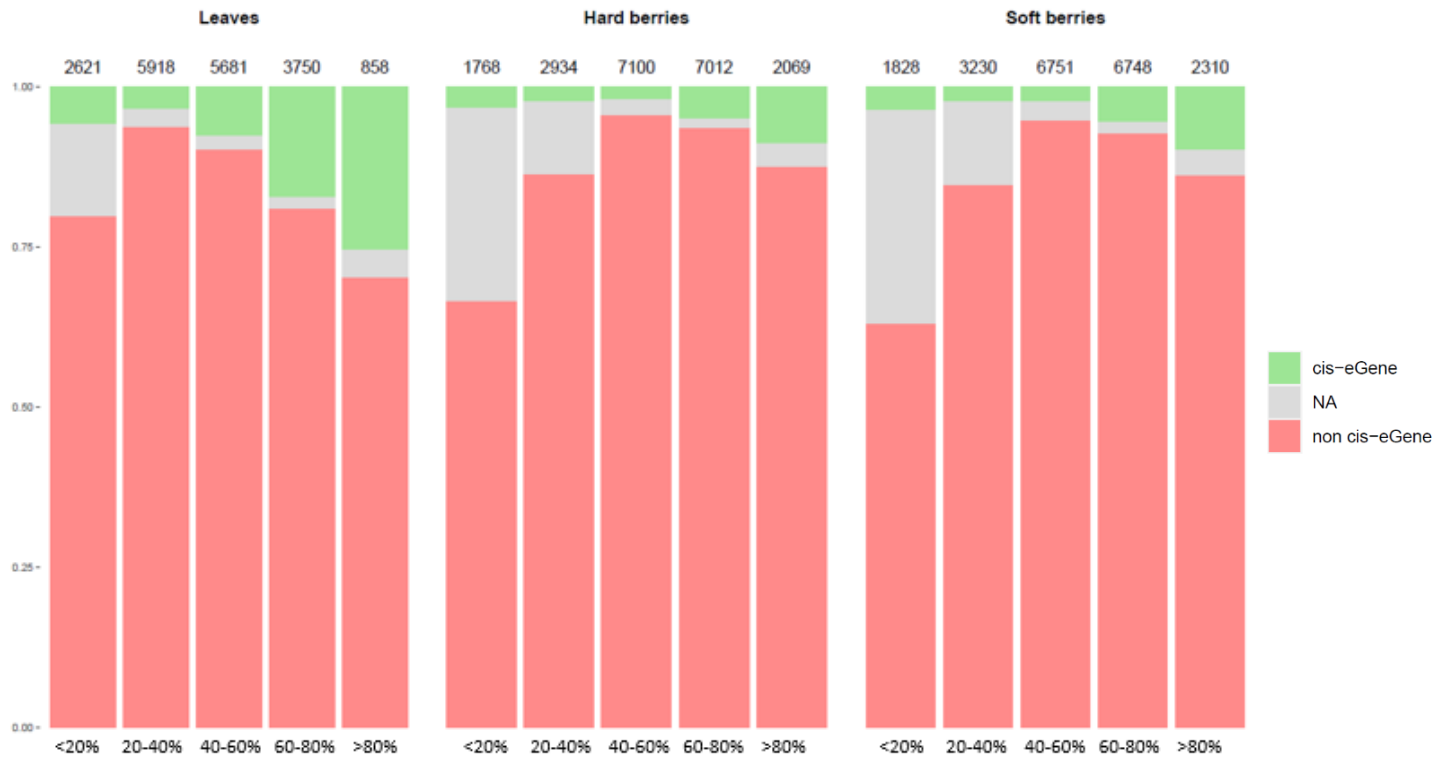

Figure 4.10 - *Partition of genes in categories from ASE analysis and eQTL analysis. The different bars in the three tissue contain genes divided according to how many cultivars that gene recorded significant AI (percentage of significant Ai cultivar on total ALLIM record for the gene plus homozygous cases). The numerosity of the group is indicated at the top of the bar. The colors of the bars represent the eQTL classification between* cis*-eGene, non* cis*-eGene (that means that have no eQTLs or only trans-eQTLs) and are not analyzed by matrix-eQTL (low expression gene).*

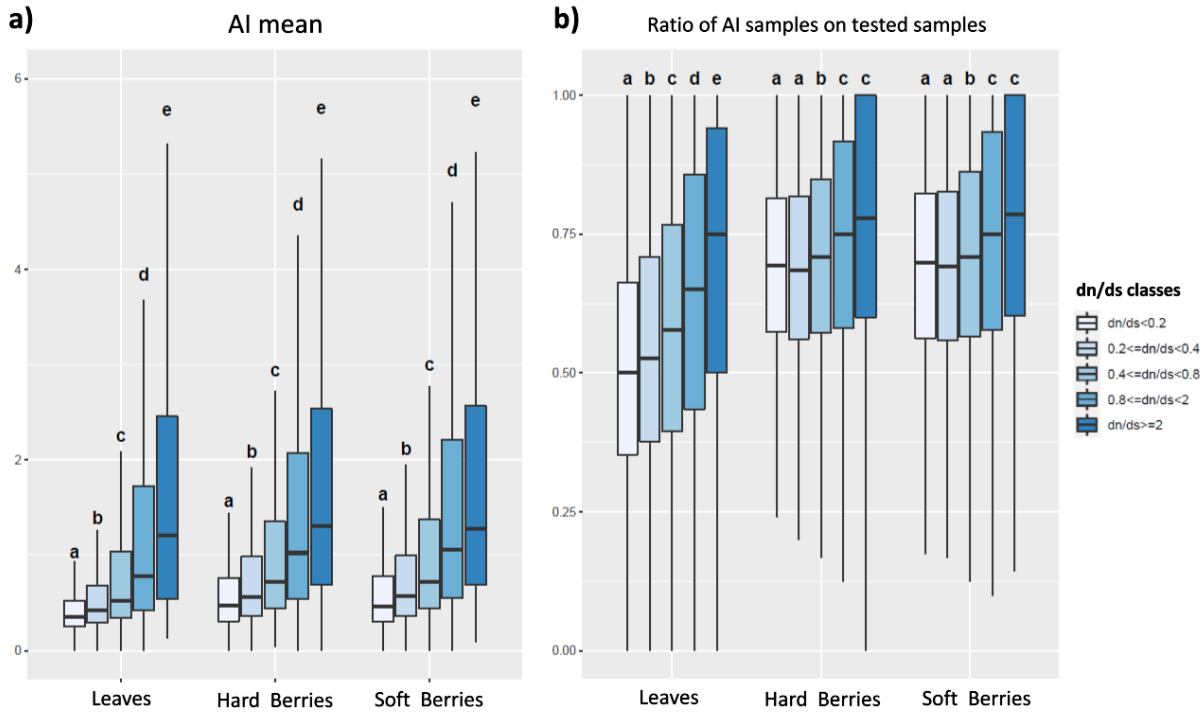

Figure 4.11 – *boxplots of AI genes metrics observed in 5 groups of genes divided according to d_N/d_S values. a) AI mean of the gene in all the observed AI significant value across the cultivar. b) ratio between the number of sample where the gene shows a significant AI on the total of sample tested for that gene (samples analyzed by allim + sample homozygous for that gene). In both a and b different letters between bars indicate a significant difference between distributions calculated with a Wilcoxon test (p-value < 0.01)*

4.4. Analysis of the allelic variant population of the genes

If we aim to understand how the expression of a gene is regulated by genetic variation in a population, both the eQTL and ASE analyses are, as presented, useful to investigate the phenomenon. Nevertheless, both methods have some shortcomings.

With eQTLs mapping we are able to investigate the different regulations of a gene through different samples, catching a measure of its variation in the population for the selected tissue. On the other hand, we can observe the variation of the gene only in relation to a singular biallelic variant at a time, independently from the others. As a result, we can't detect the effect of the combination of more than one variant on the expression of the gene.

With the analysis of allelic-specific expression of the genes we take into consideration all the differences between two haplotypes that have an effect on the two alleles of a gene, but it can be done only one sample at a time, comparing a single pair of alleles. Consequently, it's difficult to make assumptions about the regulation of the gene in the population.

To overcome these drawbacks, we investigated the ASE results taking into consideration one gene at a time, across all the individuals analysed for each tissue. Our aim is to estimate the relative *cis*-regulatory value of each gene allele observed in the population of analysed individuals by combining the information obtained from all the heterozygous combinations in which that haplotype is observed. This would allow us to infer the number of different *cis*-regulatory alleles present in the population and quantify their relative expression differences. An important assumption behind this analysis that may not always be met is that when the same genic allele is found in different individuals it is always linked with the same set of cis-regulatory variants, i.e. that LD extends well beyond the genic region to the flanking regulatory regions. This assumption is somewhat supported by the very frequent observation of *cis*-eSNPs over long distances (see also the previous discussion on the high frequency of *trans*-eQTLs *intra* chromosome). The estimation of the *cis*-regulatory value of each allele would also allow us to extend population genetics concepts such as those of expected and observed heterozygosity that are applied to sequence variation to regulatory variation.

4.4.1 Estimate of the impact of haplotype variants on expression

With the data obtained from the haplotypes of each cultivar, we compiled a comprehensive list of all allelic variants present for each gene in the population and coupled every cultivar with its two alleles (or singular allele if homozygous). We were able to obtain a haplotype collection for 26 336 genes, with a median of 22 alleles per gene (minimum 2, maximum 164).

Integrating this information with the ASE data, we measured the average relative contribution of each haplotype to the expression of the gene. The procedure is briefly described in paragraph 3.5.2. Figure 4.12a

shows the estimated net contribution (from now on ENC) of the individual alleles on an example gene. As we can see, we obtain a ranking of all the allelic variants and are now able to compare haplotypes that are never observed together in a genotype. To assess the robustness of the model we can calculate the correlation between the values of AI measured with ALLIM and the values expected for that pair of alleles using their ENCs (figure 4.12b). This example shows a high level of variation in expression among haplotypes with the most extreme haplotypes showing an approximately 64 fold difference (2^6) in expression and a range of haplotypes with intermediate levels.

In some cases, we saw that the correlation values were low and our model not properly fitting the gene's ALLIM data. We will now briefly discuss these cases, in order to identify the possible causes.

The first common case is the one represented in figure 4.13a. It is noticeable that in the correlation's scatterplot multiple observed AI values correspond to a single estimated AI value, and that these observed AI have high variability, resulting in several points vertically aligned. Such a situation occurs when we observe in the population many times the same genotype, but the AI values from those samples have high variability. This could be caused by one or more of the following: a) the two haplotypes are identical in the known sequence, but have in various samples different ASE levels determined by a variant that hasn't any

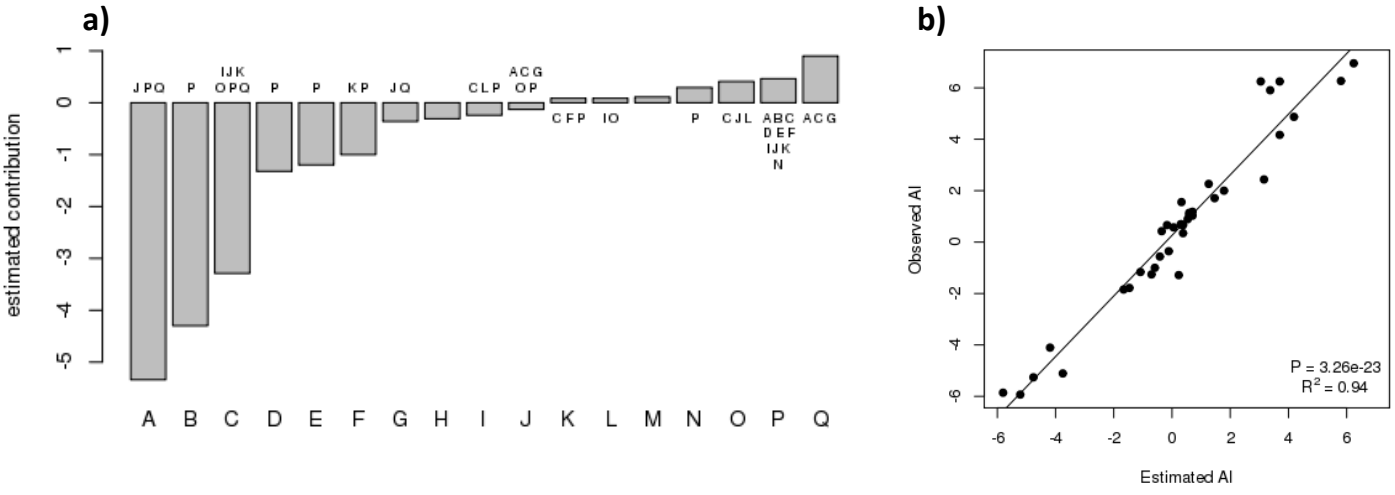

Figure 4.12 - *Graphical representations of the alleles' ENC for gene VIT_205s0102g01070 in leaves*
a) *Estimated relative levels of the contribution of each allele to the expression. The scale on the Y axis is a log2 scale where the position of the 0 is arbitrary. The letters along the X axis identify the different alleles for which the expression contribution is estimated. The letters along the basis of each bar indicate the heterozygous genotypes involving that allele where significant AI was detected.* b) *Scatterplot of the linear model assessing the quality of the alleles' regulatory value estimate. Each dot corresponds to a different heterozygous genotype analysed. Values on the Y axis represent observed AI values in every genotype, while values along the x axis are the estimated AI values obtained from the model.*

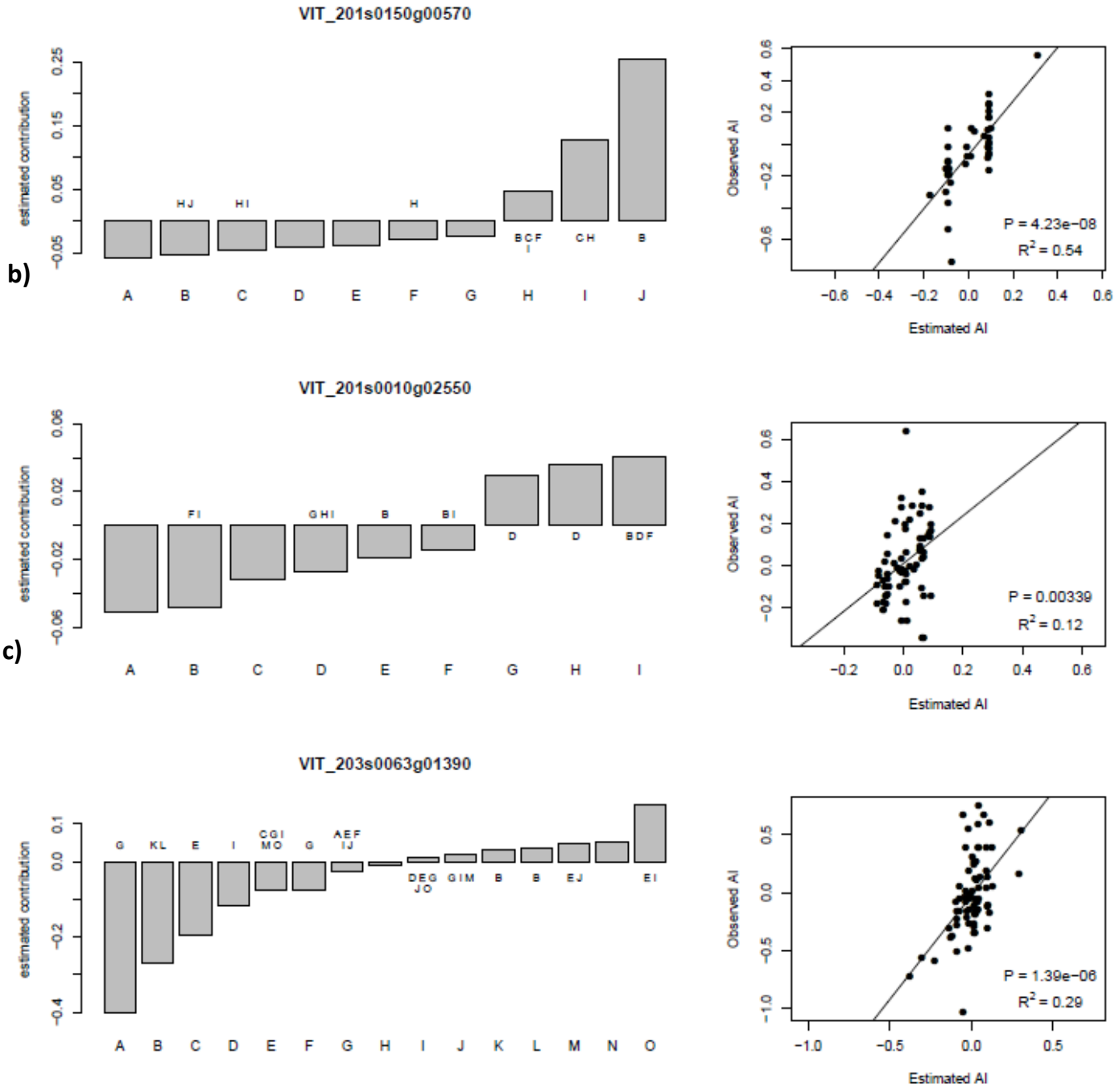

Figure 4.13 - *Graphical representation of the AI estimates values for three genes (left) and linear correlation between observed and estimated AI (right). The analyzed genes are three examples where we obtain low correlation values.*

proxy SNP in the gene exons, or which proxies are in introns, flanking regions or are not correctly identified in the SNP calling phase. In other words, those samples have not in fact all the same genotype, but two or more allelic variants were mistakenly assigned to the same allele in our model. b) The samples have in fact all the same genotype and *cis*-variants in the considered locus, but the different ASE levels are caused by some variant that acts in *trans*, through a diffusible factor, that regulates differently the expression on the two haplotypes of the individuals (figure 4.14). Since this hypothetical *trans*-variant is not present in all the samples, the ASE values differ significantly among them This phenomenon was observed by Magris *et al.* (2019) on a small fraction of *Vitis vinifera* genes. c) An error, that occurred in the haplotype calling step, resulted in the calling of the wrong allele for the samples, that were so assigned to the wrong category.

The second example of bad fitting of the model is represented in figure 4.13b. In these cases, the magnitude of the observed AI is probably too little and the variability between observed AI between samples with the same genotypes, even if small in absolute terms, impact significantly on the calculation of ENCs, thus invalidating the model.

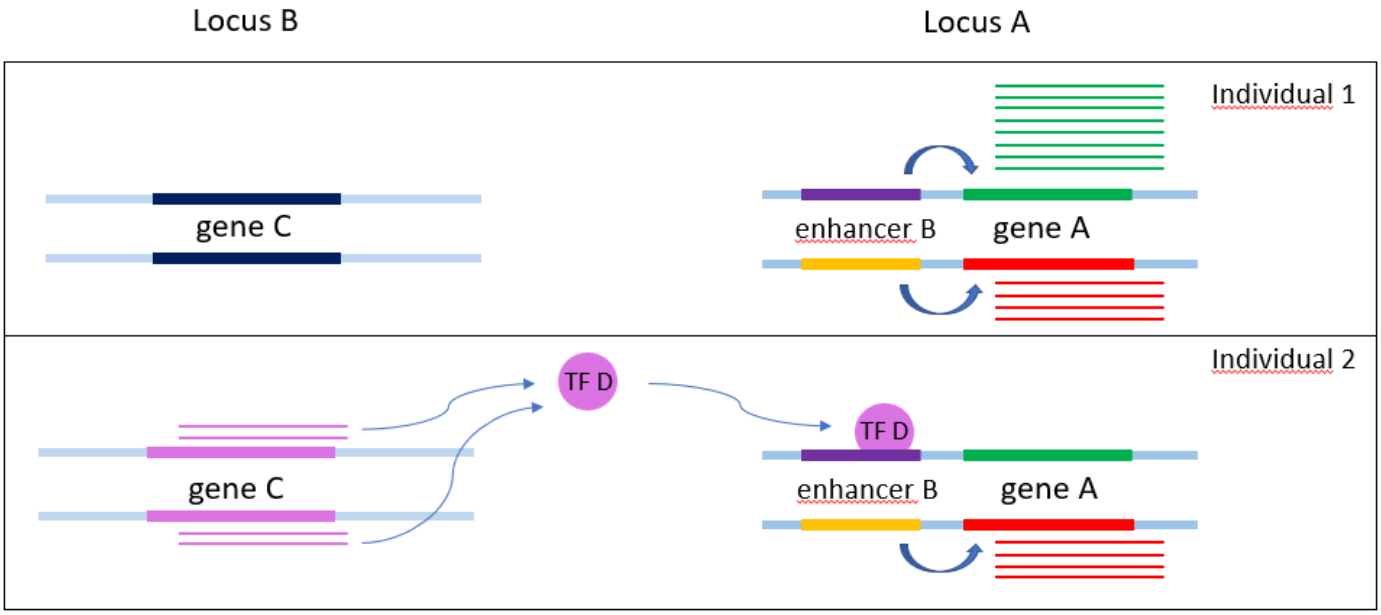

Figure 4.14 - *Graphical representation of a hypothetical case where two individuals with identical genotypes in a locus have different AI levels caused by a variant acting in trans. Individuals 1 and 2 have identical genotypes in Locus A, in individual 1 the two different enhancer B variants cause different expression levels of Gene A, resulting in the green allele being more expressed than the red one. On the other hand, in individual 2 the gene C, situated in a different locus, produces a diffusible factor (transcription factor D) that binds only with the purple variant of enhancer B, inhibiting its enhancer function and thus acting as a repressor of Gene A's expression. The result is that in individual B the red allele of Gene A is more expressed than the green one*

Figure 4.13c depicts the example of a gene where the model wasn't successful, but from the data, we cannot indicate with confidence a probable cause. We can speculate that in these cases some errors in the upstream steps have occurred. As said, it's possible that some SNPs weren't properly called, or that not all the haplotypes were correctly inferred. Another possibility is that in cases of a family of genes with similar sequences some reads were assigned with the wrong gene, conditioning the AI calculations.

In all the described cases we cannot trust the ENCs values, so for the remaining discussion of the results, we will not consider genes with low values of the correlation test, putting a threshold on Pearson's coefficient of 0.6. However, if the correlation values are regarded as indicators of the quality of the estimated AI in the genes, we must take into consideration some factors that could inflate them, invalidating our assumptions.

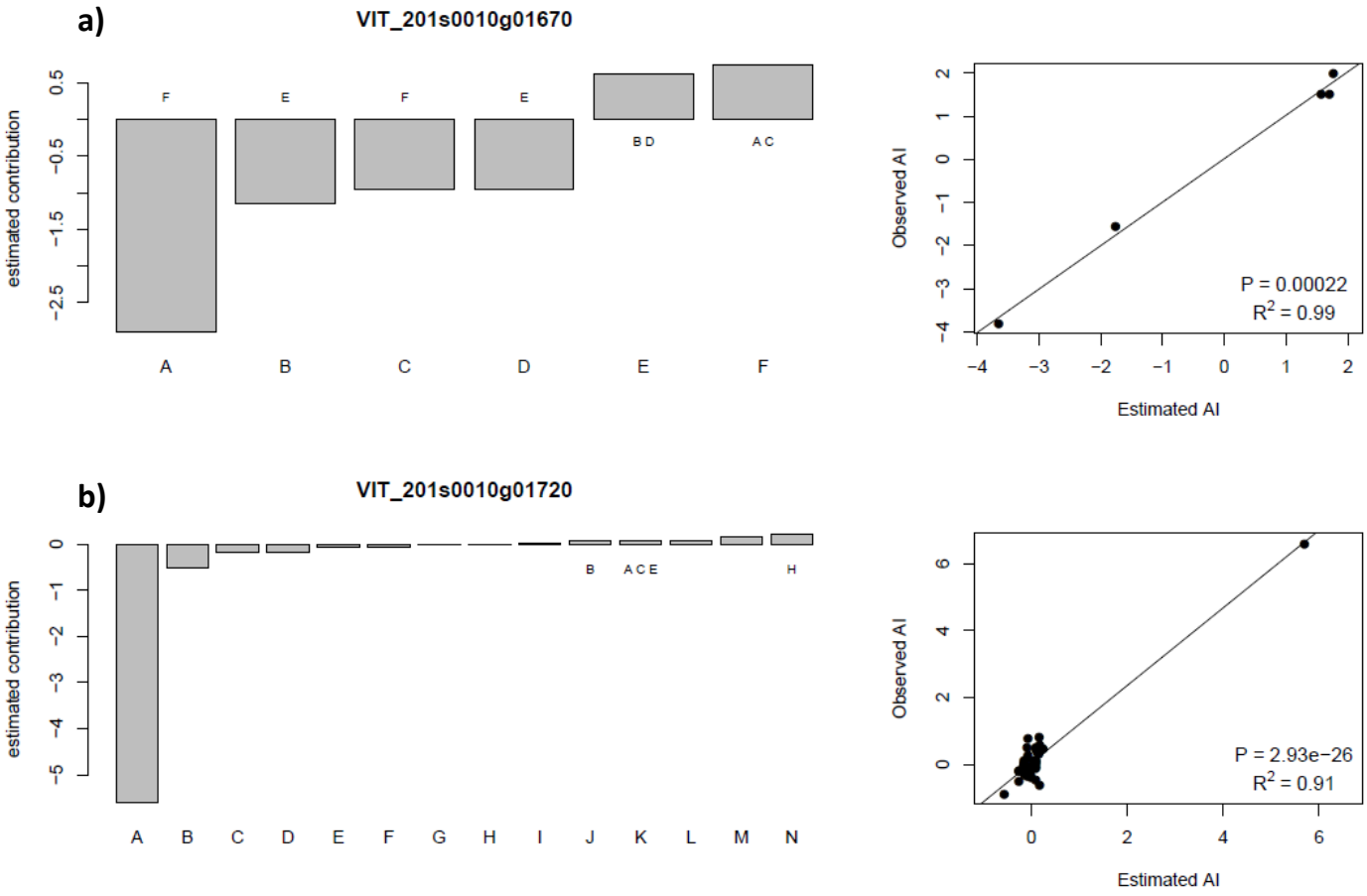

Figure 4.15 - *Graphical representation of the AI estimates values for two genes (left) and linear correlation between observed and estimated AI (right). The analyzed genes are two examples where the correlation values could be inflated.*

First, in case of a great number of alleles observed in the genes and few cultivars with ALLIM output to put in the model, it may occur that one haplotype is observed only once, and in this case the expected relative expression will by definition coincide with the observed relative expression, ultimately inflating the r^2 estimate (Figure 4.15a). Since this is a problem relative to the sample size of the analysis, the only possibility to preserve the accuracy is to filter out the genes with the ratio between the number of estimated alleles in the population and the number of samples used in the model lower than 1.5.

A second inflation factor is given by the outlier for observed AI values in a population and is well depicted in figure 4.15b. As we can see in the example given, the allele A, observed one single time, gave an ENC far lower than all the other alleles. As it is well known, Pearson's correlation coefficient is very sensitive to outliers and this can lead to false results (de Winter *et al*., 2016). A possible solution is to use Spearman's correlation coefficient when assessing the robustness of our model, since is less prone to outlier-driven inflation (*ibidem*).

In conclusion, we selected for every tissue all the genes that follow these criteria: the number of samples used for the AI estimates higher than 1.5 times the total of alleles present in the population for the gene, Pearson's correlation coefficient higher than 0.6, and Spearman's correlation coefficient higher than 0.7. We retained 4488 genes in leaves, 3210 in hard berries, and 3562 in soft berries. The lower number of genes in berries tissues is expected since the starting cultivar population in these tissues is smaller.
A total of 1165 genes pass the filter in all three tissues and comparing them could give us an indication about the behaviour of the gene alleles in different tissues. First, we compared the magnitude of the ENCs of the alleles: for every gene, we calculate the difference between the lower and the higher ENCs in every tissue and then compare them. We counted the genes with a difference of more than 0.5 between the two ENC range (that means that the range displayed by one is 1.4 times the other, since the ENC are in a log scale), obtaining 532 genes with different ENC range between leaves and hard berries, 539 between leaves and soft berries, 260 between the two berries.

For a more accurate comparison between ENCs of the same alleles in different tissues, we performed a Pearson correlation test for every gene comparing the ENCs values for the alleles present in all the tissues and then counted the genes with a coefficient higher than 0.5 and with p-value < 0.05. Among the 1165 genes, 635 pass these thresholds in the comparison leaves vs hard berries, 624 in leaves vs soft berries, and 1132 in hard vs soft berries. The low number of genes with a similar pattern of ENCs in leaves and berries tissues is probably partly caused by the difference of allele population between the tissues but is fully consistent with the tissue-dependent variability of ASE previously observed in this work. In figure 4.16 we present an example of a gene with substantial differences between ENCs in the tissues, even if the allele

population is identical among the three. It's difficult to estimate the frequency of the phenomenon, due to the drawbacks previously described and all the filters applied to eliminate the possibility of false positives but browsing the results we can state that significant shifting of an ENC value of an allele between tissues is common

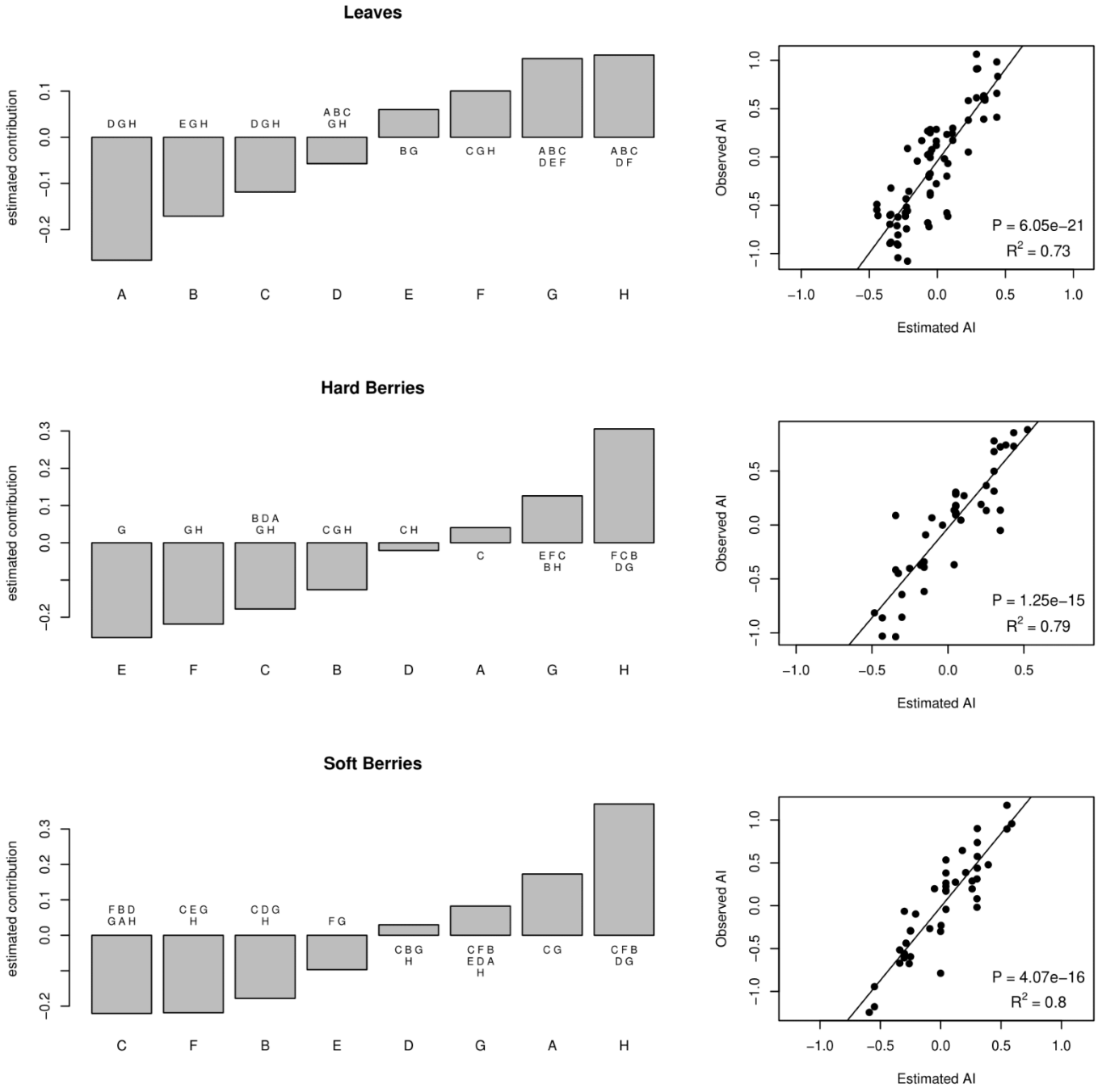

Figure 4.16: *ENC values for the gene VIT_218s0001g10460 in the three tissues. Some alleles show a strong tissue-dependent expression. For example, allele A is the lowest in leaves but with is among the highes ENC values in berries, while E and F have an opposite pattern.*

4.4.2 Identification of the genes with an abnormal distribution of low or high AI among the sample population

A common analysis in population genetics is to test a gene sequence variant for deviations from the Hardy-Weinberg proportion. Such deviations can be the result of a number of causes: deviations from random mating behaviour such as those due to inbreeding or assortative mating, and selection among others. One of the possible explanations for the deviation of the observed genotype frequencies from the expected ones in a gene can be that one of its sequence variants influences the fitness of the individuals, and therefore is subject to selection (Wang and Shete, 2012). We decided to use the results of the ENC estimation for individual haplotypes to compare the observed distribution of *cis*-regulatory variants in genotypes with that expected under a random assortment of individual haplotypes in genotypes. In other terms, we want to assess ifthe selection is favouring either the presence of individuals with large differences in expression among alleles or vice versa of individuals with similarity in the expression among alleles. Unlike the traditional Hardy Weinberg equilibrium analysis that considers sequence variants per se without any implications for their effects on phenotypes, here we are explicitly considering the *cis*-regulatory value for each haplotype in the analysis.

An advantage of having an estimate for the net contribution to the expression of a singular allele is that we can now have an estimate of what would be the expression level of a gene of a hypothetical individual with two alleles never observed together in our population. This enables us to use the traditional methods used in population genetics replacing the study of the expression levels of a gene in a population with its AI value. In particular, we tried to assess if the AI determined by a genotype could have an impact on the frequency of that genotype in the population. With a method explained in paragraph 3.5.3, we compared the AI level of the population of the genes with similar AI levels of a null population, searching for genes with unexpected distributions of genotypes, that could justify the hypothesis of selection favouring either high-AI genotypes or, on the other end, low-AI genotypes. In each cultivar we found genes in both these categories: on a total of 4488 genes tested in leaves, 3210 in hard berries, and 3562 in soft berries lower than expected AI genotypes were found in 577 genes in leaves, 291 in hard berries, and 306 in soft berries (these are cases that represent lower observed regulatory heterozygosity than expected), while higher than expected AI genotypes were found in 41 genes in leaves, 40 genes in hard berries and 41 in soft berries.

5. CONCLUSIONS

Vitis vinifera is a cultivated species with a high level of heterozygosity and this was reflected in the mapping of a high number of eQTL in our work in all three tissues, while the high level of LD observed in this species was probably the cause of the high number of significant associations per eGene. The difference in the number of significant associations found in the three tissues may be due to the difference in the sample size, and hence power, in the tissues. Consistently with previous findings, the eGenes showed more variability in their expression levels and less selective constraint. This resonates with other studies, and with the idea that non-coding regions with low selective constraints can introduce variability, acting preferably on dispensable genes than on core, essential ones. After the functional annotation of the eGenes, we found that many enriched GO categories belonged to response mechanisms to biotic and abiotic stimuli, a set of genes known for their variability.

The Allele-specific expression analysis, together with the identification of the homozygous expressed genes in every cultivar, enabled the development of an index that we called “regulatory heterozygosity”, a measure of how much the transcriptome of an individual is regulated by the allele-specific expression of his genes. Levels of regulatory heterozygosity varied across cultivars and were generally higher in berries than in leaves. These results are coherent with data from previous projects of our group and indicate that the genes responsible for the AI levels have a tissue-specific expression, or that allele-specific expression can be a tissue-dependent mechanism of regulation. Our results validate the latter hypothesis, showing that the majority of genes with AI have different expression patterns depending on the tissues. On the other hand, contiguous genes with physical proximity tend to have similar ASE patterns, introducing the idea of “islands” of cis-acting regulation.

The results of ASE analysis and eQTL mapping are overall coherent with each other, and covering the reciprocal blind spots, gave us a more comprehensive scenario of the *cis*-regulation in *Vitis*. We found that genes showing ASE more frequently in the population are more likely to have a *cis*-eQTL. Moreover, genes with high levels of allele-specific expression or with AI in many samples, tend to show lower selective constraints, like the eGenes.

In conclusion, with this project we gathered and organized a complete set of information about gene expression regulation due to genomic variants. This dataset can be investigated from three different perspectives: considering the tree tissues, we defined in each one the set of genes more prone to be controlled by genomic variants. In addition, we measured the level of regulatory heterozygosity in the different tissues of each cultivar. Finally, looking at the single genes, we gathered a series of interconnected information: if it can be considered an eGene, in which cultivar it shows AI, if this behaviour changes in

different tissues, its number of allelic variants across the population and the difference in expression between these allelic variants. This is the first comprehensive description in *Vitis vinifera* on how variants in the genome can shape the regulation of its transcriptome and can be of assistance in the studies of pathways of interest and the basis for further investigation on the topic.

6. BIBLIOGRAPHY

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Astle, W., & Balding, D. J. (2009). Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, *24*(4), 451–471. https://doi.org/10.1214/09-STS307

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188.

Bouchereau, A., Guénot, P., & Larher, F. (2000). Analysis of amines in plant materials. *Journal of Chromatography B: Biomedical Sciences and Applications*, *747*(1), 49–67. https://doi.org/10.1016/S0378-4347(00)00286-3

Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, *108*(10), 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, *8*(12), e1002822. https://doi.org/10.1371/journal.pcbi.1002822

Catlin, N. S., & Josephs, E. B. (2022). The important contribution of transposable elements to phenotypic variation and evolution. *Current Opinion in Plant Biology*, *65*, 102140. https://doi.org/10.1016/j.pbi.2021.102140

Cheng, S.-P., Jia, K.-H., Liu, H., Zhang, R.-G., Li, Z.-C., Zhou, S.-S., Shi, T.-L., Ma, A.-C., Yu, C.-W., Gao, C., Cao, G.-L., Zhao, W., Nie, S., Guo, J.-F., Jiao, S.-Q., Tian, X.-C., Yan, X.-M., Bao, Y.-T., Yun, Q.-Z., … Mao, J.-F. (2021). Haplotype-resolved genome assembly and allele-specific gene expression in cultivated ginger. *Horticulture Research*, *8*, 188. https://doi.org/10.1038/s41438-021-00599-8

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., & Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, *13*(12), Art. 12. https://doi.org/10.1038/nmeth.4035

Choi, Y., Chan, A. P., Kirkness, E., Telenti, A., & Schork, N. J. (2018). Comparison of phasing strategies for whole human genomes. *PLoS Genetics*, *14*(4), e1007308. https://doi.org/10.1371/journal.pgen.1007308

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Cubillos, F. A., Coustham, V., & Loudet, O. (2012). Lessons from eQTL mapping studies: Non-coding regions and their role behind natural phenotypic variation in plants. *Current Opinion in Plant Biology*, *15*(2), 192–198. https://doi.org/10.1016/j.pbi.2012.01.005

de Winter, J. C. F., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, *21*, 273–290. https://doi.org/10.1037/met0000079

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics (Oxford, England)*, *25*(24), 3207–3212. https://doi.org/10.1093/bioinformatics/btp579

Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F., & Marchini, J. (2013). Haplotype Estimation Using Sequencing Reads. *The American Journal of Human Genetics*, *93*(4), 687–696. https://doi.org/10.1016/j.ajhg.2013.09.002

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, *10*(1), Art. 1. https://doi.org/10.1038/s41467-019-13225-y

di Iulio, J., Bartha, I., Wong, E. H. M., Yu, H.-C., Lavrenko, V., Yang, D., Jung, I., Hicks, M. A., Shah, N., Kirkness, E. F., Fabani, M. M., Biggs, W. H., Ren, B., Venter, J. C., & Telenti, A. (2018). The human noncoding genome defined by genetic diversity. *Nature Genetics*, *50*(3), 333–337. https://doi.org/10.1038/s41588-018-0062-7

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Edge, P., Bafna, V., & Bansal, V. (2016). HapCUT2: Robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research*, gr.213462.116. https://doi.org/10.1101/gr.213462.116

Elkon, R., & Agami, R. (2017). Characterization of noncoding regulatory DNA in the human genome. *Nature Biotechnology*, *35*(8), 732–746. https://doi.org/10.1038/nbt.3863

Eyre-Walker, A. (2010). Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences*, *107*(suppl_1), 1752–1756. https://doi.org/10.1073/pnas.0906182107

Fabbro, C. D., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLOS ONE*, *8*(12), e85024. https://doi.org/10.1371/journal.pone.0085024

Flutre, T., Wen, X., Pritchard, J., & Stephens, M. (2013). A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLoS Genetics*, *9*(5), e1003486. https://doi.org/10.1371/journal.pgen.1003486

Foria, S., Magris, G., Jurman, I., Schwope, R., De Candido, M., De Luca, E., Ivanišević, D., Morgante, M., & Di Gaspero, G. (2022). Extent of wild–to–crop interspecific introgression in grapevine (Vitis vinifera) as a consequence of resistance breeding and implications for the crop species definition. *Horticulture Research*, *9*, uhab010. https://doi.org/10.1093/hr/uhab010

Fu, P., Wu, W., Lai, G., Li, R., Peng, Y., Yang, B., Wang, B., Yin, L., Qu, J., Song, S., & Lu, J. (2020). Identifying Plasmopara viticola resistance Loci in grapevine (Vitis amurensis) via genotyping-by-sequencing-based QTL mapping. *Plant Physiology and Biochemistry: PPB*, *154*, 75–84. https://doi.org/10.1016/j.plaphy.2020.05.016

Gambino, G., Dal Molin, A., Boccacci, P., Minio, A., Chitarra, W., Avanzato, C. G., Tononi, P., Perrone, I., Raimondi, S., Schneider, A., Pezzotti, M., Mannini, F., Gribaudo, I., & Delledonne, M. (2017). Whole-genome sequencing and SNV genotyping of 'Nebbiolo' (Vitis vinifera L.) clones. *Scientific Reports*, *7*(1), 17294. https://doi.org/10.1038/s41598-017-17405-y

Gilad, Y., Rifkin, S. A., & Pritchard, J. K. (2008). Revealing the architecture of gene regulation: The promise of eQTL studies. *Trends in genetics : TIG*, *24*(8), 408–415. https://doi.org/10.1016/j.tig.2008.06.001

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talón, M., Dopazo, J., & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, *36*(10), 3420–3435. https://doi.org/10.1093/nar/gkn176

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), Art. 7675. https://doi.org/10.1038/nature24277

GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science (New York, N.Y.)*, *369*(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776

Hammond, J. P., Mayes, S., Bowen, H. C., Graham, N. S., Hayden, R. M., Love, C. G., Spracklen, W. P., Wang, J., Welham, S. J., White, P. J., King, G. J., & Broadley, M. R. (2011). Regulatory hotspots are associated with plant gene expression under varying soil phosphorus supply in Brassica rapa. *Plant Physiology*, *156*(3), 1230–1241. https://doi.org/10.1104/pp.111.175612

Hasin-Brumshtein, Y., Hormozdiari, F., Martin, L., van Nas, A., Eskin, E., Lusis, A. J., & Drake, T. A. (2014). Allele-specific expression and eQTL analysis in mouse adipose tissue. *BMC Genomics*, *15*(1), 471. https://doi.org/10.1186/1471-2164-15-471

Hellwege, J. N., Keaton, J. M., Giri, A., Gao, X., Velez Edwards, D. R., & Edwards, T. L. (2017). Population Stratification in Genetic Association Studies. *Current Protocols in Human Genetics*, *95*(1), 1.22.1-1.22.23. https://doi.org/10.1002/cphg.48

Hernandez, R. D., Uricchio, L. H., Hartman, K., Ye, C., Dahl, A., & Zaitlen, N. (2019). Ultrarare variants drive substantial cis heritability of human gene expression. *Nature Genetics*, *51*(9), Art. 9. https://doi.org/10.1038/s41588-019-0487-7

Holland, J. B., Nyquist, W. E., & Cervantes-Martínez, C. T. (2002). Estimating and Interpreting Heritability for Plant Breeding: An Update. In *Plant Breeding Reviews* (pp. 9–112). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470650202.ch2

Huang, Q. Q., Ritchie, S. C., Brozynska, M., & Inouye, M. (2018). Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic Acids Research*, *46*(22), e133. https://doi.org/10.1093/nar/gky780

Jeffares, D. C., Tomiczek, B., Sojo, V., & dos Reis, M. (2015). A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods in Molecular Biology (Clifton, N.J.)*, *1201*, 65–90. https://doi.org/10.1007/978-1-4939-1438-8_4

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., Dunham, I., Elnitski, L. L., Farnham, P. J., Feingold, E. A., Gerstein, M., Giddings, M. C., Gilbert, D. M., Gingeras, T. R., Green, E. D., … Hardison, R. C. (2014). Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(17), 6131–6138. https://doi.org/10.1073/pnas.1318948111

Khansefid, M., Pryce, J. E., Bolormaa, S., Chen, Y., Millen, C. A., Chamberlain, A. J., Vander Jagt, C. J., & Goddard, M. E. (2018). Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle. *BMC Genomics*, *19*(1), 793. https://doi.org/10.1186/s12864-018-5181-0

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)*, *308*(5720), 385–389. https://doi.org/10.1126/science.1109557

Kliebenstein, D. (2009). Quantitative Genomics: Analyzing Intraspecific Variation Using Global Gene Expression Polymorphisms or eQTLs. *Annual Review of Plant Biology*, *60*(1), 93–114. https://doi.org/10.1146/annurev.arplant.043008.092114

Kryazhimskiy, S., & Plotkin, J. B. (2008). The Population Genetics of dN/dS. *PLoS Genetics*, *4*(12), e1000304. https://doi.org/10.1371/journal.pgen.1000304

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., … Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506–511. https://doi.org/10.1038/nature12531

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & L Price, A. (2016). Reference-based phasing

using the Haplotype Reference Consortium panel. *Nature Genetics*, *48*(11), 1443–1448. https://doi.org/10.1038/ng.3679

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Magris, G., Di Gaspero, G., Marroni, F., Zenoni, S., Tornielli, G. B., Celii, M., De Paoli, E., Pezzotti, M., Conte, F., Paci, P., & Morgante, M. (2019). Genetic, epigenetic and genomic effects on variation of gene expression among grape varieties. *The Plant Journal: For Cell and Molecular Biology*, *99*(5), 895–909. https://doi.org/10.1111/tpj.14370

Magris, G., Jurman, I., Fornasiero, A., Paparelli, E., Schwope, R., Marroni, F., Di Gaspero, G., & Morgante, M. (2021). The genomes of 204 Vitis vinifera accessions reveal the origin of European wine grapes. *Nature Communications*, *12*(1), Art. 1. https://doi.org/10.1038/s41467-021-27487-y

Mähler, N., Wang, J., Terebieniec, B. K., Ingvarsson, P. K., Street, N. R., & Hvidsten, T. R. (2017). Gene co-expression network connectivity is an important determinant of selective constraint. *PLOS Genetics*, *13*(4), e1006402. https://doi.org/10.1371/journal.pgen.1006402

Marroni, F., Pinosio, S., & Morgante, M. (2014). Structural variation and genome complexity: Is dispensable really dispensable? *Current Opinion in Plant Biology*, *18*, 31–36. https://doi.org/10.1016/j.pbi.2014.01.003

Martínez-García, P. J., Mas-Gómez, J., Wegrzyn, J., & Botía, J. A. (2022). Bioinformatic approach for the discovery of cis-eQTL signals during fruit ripening of a woody species as grape (Vitis vinifera L.). *Scientific Reports*, *12*(1), 7481. https://doi.org/10.1038/s41598-022-11689-5

Minio, A., Cochetel, N., Vondras, A. M., Massonnet, M., & Cantu, D. (2022). Assembly of complete diploid-phased chromosomes from draft genome sequences. *G3 Genes|Genomes|Genetics*, *12*(8), jkac143. https://doi.org/10.1093/g3journal/jkac143

Minio, A., Lin, J., Gaut, B. S., & Cantu, D. (2017). How Single Molecule Real-Time Sequencing and Haplotype Phasing Have Enabled Reference-Grade Diploid Genome Assembly of Wine Grapes. *Frontiers in Plant Science*, *8*. https://www.frontiersin.org/articles/10.3389/fpls.2017.00826

Minio, A., Massonnet, M., Figueroa-Balderas, R., Castro, A., & Cantu, D. (2019). Diploid Genome Assembly of the Wine Grape Carménère. *G3 Genes|Genomes|Genetics*, *9*(5), 1331–1337. https://doi.org/10.1534/g3.119.400030

Minio, Andrea, & Cantu, Dario. (2022). *Grapegenomics.com: A web portal with genomic data and analysis tools for wild and cultivated grapevines* (Versione 202208). Zenodo. https://doi.org/10.5281/ZENODO.7027886

Mohammadi, P., Castel, S. E., Brown, A. A., & Lappalainen, T. (2017). Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Research*, *27*(11), 1872–1884. https://doi.org/10.1101/gr.216747.116

Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli, T., Davis, C. A., Dobin, A., Kaul, R., Halow, J., Van Nostrand, E. L., Freese, P., Gorkin, D. U., Shen, Y., He, Y., Mackiewicz, M., Pauli-Behn, F., Williams, B. A., … Weng, Z. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, *583*(7818), Art. 7818. https://doi.org/10.1038/s41586-020-2493-4

Morgante, M. (2006). Plant genome organisation and diversity: The year of the junk! *Current Opinion in Biotechnology*, *17*(2), 168–173. https://doi.org/10.1016/j.copbio.2006.03.001

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1620), 20120362. https://doi.org/10.1098/rstb.2012.0362

Nowak, R. (1994). Mining treasures from «junk DNA». *Science (New York, N.Y.)*, *263*(5147), 608–610. https://doi.org/10.1126/science.7508142

Ohno, S. (1972). So much «junk» DNA in our genome. *Brookhaven Symposia in Biology*, *23*, 366–370.

Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2021). *Biostrings: Efficient manipulation of biological strings* (2.62.0). Bioconductor version: Release (3.14). https://doi.org/10.18129/B9.bioc.Biostrings

Pandey, R. V., Franssen, S. U., Futschik, A., & Schlötterer, C. (2013). Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Molecular Ecology Resources*, *13*(4), 740–745. https://doi.org/10.1111/1755-0998.12110

Pastinen, T. (2010). Genome-wide allele-specific analysis: Insights into regulatory variation. *Nature Reviews Genetics*, *11*(8), Art. 8. https://doi.org/10.1038/nrg2815

Peters, J. E., Lyons, P. A., Lee, J. C., Richard, A. C., Fortune, M. D., Newcombe, P. J., Richardson, S., & Smith, K. G. C. (2016). Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. *PLOS Genetics*, *12*(3), e1005908. https://doi.org/10.1371/journal.pgen.1005908

Pierce, B. L., Tong, L., Chen, L. S., Rahaman, R., Argos, M., Jasmine, F., Roy, S., Paul-Brutus, R., Westra, H.-J., Franke, L., Esko, T., Zaman, R., Islam, T., Rahman, M., Baron, J. A., Kibriya, M. G., & Ahsan, H. (2014). Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 South Asians. *PLOS Genetics*, *10*(12), e1004818. https://doi.org/10.1371/journal.pgen.1004818

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Auwera, G. A. V. der, Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2018). *Scaling accurate genetic variant discovery to tens of thousands of samples* (p. 201178). bioRxiv. https://doi.org/10.1101/201178

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Qu, W., Gurdziel, K., Pique-Regi, R., & Ruden, D. M. (2018). Lead Modulates trans- and cis-Expression Quantitative Trait Loci (eQTLs) in Drosophila melanogaster Heads. *Frontiers in Genetics*, *9*. https://www.frontiersin.org/articles/10.3389/fgene.2018.00395

Ramakrishnan, A. P. (2013). Linkage Disequilibrium. In S. Maloy & K. Hughes (A c. Di), *Brenner's Encyclopedia of Genetics (Second Edition)* (pp. 252–253). Academic Press. https://doi.org/10.1016/B978-0-12-374984-0.00870-6

Ramos-Madrigal, J., Runge, A. K. W., Bouby, L., Lacombe, T., Samaniego Castruita, J. A., Adam-Blondon, A.-F., Figueiral, I., Hallavant, C., Martínez-Zapater, J. M., Schaal, C., Töpfer, R., Petersen, B., Sicheritz-Pontén, T., This, P., Bacilieri, R., Gilbert, M. T. P., & Wales, N. (2019). Palaeogenomic insights into the origins of French grapevine diversity. *Nature Plants*, *5*(6), 595–603. https://doi.org/10.1038/s41477-019-0437-5

Reshef, N., Karn, A., Manns, D. C., Mansfield, A. K., Cadle-Davidson, L., Reisch, B., & Sacks, G. L. (2022). Stable QTL for malate levels in ripe fruit and their transferability across Vitis species. *Horticulture Research*, *9*, uhac009. https://doi.org/10.1093/hr/uhac009

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, *270*(5235), 467–470. https://doi.org/10.1126/science.270.5235.467

Schmidt, P., Hartung, J., Rath, J., & Piepho, H.-P. (2019). Estimating Broad-Sense Heritability with Unbalanced Data from Agricultural Cultivar Trials. *Crop Science*, *59*(2), 525–536. https://doi.org/10.2135/cropsci2018.06.0376

Shabalin, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics (Oxford, England)*, *28*(10), 1353–1358. https://doi.org/10.1093/bioinformatics/bts163

Shi, X. M. (A c. Di). (2020). *eQTL Analysis: Methods and Protocols* (Vol. 2082). Springer US. https://doi.org/10.1007/978-1-0716-0026-9

Single, R. M., & Thomson, G. (2016). Linkage Disequilibrium: Population Genetics of Multiple Loci. In R. M. Kliman (A c. Di), *Encyclopedia of Evolutionary Biology* (pp. 400–404). Academic Press. https://doi.org/10.1016/B978-0-12-800049-6.00030-5

Smit, A., & Hubley, R. (s.d.). *RepeatModeler Open-1.0. 2008-2015*. http://www.repeatmasker.org. Recuperato 13 novembre 2022, da https://www.repeatmasker.org/

Subki, A., Abidin, A. A. Z., Yusof, Z. N. B., Subki, A., Abidin, A. A. Z., & Yusof, Z. N. B. (2018). The Role of Thiamine in Plants and Current Perspectives in Crop Improvement. In *B Group Vitamins—Current Uses and Perspectives*. IntechOpen. https://doi.org/10.5772/intechopen.79350

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), Art. 8. https://doi.org/10.1038/s41576-019-0127-1

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., … ON BEHALF OF THE NHLBI EXOME SEQUENCING PROJECT. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*, *337*(6090), 64–69. https://doi.org/10.1126/science.1219240

The French–Italian Public Consortium for Grapevine Genome Characterization. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*(7161), 463–467. https://doi.org/10.1038/nature06148

Tian, J., Keller, M. P., Broman, A. T., Kendziorski, C., Yandell, B. S., Attie, A. D., & Broman, K. W. (2016). The Dissection of Expression Quantitative Trait Locus Hotspots. *Genetics*, *202*(4), 1563–1574. https://doi.org/10.1534/genetics.115.183624

Trenti, M., Lorenzi, S., Bianchedi, P. L., Grossi, D., Failla, O., Grando, M. S., & Emanuelli, F. (2021). Candidate genes and SNPs associated with stomatal conductance under drought stress in Vitis. *BMC Plant Biology*, *21*(1), 7. https://doi.org/10.1186/s12870-020-02739-z

Van der Most, P. J., Küpers, L. K., Snieder, H., & Nolte, I. (2017). QCEWAS: Automated quality control of results of epigenome-wide association studies. *Bioinformatics*, *33*(8), 1243–1245. https://doi.org/10.1093/bioinformatics/btw766

Van Dyke, K., Lutz, S., Mekonnen, G., Myers, C. L., & Albert, F. W. (2021). Trans-acting genetic variation affects the expression of adjacent genes. *Genetics*, *217*(3), iyaa051. https://doi.org/10.1093/genetics/iyaa051

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., Pindo, M., FitzGerald, L. M., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Coppola, G., Wardell, B., Micheletti, D., Macalma, T., Facci, M., Mitchell, J. T., Perazzolli, M., … Viola, R. (2007). A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. *PLoS ONE*, *2*(12), e1326. https://doi.org/10.1371/journal.pone.0001326

Velez-Irizarry, D., Casiro, S., Daza, K. R., Bates, R. O., Raney, N. E., Steibel, J. P., & Ernst, C. W. (2019). Genetic control of longissimus dorsi muscle gene expression variation and joint analysis with phenotypic quantitative trait loci in pigs. *BMC Genomics*, *20*, 3. https://doi.org/10.1186/s12864-018-5386-2

Vitulo, N., Forcato, C., Carpinelli, E. C., Telatin, A., Campagna, D., D'Angelo, M., Zimbello, R., Corso, M., Vannozzi, A., Bonghi, C., Lucchin, M., & Valle, G. (2014). A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. *BMC Plant Biology*, *14*(1), 99. https://doi.org/10.1186/1471-2229-14-99

Wang, J., & Shete, S. (2012). Testing departure from Hardy-Weinberg proportions. *Methods in Molecular Biology (Clifton, N.J.)*, *850*, 77–102. https://doi.org/10.1007/978-1-61779-555-8_6

Wang, J., Yu, H., Xie, W., Xing, Y., Yu, S., Xu, C., Li, X., Xiao, J., & Zhang, Q. (2010). A global analysis of QTLs for expression variations in rice shoots at the early seedling stage. *The Plant Journal*, *63*(6), 1063–1074. https://doi.org/10.1111/j.1365-313X.2010.04303.x

Wang, X., Chen, Q., Wu, Y., Lemmon, Z. H., Xu, G., Huang, C., Liang, Y., Xu, D., Li, D., Doebley, J. F., & Tian, F. (2018). Genome-wide Analysis of Transcriptional Variability in a Large Maize-Teosinte Population. *Molecular Plant*, *11*(3), 443–459. https://doi.org/10.1016/j.molp.2017.12.011

Wang, Y., Xin, H., Fan, P., Zhang, J., Liu, Y., Dong, Y., Wang, Z., Yang, Y., Zhang, Q., Ming, R., Zhong, G.-Y., Li, S., & Liang, Z. (2021). The genome of Shanputao (Vitis amurensis) provides a new insight into cold tolerance of grapevine. *The Plant Journal: For Cell and Molecular Biology*, *105*(6), 1495–1506. https://doi.org/10.1111/tpj.15127

Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., & Kinzler, K. W. (2002). Allelic Variation in Human Gene Expression. *Science*, *297*(5584), 1143–1143. https://doi.org/10.1126/science.1072545

Yang, S., Liu, Y., Jiang, N., Chen, J., Compton, L., Luo, Z., & Wang, M. (2014). Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals. *BMC genomics*, *15*, 13. https://doi.org/10.1186/1471-2164-15-13

Yang, Z., & Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, *15*(12), 496–503. https://doi.org/10.1016/S0169-5347(00)01994-7

Zhang, L., Yu, Y., Shi, T., Kou, M., Sun, J., Xu, T., Li, Q., Wu, S., Cao, Q., Hou, W., & Li, Z. (2020). Genome-wide analysis of expression quantitative trait loci (eQTLs) reveals the regulatory architecture of gene expression variation in the storage roots of sweet potato. *Horticulture Research*, *7*(1), Art. 1. https://doi.org/10.1038/s41438-020-0314-4

Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., Mudivarti, P. A., Wyatt, P. W., Bharadwaj, R., Makarewicz, A. J., Li, Y., Belgrader, P., Price, A. D., Lowe, A. J., Marks, P., … Ji, H. P. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, *34*(3), 303–311. https://doi.org/10.1038/nbt.3432

Zuckerkandl, E. (1992). Revisiting junk DNA. *Journal of Molecular Evolution*, *34*(3), 259–271. https://doi.org/10.1007/BF00162975

7. SUPPLEMENTARY MATERIALS

Table S1 – *sequencing and alignment metrics for leaves samples, divided according the two replicates. The four column are: "output reads": the number of sequenced reads; "trimmed pairs": number reads that successfully passed the quality filter described in paragraph 3.2. This number refer to the pair of reads as the unpaired ones were discarded; "mapped pairs": number of reads pair uniquely mapped to the reference; "number genes": number of genes with a minimum of 5 reads uniquely mapped on.*

LEAVES SAMPLES	Replicate 1				Replicate 2			
	output reads	trimmed pairs	mapped pairs	number genes	output reads	trimmed pairs	mapped pairs	number genes
Agadai	32 620 130	14 449 212	13 573 193	18 943	25 570 760	12 167 262	11 368 342	18 943
Airen	4 402 800	2 007 239	1 798 714	21 194	96 813 304	45 886 318	42 951 954	21 194
Alexandroouli	33 291 254	15 736 009	14 741 485	18 201	21 083 306	10 074 007	9 312 660	18 201
Ansonica	36 229 330	17 132 032	16 129 478	20 191	57 371 514	26 968 781	25 028 146	20 191
Ararati	33 184 918	16 202 802	15 376 910	18 057	23 740 156	11 214 010	10 264 024	18 057
Assyrtiko	30 866 730	15 066 929	14 275 650	18 797	25 670 356	12 066 933	11 291 620	18 797
Asyl Kara	36 710 412	17 772 361	16 825 963	20 439	62 896 086	29 364 346	27 330 036	20 439
Barbera	33 508 268	16 097 916	15 113 921	19 733	37 786 178	17 840 198	16 597 612	19 733
Bayan Shirei	33 148 052	15 905 598	14 904 052	20 335	58 580 436	28 155 208	26 385 695	20 335
Berzamino	40 816 948	18 527 359	17 426 373	19 608	--	--	--	--
Bombino Bianco	34 597 402	16 745 886	15 743 270	20 610	61 839 232	29 832 348	27 543 847	20 610
Cabernet Franc	30 969 004	14 803 422	13 912 973	14 394	--	--	--	--
Cabernet Sauvignon	36 098 312	17 462 932	16 441 975	20 413	68 197 912	32 492 427	30 441 232	20 413
Carignan	32 848 410	15 355 174	14 382 059	14 012	--	--	--	--
Catarratto B.C.	35 589 934	16 894 830	15 791 325	17 378	11 843 098	5 641 971	5 258 643	17 378
Cesanese d'Affile	33 651 760	16 163 330	15 251 201	17 962	21 065 830	10 023 859	9 202 506	17 962
Chaouch Blanc	128 703 158	62 224 367	58 654 424	20 983	--	--	--	--
Chasselas Blanc	35 369 058	16 705 035	15 552 020	15 551	--	--	--	--
Coarna Alba	37 644 790	18 181 157	17 071 143	18 856	22 096 334	10 592 738	9 738 732	18 856
Daphnia	30 300 098	14 501 988	13 626 099	19 777	62 208 780	29 764 286	27 701 352	19 777
Falanghina	32 744 878	15 608 447	14 556 801	20 078	50 213 050	24 058 146	22 113 130	20 078
Fiano	76 863 950	36 988 118	34 468 035	17 390	28 464 362	13 381 832	12 220 489	17 390
Fumat	33 406 276	14 957 278	13 224 544	17 940	17 954 412	8 446 203	7 848 405	17 940
Garganega	37 304 516	17 952 967	16 937 175	16 994	--	--	--	--
Garnacha	29 282 916	13 936 064	12 967 480	19 227	36 653 636	17 644 171	16 424 241	19 227
Glera	31 621 810	14 638 887	13 817 405	13 666	--	--	--	--
Gordin Verde	30 233 854	14 031 368	12 860 582	16 806	10 811 866	5 141 894	4 791 886	16 806
Gorula	32 330 250	15 275 035	14 317 629	20 832	100 951 022	47 138 755	43 306 360	20 832
Gyulyabi Dagestanskii	30 949 294	14 296 330	13 403 213	19 133	26 761 920	12 958 887	12 181 409	19 133
Harslevelue	36 098 512	17 101 060	15 901 411	20 709	68 712 940	32 895 838	30 481 999	20 709
Henab Turki	27 996 078	13 315 223	12 473 108	19 127	34 748 994	16 672 080	15 395 948	19 127
Heunisch Weiss	31 832 384	15 073 886	14 128 596	18 569	24 347 490	11 728 735	10 872 168	18 569

Italia	35 967 570	17 264 395	16 346 782	18 870	25 961 822	12 486 027	11 662 578	18 870
Kadarka	31 674 678	14 809 178	13 661 955	19 340	25 244 352	11 890 475	11 024 914	19 340
Kandahari Siah	50 478 944	24 010 349	22 541 761	19 928	--	--	--	--
Kishmish Vatkana	25 939 070	12 317 487	11 535 575	19 638	39 411 508	18 634 208	17 182 444	19 638
Lambrusco di Sorbara	30 017 290	14 129 864	13 387 223	17 205	14 350 576	6 547 558	6 162 417	17 205
Lambrusco Grasparossa	27 032 590	12 724 655	11 495 087	14 261	--	--	--	--
Limnio	37 312 962	17 816 923	16 566 822	18 199	15 914 698	7 646 351	6 986 593	18 199
Malvasia Bianca	32 048 124	15 516 192	14 572 304	17 872	17 269 654	7 838 233	7 269 791	17 872
Malvasia Istriana	32 709 566	15 559 660	14 640 099	16 734	33 147 122	15 964 694	14 902 611	16 734
Mauzac Blanc	30 699 834	13 970 226	12 517 361	20 133	54 405 754	26 365 816	24 669 102	20 133
Mavrodaphni	30 568 736	14 768 581	13 706 110	19 694	46 658 444	22 386 380	20 672 500	19 694
Merlot Noir	33 545 228	16 181 298	15 291 015	20 777	97 867 170	46 760 987	43 298 063	20 777
Montepulciano	35 556 060	13 032 304	11 117 990	20 253	63 564 830	30 150 312	27 951 772	20 253
Mtsvane Kachuri	32 355 900	15 464 253	14 032 323	19 652	38 662 774	18 658 052	17 261 731	19 652
Muscat a Petits Grains B.	32 686 922	15 769 320	14 955 209	19 230	31 028 666	14 914 907	13 854 760	19 230
Narma	27 891 196	13 271 201	12 514 523	19 853	43 672 504	20 952 175	19 455 479	19 853
Nebbiolo	36 121 202	17 327 359	16 370 240	20 468	90 393 982	42 267 139	39 080 833	20 468
Negro Amaro	32 132 000	15 366 573	14 494 195	17 490	11 428 252	5 462 364	5 080 252	17 490
Ojaleshi	37 668 600	18 050 631	16 527 746	20 272	61 342 394	28 948 852	26 675 184	20 272
Picolit	36 402 024	17 737 421	16 622 996	19 022	22 055 022	10 546 375	9 563 604	19 022
Pignoletto	37 624 502	18 373 445	17 259 418	19 202	34 047 702	16 276 713	15 124 748	19 202
Pinot Noir	27 627 974	13 240 359	12 548 107	18 263	21 610 264	10 032 468	9 315 507	18 263
Plechistik	35 153 522	16 706 209	15 365 739	19 035	--	--	--	--
Raboso Piave	35 926 086	16 645 404	15 514 097	19 123	--	--	--	--
Refosco P.R.	33 568 032	15 981 765	14 874 149	18 532	18 939 744	8 741 878	8 044 249	18 532
Ribolla Gialla	33 807 946	16 101 759	14 814 830	19 723	64 273 686	29 739 964	27 481 746	19 723
Riesling Weiss	34 092 132	16 386 124	15 359 268	17 730	19 127 574	9 058 080	8 408 855	17 730
Rkatsiteli	29 003 790	13 866 012	13 166 204	20 067	50 820 664	24 284 487	22 683 528	20 067
Sagrantino	31 757 388	14 893 844	13 347 152	18 717	25 776 582	12 282 586	11 394 210	18 717
Sangiovese	30 780 574	14 848 457	13 955 234	16 400	29 264 652	14 208 687	13 481 833	16 400
Sauvignon Blanc	29 042 282	13 781 960	12 975 887	19 395	33 653 076	16 165 180	15 158 847	19 395
Savagnin Blanc	34 806 104	16 586 359	15 711 937	16 388	--	--	--	--
Schiava Grossa	30 855 098	14 043 623	13 234 487	19 508	32 052 480	15 467 553	14 451 191	19 508
Schioppettino	29 588 666	14 023 413	13 273 204	16 778	9 271 548	4 464 430	4 129 635	16 778
Sirgula	77 746 350	37 659 186	35 264 103	17 038	12 940 326	6 140 902	5 680 259	17 038
Sultanina	32 827 460	15 435 634	14 541 900	18 011	20 850 960	9 899 122	9 130 711	18 011
Tagobi	34 402 892	16 693 205	15 812 833	19 519	46 747 830	22 045 541	20 455 197	19 519
Taifi Rozovyi	31 707 338	14 366 568	12 641 289	14 330	32 615 228	15 518 431	14 484 601	14 330
Tavkveri	37 371 064	18 068 130	17 099 364	15 725	6 630 312	3 206 423	3 006 199	15 725
Tebrizi	36 972 000	18 007 830	16 918 338	20 602	59 715 736	28 189 704	26 426 799	20 602
Terbash	34 397 752	16 249 979	15 109 169	17 385	16 755 818	7 941 969	7 161 640	17 385
Tocai Friulano	34 671 086	16 427 958	15 338 768	17 849	21 603 588	10 087 315	9 480 918	17 849
Trebbiano Toscano	35 989 688	17 462 411	16 322 566	21 109	118 946 492	56 115 076	51 906 536	21 109
Tribidrag	45 188 358	22 112 020	20 959 934	18 387	59 638 116	21 896 629	18 568 466	18 387

Tschvediansis Tetra	31 124 750	13 656 941	12 004 373	19 340	29 041 234	13 888 117	12 966 522	19 340
Uva di Troia	33 390 816	15 820 511	14 758 719	19 834	46 857 622	21 818 467	20 208 094	19 834
V267	32 839 498	15 876 774	14 917 275	19 469	32 015 926	15 344 972	14 193 853	19 469
V278	32 183 932	15 001 915	14 038 428	12 492	--	--	--	--
V292	38 838 408	18 799 036	17 644 343	19 982	52 774 534	25 176 348	23 357 166	19 982
V294	42 223 338	20 534 222	19 435 411	19 078	40 739 488	19 562 383	18 293 707	19 078
V385	30 487 152	14 721 100	13 894 305	19 592	46 338 596	22 307 399	20 882 363	19 592
V389	37 952 834	18 345 090	17 380 433	19 159	37 593 386	17 930 103	16 588 991	19 159
V410	30 199 376	14 175 381	13 244 377	17 813	17 192 836	8 254 191	7 740 221	17 813
V411	34 141 752	16 432 465	15 405 411	18 854	22 751 914	10 921 572	10 227 552	18 854
Verdicchio Bianco	35 531 162	16 874 358	15 931 114	19 849	55 836 670	26 695 262	24 871 763	19 849
Verduzzo Friulano	30 920 864	14 621 638	13 555 227	15 802	--	--	--	--
Vermentino	32 341 910	14 281 917	12 545 334	19 016	43 236 466	20 474 412	18 824 268	19 016
Vernaccia S.G.	26 880 474	12 518 468	11 715 446	14 950	--	--	--	--

Table S2 – *sequencing and alignment metrics for hard berries samples, divided according the two replicates. The four column are: "output reads": the number of sequenced reads; "trimmed pairs": number reads that successfully passed the quality filter described in paragraph 3.2. This number refer to the pair of reads as the unpaired ones were discarded; "mapped pairs": number of reads pair uniquely mapped to the reference; "number genes": number of genes with a minimum of 5 reads uniquely mapped on.*

HARD BERRIES SAMPLES	Replica 1				Replica 2			
	output reads	trimmed pairs	mapped pairs	number genes	output reads	trimmed pairs	mapped pairs	number genes
Aglianico	43 929 004	21 310 811	16 966 826	19 343	50 943 118	24 586 201	18 701 408	19 268
Airen	76 373 490	37 279 265	33 765 558	20 875	42 549 360	20 708 019	18 725 906	19 894
Alexandroouli	47 364 728	23 441 125	21 186 705	22 155	40 717 686	20 150 199	18 268 662	20 673
Ansonica	42 403 838	20 786 599	18 629 018	19 635	42 759 750	20 938 559	18 753 739	19 671
Barbera	40 644 928	19 686 211	16 987 547	19 544	43 646 428	21 144 276	18 182 424	19 627
Bayan Shirei	40 085 226	19 464 594	17 439 301	18 356	211 607 014	102 743 060	88 872 275	22 475
Berzamino	42 983 788	21 107 445	19 459 738	19 729	46 255 450	22 711 260	20 841 905	19 921
Bombino Bianco	45 602 954	22 123 680	19 294 630	18 912	44 680 752	21 549 400	18 605 027	19 194
Cabernet Franc	42 228 350	20 583 815	18 136 672	18 662	41 639 752	20 232 543	17 383 542	18 490
Cabernet Sauvignon	54 512 998	26 325 842	23 120 812	19 786	40 041 264	18 781 678	16 242 700	18 643
Cesanese d'Affile	54 556 198	26 209 538	20 223 894	18 473	46 399 726	22 618 189	19 462 955	19 927
Chasselas Blanc	39 593 700	19 513 400	17 861 607	19 451	38 737 526	18 982 341	17 226 130	19 865
Falanghina	76 433 796	37 513 594	33 589 201	20 831	43 302 490	21 017 406	18 272 834	19 849
Fiano	45 562 998	22 344 846	19 358 256	18 897	50 216 740	24 561 308	21 371 057	19 638
Fumat	63 501 306	30 950 901	27 499 170	19 954	97 963 066	47 363 332	41 803 881	21 617
Garganega	39 947 050	19 634 276	17 622 459	18 840	44 121 480	21 752 938	19 555 944	19 820
Garnacha	42 707 966	20 924 453	18 002 384	19 503	52 974 038	26 018 366	22 892 379	20 335
Gordin Verde	123 418 342	60 504 328	55 207 673	22 448	42 405 086	20 887 232	19 022 138	20 058
Grignolino	43 033 934	20 747 986	18 631 887	18 693	145 556 426	69 274 627	61 996 938	21 520
Heunisch Weiss	85 312 374	41 799 659	36 280 592	20 690	168 191 178	82 463 208	71 476 388	24 440
Kadarka	52 375 392	25 889 613	23 721 276	20 448	49 538 276	24 456 515	22 399 608	20 129
Kölner Blau	41 870 030	20 285 115	17 572 577	18 236	39 170 188	18 977 146	17 577 873	18 241
Lambrusco di Sorbara	51 913 452	25 328 106	21 262 068	18 999	44 632 142	21 738 308	18 489 900	19 346
Lambrusco Grasparossa	47 995 070	23 453 576	20 529 622	19 767	43 271 808	21 129 944	17 639 669	18 971
Malvasia Istriana	38 877 758	19 053 321	17 040 254	19 479	41 646 040	20 466 771	18 387 715	20 019
Merlot Noir	45 781 604	22 370 672	19 689 423	19 573	37 142 796	18 184 810	15 938 924	18 948
Montepulciano	44 349 988	21 734 487	18 942 206	19 723	51 673 872	25 036 697	20 526 576	19 904
Muscat a Petits Grains B.	54 788 062	26 800 679	24 006 154	20 070	41 678 008	20 510 294	18 474 762	19 737
Nebbiolo	--	--	--	--	43 226 028	20 922 523	18 471 951	19 337
Negro Amaro	54 040 428	26 178 896	22 175 666	19 672	55 242 762	26 604 347	23 257 829	20 158
Pecorino	50 786 872	24 267 048	20 161 442	19 748	50 732 736	23 737 812	20 380 267	19 717
Picolit	40 666 954	19 938 346	17 475 355	19 517	38 860 400	19 123 304	17 086 084	19 611
Pignoletto	43 446 782	21 275 932	18 550 316	18 882	46 922 806	22 947 836	20 808 500	19 788
Pinot Noir	51 351 734	25 402 717	23 799 364	19 663	52 358 820	25 954 643	24 447 384	19 724
Refosco P.R.	59 184 334	29 087 120	25 478 410	20 212	47 634 008	23 285 044	19 880 906	19 541
Ribolla Gialla	41 309 582	20 235 678	18 332 147	19 708	38 810 182	19 068 198	17 162 785	19 169

Riesling Weiss	43 941 828	21 576 066	18 864 917	20 031	40 543 484	19 960 204	17 744 118	20 013
Rkatsiteli	49 986 798	24 418 092	19 466 041	19 720	62 069 892	30 430 496	23 473 389	18 739
Sagrantino	51 205 122	25 067 915	22 322 414	19 855	58 530 872	28 663 228	25 690 141	20 125
Sangiovese	48 598 248	23 675 934	20 134 077	19 481	39 655 000	19 403 892	16 535 640	19 176
Sauvignon Blanc	46 780 800	22 811 973	18 719 853	19 212	41 253 482	20 224 919	17 898 043	19 193
Savagnin Blanc	45 866 890	22 677 538	20 761 551	19 881	39 904 202	19 630 302	17 749 766	20 023
Schiava Grossa	45 025 040	21 864 865	17 510 430	19 611	53 930 496	26 387 509	22 278 142	19 643
Schioppettino	--	--	--	--	85 218 728	41 757 452	36 541 664	20 527
Sciavtsitska	42 487 332	20 810 498	18 527 522	20 020	50 576 418	24 691 419	21 612 863	20 409
Sultanina	41 929 316	20 570 044	18 485 329	18 282	45 194 170	22 041 381	18 590 397	17 960
Tavkveri	39 107 242	19 066 912	17 164 173	19 806	44 623 896	21 920 495	19 955 522	20 895
Terrano	39 537 866	19 340 741	17 145 399	18 742	42 843 132	20 828 482	17 992 364	19 263
Tocai Friulano	65 074 566	31 942 108	27 312 223	21 649	39 068 572	19 112 077	17 112 669	19 725
Trebbiano Toscano	42 002 868	20 517 502	17 987 131	19 330	45 126 570	22 008 461	18 513 866	19 642
Tribidrag	38 911 986	19 197 060	17 421 252	19 587	40 967 178	20 152 000	17 902 050	20 455
V294	114 874 540	55 916 257	42 685 984	25 121	64 717 098	31 611 455	27 139 824	20 257
V411	--	--	--	--	84 984 096	41 502 819	35 271 499	20 461
Verdicchio Bianco	43 070 564	20 746 847	18 143 429	18 016	53 134 776	25 756 193	22 612 241	19 435
Verduzzo Friulano	46 209 082	22 120 404	17 885 784	18 909	40 847 602	19 993 849	17 934 966	19 727
Vermentino	44 684 564	21 899 956	19 443 588	19 087	57 955 668	28 478 940	25 533 706	19 932
Vernaccia S.G.	144 817 356	70 542 453	62 734 620	21 972	120 638 456	58 644 173	51 750 586	21 399
Welschriesling	39 632 572	19 428 545	17 525 921	20 187	40 700 474	19 967 144	18 262 491	19 807

Table S3 – *sequencing and alignment metrics for soft berries samples, divided according the two replicates. The four column are: "output reads": the number of sequenced reads; "trimmed pairs": number reads that successfully passed the quality filter described in paragraph 3.2. This number refer to the pair of reads as the unpaired ones were discarded; "mapped pairs": number of reads pair uniquely mapped to the reference; "number genes": number of genes with a minimum of 5 reads uniquely mapped on.*

SOFT BERRIES SAMPLES	Replica 1				Replica 2			
	output reads	trimmed pairs	mapped pairs	number genes	output reads	trimmed pairs	mapped pairs	number genes
Aglianico	44 360 498	21 611 641	18 224 352	19 300	46 723 338	22 732 442	18 414 420	19 286
Airen	42 857 404	20 969 555	18 863 384	19 352	56 726 572	27 759 899	24 883 312	19 785
Alexandroouli	53 435 940	6 595 383	6 121 541	19 532	48 517 646	23 994 137	21 956 400	19 130
Ansonica	39 690 788	19 557 992	17 938 763	19 021	47 609 816	23 445 964	21 423 244	19 513
Barbera	50 102 786	24 329 531	21 341 136	18 878	50 031 604	24 319 647	21 098 790	19 437
Bayan Shirei	96 111 668	46 145 984	40 515 441	20 186	170 249 516	82 233 119	71 974 682	21 417
Berzamino	47 220 988	23 255 128	21 507 531	20 018	50 586 646	24 887 124	22 837 481	19 852
Bombino Bianco	58 020 834	28 207 115	26 011 495	19 113	40 180 810	19 349 514	16 604 608	18 430
Cabernet Franc	38 621 790	18 854 392	16 688 453	17 916	35 339 626	17 271 275	14 980 310	17 875
Cabernet Sauvignon	--	--	--	--	45 737 816	22 282 938	19 641 512	19 125
Cesanese d'Affile	52 116 848	25 676 756	22 808 736	19 860	43 254 916	21 105 743	17 781 479	18 551
Chasselas Blanc	42 480 782	20 814 521	19 252 108	19 488	40 411 606	19 864 961	18 092 224	19 215
Falanghina	--	--	--	--	45 579 440	22 276 245	1 896 853	19 164
Fiano	47 173 668	23 132 776	20 343 723	19 089	50 212 308	24 685 725	22 419 776	18 992
Fumat	45 865 760	22 295 509	20 106 812	19 001	101 478 066	48 831 891	42 392 460	20 743
Garganega	50 284 810	24 913 197	22 977 680	19 326	51 461 894	25 425 835	22 939 209	18 973
Garnacha	39 475 810	19 447 536	17 406 204	18 690	41 321 810	20 279 159	18 232 683	19 016
Gordin Verde	47 287 594	23 429 026	21 502 644	19 338	51 216 174	25 375 497	23 363 229	19 694
Greco di Tufo	179 976 284	88 196 231	75 592 048	24 931	78 851 510	38 906 944	32 970 760	20 054
Grignolino	106 253 872	51 655 615	45 920 506	20 956	118 373 436	57 687 702	51 527 485	21 141
Heunisch Weiss	186 621 988	90 737 302	70 806 563	26 225	57 924 358	28 419 051	24 977 348	20 383
Kadarka	42 363 102	20 921 414	19 148 571	19 523	51 427 152	25 473 483	23 825 883	20 163
Kölner Blau	104 923 188	50 162 043	43 840 240	20 629	112 750 928	54 019 346	48 755 049	20 554
Lambrusco di Sorbara	52 714 666	25 668 385	21 600 698	18 389	43 761 318	21 477 471	18 261 603	18 483
Lambrusco Grasparossa	45 784 240	22 438 183	19 175 896	19 263	48 899 178	24 015 497	20 767 434	18 689
Malvasia Istriana	38 717 848	18 992 307	17 123 608	19 136	44 430 930	21 818 854	19 618 738	19 471
Merlot Noir	44 761 942	21 963 463	19 462 706	19 027	36 224 368	17 788 961	15 589 807	18 265
Montepulciano	65 664 644	32 069 128	28 539 605	19 649	46 430 504	22 684 310	20 363 843	19 175
Muscat a Petits Grains B.	46 557 010	22 985 124	21 134 454	19 473	51 871 942	25 510 276	23 430 433	19 503
Negro Amaro	56 758 730	27 512 975	22 576 620	19 692	45 597 716	22 234 195	18 107 236	19 086
Pecorino	46 129 714	22 602 080	19 499 326	17 866	34 798 196	17 125 702	15 080 686	18 499
Picolit	50 512 280	24 786 999	21 650 905	19 445	62 819 900	30 885 180	27 279 371	20 491
Pignoletto	42 008 410	20 584 914	18 181 244	19 450	43 972 876	21 603 166	19 237 272	19 519
Pinot Noir	41 304 698	20 449 833	19 219 590	17 841	42 817 778	13 737 023	12 642 774	19 303
Refosco P.R.	55 790 722	27 593 473	25 288 907	19 916	52 876 856	26 049 149	23 328 085	19 721

Ribolla Gialla	39 285 084	19 237 854	16 985 261	18 627	41 823 740	20 494 462	18 520 546	19 024
Riesling Weiss	40 717 102	20 015 452	17 275 404	19 293	43 140 784	21 233 034	18 779 470	19 491
Rkatsiteli	48 199 900	23 749 716	20 655 000	19 182	58 019 382	28 619 824	25 487 553	19 645
Sagrantino	47 148 954	23 214 224	20 880 383	19 040	50 636 538	24 833 509	22 129 219	19 494
Sangiovese	46 213 766	22 437 440	19 011 127	19 690	46 981 416	22 912 017	20 015 079	19 665
Sauvignon Blanc	40 530 358	19 918 120	17 544 592	18 467	224 483 232	109 212 468	94 330 079	21 574
Savagnin Blanc	50 069 436	24 778 043	22 641 247	18 472	56 987 310	28 059 121	24 935 086	20 920
Schiava Grossa	44 291 248	21 515 144	18 025 612	19 148	52 093 394	25 424 879	21 624 674	19 605
Schioppettino	76 954 910	37 884 956	34 156 939	20 205	76 469 696	37 552 324	32 265 327	20 086
Sciavtsitska	48 803 692	9 424 526	8 608 061	19 836	48 339 704	23 941 755	21 465 460	20 183
Sultanina	48 864 542	24 060 852	21 766 171	18 352	--	--	--	--
Tavkveri	59 729 760	29 294 011	26 880 399	20 381	67 234 080	32 816 080	29 177 214	20 463
Terrano	43 957 248	21 640 945	19 429 861	19 071	50 822 770	24 835 496	20 789 136	19 046
Tocai Friulano	42 328 070	20 788 920	18 551 997	19 502	42 321 548	20 756 849	18 785 545	19 314
Trebbiano Toscano	43 058 642	21 106 262	18 274 667	19 080	43 162 682	21 046 140	18 780 721	18 762
Tribidrag	38 906 600	19 232 234	17 574 573	19 814	43 540 634	21 497 380	19 210 053	19 344
V411	75 316 166	37 051 742	33 005 761	20 095	105 881 934	52 236 569	46 422 037	21 349
Verdicchio Bianco	45 204 774	22 122 047	20 321 871	19 224	49 622 676	24 339 203	22 409 929	19 676
Verduzzo Friulano	52 793 072	25 741 787	23 510 964	19 491	69 233 452	34 018 616	29 930 011	20 078
Vermentino	42 588 456	21 023 613	18 902 406	18 492	47 361 012	23 434 557	21 544 559	19 034
Vernaccia S.G.	127 407 800	60 860 435	54 730 166	21 085	157 965 058	75 503 764	68 457 689	21 558
Welschriesling	68 150 132	33 773 284	31 186 528	20 513	51 747 310	25 617 929	23 624 489	20 019

Figure S1: *Scatterplot of relation between p-value of eQTLs and distance between eVariant and eGene. The y coordinates fo the point are the -log10 of eQTL pvalue, while in the x-axis are the distances in bp, with the 0 centered on the eGene TSS. The red lines mark the -1e5 and 1e5 distances.*

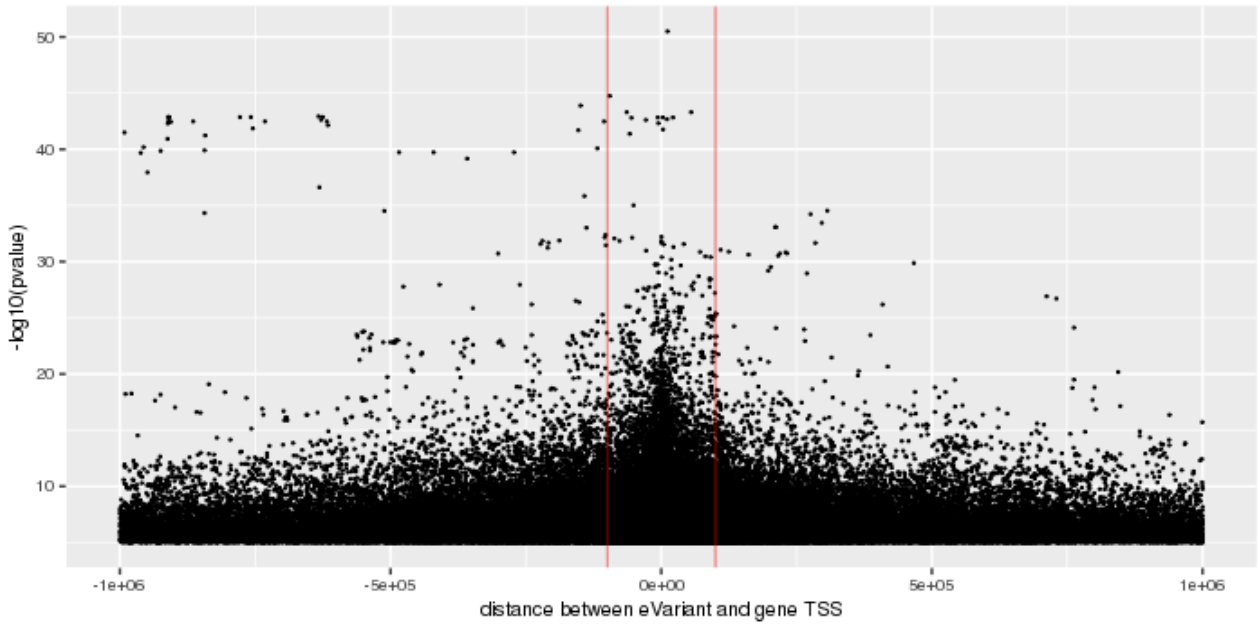

Table S4: *Gene Ontology category enriched in eGenes obtained in eQTLs analysis in leaves. The first two column are the GO tag of the category and its description. "Annotated" is the number of genes belonging to that category present in the reference group, "eGenes" is the number of gene of that category present in our eGene list and "expected" the number of genes expected if the category was not enriched. The last column is the p-value of the enrichment test (fisher test).*

GO.ID	Term	Annotated	eGenes	Expected	classicFisher
GO:0006952	defence response	1898	450	263.78	< 1e-30
GO:0008299	isoprenoid biosynthetic process	229	53	31.83	1.00E-04
GO:0006720	isoprenoid metabolic process	253	57	35.16	1.20E-04
GO:0006721	terpenoid metabolic process	213	49	29.6	2.10E-04
GO:0016114	terpenoid biosynthetic process	192	45	26.68	2.40E-04
GO:0009863	salicylic acid mediated signalling pathways	78	23	10.84	2.70E-04
GO:0043207	response to external biotic stimulus	1108	194	153.99	2.70E-04
GO:0051707	response to other organism	1108	194	153.99	2.70E-04
GO:2000031	regulation of salicylic acid mediated signalling pathway	46	16	6.39	2.90E-04
GO:0016045	detection of bacterium	22	10	3.06	3.40E-04
GO:0071732	cellular response to nitric oxide	12	7	1.67	4.10E-04
GO:0002229	defence response to oomycetes	116	30	16.12	4.40E-04
GO:0071236	cellular response to antibiotic	97	26	13.48	5.80E-04
GO:0051704	multi-organism process	1662	275	230.98	7.00E-04
GO:0071731	response to nitric oxide	13	7	1.81	7.90E-04
GO:0009607	response to biotic stimulus	1159	198	161.07	8.60E-04
GO:1900150	regulation of defence response to fungus	65	19	9.03	1.00E-03
GO:0006857	oligopeptide transport	51	16	7.09	1.07E-03
GO:0006268	DNA unwinding involved in DNA replication	21	9	2.92	1.15E-03
GO:0009861	jasmonic acid and ethylene-dependent systemic resistance	21	9	2.92	1.15E-03
GO:0046777	protein autophosphorylation	277	57	38.5	1.33E-03
GO:0008655	pyrimidine-containing compound salvage	14	7	1.95	1.39E-03
GO:0009605	response to external stimulus	1529	252	212.5	1.47E-03
GO:0060249	anatomical structure homeostasis	77	21	10.7	1.48E-03
GO:0019264	glycine biosynthetic process from serine	11	6	1.53	1.76E-03
GO:0071407	cellular response to organic cyclic compound	199	43	27.66	1.88E-03
GO:0006563	L-serine metabolic process	35	12	4.86	1.89E-03
GO:0035672	oligopeptide transmembrane transport	49	15	6.81	1.99E-03
GO:0002833	positive regulation of response to biotic stimulus	74	20	10.28	2.14E-03
GO:0031347	regulation of defence response	295	59	41	2.23E-03
GO:0045088	regulation of innate immune response	123	29	17.09	2.58E-03
GO:1900426	positive regulation of defence response to bacterium	60	17	8.34	2.61E-03
GO:0071281	cellular response to iron ion	28	10	3.89	3.17E-03
GO:0032103	positive regulation of response to external stimulus	77	20	10.7	3.55E-03
GO:0009130	pyrimidine nucleoside monophosphate biosynthetic process	20	8	2.78	3.64E-03
GO:0034614	cellular response to reactive oxygen species	57	16	7.92	3.81E-03
GO:0006545	glycine biosynthetic process	13	6	1.81	5.11E-03
GO:0098781	ncRNA transcription	17	7	2.36	5.38E-03
GO:0000723	telomere maintenance	70	18	9.73	6.16E-03
GO:0016098	monoterpenoid metabolic process	26	9	3.61	6.39E-03

GO:0043173	nucleotide salvage	26	9	3.61	6.39E-03
GO:0043902	positive regulation of multi-organism process	81	20	11.26	6.55E-03
GO:1900424	regulation of defence response to bacterium	87	21	12.09	7.14E-03
GO:0009129	pyrimidine nucleoside monophosphate metabolic process	22	8	3.06	7.16E-03
GO:0009870	defence response signalling pathway	18	7	2.5	7.76E-03
GO:0007004	telomere maintenance via telomerase	36	11	5	7.77E-03
GO:0030048	actin filament-based movement	36	11	5	7.77E-03

Table S5 - *Gene Ontology category enriched in eGenes obtained in eQTLs analysis in hard berries. The first two column are the GO tag of the category and its description. "Annotated" is the number of genes belonging to that category present in the reference group, "eGenes" is the number of gene of that category present in our eGene list and "expected" the number of genes expected if the category was not enriched. The last column is the p-value of the enrichment test (fisher test).*

GO.ID	Term	Annotated	eGenes	Expected	classicFisher
GO:0051704	multi-organism process	1761	166	129.01	3.20E-04
GO:1903047	mitotic cell cycle process	341	43	24.98	3.40E-04
GO:0022402	cell cycle process	556	63	40.73	3.50E-04
GO:0050896	response to stimulus	6198	510	454.05	4.00E-04
GO:0006268	DNA unwinding involved in DNA replication	21	7	1.54	5.20E-04
GO:0000724	double-strand break repair via homologous recombination	96	17	7.03	5.60E-04
GO:0016045	detection of bacterium	22	7	1.61	7.10E-04
GO:0009820	alkaloid metabolic process	42	10	3.08	7.30E-04
GO:0000725	recombinational repair	99	17	7.25	8.00E-04
GO:0000278	mitotic cell cycle	380	45	27.84	9.60E-04
GO:0033554	cellular response to stress	1164	113	85.27	1.11E-03
GO:0043207	response to external biotic stimulus	1174	112	86	2.11E-03
GO:0007049	cell cycle	835	83	61.17	2.62E-03
GO:0009856	pollination	351	40	25.71	3.57E-03
GO:0044706	multi-multicellular organism process	351	40	25.71	3.57E-03
GO:0006302	double-strand break repair	144	20	10.55	4.19E-03
GO:0098542	defence response to other organism	936	90	68.57	4.53E-03
GO:0007004	telomere maintenance via telomerase	37	8	2.71	4.62E-03
GO:0009607	response to biotic stimulus	1228	114	89.96	4.68E-03
GO:0009605	response to external stimulus	1599	144	117.14	4.77E-03
GO:0051646	mitochondrion localization	17	5	1.25	6.15E-03
GO:0010389	regulation of G2/M transition of mitotic cell cycle	31	7	2.27	6.16E-03
GO:0006432	phenylalanyl-tRNA aminoacylation	11	4	0.81	6.23E-03
GO:0043162	ubiquitin-dependent protein catabolic process via the multivesicular body sorting pathway	11	4	0.81	6.23E-03
GO:0042742	defence response to bacterium	533	55	39.05	6.23E-03
GO:0051259	protein complex oligomerization	91	14	6.67	6.26E-03
GO:0000723	telomere maintenance	74	12	5.42	7.21E-03
GO:0010833	telomere maintenance via telomere lengthening	40	8	2.93	7.58E-03
GO:0034614	cellular response to reactive oxygen species 57	57	10	4.18	7.84E-03
GO:0000086	G2/M transition of mitotic cell cycle	33	7	2.42	8.81E-03
GO:0006278	RNA-dependent DNA biosynthetic process	41	8	3	8.82E-03
GO:0051321	meiotic cell cycle	217	26	15.9	9.01E-03
GO:0009814	defence response, incompatible interaction	218	26	15.97	9.54E-03
GO:0007093	mitotic cell cycle checkpoint	50	9	3.66	9.64E-03
GO:0016973	poly(A)+ mRNA export from nucleus	26	6	1.9	9.86E-03
GO:0080186	developmental vegetative growth	19	5	1.39	1.02E-02
GO:0006739	NADP metabolic process	34	7	2.49	1.04E-02
GO:0007215	glutamate receptor signalling pathway	27	6	1.98	1.19E-02

GO:0045144	meiotic sister chromatid segregation	13	4	0.95	1.20E-02
GO:1902749	regulation of cell cycle G2/M phase transition	35	7	2.56	1.22E-02
GO:0006468	protein phosphorylation	1522	134	111.5	1.32E-02
GO:0009617	response to bacterium	626	61	45.86	1.35E-02
GO:0000070	mitotic sister chromatid segregation	71	11	5.2	1.38E-02
GO:0060249	anatomical structure homeostasis	81	12	5.93	1.45E-02
GO:0007346	regulation of mitotic cell cycle	141	18	10.33	1.48E-02

Table S6 - *Gene Ontology category enriched in eGenes obtained in eQTLs analysis in soft berries. The first two column are the GO tag of the category and its description. "Annotated" is the number of genes belonging to that category present in the reference group, "eGenes" is the number of gene of that category present in our eGene list and "expected" the number of genes expected if the category was not enriched. The last column is the p-value of the enrichment test (fisher test)*

GO.ID	Term	Annotated	eGenes	Expected	classicFisher
GO:0016973	poly(A)+ mRNA export from nucleus	25	9	2.14	1.40E-04
GO:1903047	mitotic cell cycle process	339	47	28.96	6.40E-04
GO:0042136	neurotransmitter biosynthetic process	30	9	2.56	6.50E-04
GO:0000278	mitotic cell cycle	378	51	32.29	7.30E-04
GO:0050829	defence response to Gram-negative bacterium	44	11	3.76	9.30E-04
GO:0006268	DNA unwinding involved in DNA replication	21	7	1.79	1.30E-03
GO:0043162	ubiquitin-dependent protein catabolic process via the multivesicular body sorting pathway	11	5	0.94	1.35E-03
GO:0030048	actin filament-based movement	34	9	2.9	1.74E-03
GO:0016045	detection of bacterium	22	7	1.88	1.77E-03
GO:0051646	mitochondrion localization	17	6	1.45	2.09E-03
GO:0022402	cell cycle process	552	67	47.16	2.12E-03
GO:0055114	oxidation-reduction process	1827	188	156.07	3.22E-03
GO:0006950	response to stress	3930	378	335.72	3.58E-03
GO:0042133	neurotransmitter metabolic process	52	11	4.44	3.90E-03
GO:0006405	RNA export from nucleus	67	13	5.72	3.95E-03
GO:0007049	cell cycle	825	92	70.48	4.70E-03
GO:0071166	ribonucleoprotein complex localization	62	12	5.3	5.63E-03
GO:0071426	ribonucleoprotein complex export from nucleus	62	12	5.3	5.63E-03
GO:0000463	maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	21	6	1.79	6.80E-03
GO:0051704	multi-organism process	1730	175	147.79	8.68E-03
GO:0080147	root hair cell development	50	10	4.27	8.74E-03
GO:0048768	root hair cell tip growth	16	5	1.37	8.87E-03
GO:0031503	protein-containing complex localization	74	13	6.32	9.36E-03
GO:0010193	response to ozone	29	7	2.48	9.57E-03
GO:0007004	telomere maintenance via telomerase	36	8	3.08	9.67E-03
GO:0006545	glycine biosynthetic process	11	4	0.94	1.07E-02
GO:0042138	meiotic DNA double-strand break formation	11	4	0.94	1.07E-02
GO:0015740	C4-dicarboxylate transport	23	6	1.96	1.09E-02
GO:0010389	regulation of G2/M transition of mitotic cell cycle	30	7	2.56	1.16E-02
GO:0030001	metal ion transport	355	43	30.33	1.25E-02
GO:0048588	developmental cell growth	120	18	10.25	1.32E-02
GO:0009820	alkaloid metabolic process	38	8	3.25	1.34E-02
GO:0000281	mitotic cytokinesis	147	21	12.56	1.35E-02
GO:0006406	mRNA export from nucleus	46	9	3.93	1.45E-02
GO:0071427	mRNA-containing ribonucleoprotein complex export from nucleus	46	9	3.93	1.45E-02
GO:0000724	double-strand break repair via homologous recombination	95	15	8.12	1.45E-02
GO:0006611	protein export from nucleus	70	12	5.98	1.48E-02

GO:0006544	glycine metabolic process	18	5	1.54	1.51E-02
GO:0019742	pentacyclic triterpenoid metabolic process	18	5	1.54	1.51E-02
GO:0051321	meiotic cell cycle	214	28	18.28	1.55E-02
GO:0010833	telomere maintenance via telomere lengthening	39	8	3.33	1.57E-02
GO:1901991	negative regulation of mitotic cell cycle phase transition	39	8	3.33	1.57E-02
GO:0051168	nuclear export	79	13	6.75	1.59E-02
GO:0000086	G2/M transition of mitotic cell cycle	32	7	2.73	1.65E-02
GO:0006563	L-serine metabolic process	32	7	2.73	1.65E-02
GO:0061640	cytoskeleton-dependent cytokinesis	150	21	12.81	1.67E-02
GO:0001505	regulation of neurotransmitter levels	63	11	5.38	1.68E-02
GO:0006812	cation transport	606	67	51.77	1.71E-02
GO:0006278	RNA-dependent DNA biosynthetic process	40	8	3.42	1.82E-02
GO:0000725	recombinational repair	98	15	8.37	1.89E-02
GO:0071804	cellular potassium ion transport	81	13	6.92	1.93E-02
GO:0010014	meristem initiation	33	7	2.82	1.94E-02
GO:0051640	organelle localization	162	22	13.84	2.01E-02
GO:0000723	telomere maintenance	73	12	6.24	2.02E-02
GO:0030007	cellular potassium ion homeostasis	13	4	1.11	2.03E-02
GO:0045144	meiotic sister chromatid segregation	13	4	1.11	2.03E-02

Figure S2 - *AI levels, divided in categories, in Hard and Soft berries of genes wich display a monoallelic expression in leaves, in every cultivar. The categories are "MAG" (monoallelic), "AI > 2", "1 < AI <2" and "AI < 1" following the AI level, and are divided in "agreeing" and "disagreeing". This indicate if the haplotype more expressed in the tissue is the same allele expressed in leaves, or if it is the silenced one. "low informative reads" refers to the cases with too few reads to be analyzed.*

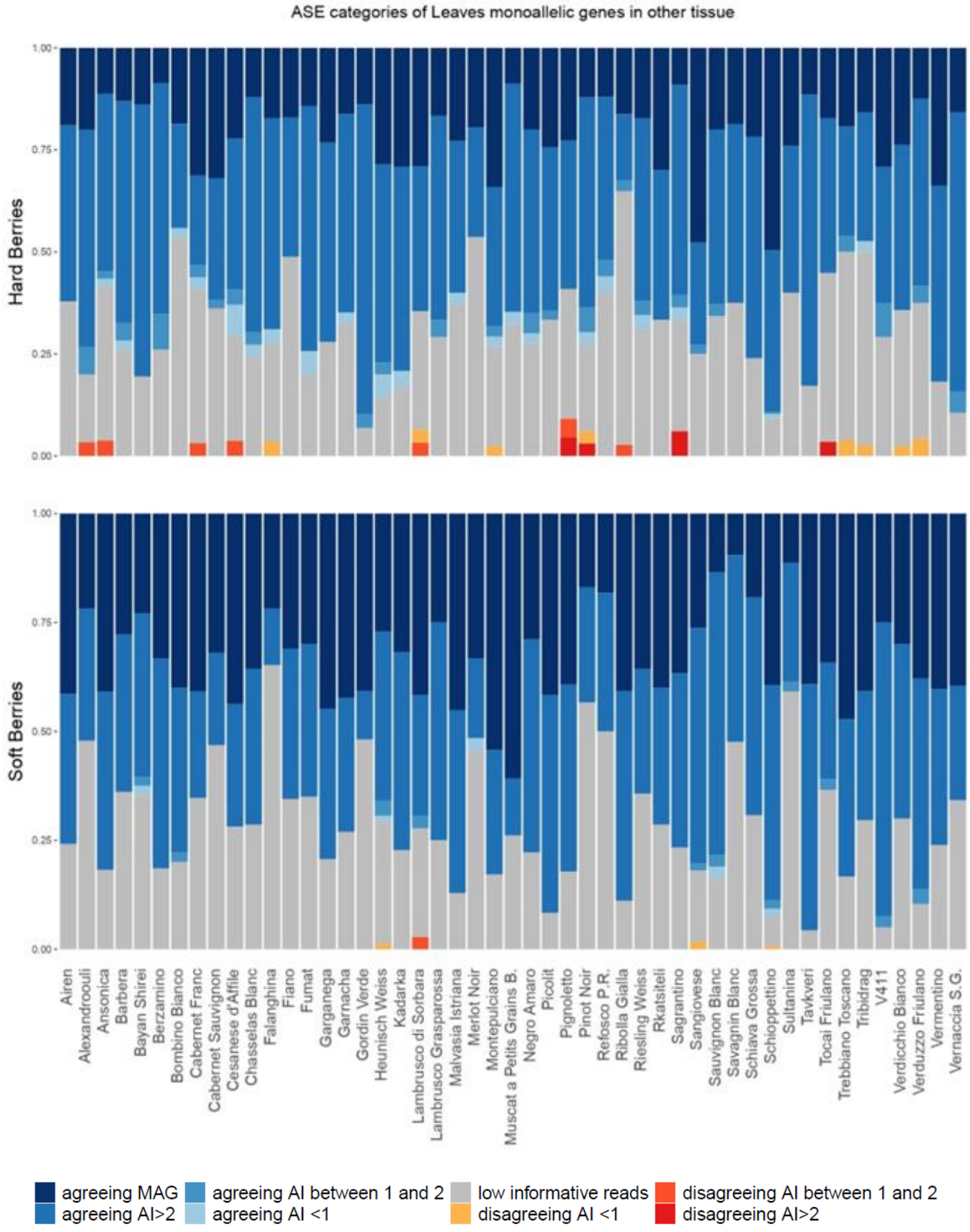

Figure S3 - *AI levels, divided in categories, in leaves and soft berries of genes which display a monoallelic expression in hard berries, in every cultivar. The categories are "MAG": monoallelic, "AI > 2", "1 < AI <2" and "AI < 1" following the AI level, and are divided in "agreeing" and "disagreeing". This indicate if the haplotype more expressed in the tissue is the same allele expressed in hard berries, or if it is the silenced one. "low informative reads" refers to the cases with too few reads to be analyzed.*

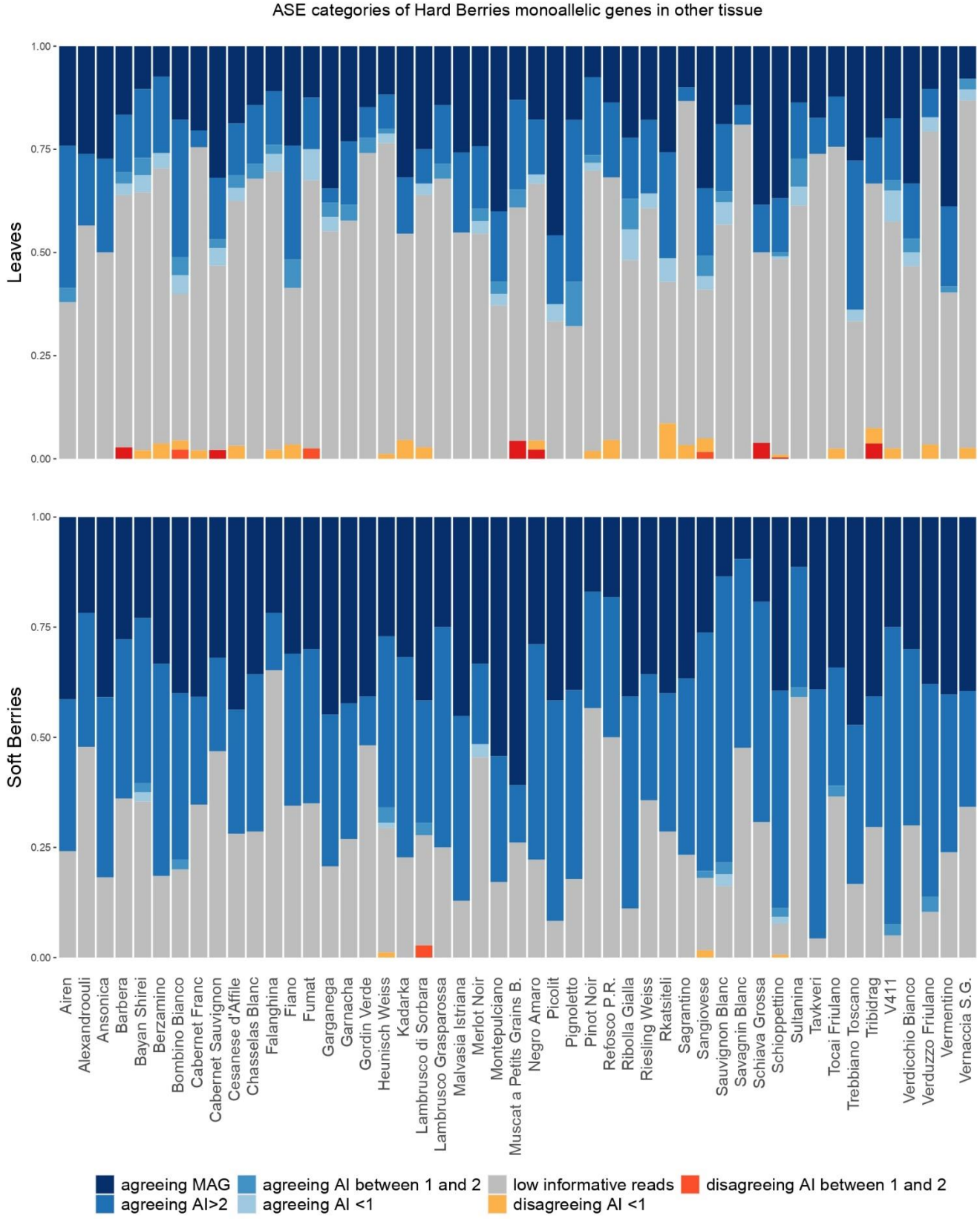

Figure S4 - *AI levels, divided in categories, in leaves and hard berries of genes which display a monoallelic expression in soft berries, in every cultivar. The categories are "MAG": monoallelic, "AI > 2", "1 < AI <2" and "AI < 1" following the AI level, and are divided in "agreeing" and "disagreeing". This indicate if the haplotype more expressed in the tissue is the same allele expressed in soft berries, or if it is the silenced one. "low informative reads" refers to the cases with too few reads to be analyzed.*

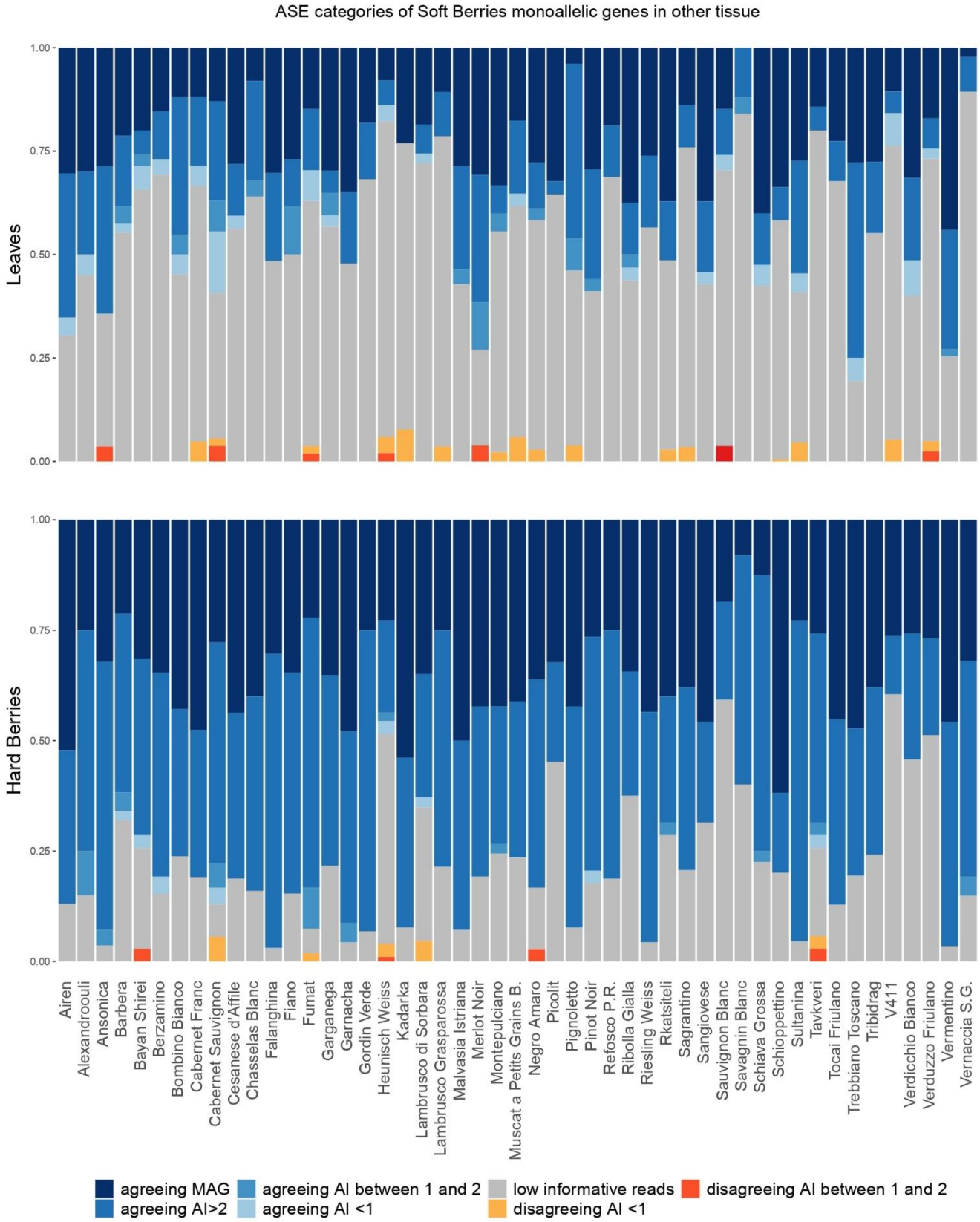

Table S7 - *Gene Ontology category enriched in genes with allelic imbalance in more than 80% of the cultivars tested in leaves. The first two columns are the GO tag of the category and its description. "Annotated" is the number of genes belonging to that category present in the reference group, "Significant" is the number of gene of that category present in our gene list and "expected" the number of genes expected if the category was not enriched. The last column is the p-value of the enrichment test (fisher test).*

Gene Ontology ID	Category description	Annotated	Significant	Expected	Fisher p-value
GO:0019722	calcium-mediated signaling	44	9	2.06	0.00017
GO:0006952	defense response	1562	103	73.09	0.00017
GO:0030001	metal ion transport	286	28	13.38	0.00018
GO:0006950	response to stress	3136	183	146.73	0.00039
GO:0043200	response to amino acid	49	9	2.29	0.00039
GO:0016051	carbohydrate biosynthetic process	315	29	14.74	0.0004
GO:0006857	oligopeptide transport	51	9	2.39	0.00053
GO:0070588	calcium ion transmembrane transport	87	12	4.07	0.00071
GO:0006816	calcium ion transport	88	12	4.12	0.00078
GO:0006811	ion transport	763	55	35.7	0.0009
GO:0071417	cellular response to organonitrogen compound	67	10	3.13	0.00104
GO:0009311	oligosaccharide metabolic process	79	11	3.7	0.00108
GO:0005975	carbohydrate metabolic process	883	61	41.32	0.00136
GO:1901699	cellular response to nitrogen compound	95	12	4.45	0.00155
GO:1901615	organic hydroxy compound metabolic process	283	25	13.24	0.00175
GO:0098662	inorganic cation transmembrane transport	364	30	17.03	0.00191
GO:0009250	glucan biosynthetic process	140	15	6.55	0.00232
GO:0098655	cation transmembrane transport	402	32	18.81	0.00238
GO:0019740	nitrogen utilization	13	4	0.61	0.00242
GO:0030244	cellulose biosynthetic process	88	11	4.12	0.00261
GO:0010025	wax biosynthetic process	41	7	1.92	0.00266
GO:0009820	alkaloid metabolic process	31	6	1.45	0.00277
GO:0055085	transmembrane transport	1039	68	48.62	0.00287
GO:0006629	lipid metabolic process	912	61	42.67	0.00289
GO:0010166	wax metabolic process	42	7	1.97	0.00307
GO:0034637	cellular carbohydrate biosynthetic process	219	20	10.25	0.00332
GO:0044262	cellular carbohydrate metabolic process	402	31	18.81	0.00442
GO:0009699	phenylpropanoid biosynthetic process	122	13	5.71	0.00463
GO:0006082	organic acid metabolic process	1024	66	47.91	0.0047
GO:0098660	inorganic ion transmembrane transport	407	31	19.04	0.00528
GO:0034308	primary alcohol metabolic process	16	4	0.75	0.00551
GO:0009607	response to biotic stimulus	995	64	46.56	0.00559
GO:0010038	response to metal ion	262	22	12.26	0.00585
GO:0051274	beta-glucan biosynthetic process	98	11	4.59	0.00601
GO:0006749	glutathione metabolic process	36	6	1.68	0.00601
GO:0043436	oxoacid metabolic process	1019	65	47.68	0.00631
GO:0000302	response to reactive oxygen species	141	14	6.6	0.00632
GO:0010035	response to inorganic substance	656	45	30.69	0.0064
GO:0030243	cellulose metabolic process	99	11	4.63	0.00648

GO:1901570	fatty acid derivative biosynthetic process	48	7	2.25	0.00658
GO:0009067	aspartate family amino acid biosynthetic process	38	6	1.78	0.00788
GO:0044281	small molecule metabolic process	1566	93	73.27	0.00872
GO:0006833	water transport	28	5	1.31	0.00886
GO:0072511	divalent inorganic cation transport	134	13	6.27	0.01003
GO:0005984	disaccharide metabolic process	52	7	2.43	0.01019
GO:0006081	cellular aldehyde metabolic process	78	9	3.65	0.0104
GO:0006766	vitamin metabolic process	92	10	4.3	0.0107
GO:0000097	sulfur amino acid biosynthetic process	41	6	1.92	0.0114
GO:0000281	mitotic cytokinesis	108	11	5.05	0.01218
GO:0009086	methionine biosynthetic process	20	4	0.94	0.01264
GO:0005986	sucrose biosynthetic process	11	3	0.51	0.01269
GO:0006073	cellular glucan metabolic process	217	18	10.15	0.01342
GO:0044042	glucan metabolic process	217	18	10.15	0.01342
GO:0019752	carboxylic acid metabolic process	944	59	44.17	0.01351
GO:0009765	photosynthesis, light harvesting	31	5	1.45	0.01366
GO:0061640	cytoskeleton-dependent cytokinesis	110	11	5.15	0.01386
GO:0015979	photosynthesis	186	16	8.7	0.01392
GO:0001101	response to acid chemical	1021	63	47.77	0.01405
GO:0080167	response to karrikin	82	9	3.84	0.01419
GO:0051273	beta-glucan metabolic process	111	11	5.19	0.01475
GO:0034220	ion transmembrane transport	563	38	26.34	0.01482
GO:0046364	monosaccharide biosynthetic process	32	5	1.5	0.01558

Table S8 - *Gene Ontology category enriched in genes with allelic imbalance in more than 80% of the cultivars tested in hard berries. The first two columns are the GO tag of the category and its description. "Annotated" is the number of genes belonging to that category present in the reference group, "eGenes" is the number of gene of that category present in our gene list and "expected" the number of genes expected if the category was not enriched. The last column is the p-value of the enrichment test (fisher test).*

Gene Ontology ID	Category description	Annotated	Significant	Expected	Fisher p-value
GO:1901607	alpha-amino acid biosynthetic process	171	34	17.45	0.00011
GO:0009636	response to toxic substance	294	51	30.01	0.00011
GO:0008652	cellular amino acid biosynthetic process	193	37	19.7	0.00012
GO:0008610	lipid biosynthetic process	590	87	60.22	0.00026
GO:0009607	response to biotic stimulus	1068	143	109.01	0.00035
GO:0006629	lipid metabolic process	1003	135	102.38	0.00041
GO:0097164	ammonium ion metabolic process	63	16	6.43	0.00045
GO:0006833	water transport	30	10	3.06	0.00053
GO:0055085	transmembrane transport	1154	151	117.79	0.00067
GO:1901605	alpha-amino acid metabolic process	294	48	30.01	0.0007
GO:0017144	drug metabolic process	688	96	70.23	0.00088
GO:0019742	pentacyclic triterpenoid metabolic process	17	7	1.74	0.00088
GO:0072329	monocarboxylic acid catabolic process	73	17	7.45	0.00089
GO:0006006	glucose metabolic process	43	12	4.39	0.00093
GO:0016052	carbohydrate catabolic process	290	47	29.6	0.00093
GO:0098754	detoxification	156	29	15.92	0.00103
GO:0044282	small molecule catabolic process	230	39	23.48	0.00104
GO:0098660	inorganic ion transmembrane transport	462	68	47.16	0.00123
GO:0016053	organic acid biosynthetic process	604	85	61.65	0.00132
GO:0009070	serine family amino acid biosynthetic process	39	11	3.98	0.00136
GO:0010038	response to metal ion	280	45	28.58	0.00141
GO:0006520	cellular amino acid metabolic process	424	63	43.28	0.00145
GO:0009407	toxin catabolic process	19	7	1.94	0.0019
GO:0044255	cellular lipid metabolic process	765	103	78.09	0.00192
GO:0009404	toxin metabolic process	35	10	3.57	0.00201
GO:0009833	plant-type primary cell wall biogenesis	53	13	5.41	0.00211
GO:0046686	response to cadmium ion	172	30	17.56	0.00241
GO:0006595	polyamine metabolic process	25	8	2.55	0.00255
GO:0031407	oxylipin metabolic process	42	11	4.29	0.00262
GO:0031408	oxylipin biosynthetic process	42	11	4.29	0.00262
GO:0006812	cation transport	551	77	56.24	0.00262
GO:1901566	organonitrogen compound biosynthetic process	1472	182	150.25	0.0028
GO:0044283	small molecule biosynthetic process	792	105	80.84	0.00287
GO:0006722	triterpenoid metabolic process	26	8	2.65	0.00336
GO:0009199	ribonucleoside triphosphate metabolic process	146	26	14.9	0.00339
GO:0046395	carboxylic acid catabolic process	161	28	16.43	0.00344
GO:0006811	ion transport	849	111	86.66	0.00347
GO:0006596	polyamine biosynthetic process	16	6	1.63	0.00364
GO:0031640	killing of cells of other organism	16	6	1.63	0.00364

GO:0000097	sulfur amino acid biosynthetic process	44	11	4.49	0.00388
GO:0016054	organic acid catabolic process	163	28	16.64	0.00411
GO:0032787	monocarboxylic acid metabolic process	544	75	55.53	0.00418
GO:0009620	response to fungus	371	54	37.87	0.0047
GO:0006549	isoleucine metabolic process	12	5	1.22	0.00472
GO:0009205	purine ribonucleoside triphosphate metabolic process	135	24	13.78	0.00485
GO:0006576	cellular biogenic amine metabolic process	58	13	5.92	0.0049
GO:0019693	ribose phosphate metabolic process	275	42	28.07	0.00513
GO:0019319	hexose biosynthetic process	17	6	1.74	0.00514
GO:0030308	negative regulation of cell growth	17	6	1.74	0.00514
GO:0098662	inorganic cation transmembrane transport	415	59	42.36	0.00542
GO:0006952	defense response	1687	203	172.2	0.00554
GO:1901700	response to oxygen-containing compound	1460	178	149.03	0.00555
GO:0009259	ribonucleotide metabolic process	253	39	25.82	0.00583
GO:0009144	purine nucleoside triphosphate metabolic process	138	24	14.09	0.00643
GO:0046777	protein autophosphorylation	255	39	26.03	0.00665
GO:0008216	spermidine metabolic process	13	5	1.33	0.00703
GO:0006631	fatty acid metabolic process	265	40	27.05	0.00755
GO:1902600	proton transmembrane transport	202	32	20.62	0.00792
GO:0044272	sulfur compound biosynthetic process	133	23	13.58	0.00806
GO:0000302	response to reactive oxygen species	156	26	15.92	0.00826
GO:0009201	ribonucleoside triphosphate biosynthetic process	111	20	11.33	0.00826
GO:0016104	triterpenoid biosynthetic process	24	7	2.45	0.00828
GO:0009260	ribonucleotide biosynthetic process	180	29	18.37	0.00882
GO:0046470	phosphatidylcholine metabolic process	19	6	1.94	0.00944
GO:0046390	ribose phosphate biosynthetic process	181	29	18.48	0.00951

Table S9 - *Gene Ontology category enriched in genes with allelic imbalance in more than 80% of the cultivars tested in soft berries. The first two columns are the GO tag of the category and its description. "Annotated" is the number of genes belonging to that category present in the reference group, "eGenes" is the number of gene of that category present in our gene list and "expected" the number of genes expected if the category was not enriched. The last column is the p-value of the enrichment test (fisher test).*

Gene Ontology ID	Category description	Annotated	Significant	Expected	Fisher p-value
GO:0072329	monocarboxylic acid catabolic process	77	21	8.83	0.00011
GO:0042447	hormone catabolic process	28	11	3.21	0.00015
GO:0006952	defense response	1690	240	193.75	0.00015
GO:0009651	response to salt stress	505	85	57.89	0.00016
GO:0000272	polysaccharide catabolic process	176	37	20.18	0.00018
GO:0046244	salicylic acid catabolic process	13	7	1.49	0.00024
GO:0009205	purine ribonucleoside triphosphate metabolic process	129	29	14.79	0.00026
GO:0042737	drug catabolic process	232	45	26.6	0.00027
GO:0042493	response to drug	558	91	63.97	0.0003
GO:0044275	cellular carbohydrate catabolic process	61	17	6.99	0.00036
GO:0000302	response to reactive oxygen species	157	33	18	0.00039
GO:0009144	purine nucleoside triphosphate metabolic process	132	29	15.13	0.0004
GO:0006081	cellular aldehyde metabolic process	84	21	9.63	0.0004
GO:0009201	ribonucleoside triphosphate biosynthetic process	108	25	12.38	0.00043
GO:0009607	response to biotic stimulus	1070	157	122.67	0.00054
GO:0019362	pyridine nucleotide metabolic process	111	25	12.73	0.00066
GO:0009407	toxin catabolic process	19	8	2.18	0.00069
GO:0019336	phenol-containing compound catabolic process	15	7	1.72	0.00072
GO:0009141	nucleoside triphosphate metabolic process	156	32	17.88	0.00073
GO:0046677	response to antibiotic	314	55	36	0.00086
GO:0010193	response to ozone	24	9	2.75	0.00088
GO:0006631	fatty acid metabolic process	273	49	31.3	0.00094
GO:0044247	cellular polysaccharide catabolic process	44	13	5.04	0.00095
GO:0046034	ATP metabolic process	120	26	13.76	0.00096
GO:0042221	response to chemical	2592	344	297.15	0.00099
GO:0009142	nucleoside triphosphate biosynthetic process	114	25	13.07	0.00099
GO:0046777	protein autophosphorylation	253	46	29	0.001
GO:0072521	purine-containing compound metabolic process	253	46	29	0.001
GO:0009145	purine nucleoside triphosphate biosynthetic process	97	22	11.12	0.00123
GO:0006970	response to osmotic stress	577	90	66.15	0.00138
GO:1901700	response to oxygen-containing compound	1482	206	169.9	0.0014
GO:0006754	ATP biosynthetic process	92	21	10.55	0.00144
GO:0055086	nucleobase-containing small molecule metabolic process	444	72	50.9	0.00144
GO:0019359	nicotinamide nucleotide biosynthetic process	80	19	9.17	0.00145
GO:0098754	detoxification	163	32	18.69	0.00158
GO:0072524	pyridine-containing compound metabolic process	118	25	13.53	0.00166
GO:0006629	lipid metabolic process	996	144	114.18	0.00167
GO:0019363	pyridine nucleotide biosynthetic process	81	19	9.29	0.00169
GO:0044042	glucan metabolic process	253	45	29	0.00178

GO:0009056	catabolic process	1546	213	177.24	0.00181
GO:0055085	transmembrane transport	1149	163	131.72	0.00192
GO:0006096	glycolytic process	70	17	8.03	0.00195
GO:0019693	ribose phosphate metabolic process	270	47	30.95	0.00225
GO:0072525	pyridine-containing compound biosynthetic process	89	20	10.2	0.00225
GO:1901135	carbohydrate derivative metabolic process	747	111	85.64	0.00227
GO:0042866	pyruvate biosynthetic process	71	17	8.14	0.0023
GO:0044264	cellular polysaccharide metabolic process	314	53	36	0.00245
GO:0006733	oxidoreduction coenzyme metabolic process	128	26	14.67	0.00254
GO:0009311	oligosaccharide metabolic process	90	20	10.32	0.00259
GO:0015851	nucleobase transport	14	6	1.61	0.00299
GO:0044255	cellular lipid metabolic process	769	113	88.16	0.00302
GO:0006165	nucleoside diphosphate phosphorylation	79	18	9.06	0.00311
GO:0009135	purine nucleoside diphosphate metabolic process	73	17	8.37	0.00314
GO:0009833	plant-type primary cell wall biogenesis	50	13	5.73	0.00339
GO:0051259	protein complex oligomerization	86	19	9.86	0.0035
GO:0006753	nucleoside phosphate metabolic process	364	59	41.73	0.00373
GO:0046939	nucleotide phosphorylation	81	18	9.29	0.00413
GO:1901575	organic substance catabolic process	1375	188	157.63	0.00465
GO:0006812	cation transport	548	83	62.82	0.00467
GO:0009132	nucleoside diphosphate metabolic process	82	18	9.4	0.00474
GO:0009117	nucleotide metabolic process	361	58	41.39	0.00485
GO:0044283	small molecule biosynthetic process	789	114	90.45	0.00501
GO:0006805	xenobiotic metabolic process	11	5	1.26	0.00502

Table S10 - *Gene Ontology category enriched in genes monoallelic expression in leaves. The first two column are the GO tag of the category and its description. "Annotated" is the number of genes belonging to that category present in the reference group, "Monoallelic" is the number of gene of that category present in our monoallelic gene list list and "expected" the number of genes expected if the category was not enriched. The last column is the p-value of the enrichment test (fisher test).*

Gene Ontology ID	Category description	Annotated	Monoallelic	Expected	Fisher p-value
GO:0006418	tRNA aminoacylation for protein translat...	60	16	5.79	0.00013
GO:0043038	amino acid activation	64	16	6.18	0.00029
GO:0043039	tRNA aminoacylation	64	16	6.18	0.00029
GO:1902290	positive regulation of defense response to oomycetes	36	11	3.48	0.0004
GO:0016098	monoterpenoid metabolic process	21	8	2.03	0.00048
GO:0021700	developmental maturation	153	28	14.77	0.00068
GO:0035235	ionotropic glutamate receptor signaling pathway	28	9	2.7	0.0009
GO:0031640	killing of cells of other organism	14	6	1.35	0.00122
GO:0009699	phenylpropanoid biosynthetic process	122	23	11.78	0.0013
GO:0071236	cellular response to antibiotic	81	17	7.82	0.00162
GO:0019748	secondary metabolic process	241	38	23.27	0.00168
GO:0071695	anatomical structure maturation	132	24	12.74	0.00173
GO:0048765	root hair cell differentiation	62	14	5.99	0.00198
GO:0009820	alkaloid metabolic process	31	9	2.99	0.00201
GO:1902288	regulation of defense response to oomycetes	43	11	4.15	0.00205
GO:0071446	cellular response to salicylic acid stimulus	69	15	6.66	0.00207
GO:0035834	indole alkaloid metabolic process	11	5	1.06	0.00234
GO:0042742	defense response to bacterium	442	61	42.67	0.00265
GO:0048469	cell maturation	64	14	6.18	0.00271
GO:0002239	response to oomycetes	107	20	10.33	0.0029
GO:0048544	recognition of pollen	93	18	8.98	0.0031
GO:0009863	salicylic acid mediated signaling pathway	65	14	6.28	0.00315
GO:0007215	glutamate receptor signaling pathway	33	9	3.19	0.00322
GO:0050826	response to freezing	33	9	3.19	0.00322
GO:0009620	response to fungus	352	50	33.98	0.00346
GO:0008037	cell recognition	95	18	9.17	0.00394
GO:0071230	cellular response to amino acid stimulus	34	9	3.28	0.00401
GO:0010054	trichoblast differentiation	67	14	6.47	0.00422
GO:0009875	pollen-pistil interaction	104	19	10.04	0.00472
GO:0009698	phenylpropanoid metabolic process	143	24	13.81	0.00508
GO:0051704	multi-organism process	1400	163	135.16	0.0053
GO:0006749	glutathione metabolic process	36	9	3.48	0.00604
GO:0002229	defense response to oomycetes	99	18	9.56	0.00617
GO:0016137	glycoside metabolic process	30	8	2.9	0.00625
GO:0044550	secondary metabolite biosynthetic process	155	25	14.96	0.00722
GO:0050832	defense response to fungus	297	42	28.67	0.00752
GO:0006349	regulation of gene expression by genomic imprinting	14	5	1.35	0.00792
GO:0019722	calcium-mediated signaling	44	10	4.25	0.00798
GO:0048646	anatomical structure formation involved in morphogenesis	173	27	16.7	0.00837
GO:0016138	glycoside biosynthetic process	26	7	2.51	0.00979

GO:2000031	regulation of salicylic acid mediated signaling pathway	39	9	3.77	0.01046
GO:0071514	genetic imprinting	15	5	1.45	0.01095
GO:0019932	second-messenger-mediated signaling	46	10	4.44	0.01099
GO:0006518	peptide metabolic process	646	80	62.37	0.01154
GO:0009751	response to salicylic acid	186	28	17.96	0.01184
GO:0048571	long-day photoperiodism	40	9	3.86	0.01237
GO:0048767	root hair elongation	40	9	3.86	0.01237
GO:0009617	response to bacterium	512	65	49.43	0.01297
GO:0090627	plant epidermal cell differentiation	91	16	8.79	0.013
GO:0010053	root epidermal cell differentiation	76	14	7.34	0.01313
GO:0071407	cellular response to organic cyclic comp...	171	26	16.51	0.01319
GO:0006399	tRNA metabolic process	196	29	18.92	0.01325
GO:1900426	positive regulation of defense response ...	55	11	5.31	0.01467
GO:0051646	mitochondrion localization	16	5	1.54	0.01469

Table S11 - : *Gene Ontology category enriched in genes monoallelic expression in hard berries. The first two column are the GO tag of the category and its description. "Annotated" is the number of genes belonging to that category present in the reference group, "Monoallelic" is the number of gene of that category present in our monoallelic gene list list and "expected" the number of genes expected if the category was not enriched. The last column is the p-value of the enrichment test (fisher test).*

Gene Ontology ID	Category description	Annotated	Monoallelic	Expected	Fisher p-value
GO:0048544	recognition of pollen	96	18	7.25	0.00027
GO:0008037	cell recognition	98	18	7.4	0.00035
GO:0009875	pollen-pistil interaction	108	19	8.15	0.00043
GO:0009856	pollination	279	36	21.07	0.0011
GO:0044706	multi-multicellular organism process	279	36	21.07	0.0011
GO:0030048	actin filament-based movement	30	8	2.27	0.00134
GO:0016114	terpenoid biosynthetic process	168	24	12.69	0.00184
GO:0006721	terpenoid metabolic process	189	26	14.27	0.0021
GO:0006897	endocytosis	103	16	7.78	0.00439
GO:0050826	response to freezing	36	8	2.72	0.00464
GO:0030029	actin filament-based process	104	16	7.85	0.00483
GO:0048768	root hair cell tip growth	16	5	1.21	0.00526
GO:0030036	actin cytoskeleton organization	96	15	7.25	0.0054
GO:0051645	Golgi localization	11	4	0.83	0.00694
GO:0090436	leaf pavement cell development	11	4	0.83	0.00694
GO:0008299	isoprenoid biosynthetic process	207	26	15.63	0.00717
GO:0007015	actin filament organization	81	13	6.12	0.00738
GO:0019722	calcium-mediated signaling	39	8	2.94	0.00772
GO:0006950	response to stress	3390	289	255.97	0.00874
GO:0006720	isoprenoid metabolic process	232	28	17.52	0.00922
GO:0009310	amine catabolic process	12	4	0.91	0.00979
GO:0006468	protein phosphorylation	1199	112	90.53	0.00987
GO:0098657	import into cell	123	17	9.29	0.01098
GO:0021700	developmental maturation	173	22	13.06	0.01105
GO:0007004	telomere maintenance via telomerase	26	6	1.96	0.01135
GO:0005984	disaccharide metabolic process	59	10	4.45	0.0122
GO:0071695	anatomical structure maturation	146	19	11.02	0.01383
GO:0035235	ionotropic glutamate receptor signaling pathway	20	5	1.51	0.0145
GO:0070475	rRNA base methylation	20	5	1.51	0.0145
GO:0071554	cell wall organization or biogenesis	589	59	44.47	0.0154
GO:0016310	phosphorylation	1499	135	113.19	0.01559
GO:0008610	lipid biosynthetic process	590	59	44.55	0.0159
GO:0010833	telomere maintenance via telomere lengthening	28	6	2.11	0.01633
GO:2000031	regulation of salicylic acid mediated signaling pathway	36	7	2.72	0.01656
GO:0051090	regulation of DNA-binding transcription factor activity	14	4	1.06	0.01754
GO:0051194	positive regulation of cofactor metabolic process	14	4	1.06	0.01754
GO:0000003	reproduction	1518	136	114.62	0.01782
GO:0010091	trichome branching	37	7	2.79	0.01913
GO:0006360	transcription by RNA polymerase I	29	6	2.19	0.01931
GO:1901401	regulation of tetrapyrrole metabolic process	22	5	1.66	0.02173

GO:0010075	regulation of meristem growth	38	7	2.87	0.02197
GO:0120029	proton export across plasma membrane	15	4	1.13	0.02252
GO:0051704	multi-organism process	1506	134	113.72	0.02273
GO:0048765	root hair cell differentiation	65	10	4.91	0.02315
GO:0051193	regulation of cofactor metabolic process	47	8	3.55	0.02318
GO:0022414	reproductive process	1511	134	114.09	0.02497
GO:0009832	plant-type cell wall biogenesis	166	20	12.53	0.02537
GO:0048589	developmental growth	352	37	26.58	0.0254
GO:0050896	response to stimulus	5340	434	403.21	0.02585
GO:0035266	meristem growth	57	9	4.3	0.02607
GO:0010229	inflorescence development	23	5	1.74	0.02608
GO:0007166	cell surface receptor signaling pathway	220	25	16.61	0.02627
GO:0045229	external encapsulating structureorganization	468	47	35.34	0.02717
GO:0006928	movement of cell or subcellular component	126	16	9.51	0.02774
GO:0006629	lipid metabolic process	1003	92	75.73	0.02796
GO:0048469	cell maturation	67	10	5.06	0.02803
GO:0006221	pyrimidine nucleotide biosynthetic process	49	8	3.7	0.02914
GO:0040007	growth	566	55	42.74	0.03156

Table S12 - *Gene Ontology category enriched in genes monoallelic expression in soft berries. The first two column are the GO tag of the category and its description. "Annotated" is the number of genes belonging to that category present in the reference group, "monoallelic" is the number of gene of that category present in our monoallelic gene list and "expected" the number of genes expected if the category was not enriched. The last column is the p-value of the enrichment test (fisher test).*

Gene Ontology ID	Category description	Annotated	Monoallelic	Expected	Fisher p-value
GO:0009310	amine catabolic process	11	5	0.75	0.00047
GO:0006952	defense response	1690	148	115.08	0.00064
GO:0016045	detection of bacterium	18	6	1.23	0.0009
GO:0006468	protein phosphorylation	1193	108	81.24	0.00121
GO:0009875	pollen-pistil interaction	103	16	7.01	0.00156
GO:0035235	ionotropic glutamate receptor signaling ...	20	6	1.36	0.00166
GO:0030048	actin filament-based movement	29	7	1.97	0.00276
GO:0019748	secondary metabolic process	315	35	21.45	0.00289
GO:0048544	recognition of pollen	92	14	6.26	0.00363
GO:0009699	phenylpropanoid biosynthetic process	155	20	10.55	0.00429
GO:0008037	cell recognition	94	14	6.4	0.00442
GO:0050829	defense response to Gram-negative bacter...	40	8	2.72	0.0049
GO:0006576	cellular biogenic amine metabolic process	58	10	3.95	0.00536
GO:0007215	glutamate receptor signaling pathway	25	6	1.7	0.00567
GO:0044550	secondary metabolite biosynthetic process	194	23	13.21	0.00661
GO:1900457	regulation of brassinosteroid mediated signaling pathway	12	4	0.82	0.0068
GO:0009698	phenylpropanoid metabolic process	184	22	12.53	0.00707
GO:0016310	phosphorylation	1490	125	101.46	0.0074
GO:0071230	cellular response to amino acid stimulus	27	6	1.84	0.00843
GO:0009856	pollination	268	29	18.25	0.00902
GO:0044706	multi-multicellular organism process	268	29	18.25	0.00902
GO:0043090	amino acid import	13	4	0.89	0.0093
GO:0070475	rRNA base methylation	20	5	1.36	0.00951
GO:0043038	amino acid activation	63	10	4.29	0.00967
GO:0010584	pollen exine formation	28	6	1.91	0.01013
GO:0030036	actin cytoskeleton organization	96	13	6.54	0.01307
GO:0006749	glutathione metabolic process	47	8	3.2	0.01316
GO:0040011	locomotion	97	13	6.61	0.01418
GO:0048767	root hair elongation	39	7	2.66	0.01502
GO:0055114	oxidation-reduction process	1635	133	111.33	0.01539
GO:0035834	indole alkaloid metabolic process	15	4	1.02	0.01592
GO:0044106	cellular amine metabolic process	68	10	4.63	0.01622
GO:0010927	cellular component assembly involved in morphogenesis	40	7	2.72	0.01717
GO:0010229	inflorescence development	23	5	1.57	0.01742
GO:0006418	tRNA aminoacylation for protein translation	59	9	4.02	0.01761
GO:0043200	response to amino acid	41	7	2.79	0.01952
GO:0003002	regionalization	146	17	9.94	0.02096
GO:0007015	actin filament organization	81	11	5.52	0.021
GO:0006935	chemotaxis	51	8	3.47	0.021
GO:0018105	peptidyl-serine phosphorylation	52	8	3.54	0.02339

GO:0042330	taxis	52	8	3.54	0.02339
GO:0030029	actin filament-based process	104	13	7.08	0.02412
GO:0006595	polyamine metabolic process	25	5	1.7	0.02459
GO:0016132	brassinosteroid biosynthetic process	25	5	1.7	0.02459
GO:0043009	chordate embryonic development	17	4	1.16	0.02491
GO:0010208	pollen wall assembly	34	6	2.32	0.02559
GO:0048646	anatomical structure formation involved in morphogenesis	196	21	13.35	0.02611
GO:0009308	amine metabolic process	106	13	7.22	0.02773
GO:0010183	pollen tube guidance	44	7	3	0.02792
GO:0034754	cellular hormone metabolic process	44	7	3	0.02792
GO:0048764	trichoblast maturation	64	9	4.36	0.02858
GO:0048765	root hair cell differentiation	64	9	4.36	0.02858
GO:0009792	embryo development ending in birth or eg...	18	4	1.23	0.03034
GO:0009626	plant-type hypersensitive response	176	19	11.98	0.03099
GO:0034050	host programmed cell death induced by symbiont	176	19	11.98	0.03099
GO:0018209	peptidyl-serine modification	65	9	4.43	0.03125
GO:0008215	spermine metabolic process	11	3	0.75	0.03438
GO:0008593	regulation of Notch signaling pathway	11	3	0.75	0.03438