Transcriptomic profiling to study the role of MYBA1 and MYB24 in the *Vitis vinifera* variegation

2021

Arnau Peris Cuesta

Tutor: Dr. José Tomás Matus
Institute for Integrative Systems Biology (I2Sysbio)
University of Valencia - Spanish National Research Council (CSIC)

Msc of Bioinformatics
Università di Bologna

INDEX

1. Introduction

1.1 Berry coloration in grapevine

Being one of the most remarkable crops that has been domesticated worldwide, the grapevine (*Vitis vinifera L.*) has shown a broad phenotypic diversity that leads to economic interesting traits during the last 6.000 - 8.000 years [1-3]. Among others, the color and flavor of the berry has been studied from the genetic point of view, leaving several open questions that remain unknown [4-5]. The "Parallel Model" [6] and "Sequential Model" [7] have been suggested to explain the evolutionary events on the formation of berry color somatic variants, but these two models just describe the specific cultivars and have not been generally evaluated considering the highly diversified phenotypic variation of grapevine. Meanwhile, the flavor changes, corresponding to the composition of secondary compounds, in response to the environmental factors is far from being clarified [8].

Somatic mutations can develop different phenotypes on plants of the same cultivar, commonly known as clonal polymorphism. The stratified apical meristems are composed by independent adjacent layers of dividing cells that conform the different organ tissues [9]. Specifically, the shoot apical meristem (SAM) is divided into the epidermis (L1) and the other parts of the plant (L2) [10]. In that sense, somatic variations in a single cell can propagate over the whole layer, evolving in a mutated section and rising to a chimera. An example of the color variation as a consequence of the chimerism is observed in the variegation of the berries, a phenomenon by which the same fruit presents both white and red skin sections.

The variegation has been observed in different cultivars, for instance the cv. 'Pinot Gris' PG52 certified clone [11]. Occasionally, a red skin section was identified as tri-allelic at VVS2 microsatellite loci (chimeric structure consisting of two cell layers of specific allele for each one plus a combined genotype with both alleles), and a diallelic white skin section. According to [12] a displacement from the L2 layer to the L1 could produce the white sections on the cultivar, affecting a few berries or even a whole group, depending on the moment of the development stage that suffers from the displacement.

Indeed, white sections are associated to the absence of anthocyanins that have been related to mutations on MYBA1 and MYB2 genes, for instance the insertion of the *Gret1* retrotransposon in the promoter region of MYBA1 gene, in addition with two non-conservative mutations in MYBA2, both involved in the transcription of UFGT (UDP-glucose flavonoid 3-O-glucosyl

transferase), thus interrupting the anthocyanin synthesis pathway [13-14]. Recently, a shorter insertion in the intron of MYBA1 has been discovered to explain a weakly colored berry by reducing its transcripts. Other mutations have been described in previous studies [7,15-16].

The fruit development is divided into three distinguished phases, having a double sigmoid growth phase separated by a lag phase. The core for this project has been situated in the end of the lag phase, when the onset of berry ripening occurs in the so-called procedure veraison. At this point, the berry pigmentation initiates and it is determined by the biosynthesis of phenolic compounds (mainly flavonols and anthocyanins) in the cultivars [17-18]. Flavonols provide the yellow pigmentation that is directly related to the white and it is masked by anthocyanins in the red sections [4].

1.2 Anthocyanins depletion: R2R3-MYB

Flavonoids, which compresses flavonols and anthocyanins, are synthesized through the phenylpropanoid pathway under the combination of several regulators grouped in three classes: myeloblastosis (R2R3-MYB), basic helix-loop-helix (bHLH) and tryptophan-aspartic acid repeat (WDR/WD40) [19]. Focusing on the first class, the R2R3-MYB transcription factor family corresponds to proteins that present two DNA binding domains and a variant number of C-terminal motifs, comprising an estimated number of 134 genes that have been related to the phenylpropanoid pathway and have probably expanded by gene duplication [20].

Anthocyanin synthesis is also regulated by several members of the R2R3-MYB family, highlighting the MYBA1 and MYBA2 genes located at Chromosome 2, as commented above [14]. Not only the UFGT but also the anthocyanin 3-O-glucoside-6’’-O-acyltransferase (3AT) has shown a tight regulation during the fruit development [15,21]. The activation of the anthocyanin pathway in the fruits depends on the allelic condition of the R2R3-MYBA1/A2, as established before, suggesting an inhibition in the white skin. Interestingly, a recent publication has identified a new and uncharacterized regulator, MYB24, to be involved in the pigmentation in a different manner. The anthocyanins act as a sunscreen, therefore, the lack of those protectors increase the light or radiation that the cells receive, thus, empowering the photosynthesis and production of volatile molecules, such as terpenes. In that way, MYB24 has been suggested as a modulator of light responses including the synthesis of flavonoids (flavonols) and isoprenoids (terpenes, putatively carotenoids). The association between MYB24 and monoterpenes implies that the regulation is broadly triggered

towards the ripening and, most interestingly, the absence of anthocyanins in the white skin accelerates its activation in a dose-dependent manner due to increased radiation exposure [22].

1.3 Combinatorial tools in a bulk RNAseq analysis

The breakthrough of RNAseq has offered researchers a powerful tool with relatively low cost and high reproducibility to perform different assessment analysis of the whole transcriptome [23]. Along with the gene variant calling, gene fusion, *de novo* transcriptome construction and other applications, the differential gene expression analysis has become a standard tool to compare transcriptomic profiles between different conditions in a pairwise manner.

Additionally, when sampling has been performed continually during a time series in a dynamic biological process, a time-course experiment analysis might provide useful information regarding the gene expression trajectories [24]. Raising the complexity of the input data in one dimension by considering the time as an additional variable requires the use of specific statistical tools , however the preprocessing and quality control is common for any differential gene expression (DEG) approach.

Once identified the significant differentially expressed genes, the downstream analysis includes clustering algorithms and co-expression network construction software that attempts to overcome the initial biological hypothesis to be tested. The assumption behind co-expression analysis is that genes involved in the same pathway should present similar expression patterns [25]. Clustering coupled with functional enrichment analysis provide insights into complex data by establishing putative candidates that share a biological function using annotated databases, such as Gene Ontology [26] or MapMan [27].

Integrating the data obtained in the pairwise and time-course approaches along with the downstream analysis might bring stronger validation of the results as well as staking out new hypotheses and alternative ways to move forward.

1.4 Experimental design

Returning to the onset of the ripening, variegated berries have been sampled from the same black-skinned plant of cultivar (cv.) 'Béquignol' at 0,3,6 and 9 weeks after veraison (WAV). For each of the four time points three biological replicates were extracted from both red and white skin sections, arranging a total of 24 samples (*Figure 1)*. Sampling sections from berries that belong to

the same plant reduces significantly the variability since the transcriptomic profiles should be practically identical for both white and red skin, except those regulators and activated genes associated with the differences in the pigmentation. The extraction and purification of the mRNA and the subsequent sequencing was performed according to the Illumina TrueSeq protocol (*See 2.Methodology)*. The data analysis was performed using the computing server Garnatxa located at I^{2}SysBio, a multicomputer cluster with a global archive system with 3,2 Pb of capacity, total RAM over 17 Tb and 648 cores, which distributes the tasks based on the SLURM queueing system.

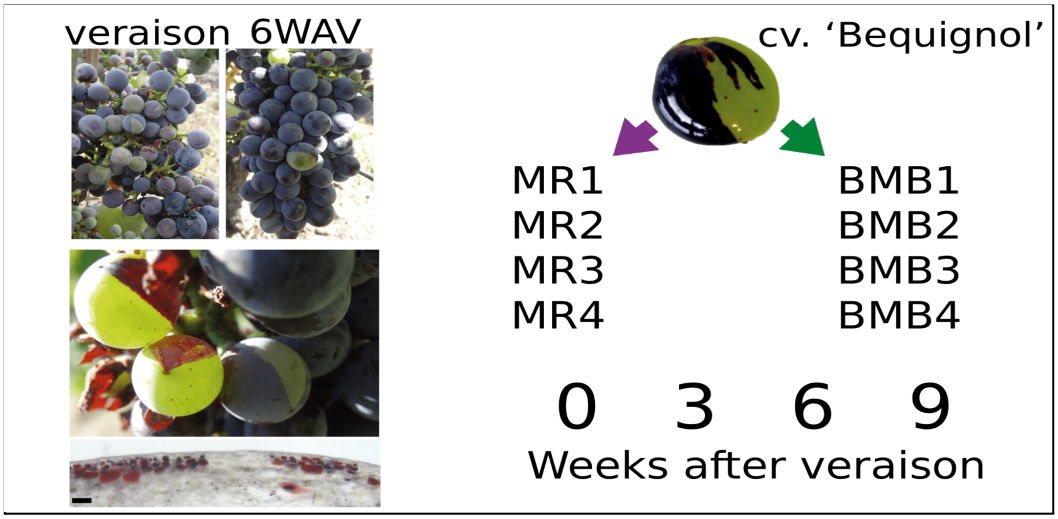

Figure 1. Experimental design representation. White (BMB) and red (MR) skin sections were sampled at four time points after the veraison: 1(0 weeks), 2 (3 weeks), 3 (6 weeks) and 4(9 weeks). On the left it is noticeable the variegation effect and the differences in the pigmentation of the skin sections at 6 weeks after veraison.

Here, I describe the occurence of a natural berry color variegation found in the 'Béquignol' cultivar. Red and white skin sections were compared to understand the origin and consequences of this color alteration, providing new information about the cross-regulation of phenylpropanoids and isoprenoids in response to pigment depletion, thus, establishing a transcriptional association between these two specialized metabolic pathways in plants. The variegation activates the accumulation of several metabolites, highlighting anthocyanins and flavonols, that potentially filter radiation and control the oxidative damage. Using as a gold standard the well-known MYBA1 promoter, our results point out different potential regulators, highlighting the effect of MYB24 in the white-skinned sections during the ripening. This project shows partial results of the work conducted by the Dr. Tomás Matus that was recently published [22] (*See Acknowledgements)*.

2. Methodology

As commented in the previous chapter, the experimental design was based on 4 time points and 3 biological replicates for both red and white skin. Therefore, a **Paired-end 2x150bp Stranded mRNA-Seq Illumina TrueSeq** protocol was applied with a sequencing depth of 60 million reads per sample. In total, it was obtained 48 libraries with 30 million reads per library from the sequencing provider Novogene, receiving the already demultiplexed and compressed FastQ files in PHRED+33V2 score encoding (*See 2.1.1 Quality control chapter)*. The overall analysis is composed of two main parts: **Pre-processing / Quality control (QC)** and **downstream analysis**, in which two parallel pipelines were used, a **pairwise** analysis for each time point and a **time-course** analysis for the entire trajectories over the time. In this chapter it is provided an exhaustive explanation, step by step, plus the algorithms and methods that were used, including a detailed manual for the overall downstream procedure.

2.1 Pre-processing and quality control of the raw data

Denoising and clearing the raw input dataset has become the most important step in any data mining procedure. In most of the cases it requires the major part of the efforts and computational resources and the final results will strictly depend on an accurate pre-processing of the data [28]. Like the standard operating pipeline, it was attempted to identify and remove the poor-quality reads, adapters and other artifacts that can arise from library preparation and sequencing errors. Secondly, reads were aligned into a reference genome and the positive hits were summarized to obtain a raw counts matrix as a final output of the overall procedure (*Figure 2)*.

2.1.1 Quality control with FASTQC

Before clearing the sequences, the first step of processing the RNA-seq data is to check the quality of the raw reads to inspect how accurate the sequencing execution was. For that reason, a Java-based tool was utilized, FASTQC 0.11.9, which provides a modular set of QC analysis that can be used to obtain a first overview of the problems that need to be solved prior to performing any further analysis.

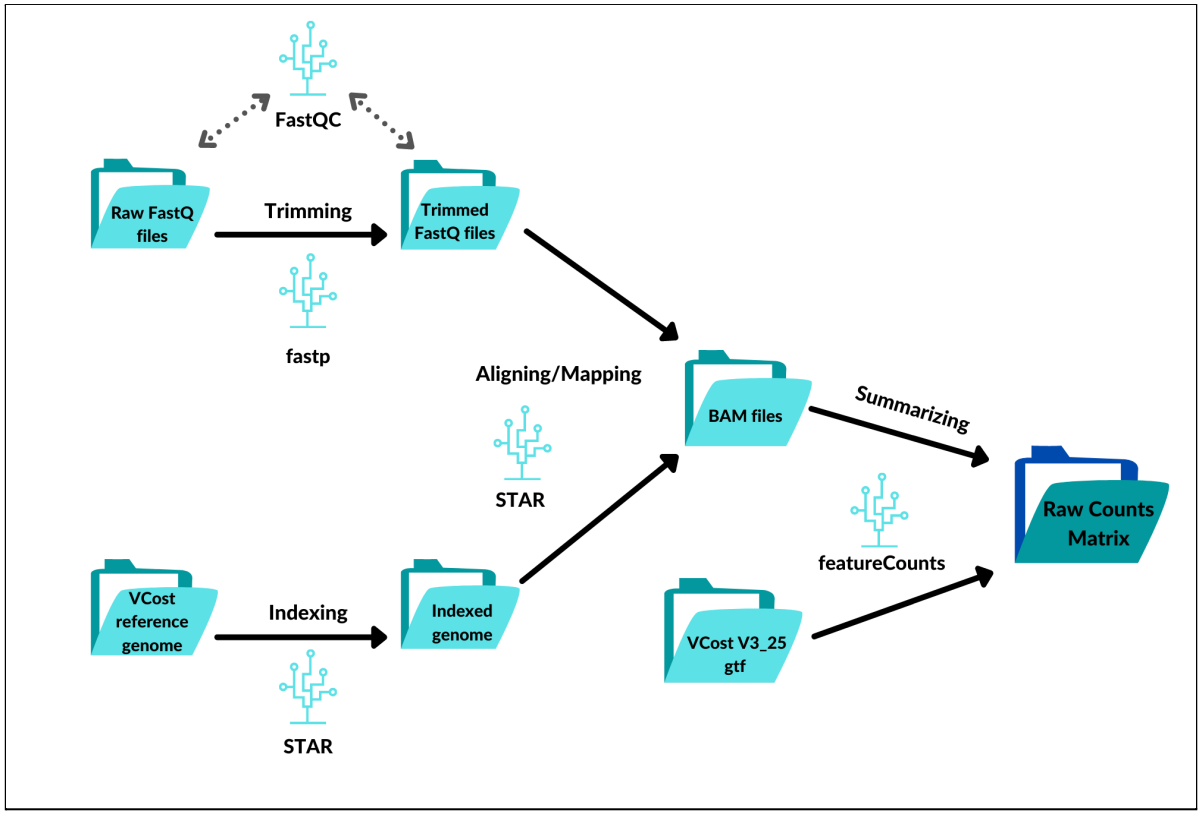

Figure 2. Data Pre-processing flowchart summarizing the methodology. Procedures are indicated with arrows while intermediate files are inside of the folder icons and software used with blue icons.

Apart from the basic statistics, in the output it is reported the per base sequence quality, per sequence quality scores, per base sequence content, per base/sequence GC content, per base N content, sequence length distribution, overrepresented sequences and presence of duplications or Kmers [29]. The quality for each base is encoded according to the Illumina PHRED+33 V2 system [30], in which the probability for each nucleotide calling is defined by the *Equation 1*:

$$Q = -10log_{10}P$$

(Eq.1)

where Q is the quality score that usually ranges from 10 to 40, whereas P is the probability of getting an error in the base calling [31]. Therefore, a PHRED score of 30 will set the probability error in 1 out of 1000.

2.1.2 Trimming reads and adapters with fastp

During the trimming step not only the low quality bases and the adapters are removed but also the short reads, chimeras and other terminal artifacts that might have been generated during the sequencing. Choosing the right tool for trimming can be

challenging and there is not a single valid solution for every approach. Nowadays, there are plenty of free resources available, highlighting *Cutadapt* [32] or the most famous *Trimmomatic* [33] that was presented in *Bioinformatics* journal in 2014. However, a recently published tool, Fastp, has shown interesting results in terms of automated unique molecular identification (UMI), per-read polyG tail trimming and multi-threading implementation (developed in C/C++ in contrast with the other Java/Python tools). Apart from the ultra-fast computing time, which ranges from 2 to 5 times faster (*Table 1)*, and the best overall cleaning of the data compared to the conventional tools [34], the main interest for this project is related to the automated adapter removal for Illumina sequencing that simplifies and speeds up the overall pipeline implementation.

Tool	**Time(min)**	**Throughput(reads/s)**
Fastp	13.3	116.750
FASTQC	25.8	60.185
Cutadapt	24.6	63.120
Trimmomatic	60.9	25.497

Table 1. Speed comparison of Fastp and other trimming software. Results extracted from [34] in which Paired-End B17NCB1 dataset was utilized to generate the results.

Therefore, the trimming was performed with Fastp v0.20.1. The corresponding GitHub repository with all the files, the manual and instructions for other installation methods are freely available at link (https://bit.ly/3cW6bBf).

All the raw demultiplexed and compressed Fastq files were stored and sorted in the same folder. The corresponding code implemented in each stage is fully available at annexes (*See Chapter 6. Annexes*).

For Paired-End data both libraries can be processed in the same run with the '-i/-I' option and specify the correspondent output with '-o/-O'. With '-w' the number of threads are specified whereas the '-l' of 20 will discard all the reads shorter than that limit threshold. In relation to the cutting parameters, '--cut_front_window_size', '--cut_front_mean_quality' and '--cut_front' will establish a window size of 1 base starting from front (5' sense) with a minimum threshold of 30 for the PHRED score, moving the sliding window from front (5') to tail (3'). The same parameters are applied to the tail, ensuring a double check of the reads. Finally, the --n_base_limit of 5 will discard the

pair of reads in which at least one of them has more than 5 ambiguous (N) bases.

At the end, with these parameters a very strict trimming process was applied, in which libraries lost 32% of reads, averaging for each sample. In that sense, the main goal was to ensure that the remaining reads correspond to real transcripts, while sequencing noise was removed.

2.1.3 STAR: genome indexing and mapping

Aiming to match millions of reads generated by high-throughput sequencing technologies that contain the information regarding non-contiguous spliced exons has become the bottleneck in most of the RNA sequencing data analysis protocols, not only in terms of mapping accuracy but, overall, in the computational resources as well. The presence of reads with mismatches and derived from spliced sequences that have to be joined together to reconstruct the full genomic RNA are the main reasons why an RNA-seq mapping can be insanely expensive, from the computational point of view [35]. Several algorithms have been developed to deal with short read technologies (smaller than 200 bases) [36-39] but they are not highly accurate and demand a lot of disk space and running time.

However, nowadays there is a plethora of available tools that work extraordinarily, which makes it difficult to decide the most appropriate software for the analysis. When choosing the right tool, the runtime must be never considered as a primary advantage due to the fact that most of the modern aligners have been implemented with the multithreading option. Therefore, the efficiency of the run is highly dependent on the available computational resources. The main part of the accuracy of an aligner can be defined as the percentage of reads aligned. For a dataset that has been sequenced with high-quality reads it is expected a high degree of mapping rate, however, it also depends on the completeness of the reference genome. Another important issue is the multi-reads detection, fragments that map into multiple regions due to repetitive sequences and must be discarded and reported by the aligner. In conclusion, the selection of the right tool is a trade-off between the available computational resources, the experimental approach to assess and the completeness of the tool, in relation to the limitations described above [40]. In that sense, STAR [41] was selected because of its excellent mapping rate, the fast runtime [40] and its option of inputting a transcriptome alongside the reference to identify putative splice sites. Bear in mind that most the aligners were

created to handle DNA reads so using an RNA-seq aligner will provide a meaningful advantage.

Prior to aligning the trimmed reads, an indexing of the reference genome must be performed. Most of the aligners implement a full-text index in Minute space (FM-Index) and its use to the Burrows - Wheeler transformation (BWT), which performs significantly well in reducing the runtime and memory usage [42]. It is based on an array of suffix rotations in which the full genome is set as the first suffix, then an iterative procedure removes the first character and appends it to the end. For example, for the genomic fragment 'ATG':

1.ATG$	*4.$AT**G***
2.TG$A	*1.ATG**$***
3.G$AT	*3.G$A**T***
4.$ATG	*2.TG$**A***

The lexicographically sorted array will provide the BWT by looking at the last column of the array (in this case, **G$TA**). Of course, for a full genome it will be encountered thousands of times the same character in a row, which allows for compressing the index, ending up in a reduced size of the genome index. In STAR, the generation of the suffix array is generated in the same way as the BWT suffix rotation array, nevertheless, the ending prefixes are not appended [40]. For instance, 'ATGAT':

1.ATGAT$	*4.AT$*
2.TGAT$	*1.ATGAT$*
3.GAT$	*3.GAT$*
4.AT$	*5.T$*
5.T$	*2.TGAT$*
6.$	

The suffixes starting with the same genetic fragment will appear consecutively in the array, providing a very fast scanning when finding exact matches of a read. In this example, looking for mapping the read 'AT' will create two parallel alignments against 'AT$' and 'ATGAT$'. The main advantage of this method compared to FM-Index aligners is avoiding the reconversion of BWT back into the reference genome when aligning the reads. However, using STAR will require a huge amount of disk memory to load all the suffixes, which might present a problem for large genomes on systems down to 32GB of RAM.

The last release for *Vitis vinifera* was used as the reference, the 12X.V2 chromosome assembly which is based on the scaffolds of the grapevine reference genome build (FN594950-FN597014, EMBL release

102; *Vitis vinifera cv. PN40024*) [43]. All the corresponding data is available at (https://urgi.versailles.inra.fr/Species/Vitis/Annotations).

All the files required for compilation, the correspondent documentation and miscellaneous files and scripts are available at the author GitHub repository (https://github.com/alexdobin/STAR).

The 'genomeGenerate' runmode is applied for the indexing. The directory that contains the reference genome will be the output location by default and the name of the file in fasta format must be supplied. In the indexing, 8 threads were used for the parallelization.

Once the genome has been indexed, mapping of the reads is initiated and the BAM files are generated. The sequence alignment map (SAM) files are tab-delimited text format with a header and an alignment section [44]. Its binary equivalent alignment map (BAM) files are a compressed representation that stores the same data and the desired output. In order to get a better compression ratio, grouping similar sequences together and leaving the unmapped reads at the end of the file are referred to as sorted BAM files. Both binary compression and sorting are performed in one implementation with STAR mapping with the option '--outSAMtype BAM SortedByCoordinate'. The run mode 'alignReads' is set for mapping and 16 threads were used for parallelizing this step, which is the most expensive in terms of computational resources.

At the end, all the BAM files corresponding to the 24 samples are generated and all the trimmed Fastq files and control logs reports are removed.

2.1.4 Summarization and assignment to genomic features

The final step in the preprocessing is the read assignment to genomic features, in this case genes according to the VCost.v3 [43] annotation in *gff3* format, the General Feature Format, which consist of a tab-delimited text file that contains the information that can be used to translate any kind of feature, such as CDS, exons, microRNAs, binding domains, ORFs and others (the official documentation of *gff3* can be found at https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md).

For summarizing the gene counts *featureCounts 2.0.1* was utilized, a tool suitable for either RNA or genomic DNA approaches (single or paired-end reads) that provides a very fast and accurate tool in terms of feature assignments. The implementation is based on two

main stages: firstly, a hash table is constructed for the reference sequence names, facilitating the match between BAM and *gff* annotation. Secondly, a hierarchical data structure is applied to the features in each reference sequence. The features are assigned to blocks and blocks are assigned to bins according to each reference sequence starting position [45]. Allowing a hierarchical search is the clue for the rapid assignment of the algorithm, which works faster by an order of magnitude for gene-level summarization than other tools like *HT-seq* [46] or *BedTools* [47]. For the required *featureCounts* input, the annotations need to be formatted from *gff3* to *gtf* (Gene transfer format), which is identical but contains additional conventions specific to gene information [48].

The software is available under GNU General Public License as part of the Subread (http://subread.sourceforge.net) and there is an R adaption accessible through Bioconductor project (http://www.bioconductor.org).

The '-p' option specifies that the input data contain paired-end reads, '-T' assess the number of threads, '-f' performs the summarization at feature level, '-C' avoids counting chimeric fragments (two ends aligned to different locations), '-a' provides the pathing for the annotation file and '-o' gives the name of the output file.

2.2 Pairwise differential expression analysis (DEA) for each time point

The downstream analysis was divided into the pairwise DEA (*detailed explanation in this chapter*) and the time-course DEA (*chapter 2.3)*, which were computed simultaneously in order to obtain different biological conclusions (*Figure 3*).

In order to identify the putative differentially expressed genes for each time point when comparing the red berry skin against the white berry skin, the Bioconductor R package edgeR [49] was applied to the preprocessed reads. Designed for the analysis of digital gene expression data produced by RNA sequencing, several statistical methods are implemented in the package.

2.2.1 EdgeR insights

Over the past years, we have learned that the power to identify differential expression can be improved while reducing the false discoveries by sharing information across all samples. For example, in the limma [50] package, an empirical Bayes model is used to obtain the moderated variances for each gene, which will replace the probe-wise variances in the t- and F-statistic test.

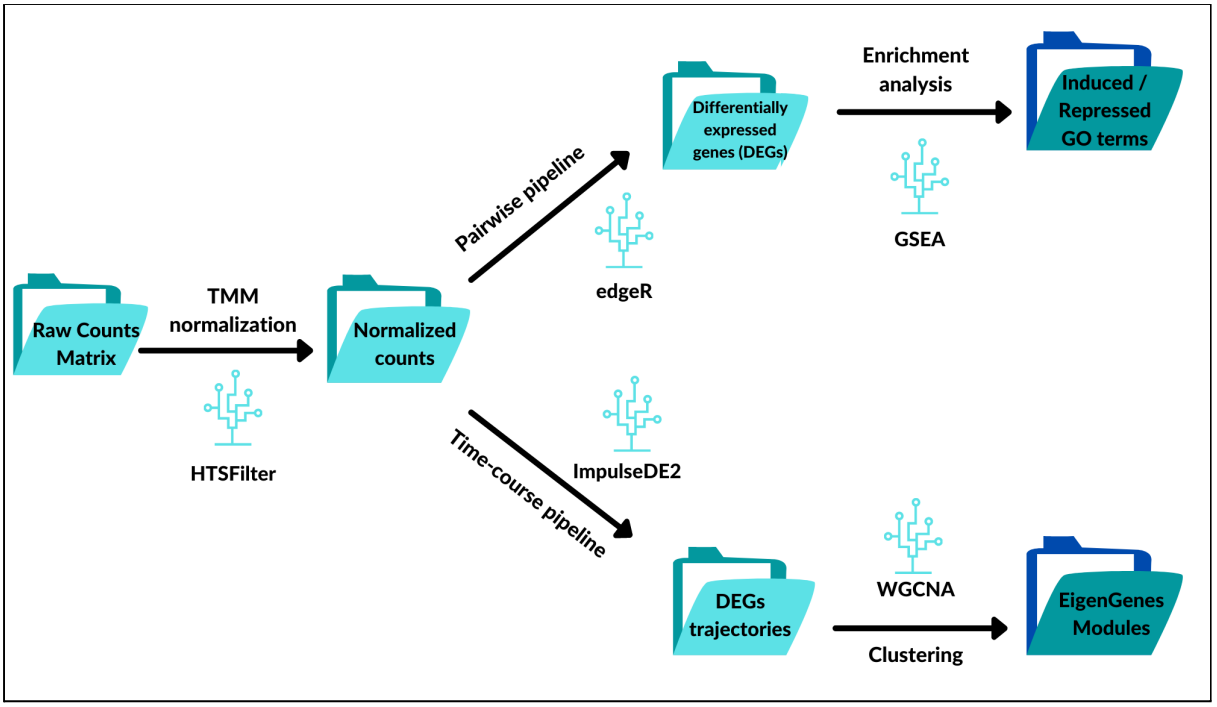

Figure 3. Analysis flowchart for both DEA pipelines. Processes are indicated with arrows, software used with blue icons and the intermediate files are stored inside the folder icons.

On the other hand, with edgeR, the model counts data using an overdispersed Poisson model, and it moderates the degree of overdispersion across the genes with an empirical Bayes model. Data is collected into a table of raw gene counts (or exons or transcripts) and the column names correspond to sample labels. Then, data is modeled as negative binomial (NB) distributed, represented in the following expression (*Equation 2*):

$$Y_{gi} \sim NB(M_i P_{gj}, \phi_g) \qquad (Eq.2)$$

where for each gene (g) and sample (i), M_i represents the total number of reads, Φ_g the dispersion and P_{gj} is the relative abundance for that gene in the experimental group (j) to which the sample belongs. Therefore, for differential expression analysis, P_{gj} is the parameter of interest. Considering that the NB distribution acts as Poisson when ϕ_g=0, many DGE approaches can treat the technical variation as Poisson. In that sense, ϕ_g represents the coefficient of biological variation between the samples, having a model that is able to separate biological from technical variation [49].

Finally, for each gene, dispersions are estimated through conditional maximum likelihood (ML), having a direct effect on the total count for that gene [51]. Then, a consensus value is applied to contract these dispersion outcomes by an empirical Bayes

procedure, aiming to take the information between genes [52]. At the end, the differential expression is performed with an exact analogous Fisher's test adapted to overdispersed data [53].

EdgeR works particularly well when dealing with the estimation of biological variation between replicate libraries, even minimal numbers of replicates [49]. In that sense, considering the experimental design, this package should outperform other options such as DESeq2 [54].

2.2.2 Raw counts normalization with HTSFilter

Before proceeding with the statistical analysis, it is required to normalize the raw counts matrix. In that sense, the **TMM (Trimmed Mean of M-values)** normalization was applied through the HTSfilter Bioconductor R package [55].

The selection of the most appropriate normalization method strictly depends on the experimental design. In this case, the main goal is to compare the expression between conditions, in other words, between-sample contrast, in which the library depth will be highly relevant and the expression comparison between genes is not being considered. Secondly, the differences in the expression between the red and white skin will rely on a relatively small group of genes that regulates the pigmentation, highlighting the importance of the RNA composition. Considering that within-sample normalization is not relevant for this approach and algorithms such as GeTMM [56] perform similarly to TMM , the transcript length was not taken into account. Therefore, the TMM normalization is the most suitable for the current approach.

TMM assumes that most of the genes are not differentially expressed in a between-sample normalization method unlike within-sample normalization methods (RPM, TPM and RPKM/FPKM). It is recommended to remove the batch effects while comparing the samples from very different tissues or in cases where RNA reads population will be significantly different among the samples [57].

More in detail, to calculate the TMM, it is considered the library size normalized read count for each gene in each sample and calculate the log2 fold change between the two samples (M-value, *Eq.3*):

$$M = log_2 \frac{treated\ sample\ count}{control\ sample\ count}$$

(Eq.3)

Then, the absolute expression count is obtained (A-value, *Eq.4*):

$$A = log_2 \frac{treated\ sample\ count + log_2(control\ sample\ count)}{2} \quad (Eq.4)$$

Now, it is double trimmed the upper and lower percentages of the data (M values by 30% and A values by 5%). Finally, the weighted mean of M is obtained after trimming and the normalization factor is calculated [57].

In this case, the reference tissue will be the white skin, therefore, the treated sample count will be the red skin sample counts and control sample count will be the white skin sample counts.

With HTSFilter not only a normalization but also a filtering procedure is applied based on a global Jaccard similarity index (which measures the overlap of two sets), in order to discard genes with low or constant levels of expression across the experimental conditions. Given two binary vectors of length *n*, Jaccard index (J) is defined as (*Equation 5)*:

$$J = \frac{a}{a+b+c}, \text{ where } a + b + c + d = n \quad (Eq.5)$$

Having 'a' as the number of attributes with a value of 1 in both vectors, 'b' with value of 1 in the first vector and 0 in the second, 'c' is opposite to 'b' and 'd' with 0 for both vectors [58]. It is assessed 1 as similar and 0 as dissimilar and the two vectors will be translated to normalized read counts in a pair of experimental conditions in a sample. Because of having different replicates for multiple samples the *Equation 5* is extended to a global Jaccard index by averaging the indices calculated over all the pairs in each condition [55]. Afterwards, a cutoff value '*s*' is defined, which corresponds to the greatest similarity (based on the global Jaccard index) among replicates (*Equation 6*).

$$s = argmax\, J(y) \quad (Eq.6)$$

At the end, the value of the global Jaccard index is plotted for a fixed set of threshold values and the data is fitted with a loess curve [59]. Therefore, the final value of '*s*' will be the maximum value of the fitted curve. This value will be used as a maximum-based filter, discarding normalized counts below the established threshold, in this case '*s = 109.31*' (*Figure 4*).

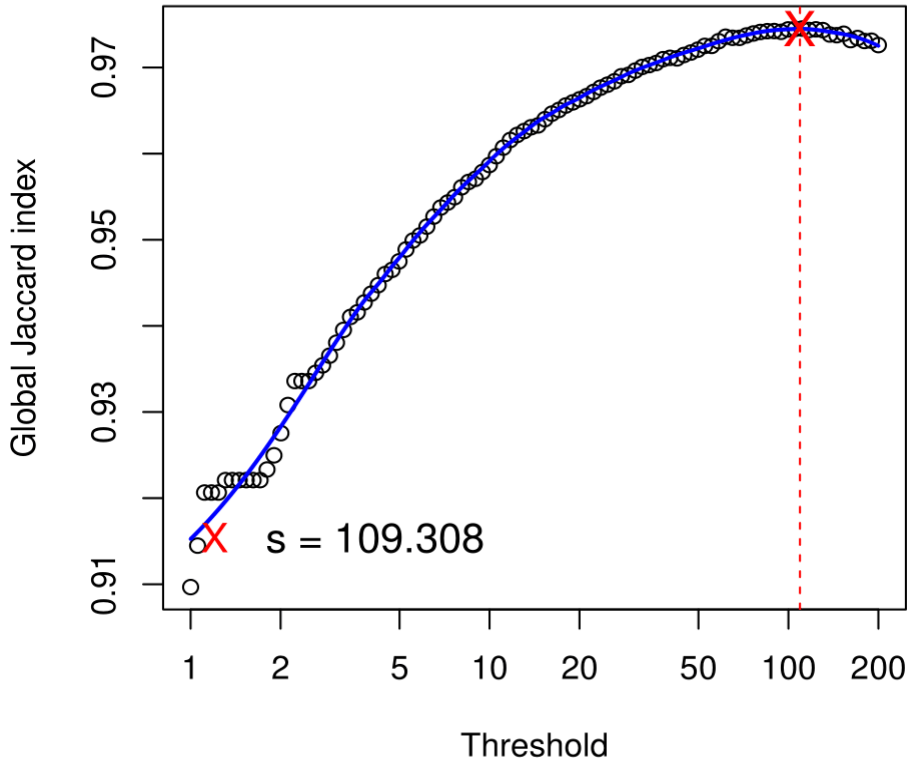

Figure 4. Threshold 's' applied to the filter based on a fitted loess curve (blue), having on the y-axis the Global Jaccard index similarity values and on the x-axis a set of 's' threshold values. The maximum value that was used for filtering is highlighted in red.

2.2.3 Data exploration with the Principal Component Analysis

Before analyzing each time point, in order to bring out strong patterns from the complex biological dataset and to ensure that the labeling has been correct when sequencing the samples, a Principal Component Analysis (PCA) [60] plot was generated (*Figure 7*).

2.2.4 EdgeR analysis for each time point

Once the counts have been normalized, the next step is to execute the statistical analysis for each of the four time points independently (0WAV, 3WAV, 6WAV and 9WAV). The full implementation of the code is available at the annexes (*See Chapter 6.Annexes)*.

As a result, it is obtained a table with six columns [49], having from left to right (*Table 2)*:

- VCost.V3 annotation system identifier
- logFC: $\log_2$ fold change between the conditions. E.g. value 2 means that the expression has increased 4-fold.
- logCPM: the average $\log_2$ counts-per-million.

- LR: likelihood ratio statistics.
- P-value: the two-sided p-value.
- FDR: adjusted p-value with the False Discovery Rate.

The functions from edgeR applied during the analysis attempts to extract the list of differentially expressed genes (DGE). In summary, the first step is to create a DGEList object that contains the counts and relevant associated information, such as normalization method applied, library size and samples metadata, conforming a more suitable format for the subsequent analysis. The makeContrasts constructs a numerical matrix that includes the contrasts between red and white samples using the normalized expressions previously fitted in the model. At the same time, calcNormFactors calculates the normalization factors to scale the library sizes, taking into account the normalization method applied, in that case TMM. Finally, estimateDisp function maximizes the negative binomial likelihood to provide the estimate of the common, trended and sample-wise dispersions across the whole data. Once the DGEList object has been updated with all the above information, the data is fitted into a negative binomial generalized log-linear model to the read counts for each gene with glmFit while conducting gene-wise likelihood ratio for the coefficients in the linear model with glmLRT. Results are displayed with the topTags function.

2.2.5 Volcano plot

The volcano plot represents the relationship between the False Discovery Rate (FDR) of the adjusted p-values and the difference in the expression of the samples in the two compared conditions, expressed in Fold Change (FC). On the y-axis, the $-\log_{10}$ FDR values are represented while on the x-axis it is plotted the $\log_{10}$ FC values. Therefore, the larger the difference in expression of a gene, the more extreme its point will lie on the x-axis, both for overexpression and repression. On the other hand, for the y-axis the more significant the difference in expression, the smaller FDR and thus the higher the -log value will be, locating the significant points in the top of the graph [61]. Combining both axes, there are the most significant genes that are differently expressed in the red and white skin of the berry in the top and left/right corners of the plot.

2.2.6 MA plot

Commonly used to represent the relation between average expression of a gene and its differential expression between two conditions. As in the volcano plot a scatter plot is generated, but having in

the x-axis the $\log_{10}$ of the Counts Per Million (CPM) values. For each gene (i), the CPM is the count of sequenced reads mapping to the gene feature, scaled by the total number of reads times one million (*Equation 7)*.

$$CPM = \frac{Number\ of\ reads\ mapped\ for\ gene(i)}{Total\ number\ of\ mapped\ reads} \cdot 10^6 \quad [62]\ (Eq.7)$$

On the y-axis the logFC is represented and the data points with extreme repression or overexpression will appear in the upper and lower regions of the graph, representing the genes that have highly differential expression levels, but not necessarily differentially expressed since the figure does not display any kind of statistical significance measure on the axes [49].

Therefore, having each datapoint representing a single gene, it is easier to visualize relevant biological information. An MA plot with a large number of points falling above or below the 0.5 FC threshold will indicate a more significant number of genes being either upregulated or downregulated, respectively. In the case in which most of the data lies close to 0 along the y-axis, it should be concluded that the two conditions present similar expression patterns.

2.2.7 Bar plot summary

Attempting to summarize the gene counts for each time point and to extract generalized biological conclusions, a bar plot was constructed for the total genes, the significant ones based on the FDR values and the up/down-regulated genes in relation to the FC values.

2.2.8 GSEA plot

The Gene Set Enrichment Analysis (GSEA) is the computational implementation of an algorithm that establishes whether an *a priori* group of genes, related to an specific metabolic pathway or biological function, presents statistically significant differences between two biological conditions or states. Mainly, GSEA focuses on genome-wide expression profiles belonging to two classes. All the genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric (for example, a specific phenotype or a metabolic pathway). The overall methodology can be summarized in three steps [63-64]:

1) **Calculation of an enrichment score (ES):** it represents the degree to which a group of phenotypical related genes (S) are overrepresented at the extremes (top or bottom) of the whole ranked genes list (L). While scanning down the list L , the

ES score is calculated through a running-sum statistic when a gene in S is encountered, while encountering a gene not in S will decrease the value. The quantification of the variation in ES score depends on the correlation of the particular gene with the phenotype that represents S. Therefore, the ES is the maximum deviation from zero encountered in the random scanning of L and it corresponds to a weighted Kolmogorov-Smirnov-like statistic.

2) **Estimation of significance level of ES:** calculated with an empirical phenotype-based permutation test that provides a nominal P value. More into detail, the phenotype labels are permuted and the ES is recomputed for a gene set (S), which generates a null distribution for the ES. Therefore, the nominal P value is calculated relative to this null distribution. Last but not least, the permutation of the labels preserves gene-gene correlations, attempting to represent a more biologically reasonable assessment of significance.
3) **Adjustment for multiple hypothesis testing:** firstly, the ES score for each set (S) is normalized (NES) in relation to the size of the set. Then, the false discovery rate is calculated estimating the probability that a set with given NES represents a false positive identification (calculated by comparing the tails of the empirical and null distributions for the NES).

Therefore, a GSEA analysis was performed for common up/down-regulated genes among all the time points, for the same dataset but excluding exclusive 0WAV genes and for MYBA1 / MYB24 EigenGenes clusters (*See Chapter 2.3.2 WGCNA*). In that sense, the R packages *gprofiler2* [65] version 0.2.1 and *clusterProfiler* 3.8 [66] were used to perform the analysis.

The annotation was based on the MapMan tool [67] which is composed of two main modules: Scavenger and ImageAnnotator. The Scavenger distributes measured parameters into functional categories based on text search algorithms and manual curation. The module architecture compresses genes grouped in bins which are themselves split into sub-bins. On the other hand, ImageAnnotator utilizes the functional categories to map the experimental data and display a graphical representation that suits the user's input dataset [27].

2.2.8 Venn diagram

In order to compare the significantly differentially expressed genes among the different time points, Venn diagrams were constructed for both up-regulated (*Figure 10*) and down-regulated

(*Figure 11*) outcomes with the interactive tool Venny [68]. Additional comparisons between the differentially expressed genes from *edgeR* pipeline and an external *limma* pipeline were also performed (*Figure 15*).

2.3 Time-course differential expression analysis for overall trajectories

Attempting to identify genes that have different expression trajectories over the time between the red berry skin and the white berry skin, it was taken advantage of the Bioconductor R package ImpulseDE2 1.8.0 [69].

Designed for analyzing longitudinal count data sets coming from sequencing experiments (RNA-seq, ChIP-seq, ATAC-seq and DNAsel-seq), the algorithm is based on a negative binomial noise model. The main advantage for this experimental approach is the capability of dispersion trend smoothing that uses a model to constrain the mean expression trajectory for each time point. Even when the sampling over the time is limited (as in this case, in which there are only four time points) the trajectory remains stable and concordant to the predictive model.

The program can be used, mainly, for two different approaches: case-only and case-control analysis. The first one will test whether the expression level of a gene changes significantly over time while the second one (which fits with the experimental design) will compare the expression trajectories of a gene over the time between samples from two different conditions.

2.3.1 ImpulseDE2 description

The software that was used for pairwise analysis, the standard differential expression algorithms, might also be utilized for time-course analysis in which they treat time as a categorical variable through generalized linear models (for example, edgeR and limma). Contrastingly, there are other methods that attempt to model the dependence between the different experimental time points as a continuous function of time by using non-linear models (for example, ImpulseDE2).

The main advantage resides in statistical testing power in which the categorical linear models experience a loss of statistical power when many time points are observed, specially when the two conditions to compare are sampled in different time points. In that way, continuous models overcome all these issues by comparing fitted values in unmeasured time points implicitly.

Therefore, it is implemented a time-course differential expression analysis with the ImpulseDE2 package that employs a noise model specific to count reads from the different batches and combines it with a likelihood ratio test. Its algorithm has shown the fastest and most accurate inference, compared to the old ImpulseDE, as well as the linear models of DESeq2 and the continuous models from limma and edgeR [70].

More specifically, the normalized count data is fitted to an **impulse model** [71] and the differential expression analysis is implemented according to the model matches with a **log-likelihood ratio test.**

More in detail, in the impulse model the expression level of a gene is represented as a function of the time f(t) and it corresponds to the scaled product of two sigmoid functions (*Eq.8*, [72]). It considers three state-specific expression values: initial, peak and steady state, in which the transitions are represented by the two sigmoids.

$$f(t) = \frac{1}{h_1}(h_0 + (h_1 - h_0)\frac{1}{1+e^{-\beta(t-t_1)}}) * (h_2 + (h_1 - h_2)\frac{1}{1+e^{\beta(t-t_2)}}) \qquad (Eq.8)$$

where the steady state expression is modeled as $h_0 = f(t \to -\infty)$, $h_2 = f(t \to \infty)$, h_1 refers to the intermediate expression, t_1 and t_2 are the state transition times and $\boldsymbol{\beta}$ is the slope parameter for both sigmoid functions. Note that a shared $\boldsymbol{\beta}$ is used instead of individual slope parameters for each function, reducing the overall number of parameters in the model and accelerating the total computing time.

Defining the likelihood function, the number of reads x generated from μ transcripts is supposed to follow a negative binomial distribution (*Eq.9*). The likelihood value $L(x_{i,\cdot} \mid \mu_{i,\cdot}, \phi_i)$ of the count data x_i for gene i observed in J samples at time points t_j is represented as:

$$L(x_{i,\cdot}, t \mid \mu_i, C_{i,\cdot}, \phi_i, s) = \prod_{j=1}^{J} L_{NB}(x_{i,j} \mid \mu_i(t_j) * exp(< X_{j,\cdot}, C_{i,\cdot} >) * \check{s}_j, \phi^*_i) \qquad (Eq.9)$$

where L_{NB} is the negative binomial likelihood:

$$L_{NB}(x|\mu, \phi) = \frac{(\phi+x)}{x!\ (\phi)}\left(\frac{\mu}{\phi+\mu}\right)^x \left(\frac{\phi}{\phi+\mu}\right)^{\phi} \qquad (Eq.10)$$

Thus, the mean expression for each time point $\mu_i(t_j)$ is calculated by a fit of the impulse model $f(t)$. The sample-specific size factor $\check{s}_j$

corrects the library size while the gene -specific batch correction factor C_i models a matrix that predetermines batch assignments or other covariates related to the samples, for example guanine-cytosine content bias. In any case, as no count normalization method is implemented in ImpulseDE2, the negative binomial distribution assumption can not be infringed with the selected normalization method [73]. Finally, ϕ^*_i corresponds to the dispersion factor that correlates the mean of the negative binomial distribution to its variance (*Eq.11*). It is implemented as a constant hyperparameter for each gene through the DESeq2 package [74].

$$\sigma_i(t_j)^2 = \mu_i(t_j) + \phi^*_i \cdot \mu_i(t_j)^2 \qquad \textit{(Eq.11)}$$

The final likelihood value for the data is calculated as the product of the gene-wise likelihoods. The estimation of the parameters for the impulse model $\{h_0, h_1, h_2, t_1, t_2, \beta\}$ is set according to the Broyden-Fletcher-Goldfarb-Shanno algorithm [75]. As established before, the ϕ^*_i hyperparameter is estimated with DESeq2 and treated as a constant during the impulse model fitting. The log-likelihood ratio test is computed by comparing the null likelihood with the alternative model using a X^2 - distributed deviance test statistic [70].
The manual guide for the users is available online at link(https://bit.ly/30IUziB).

Digging into the practice, it is needed the TMM normalized counts table (in matrix format) and the metadata declaring the different experimental conditions (in this case, red and white for the berry skin). For convenient usage of the package, it is required to label as 'case' and 'control' the metadata for the conditions, thus, the red skin was selected as the 'control' layer.

The analysis was run in 'case-control' mode with a significance threshold ('scaQThres') of 0.05. As a result, the gene-wise trajectories for the dataset are displayed. In order to plot a specific vector of genes, it can be used the 'plotGenes' function.

Finally, graphical modifications were applied for improving the visual experience using the 'ggplot2' package. Results for the MYBA1 and MYB24 genes are available in the results section (*Figure 12*).

2.3.2 Weighted gene co-expression network analysis

Attempting to describe the correlation patterns between expression

levels across different conditions, weighted gene co-expression network analysis (WGCNA) has been widely used during the last decade. The standard downstream pipeline is based on finding clusters (modules) of highly correlated genes, recapping such clusters through the module eigengene concept (explained further below) and establishing a relationship between a module and a phenotypic trait or condition. Despite the fact that its major usage is related to identifying putative biomarkers or therapeutic targets, it can be used to find candidate genes that participate in a metabolic pathway or a specific ontology [76-77]. In that sense, focusing on the ontologies from the genes that belong to the MYBA1/MYB24 modules will give us a useful insight about their metabolic role.

Therefore, the WGCNA R package [76] was utilized for performing a weighted correlation network analysis over the **exclusive differentially expressed** genes from **ImpulseDE2 output**. In other words, a co-expression network was constructed only for the genes that showed significant differences in the expression trajectories between red and white skin sections.

The background behind the analysis relies on the correlation networks construction. There are different applications implemented in the package, such as identification of interconnected nodes clusters [77], detection of 'significant' modules [78], fuzzy measures to annotate nodes closely related to already identified clusters [79], network neighborhood construction for a given set of nodes [80] or contrast one network with another one [81]. Regarding this project, scale-free co-expression networks based on soft thresholding (β) have been constructed. In that sense, instead of establishing an absolute threshold that splits the correlations into unconnected and connected nodes (hard threshold), in the soft definition the correlations are exponentiated to a power (*Equation 12-13)* that emphasizes more on stronger associations.

Constructing a weighted gene network entails the choice of soft thresholding power (β) to which co-expression similarity is raised to calculate adjacency. For that reason, a vector of powers is created and the network topology analysis function 'pickSoftThreshold' is called with the 'signed' network type option. Bear in mind that for a correct clusterization the data should not be previously filtered with respect to a sample trait. Particularly, in this project the input data correspond to the DEG from the time-course pipeline, which hinders the finding of an appropriate index for the scale-free topology state. In fact, when plotting the scale-free model indexes, it was not possible to observe a plateau in the curve no matter how much the power vector

was increased (*Figure 6)*. Therefore, following the instructions in the manual (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/faq.html), when the lack of scale-free topology fit turns out to be caused by a biological restraint, it is recommended to assess a power of 16 for signed networks with 20-30 samples.
The applicability for describing the pairwise relationships among gene transcripts have been demonstrated several times [82-87]. It should be highlighted that the standard network terminology presents few connotations for gene co-expression networks. Corresponding to undirected, weighted networks, the **nodes** refer to gene expression profiles (such as the TPM values for a particular gene in all the samples) while the edges between nodes are the pairwise correlations between gene expressions. By increasing the absolute value of the correlation to a power $\beta \geq 1$ (soft thresholding), the algorithm will underline high correlations at the expense of low correlations. In that sense, the adjacency matrix that establishes the correlations among the gene expressions (for each gene 'i' and 'j') is calculated, either with:

$$a_{ij} = \left|cor(x_i, x_j)\right|^{\beta} \qquad \textit{in case of unsigned networks; (Eq.12)}$$

or by:

$$a_{ij} = \left|\frac{(1+cor(x_i, x_j))}{2}\right|^{\beta} \qquad \textit{in case of signed networks; (Eq.13)}$$

The input gene expressions derive from the significant time-course analysis outcomes by ImpulseDE2. In that sense, in case that an unsigned adjacency definition is applied it might be found clustered together genes that show an specific trajectory, but also the negative mirrored ones. For example, a gene whose expression increases exponentially from 0WAV to 9WAV will cluster with another gene that is repressed exponentially from the same time points. To solve the situation and improve the clustering performed the signed networking type was selected to calculate the correlations (*Equation 13*).

WGCNA requires the input data to be normalized in *transcripts per million(TPM)*. Therefore, it is required to translate the initial counts matrix from raw counts to TPM, taking into account the transcript length. In that sense, an extra column containing the

transcripts lengths was added for each identifier. After converting the raw counts into TPM, only genes that were identified as significant with ImpulseDE2 were maintained.

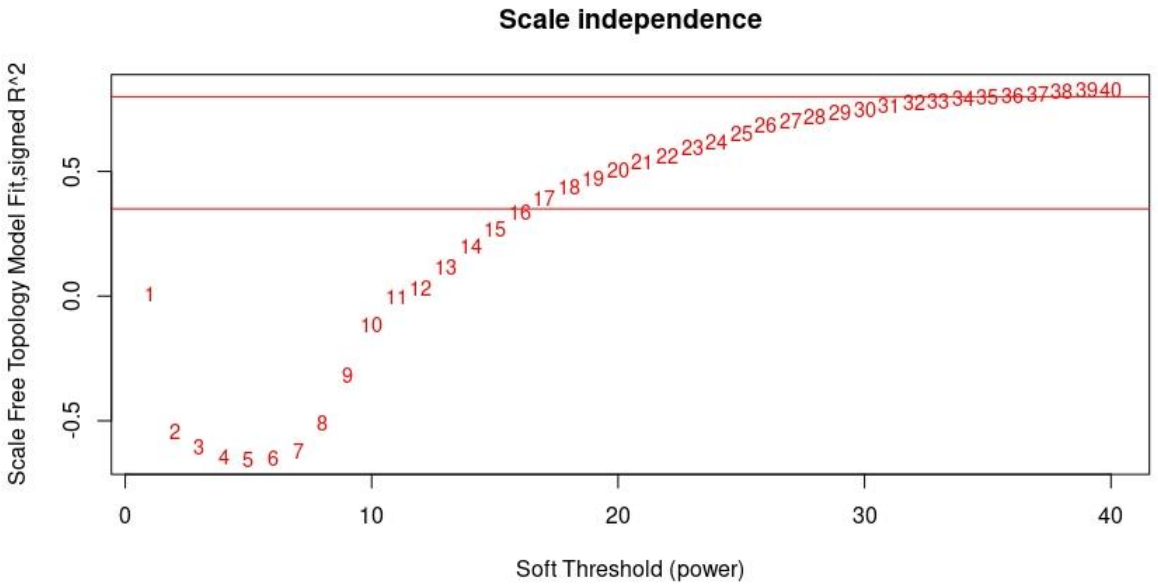

Figure 5. Scale independence with a free topology model. On the x-axis the soft threshold power (β) is represented whereas the signed R^2 for the model fitting is on the y-axes. The cut height for the power 16 and the hypothetical plateau (power of 36) is indicated with a red line.

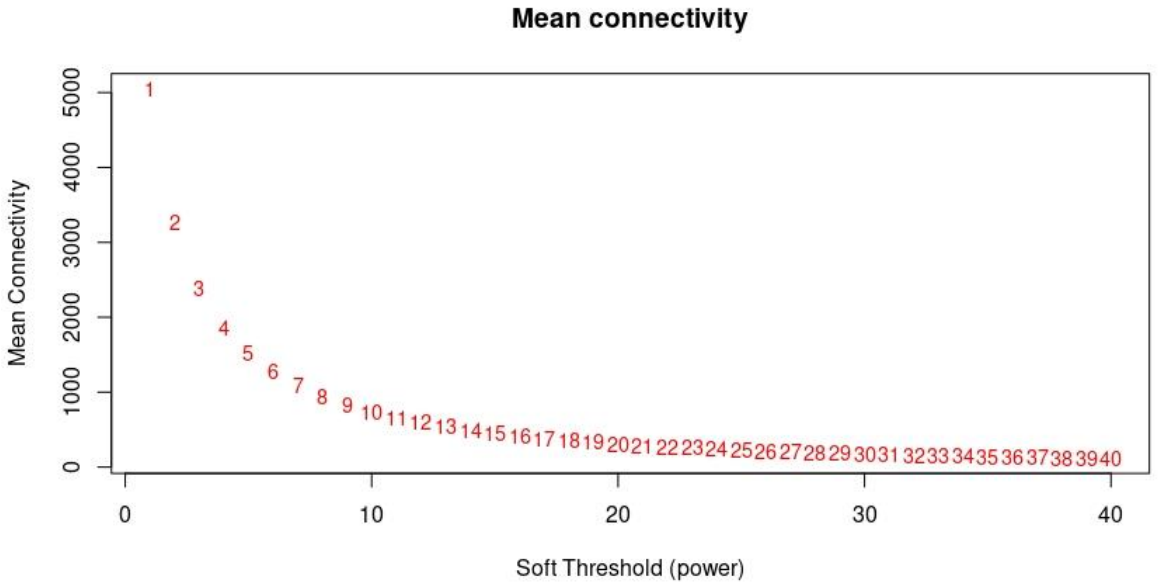

Figure 6. Mean connectivity of the entire dataset with sequential soft threshold power (β) values.

To simultaneously construct the network and perform the eigengene module detection in a block-wise manner the 'blockwiseModules' function is utilized, in which there are different parameters that

affect the clustering algorithm. The soft-thresholding **power** β (already explained above). The **maxBlockSize** establishes the maximum number of genes that a block can contain in the hierarchical clustering for module detection and, in case the input dataset is bigger than the chosen value, genes will be pre-clustered into blocks that satisfy that threshold. The **minModuleSize** is the minimum size allowed for a module, avoiding having multiple clustering of very few genes which are not useful for extracting biological conclusions in most of the cases. **DeepSplit** ranges from 0 to 4 and is a simplified control over the module detection sensitivity, providing the most sensible clustering at 4 and the opposite for 0. Regarding the cut height in which the dendrogram will split the different branches in the module merging step, it is available the **mergeCutHeight** parameter (*See WGCNA R manual)*. Both deepSplit and mergeCutHeight will be the main parameters involved in the algorithm sensitivity and will provide different amounts of total modules and number of genes per module, depending on the values that are utilized. In that sense, there is no better option for the analysis, it depends on the biological questions that you have to answer and what is more meaningful for each approach. In this case, several tests were performed with different options for both parameters and selected the one that was more reasonable for us.

In order to observe how the expression of each gene behaves with respect to the average value of expression across all the samples, the z-score normalization was applied to the dataset. For each module eigengene (ME) cluster, it was generated a sub-matrix of TPM normalized counts for the genes that belong to the correspondent cluster. Therefore, the Z-score was calculated for each sub-matrix and each sample, providing an average expression value for all the genes. At the end, a heatmap with the Z-score values was generated allowing an overview of the average expression variation for all the identified clusters across the different biological replicates, time points and conditions (*Figure 14*).

3. Results

3.1 Pairwise analysis pipeline

In order to observe whether the expression data shows trends or raw correlationship according to any variable it has been generated a principal component analysis (PCA) plot.

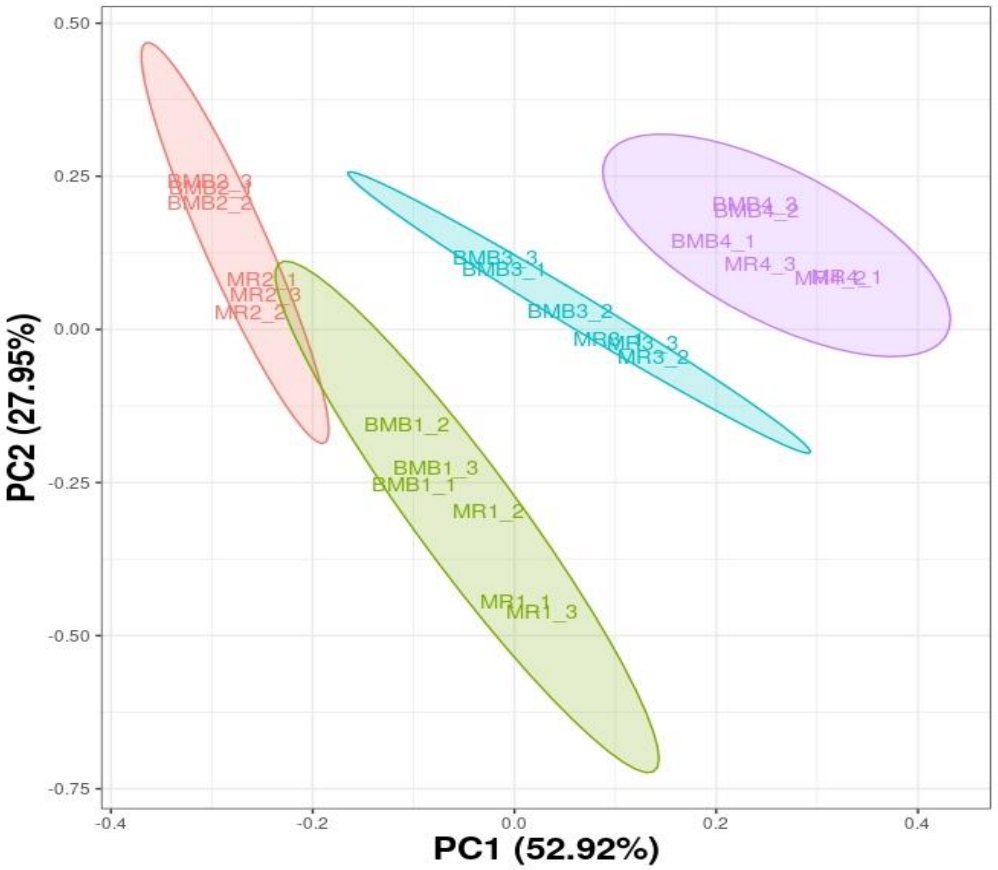

Figure 7. PCA plot for all the samples. The first number in the labeling corresponds to the time point, being the 1 (WAV0), 2 (WAV3), 3(WAV6) and the 4(WAV9). The second number in the labeling corresponds to the number of the biological replicate. Color areas according to each time point. BMB and MR correspond to the white and red skin section samples, respectively.

As expected, the samples corresponding to the different time points clustered together. It seems that the data variation has to be explained through different components. Despite the fact that the PC1 separates correctly the experimental groups, it is required the combination of PC1 and PC2 to split the WAV0 and WAV6

data. Note that inside of each time point group there is a separation between the white and red skin samples according to PC2 (*Figure 7)*. In that sense, it could be declared that the component 1 might be related to the time variable and the component 2 might be correlated to the phenotype.

As a result from edgeR, a table with different statistical outcomes is released *(Table 2)*. The main focus was on the False Discovery Rate (FDR) to assess the significance for each gene, looking for values lower than 0.05. The number of differentially expressed genes showed a noticeable increase (*Table 3,4)* from the 3WAV time point (~7000 genes), according to the hypothesis that supports the regulation of the expression in the secondary metabolism involving genes related to the synthesis of anthocyanins and flavonoids.

VCost ID	logFC	logCPM	LR	P-Value	FDR
Vitvi04g00259	-1.69	6.49	907.33	2.50E-199	9.44E-197
Vitvi13g01911	-1.45	7.59	893.81	2.18E-196	8.01E-194
Vitvi10g00020	1.4	5.99	847.68	2.32E-186	8.32E-184
Vitvi16g00139	-1.55	6.20	842.35	3.36E-185	1.17E-182
Vitvi18g00899	-1.85	4.98	798.59	1.09E-175	3.74E-173
Vitvi02g00717	-1.19	7.00	779.78	1.35E-171	4.49E-169
Vitvi02g01224	-1.47	5.44	740.85	3.92E-163	1.28E-160
Vitvi02g00983	-1.45	5.94	669.89	1.06E-147	3.36E-145
Vitvi18g00376	-1.47	5.45	669.43	1.33E-147	4.15E-145

Table 2. Fragment of the resulting table from edgeR analysis for 0WAV time point.

Assigning whether a gene was up or down-regulated depends on the logFC threshold selected. There is not a standard recommendation to select a cut-off value and strictly depends on the experimental design. In this case, there were thousands of genes with a negative logFC value very close to 0, therefore, establishing a threshold rigorously at 0 is senseless for this approach.

	0WAV	3WAV	6WAV	9WAV
Significant	5011	6988	7761	8373
Up-Regulated	773	897	1273	1528

Down-Regulated	4238	6091	6488	6845

Table 3. Gene count for each time point. It is considered the logFC threshold strictly greater or smaller than 0 for the up/down-regulated assessment.

In that sense, *Table 4* represents the up and down-regulated genes that have a logFC higher than 0.5 or lower than -0.5 respectively. It provides a reasonable outcome for us to establish a gene expressed 2-fold more in one condition to set it as overexpressed or, oppositely, repressed. When being more restrictive in the logFC threshold, the tendency of a pronounced increase in the number of significant genes is observed from 3WAV as well, concording with the latest part of the berry ripening.

Collecting the results for 3WAV, it is noticeable how MYBA1 is significantly repressed in the white skin whereas MYB24 starts getting overexpressed, both in concordance with the initial hypothesis (*Figure 8)*. As it can be observed in the plots *(Figure 8,9)*, there are more genes getting expressed significantly in the red skin, indicating the activation of many secondary metabolism related pathways. Note that for assessing the significance in the Volcano and MA plot it has been established a logFC threshold of -0.5/0.5 while in the barplot it is displayed all the genes with a logFC greater or lower than 0.

	0WAV	**3WAV**	**6WAV**	**9WAV**
Up-Regulated	671	796	1083	1399
Down-Regulated	452	1244	1577	1691

Table 4. Gene count for each time point. It is considered the logFC threshold greater than 0.5 or smaller than -0.5 for the up/down-regulated assessment, respectively.

As the time passes since the veraison, the phenotypic differences observable in the variegated berries are a consequence of a differential regulation in the expression of different genes, both in white and red skin. In that way, there are 60% more overall significant genes that are differentially expressed among the starter time (0WAV) and more than 60 days later, at 9WAV (*Table 4-Figure 9)*. MYBA1 is one of the most significant genes that is down-regulated in the white skin from the very beginning while MYB24 initiates its overexpression at the mid-late stages of the fruit maturation. Actually, MYB24 is not overexpressed at 0WAV indicating a late regulation in the white skin sections (*See supplementary material)*.

From the onset of the grapes ripening, the greatest phenotypic changes are noticeable at 3 weeks after *veraison* (3WAV)

approximately. As commented before, most of the expression differences are condensed in the 3WAV, 6WAV and 9WAV stages. Therefore, the main focus relies on studying the ontologies corresponding to the common differentially expressed genes coming from that time points (excluding 0WAV).

In case of the up-regulated genes in the white skin, a total of 242 genes were identified in common for all the time points, while 181 common genes were found in common for 3WAV, 6WAV and 9WAV (*Figure 10*). It seems that the most significant biochemical pathway activated by these genes is the photosynthesis, specifically the photophosphorylation in the photosystems assembly and maintenance plus other enzymes that participate in intermediate reactions, for example the ferredoxin-NADP oxidoreductase (*Table 5*). In that sense, the lack of pigmentation exposes the skin cells to light and radiation, promoting light-response genes, such as photosystems and chlorophyll metabolic processes.

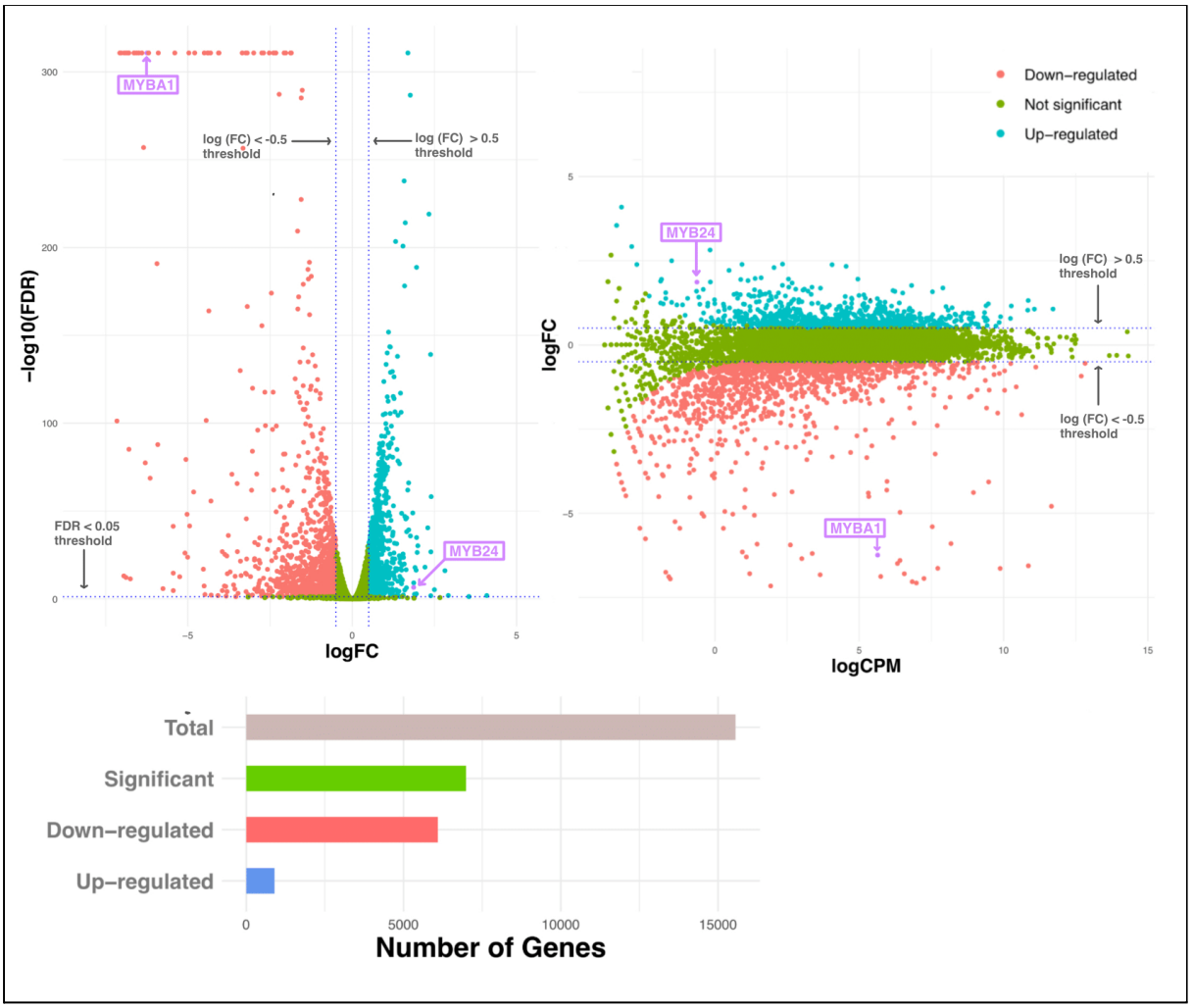

Figure 8. Representation of the pairwise analysis results for the samples belonging to the second time point (3WAV). Upper-left: volcano-plot, Upper-right: MA plot. The utilized thresholds have been indicated with

arrows, MYBA1 and MYB24 genes have been highlighted in purple. Bottom: Bar plot summary.

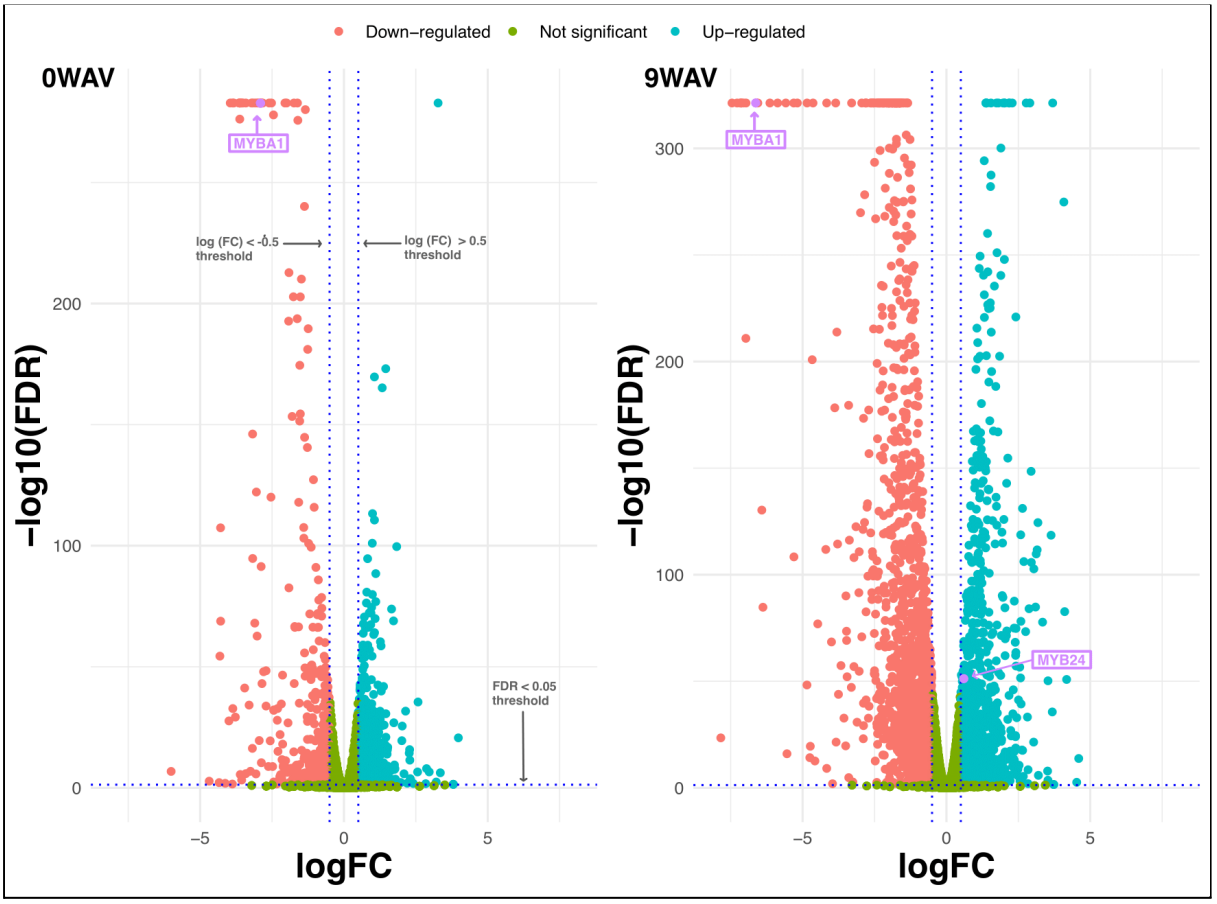

Figure 9. Volcano plot comparison between the first (0WAV) and the last (9WAV) time points. Genes MYBA1 and MYB24 have been highlighted in purple. The utilized thresholds have been indicated with arrows

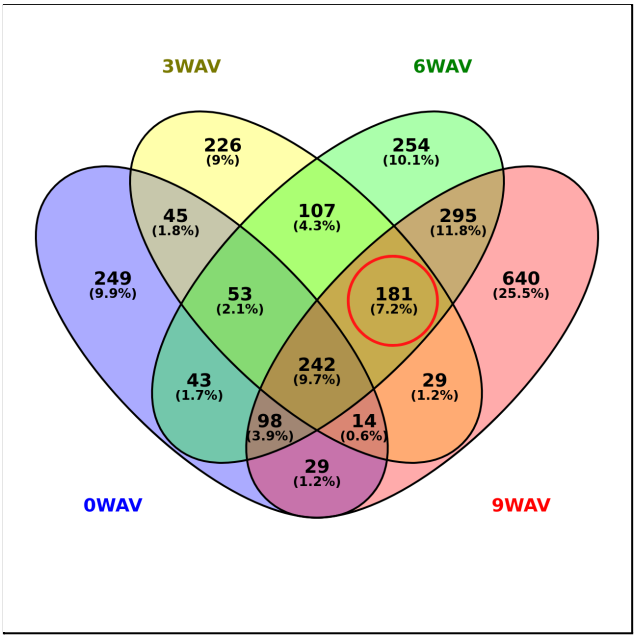

Figure 10. Venn diagram for white skin up-regulated genes. Thresholds applied correspond to FDR < 0.05 and logFC > 0.5 / logFC < -0.5. The common

genes for the latest stage of the berry ripening (3,6 and 9 WAV) is highlighted in a red circle.

Other relevant ontologies detected are related to the metabolism of the chlorophyll and basic cellular functions, such as the transport of metabolites through the fatty acids biosynthesis and the polyester biosynthesis lysophospholipase (BDG), both involved in the cell wall organization and stability.

On the other hand, only 225 white-skin down-regulated genes (or in other words, red skin up-regulated genes) were identified as differentially expressed in common with the four time points while 312 were exclusive of 3WAV, 6WAV and 9WAV *(Figure 11)*.

The greatest impact is observed in the secondary metabolism related to the phenolic pathways, highlighting the flavonoids biosynthesis and the p-Coumaroyl-CoA activation, which is found in the chalcones and catalyzes many clue reactions in the flavonoids and stilbenes metabolism (*Table 6)*. The overexpression of these genes in the red skin correspond to the fact that the flavonoids are involved in the pigmentation of the berry, bringing a red coloration when the accumulation of flavonoids is present.

Term ID	Term label	P-value
1.1	Photosynthesis photophosphorylation	1.52E-10
1	Photosynthesis	5.41E-10
1.1.1	Photosynthesis photophosphorylation photosystem II	5.69E-05
17.7.2	Protein biosynthesis organelle machinery plastidial ribosome	5.69E-05
17.7	Protein biosynthesis organelle machinery	0.0005
1.1.1.3	Photosynthesis photophosphorylation photosystem II assembly and maintenance	0.0013
1.1.5.2.1	Photosynthesis photophosphorylation linear electron flow FNR activity ferredoxin-NADP oxidoreductase	0.0013
17.7.2.1	Protein biosynthesis organelle machinery plastidial ribosome large ribosomal subunit proteome	0.0013
19.4.5.8.2.1	Protein homeostasis proteolysis metallopeptidase activities FtsHE activity FtsH plastidial protease complexes	0.0027
1.1.1.3.7	Photosynthesis photophosphorylation photosystem II assembly and maintenance:HCF244-OHP assembly factor complex	0.0027

1.1.2.9.2	Photosynthesis photophosphorylation cytochrome b6/f complex assembly CCS maturation system II	0.0027
1.1.2	Photosynthesis photophosphorylation cytochrome b6/f complex	0.0032
19.4.5.8.2	Protein homeostasis proteolysis metallopeptidase activities FtsHE activity	0.0046
15.6.2.1	RNA biosynthesis organelle machinery transcriptional regulation basal TF (Sigma)	0.0098
1.1.5.2	Photosynthesis photophosphorylation linear electron flow FNR activity	0.0098

Table 5. Result of the Gene Set Enrichment Analysis performed with *gprofiler2* for 3,6 and 9WAV common up-regulated genes. Annotation was assessed in concordance to MapMan for *Vitis vinifera VCostV3*. Top15 functional ontologies have been displayed in relation to the P-value.

Secondly, it is noticeable the overall activation of the carbohydrate metabolism in the red skin at oligosaccharide level, more specifically overexpressing the sorbitol dehydrogenase and the fermentation alcohol dehydrogenase, both involved in the ripening of the berry. Other relevant ontologies were associated with the expression of transcription factors that regulates the RNA biosynthesis and the degradation of the proteins through the ubiquitin-proteasome system. Many genes connected to different phytohormone actions were also down-regulated in white skin (*Table 6)*.

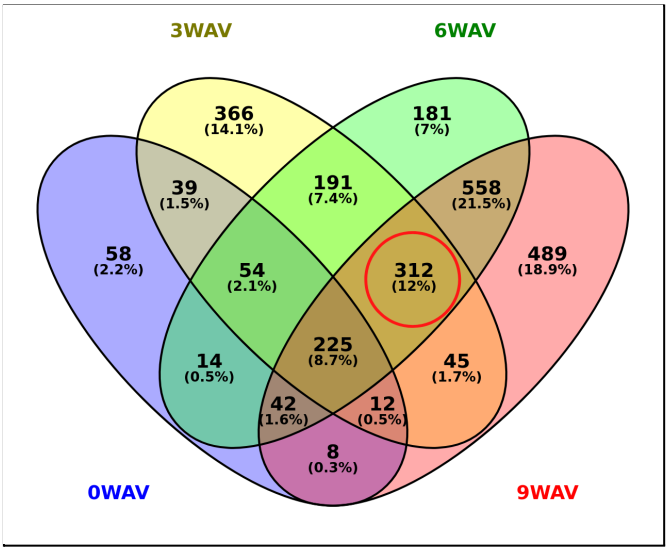

Figure 11. Venn diagram for white skin down-regulated genes. Thresholds applied correspond to FDR < 0.05 and logFC > 0.5 / logFC < -0.5. The common genes for the latest stage of the berry ripening (3,6 and 9 WAV) is highlighted in a red circle.

Term ID	Term label	P-value
9.2	Secondary metabolism phenolics	5.99E-05
11	Phytohormone action	0.0003
9.2.1	Secondary metabolism phenolics p-coumaroyl-CoA biosynthesis	0.0010
15.5	RNA biosynthesis transcriptional regulation	0.0011
3.11.1.2	Carbohydrate metabolism fermentation alcoholic fermentation alcohol dehydrogenase	0.0012
9.2.2.1.1	Secondary metabolism phenolics flavonoid biosynthesis chalcones chalcone synthase activity	0.0016
11.7	Phytohormone action jasmonic acid	0.0016
9.2.2.1.1.1	Secondary metabolism phenolics flavonoid biosynthesis chalcones chalcone synthase activity CHS	0.0016
9.2.2.1	Secondary metabolism phenolics flavonoid biosynthesis chalcones	0.0016
3.11.1	Carbohydrate metabolism fermentation alcoholic fermentation	0.0024
9.2.1.1	Secondary metabolism phenolics p-coumaroyl-CoA biosynthesis phenylalanine ammonia lyase activity	0.0029
3.11	Carbohydrate metabolism fermentation	0.0039
3	Carbohydrate metabolism	0.0055
3.5.2	Carbohydrate metabolism sorbitol metabolism sorbitol dehydrogenase	0.0063
15	RNA biosynthesis	0.0067

Table 6. Result of the Gene Set Enrichment Analysis performed with *gprofiler2* for 3,6 and 9WAV common down-regulated genes. Annotation was assessed in concordance to MapMan for *Vitis vinifera VCostV3*. Top15 functional ontologies have been displayed in relation to the P-value.

3.2 Time-course analysis pipeline

When looking at the overall trajectories, 10.031 genes showed significant differences between the white and red skin sections (*See supplementary materials*).

The down-regulation of MYBA1 in white skin is also observed in the results from ImpulseDE2 (*Figure 12*). Surprisingly, an overall repression is observed in both conditions. The trajectory shows a pronounced decrease of the read counts during the ripening in the red skin while in the white skin the transcript levels are close to 0. It seems that MYBA1 is overexpressed in the red skin before the onset of the ripening and, from that, it is not expressed anymore during the whole procedure.

MYB24 starts its expression in the middle stage of ripening, when the biosyntentical secondary metabolism is more pronounced. In spite of the fact that it is expressed in the whole berry, the white skin showed the major expression levels, in concordance with the initial hypothesis.

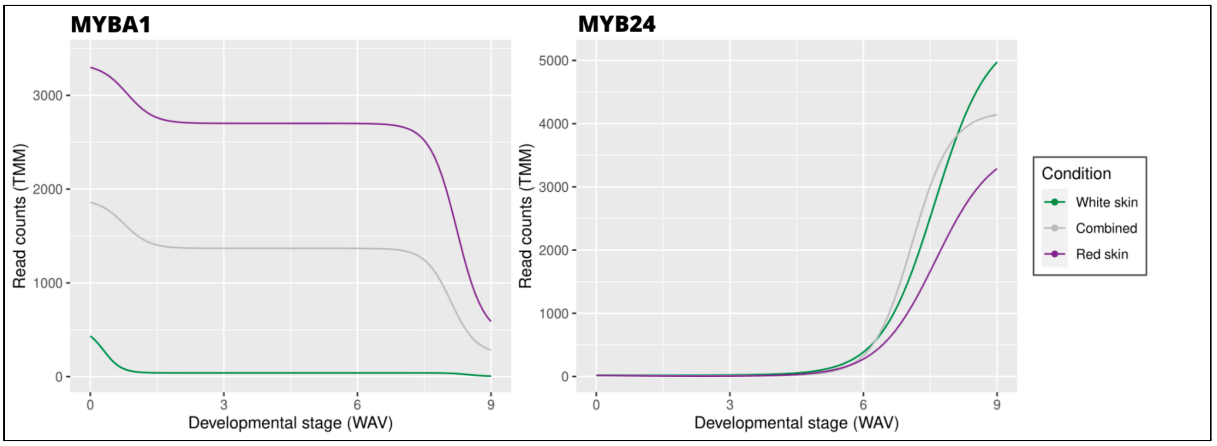

Figure 12. Trajectories for MYBA1 (left) and MYB24 (right) genes displayed with the ImpulseDE2 R package. On the y-axis it is represented the mean normalized counts for each time point, while the x-axis corresponds to the developmental stage time point, weeks after veraison. The green and purple trajectories correspond to the white and the red skin, respectively.

For further understanding of the biological functions that each trajectory is affiliated with, it has been performed a weight gene co-expression network analysis, building a dendrogram (*Figure 13*) that highlights the cut height for each eigengene module, represented in different colors.

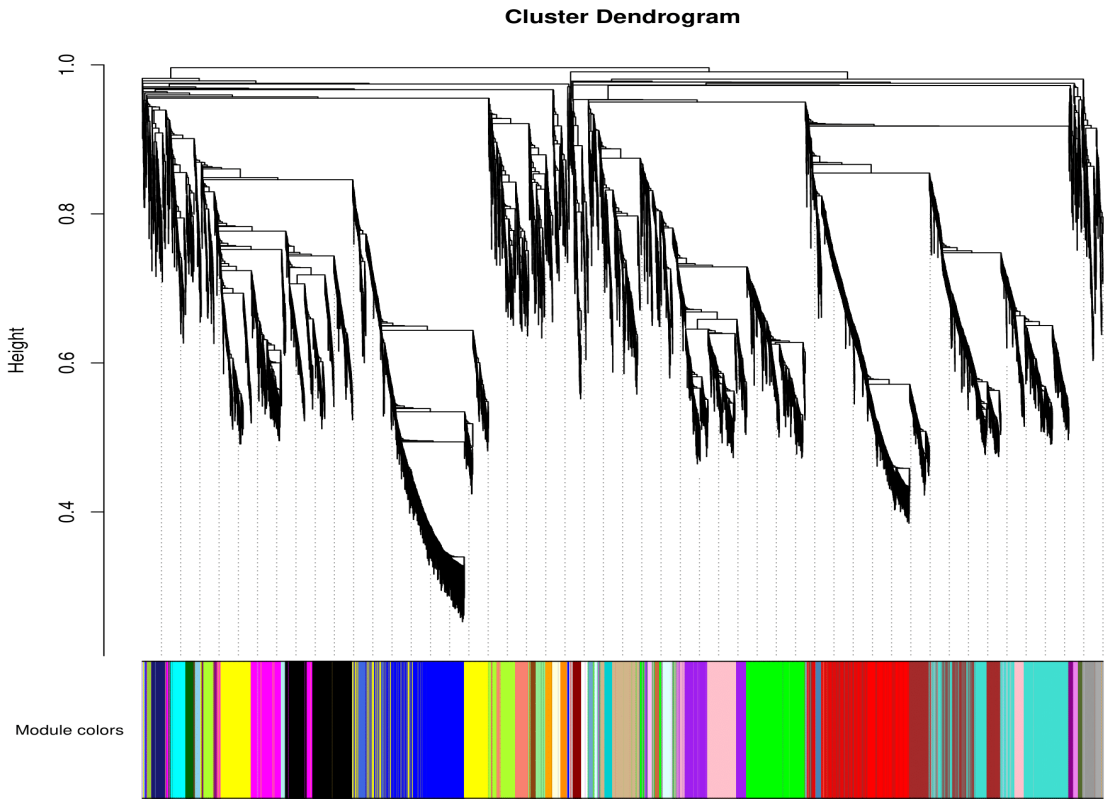

Figure 13. Clustering dendrogram of normalized expressions, with dissimilarity based on topological overlap, together with assigned module colors.

A total of 38 Module Eigengene (ME) clusters (*Table* 7) was obtained, in which the first cluster, ME0, contains all the unclustered genes, being significantly empty and supporting an appropriate clusterization.

ME0	ME1	ME2	ME3	ME4	ME5	ME6	ME7	ME8	ME9
25	946	889	884	857	738	712	679	571	456
ME10	**ME11**	**ME12**	**ME13**	**ME14**	**ME15**	**ME16**	**ME17**	**ME18**	**ME19**
437	334	319	189	151	139	131	121	114	100
ME20	**ME21**	**ME22**	**ME23**	**ME24**	**ME25**	**ME26**	**ME27**	**ME28**	**ME29**
95	89	88	85	82	80	72	71	66	60
ME30	**ME31**	**ME32**	**ME33**	**ME34**	**ME35**	**ME36**	**ME37**	**ME38**	
60	57	52	52	50	47	46	46	46	

Table 7. Number of genes for each module eigengene (ME) provided by the WGCNA analysis.

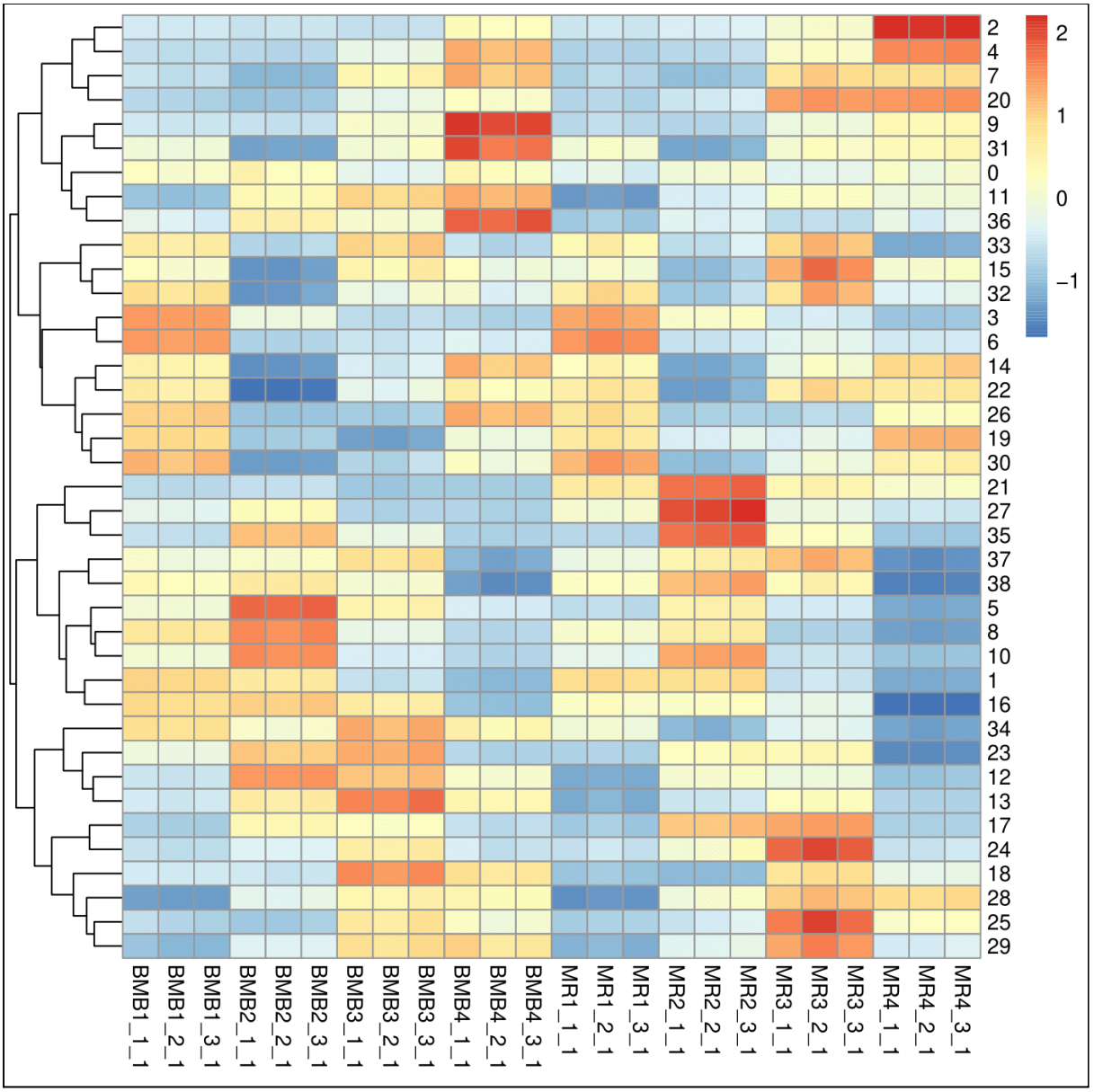

Figure 14. Heatmap with the eigengenes modules (on the y-axis) based on Z-score TPM normalized expression values. BMB: white skin, MR: red skin; first number indicates time point: 1 for 0WAV, 2 for 3WAV, 3 for 6WAV and 4 for 9WAV; second number indicates the biological replicate.

In the heatmap it is highlighted how the genes belonging to the module ME21, which includes MYBA1, are down-regulated in the white skin (Z-score < 0) and are overexpressed in the red skin, being more pronounced at 3WAV and decreasing from that time point. That kinetical behavior agrees with the fact that, as soon as the veraison starts, MYBA1 and other regulators are overexpressed in the early stages of the ripening (0WAV and 3WAV), promoting the biosynthesis and accumulation of flavonols and anthocyanins and decreasing their expression in the latest stages of the ripening (6WAV and 9WAV), when their function has already been exerted. Oppositely, in the white skin MYBA1 is tightly repressed and inhibited during the whole ripening.

In relation to the genes belonging to the module ME9, which includes MYB24, there is a continuous expression increase in the white skin that gets significantly pronounced at the latest stage of the ripening (9WAV). The red skin also presents the same kinetic behavior but far less pronounced. This is related to the sunscreen effect of flavonoids in red skin, which is absent in the white sections and promotes the activation of light-exposure genes.

Digging into the biological functions of the genes that compose the module ME21, the secondary metabolism related to the flavonols and anthocyanins biosynthesis is regulated significantly. It have been identified different enzymes involved in the phenolic biosynthesis pathway, such as the flavonoid 3-hydroxylase or the p-coumaroyl-CoA (*Table 8*). Other significant molecular functions have been related to the plant cell wall organization, specially the monolignol biosynthesis .

Interestingly, the genes conforming to the module ME9 have shown proteolytic and protein translocation functions. Above all, the ubiquitin-proteasome system is deeply activated, emphasizing in the 26S subunit and its regulatory components (*Table 9*). The protein translocation in the mitochondrial membranes have also been significant as well as the transcription factors that regulate the RNA biosynthesis (*See supplementary material*). Specific ontologies related with the photosynthetical attributes of MYB24 have not been significantly detected, probably due to the reduced size of the module and the heterogeneous landscape of their components.

Term ID	Term label	P-value
9.2	Secondary metabolism phenolics	4.64E-19
9.2.2	Secondary metabolism phenolics flavonoid biosynthesis	6.52E-14
9	Secondary metabolism	1.84E-12
9.2.2.4	Secondary metabolism phenolics flavonoid biosynthesis dihydroflavonols	1.60E-10
9.2.2.9	Secondary metabolism phenolics flavonoid biosynthesis anthocyanidins	1.55E-05
9.2.1	Secondary metabolism phenolics p-coumaroyl-CoA biosynthesis	4.12E-05
21.6.1.4	Cell wall organization lignin monolignol biosynthesis CCoA-OMT	6.58E-05

9.2.2.4.1	Secondary metabolism phenolics flavonoid biosynthesis dihydroflavonols flavonoid 3-hydroxylase	0.0002
9.2.2.4.2	Secondary metabolism phenolics flavonoid biosynthesis dihydroflavonols flavonoid 3	0.0005
21.6.1	Cell wall organization lignin monolignol biosynthesis	0.0008

Table 8. Result of the Gene Set Enrichment Analysis performed with gprofiler2 for the ModuleEigengene cluster 21 (ME21) genes, including MYBA1. Annotation was assessed in concordance to MapMan for *Vitis vinifera VCostV3*.

Term ID	Term label	P-value
19.2	Protein homeostasis ubiquitin-proteasome system	1.80E-08
19.2.6	Protein homeostasis ubiquitin-proteasome system 26S proteasome	8.43E-07
19	Protein homeostasis	1.11E-06
19.2.6.2	Protein homeostasis ubiquitin-proteasome system 26S proteasome 19S regulatory particle	2.04E-06
23.2	Protein translocation mitochondrion	3.67E-05
23.2.3	Protein translocation mitochondrion inner mitochondrion membrane TIM translocation system	0.0001
19.2.4.2	Protein homeostasis ubiquitin-proteasome system ERAD substrate extraction	0.0008
19.1.3	Protein homeostasis protein quality control mitochondrial Hsp70 chaperone system	0.0021
15.5.17	RNA biosynthesis transcriptional regulation transcription factor (NAC)	0.0091
19.2.6.2.2	Protein homeostasis ubiquitin-proteasome system 26S proteasome 19S regulatory particle non-ATPase components	0.0091

Table 9. Result of the Gene Set Enrichment Analysis performed with gprofiler2 for the ModuleEigengene cluster 9 (ME9) genes, including MYB24. Annotation was assessed in concordance to MapMan for *Vitis vinifera VCostV3*.

3.3 Pipelines comparison

In order to take advantage of the different analysis performed in the laboratory, a comparison of the significant genes list was applied against an external analysis with the limma R package [50] generated by another researcher (*see Supplementary material*).

Having a total of 10.031 significant DE genes from the time-course pipeline, 12.256 unique genes (considering the four time points) from the edgeR pairwise pipeline and 13.317 unique genes from the limma pairwise pipeline, only 56 genes have been identified as significant exclusively in the time-course pipeline, representing less than 1% of the total identified genes (*Figure 15*). When comparing both pairwise pipelines, a common core composed of more than 11.400 genes (80.9% of total DE genes) was identified.

In that sense, it provides a relatively strong validation of the results in which the major part of the genes (9789, meaning the 69.1% of the total DE genes) have been declared as significant for the three parallel pipelines. The major differences are observed between *edgeR* and *limma* tools due to their assumptions undergoing in the algorithms behind the software.

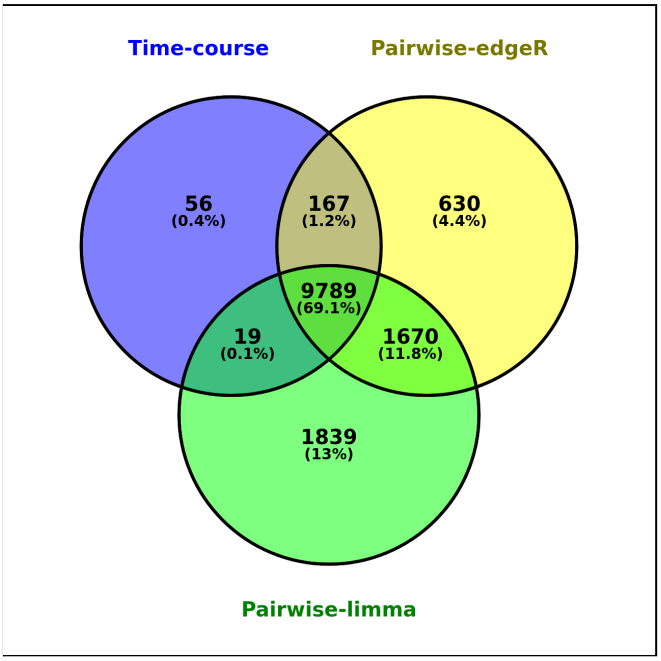

Figure 15. Venn diagram for common differentially expressed genes between Time-course and edgeR/limma Pairwise pipelines. Thresholds applied to the pairwise list correspond to FDR < 0.05 and logFC greater or lower than 0.

4. Conclusions

The naturally-occurring color mutants in plants, such as variegated cultivars, permits the identification of many regulatory genes implicated in the synthesis of discrete metabolites associated with the specialized secondary metabolism. After comparing the red and white berry skin sections at the transcriptomic level, the uncolored white sections convened non-functional alleles of the anthocyanin regulators, emphasizing *MYBA1*, and explaining the lack of pigments. Promotion of the flavonoids biosynthesis was validated from both pairwise and time-course pipelines, providing a list of putative regulators (belonging to the Module Eigengene 21) with the same behavior as *MYBA1* that showed similar biological functions in the enrichment analysis.

On the other side, the uncharacterized *MYB24* gene has been suggested as a modulator of light responses, including biological functions tightly related with the photosynthesis. This hypothesis has been corroborated by the enrichment analysis in the pairwise pipeline, but showed controversy when analyzing the batches in the time-course pipeline.

The selection of the appropriate algorithm and tool should be concordant with the experimental design and the biological questions to answer as well as the available resources, but never generalized in a bulk RNA-seq analysis.

The integration of different pipelines have been greatly useful to validate the initial hypothesis and to provide new insights into the biological functions of the R2R3-MYB transcription factors family.

5. Bibliography

1. McGovern, P., Jalabadze, M., Batiuk, S., Callahan, M.P., Smith, K.E., Hall, G.R., Kvavadze, E., Maghradze, D., Rusishvili, N., Bouby, L., Failla, O., Cola, G., Mariani, L., Boaretto, E., Bacilieri, R., This, P., Wales, N., Lordkipanidze, D.(2017). Early Neolithic wine of Georgia in the South Caucasus. Proc. Natl. Acad. Sci. 114, E10309–E10318.
2. Terral, J.-F., Tabard, E., Bouby, L., Ivorra, S., Pastor, T., Figueiral, I., Picq, S., Chevance, J.-B., Jung, C., Fabre, L., Tardy, C., Compan, M., Bacilieri, R., Lacombe, T., This, P.(2010). Evolution and history of grapevine (Vitis vinifera) under domestication: new morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient European cultivars. Ann. Bot. 105, 443–455.
3. This, P., Lacombe, T., Thomas, M.R.(2006). Historical origins and genetic diversity of wine grapes. Trends Genet. 22, 511–519.
4. Ferreira, V., Pinto, O., Castro, I.(2018). Berry color variation in grapevine as a source of diversity. Plant Physiol (Vol.132): 696-707.
5. Teixeira, A., Eiras-Dias, J., Castellarin, S.D., Gerós, H.(2013). Berry phenolics of grapevine under challenging environments. Int. J. Mol. Sci. 14, 18711–39.
6. Vezzulli, S., Leonardelli, L., Malossini, U., Stefanini, M., Velasco, R., Moser C,.(2012). Pinot blanc and Pinot gris arose as independent somatic mutations of Pinot noir. J Exp Bot. 63:6359–69.
7. Walker, AR., Lee, E., Robinson, SP.(2006). Two new grape cultivars, bud sports of Cabernet Sauvignon bearing pale-coloured berries, are the result of deletion of two regulatory genes of the berry colour locus. Plant Mol Biol. 62:623–35.
8. Ferreira, V.(2018). Molecular Characterization of Berry Skin Color Reversion on Grape Somatic Variants. Journal of berry research (Vol.8): 147–162.
9. Neilson-Jones, W., 1969. Plant chimeras, 2nd ed. Methuen, London.
10. **T**hompson, M.M., Olmo, H.P.(1963). Cytohistological Studies of Cytochimeric and Tetraploid Grapes. Am. J. Bot. 50, 901.
11. Hocquigny, S., Pelsy, F., Dumas, V., Kindt, S., Heloir, M.-C., Merdinoglu, D.(2004).Diversification within grapevine cultivars goes through chimeric states. Genome. 47, 579–589.
12. Pelsy, F., 2010. Molecular and cellular mechanisms of diversity within grapevine varieties. Heredity (Edinb). 104,331-40.

13. Kobayashi, S., Goto-Yamamoto, N., Hirochika, H.(2004). Retrotransposon-Induced Mutations in Grape Skin Color. Science (Vol.304), 982–982.
14. Walker, A.R., Lee, E., Bogs, J., McDavid, D.A.J., Thomas, M.R., Robinson, S.P.(2007). White grapes arose through the mutation of two similar and adjacent regulatory genes. Plant J. 49, 772–785.
15. Matus, J.T., Cavallini, E., Loyola, R., Höll, J., Finezzo, L., Dal Santo, S., Vialet, S., Commisso, M., Roman, F., Schubert, A., Alcalde, J.A., Bogs, J., Ageorges, A., Tornielli, G.B., Arce-Johnson, P.(2017). A group of grapevine MYBA transcription factors located in chromosome 14 control anthocyanin synthesis in vegetative organs with different specificities compared with the berry color locus. Plant J. 91,220–236.
16. Yakushiji, H., Kobayashi, S., Goto-Yamamoto, N., Tae Jeong, S., Sueta, T., Mitani, N., Azuma, A.(2006). A Skin Color Mutation of Grapevine, from Black-Skinned Pinot Noir to White-Skinned Pinot Blanc, Is Caused by Deletion of the Functional VvmybA Allele. Biosci.70, 1506–1508.
17. Hardie, W.J., Brien, T.P.O., Jaudzems, V.G.(1996). Morphology , anatomy and development of the pericarp after anthesis in grape , Vitis vinifera L . Aust. J. Grape Wine Res. 2, 97–142.
18. **R**obinson, S.P., Davies, C.(2000). Molecular biology of grape berry ripening. Aust. J. Grape Wine Res. 6, 175–188.
19. Ramsay, N.A., Glover, B.J.(2005). MYB-bHLH-WD40 protein complex and the evolution of cellular diversity. Trends Plant Sci. 10, 63–70.
20. Wong, D.C.J., Schlechter, R., Vannozzi, A., Höll, J., Hmmam, I., Bogs, J., Tornielli, G.B., Castellarin, S.D., Matus, J.T.(2016). A systems-oriented analysis of the grapevine R2R3-MYB transcription factor family uncovers new insights into the regulation of stilbene
accumulation.DNA Res. 23,451–466.
21. Rinaldo, A.R., Cavallini, E., Jia, Y., Moss, S.M.A., McDavid, D.A.J., Hooper, L.C., Robinson, S.P., Tornielli, G.B., Zenoni, S., Ford, C.M., Boss, P.K., Walker, A.R.(2015). A Grapevine Anthocyanin Acyltransferase, Transcriptionally Regulated by VvMYBA, Can Produce Most Acylated Anthocyanins Present in Grape Skins. Plant Physiol. 169, 1897–916.
22. Zhang, C., Zhanwu, D., Thilia, F., Orduña, L., Santiago, A., Peris, A., Wong, D., Matus, J. (2021). The grape MYB24 mediates the coordination of light-induced terpene and flavonol accumulation in response to berry anthocyanin sunscreen depletion. BioRxiv. Preprint.
23. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M.(2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 320(5881):1344-9.

24. Bar-Joseph Z, Gitter A, Simon I.(2012) Studying and modelling dynamic biological processes using time-series gene expression data. Nat Rev Genet. 13(8):552-64.
25. Eisen B, M. (1998). Cluster analysis and display of genome-wide expression patterns. PNAS. 25: 863-868.
26. Ashburner, M., Ball, C., Blake, J. (2000). Gene Ontology: tool for the unification of biology. Nat Genet 25, 25–29.
27. Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L. A., Rhee, S. Y., & Stitt, M. (2004). mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. In The Plant Journal (Vol. 37, Issue 6, pp. 914–939).
28. Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. In Frontiers in Energy Research (Vol. 9). Frontiers Media SA.
29. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
30. Ewing B; Hillier L; Wendl MC; Green P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*. 8 (3): 175–185.
31. Ewing B, Green P (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities". *Genome Research*. 8 (3): 186–194.
32. Martin M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J., 17, 10–12.
33. Trimmomatic: a flexible trimmer for Illumina sequence data. Bolger AM, Lohse M, Usadel B. Bioinformatics. 2014 Aug 1; 30(15):2114-20.
34. Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. In Bioinformatics (Vol. 34, Issue 17, pp. i884–i890).
35. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. In Bioinformatics (Vol. 29, Issue 1, pp. 15–21).
36. Au, K. F., Jiang, H., Lin, L., Xing, Y., & Wong, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. In Nucleic Acids Research (Vol. 38, Issue 14, pp. 4570–4578).
37. Grant, G. R., Farkas, M. H., Pizarro, A. D., Lahens, N. F., Schug, J., Brunk, B. P., Stoeckert, C. J., Hogenesch, J. B., & Pierce, E. A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). In Bioinformatics (Vol. 27, Issue 18, pp. 2518–2528).
38. Zhang, Y., Lameijer, E.-W., 't Hoen, P. A. C., Ning, Z., Slagboom, P. E., & Ye, K. (2012). PASSion: a pattern growth

algorithm-based pipeline for splice junction detection in paired-end RNA-Seq data. In Bioinformatics (Vol. 28, Issue 4, pp. 479–486).
39. Han, J., Xiong, J., Wang, D., & Fu, X.-D. (2011). Pre-mRNA splicing: where and when in the nucleus. In Trends in Cell Biology (Vol. 21, Issue 6, pp. 336–343).
40. Musich, R., Cadle-Davidson, L., & Osier, M. V. (2021). Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. In Frontiers in Plant Science (Vol. 12).
41. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.
42. Ferragina, P., and Manzini, G. (2000). "Opportunistic Data Structures With Applications" in *Proceedings 41st Annual Symposium on Foundations of Computer Science*; November 12–14, 2000; 390–398.
43. Canaguier A, Grimplet J, Di Gaspero G, Scalabrin S, Duchêne E, Choisne N, Mohellibi N, Guichard C, Rombauts S, Le Clainche I, Bérard A, Chauveau A, Bounon R, Rustenholz C, Morgante M, Le Paslier M-C, Brunel D, Adam-Blondon A-F (2017) A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). Genomic Data, 14:56-62.
44. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. In Bioinformatics (Vol. 25, Issue 16, pp. 2078–2079).
45. Liao, Y., Smyth, G. K., & Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. In Bioinformatics (Vol. 30, Issue 7, pp. 923–930).
46. Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq--a Python framework to work with high-throughput sequencing data. In Bioinformatics (Vol. 31, Issue 2, pp. 166–169).
47. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. In Bioinformatics (Vol. 26, Issue 6, pp. 841–842).
48. https://www.ensembl.org/info/website/upload/gff.html
49. Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics, 26(1), 139-140.
50. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Research*, 43(7), e47.
51. Smyth, G. K., & Verbyla, A. P. (1996). A Conditional Likelihood Approach to Residual Maximum Likelihood Estimation in Generalized Linear Models. In Journal of the Royal Statistical

Society: Series B (Methodological) (Vol. 58, Issue 3, pp. 565–572).
52. Robinson, M. D., & Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. In Bioinformatics (Vol. 23, Issue 21, pp. 2881–2887).
53. Robinson, M. D., & Smyth, G. K. (2007). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. In Biostatistics (Vol. 9, Issue 2, pp. 321–332).
54. Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, 15, 550.
55. Rau A, Gallopin M, Celeux G, Jaffrezic F (2013). "Data-based filtering for replicated high-throughput transcriptome sequencing experiments." *Bioinformatics*, 29(17), 2146-2152.
56. Smid, M., Coebergh van den Braak, R. R. J., van de Werken, H. J. G., van Riet, J., van Galen, A., de Weerd, V., van der Vlugt-Daane, M., Bril, S. I., Lalmahomed, Z. S., Kloosterman, W. P., Wilting, S. M., Foekens, J. A., IJzermans, J. N. M., Martens, J. W. M., & Sieuwerts, A. M. (2018). Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons. In BMC Bioinformatics (Vol. 19, Issue 1).
57. Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. In Genome Biology (Vol. 11, Issue 3, p. R25)
58. Jaccard, P. (1912). THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. In New Phytologist (Vol. 11, Issue 2, pp. 37–50).
59. William S. Cleveland (1979) Robust Locally Weighted Regression and Smoothing Scatterplots, Journal of the American Statistical Association, 74:368, 829-836,
60. Karl Pearson F.R.S. . (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572.
61. Cui, X., & Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. In Genome Biology (Vol. 4, Issue 4).
62. https://www.reneshbedre.com/blog/expression_units.html.
63. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. In Proceedings of the National Academy of Sciences (Vol. 102, Issue 43, pp. 15545–15550).
64. Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P.,

Spiegelman, B., Lander, E. S., Hirschhorn, J. N., … Groop, L. C. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. In Nature Genetics (Vol. 34, Issue 3, pp. 267–273).
65. Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H (2020). "gprofiler2– an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler." *F1000Research*, 9 (ELIXIR)(709).
66. Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. In OMICS: A Journal of Integrative Biology (Vol. 16, Issue 5, pp. 284–287).
67. Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. Plant Cell Environment, 32: 1211-1229.

68. Oliveros, J.C. (2007-2015) Venny. An interactive tool for comparing lists with Venn's diagrams. https://bioinfogp.cnb.csic.es/tools/venny/index.html
69. Fischer D (2019). ImpulseDE2: Differential expression analysis of longitudinal count data sets.
70. Fischer, D. S., Theis, F. J., & Yosef, N. (2018). Impulse model-based differential expression analysis of time course sequencing data. In Nucleic Acids Research.
71. Yosef, N., & Regev, A. (2011). Impulse Control: Temporal Dynamics in Gene Transcription. In Cell (Vol. 144, Issue 6, pp. 886–896).
72. Chechik, G., & Koller, D. (2009). Timing of Gene Expression Responses to Environmental Changes. In Journal of Computational Biology (Vol. 16, Issue 2, pp. 279–290).
73. Risso, D., Schwartz, K., Sherlock, G., & Dudoit, S. (2011). GC-Content Normalization for RNA-Seq Data. In BMC Bioinformatics (Vol. 12, Issue 1).
74. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. In Genome Biology (Vol. 15, Issue 12).
75. Fletcher, Roger (1987), *Practical Methods of Optimization* (2nd ed.), New York: John Wiley & Sons, ISBN 978-0-471-91547-8.
76. Langfelder, P., Zhang, B., & Horvath, S. (2007). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. In Bioinformatics (Vol. 24, Issue 5, pp. 719–720).
77. Rhrissorrakrai, K., & Gunsalus, K. C. (2011). MINE: Module Identification in Networks. In BMC Bioinformatics (Vol. 12, Issue 1).
78. Diniz, W. J. S., Mazzoni, G., Coutinho, L. L., Banerjee, P., Geistlinger, L., Cesar, A. S. M., Bertolini, F., Afonso, J., de Oliveira, P. S. N., Tizioto, P. C., Kadarmideen, H. N., &

Regitano, L. C. A. (2019). Detection of Co-expressed Pathway Modules Associated With Mineral Concentration and Meat Quality in Nelore Cattle. In Frontiers in Genetics (Vol. 10).
79. Yeh, R. T., & Bang, S. Y. (1975). FUZZY RELATIONS, FUZZY GRAPHS, AND THEIR APPLICATIONS TO CLUSTERING ANALYSIS**The research reported here is supported in part by the National Science Foundation under Grant No. GJ-31528. In Fuzzy Sets and their Applications to Cognitive and Decision Processes (pp. 125–149).
80. Li, A., & Horvath, S. (2006). Network neighborhood analysis with the multi-node topological overlap measure. In Bioinformatics (Vol. 23, Issue 2, pp. 222–231).
81. Shimamura, T., Imoto, S., Yamaguchi, R., Nagasaki, M., & Miyano, S. (2010). Inferring dynamic gene networks under varying conditions for transcriptomic network comparison. In Bioinformatics (Vol. 26, Issue 8, pp. 1064–1072).
82. Zhou, X., Kao, M.-C. J., & Wong, W. H. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. In Proceedings of the National Academy of Sciences (Vol. 99, Issue 20, pp. 12783–12788).
83. Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. In Science (Vol. 302, Issue 5643, pp. 249–255).
84. Steffen, M., Petti, A., Aach, J., D'haeseleer, P., & Church, G. (2002). In BMC Bioinformatics (Vol. 3, Issue 1, p. 34).
85. Carey, V. J., Gentry, J., Whalen, E., & Gentleman, R. (2004). Network structures and algorithms in Bioconductor. In Bioinformatics (Vol. 21, Issue 1, pp. 135–136).
86. Chuang, C.-L., Jen, C.-H., Chen, C.-M., & Shieh, G. S. (2008). A pattern recognition approach to infer time-lagged genetic interactions. In Bioinformatics (Vol. 24, Issue 9, pp. 1183–1190).
87. Cokus, S., Rose, S., Haynor, D., Grønbech-Jensen, N., & Pellegrini, M. (2006). Modelling the network of cell cycle transcription factors in the yeast Saccharomyces cerevisiae. In BMC Bioinformatics (Vol. 7, Issue 1).

6. Annexes

List of the complete results and identified genes, as well as the interactive plots are placed at https://tomsbiolab.com/transcriptomics.

The major part of the code associated with the project can be found at the GitHub repository https://github.com/apc1992/MS_thesis.

Acknowledgements

The project described here represents partial results of the work supported by the Dr. Matus group supported by the Ramón y Cajal grant RYC-2017-23645 in collaboration with the China Scholarship Council and the Slovenian Research Agency, which ended up in a recent publication, based upon work from COST Action CA 17111 INTEGRAPE, supported by COST (European Cooperation in Science and Technology). Data has been treated and uploaded in public repositories according to the FAIR principles, in accordance with the guidelines found at INTEGRAPE. Among the authors, I would like to thank ph.D Chen Zhang for the biological insights and detailed explanations, ph.D Luis Orduña for the programming support and very smart tips. Finally, Dr. J.T Matus for the continuous mentoring and the integrative attitude, creating a perfect environment in the laboratory and providing very useful knowledge from every point of view.