PhD Dissertation

Information and Communication Technologies
ICT Doctoral School - University of Trento

in joint supervision with

Ecole Doctorale SIBAGHE
INRA - Montpellier SupAgro

Grapevine acidity: SVM tool development and NGS data analyses

Lorena Leonardelli, XXVI cycle

Advisor:
Dr. Claudio Moser, Fondazione Edmund Mach

Co-Advisors:
Prof. Charles Romieu, INRA - Montpellier SupAgro
Prof. Patrice This, INRA - Montpellier SupAgro

November 2014

Abstract (English)

Single Nucleotide Polymorphisms (SNPs) represent the most abundant type of genetic variation and they are a valuable tool for several biological applications like linkage mapping, integration of genetic and physical maps, population genetics as well as evolutionary and protein structure-function studies. SNP genotyping by mapping DNA reads produced via Next generation sequencing (NGS) technologies on a reference genome is a very common and convenient approach in our days, but still prone to a significant error rate. The need of defining in silico true genetic variants in genomic and transcriptomic sequences is prompted by the high costs of the experimental validation through re-sequencing or SNP arrays, not only in terms of money but also time and sample availability. Several open-source tools have been recently developed to identify small variants in whole-genome data, but still the candidate variants, provided in the VCF output format, present a high false positive calling rate.

Goal of this thesis work is the development of a bioinformatic method that classifies variant calling outputs in order to reduce the number of false positive calls. With the aim to dissect the molecular bases of grape acidity (Vitis vinifera L.), this tool has been then used to select SNPs in two grapevine varieties, which show very different content of organic acids in the berry. The VCF parameters have been used to train a Support Vector Machine (SVM) that classifies the VCF records in true and false positive variants, cleaning the output from the most likely false positive results. The SVM approach has been implemented in a new software, called VerySNP, and applied to model and non-model organisms. In both cases, the machine learning method efficiently recognized true positive from false positive variants in both genomic and transcriptomic sequences.

In the second part of the thesis, VerySNP was applied to identify true SNPs in RNA-seq data of the grapevine variety Gora Chirine, characterized by low acidity, and Sultanine, a normal acidity variety closely related to Gora. The comparative transcriptomic analysis crossed with the SNP information lead to discover non-synonymous polymorphisms inside coding regions and, thus, provided a list of candidate genes potentially affecting acidity in grapevine.

Abstract (French)

Les polymorphismes d'un seul nucleotide (SNPs) sont le plus fréquent type de variation génétique. Ce sont des outils précieux pour divers domaines de la biologie, comme la cartographie de liaison, l'intégration des cartes physiques et génétique, la génétique des populations ainsi que les études sur l'évolution et les relations structure-fonction des protéines. De nos jours, l'identification des SNPs par alignement des données issues de Séquençage de Nouvelle Génération (NGS) sur un génome de référence constitue une approche commune et pratique, cependant elle reste sujette à un fort taux d'erreurs. Le coût élevé, en termes financier, de durée et de disponibilité des échantillons, d'une validation expérimentale par re-séquençage ou hybridation sur puce à SNP renforce la nécessité d'identifier correctement les variants génétiques in silico, dans les séquences génomiques comme les transcrits. Plusieurs logiciels open-source ont été récemment développés afin d'identifier les petits variants dans les données génomiques, mais l'on trouve encore un taux élevé de faux positifs parmi les candidats extraits des fichiers de sortie au format VCF.

L'objectif de ce travail de thése est de développer une méthode bioinformatique de tri pour réduire ce nombre de faux positifs. Cet outil a ensuite permis de détecter les SNPs dans deux cultivars de vigne (Vitis vinifera L.) aux contenus très différents en acides organiques des baies, afin d'appréhender les bases moléculaires de l'acidité du raisin. Les paramètres VCF ont été utilisés pour entrainer un Séparateur à Vaste Marge (ou Machine à Vecteur de Support, SVM) au tri des faux et vrai positifs, afin d'éliminer les faux positifs les plus probables des sorties VCF. L'approche SVM a été implémentée sous forme d'un nouveau programme, VerySNP, et appliquée à différentes espèces modèles et non modèles. Dans tous les cas, la méthode d'apprentissage automatique a permis de distinguer efficacement les vrais des faux positifs, dans les données génomiques comme transcriptomiques.

Dans la seconde partie de la thèse, VerySNP a permis d'identifier les vrais SNP à partir de données RNA-seq obtenues sur la variété Gora Chirine, qui se caractérise par une acidité insignifiante, et sur la Sultanine, une variété très proche du Gora mais d'acidité usuelle. L'analyse comparative transcriptomique croisée avec l'information SNP a permis de découvrir des polymorphismes non-synonymes au sein des régions codantes et ainsi d'établir une liste de gènes candidats potentiellement impliqués dans le contrôle de l'acidité chez la Vigne.

Keywords
NGS, SVM, variant calling, grapevine, acidity

Mots-clés
NGS, SVM, appeler variantes, vigne, acidité

Contents

List of Tables

List of Figures

This PhD work was co-supervisioned by the ICT Doctoral School, which is part of University of Trento (Italy), and the École Doctorale SIBAGHE, which is part of the INRA-Supagro in Montpellier (France); while the funding for the PhD scholarship has been entirely provided by the Fondazione Edmund Mach, in S. Michele all'Adige (Italy).

Aim

First aim of this thesis work is the development of a bioinformatic method that classifies variant calling outputs obtained by mapping DNA reads on a reference sequence, to reduce the number of false positive calls. The possibility to limit the number of false positive variants returned by the most used variant calling methods is indeed of great interest due to the high costs of experimental validation through re-sequencing or SNP-chip array. This issue is of particular relevance in the field of crop genetics, where usually only one reference genome sequence is available and the genetic diversity within the species is far from being completely known. The approach is based on a Support Vector Machine (SVM), which requires a known set of validated variants in order to train the SVM classifier, and it has been implemented in a new software called VerySNP. VerySNP was applied on genomic data of model (yeast) and non-model organisms (grapevine) and, in the second part of the thesis, also on transcriptomic data obtained from the RNA-seq experiments of two related grapevine varieties showing different acidity levels (Gora Chirine and Sultanine).

Second aim of this thesis work is the selection of candidate genes whose mutation is responsible for the low acidity phenotype observed in Gora berries. The approach has been a combination of RNA-seq data analysis, gene ontology annotation and single nucleotide polymorphism (SNP) detection to reduce the number of gene candidates from thousands to a few, which would be better candidates than the other and, thus, numerically assessable by experimental validation.

Structure of the thesis

The thesis is composed of three main chapters.

Chapter 1 provides the essential background on single nucleotide polymorphisms (SNPs) and on the different techniques available to detect them. Special attention is given to the different sequencing methodologies and to their evolution in the last years. The last part of the chapter describes the analysis of RNA-seq data from a bioinformatics point of view. The chapter does not contain any original result of the thesis.

Chapter 2 and 3 are organized in the form of scientific papers and they report the results of the thesis as self-contained documents. Chapter 2 describes the results of the development of a bioinformatics method based on a Support Vector Machine (SVM) to classify variant calling outputs. After a brief introduction on SVM and variant calling methods, it illustrates VerySNP, the method developed within the thesis and the results of its application on genomic sequencing data of model and non-model organisms.

Chapter 3 deals with the biological issue of the thesis work, namely the genetic bases of grapevine berry acidity. It starts with an introduction on the acidity metabolism and on the *Vitis* varieties under study (Gora and Sultanine). The following section is about the methods and is divided in four parts: (i) the grape berry sampling and their acid/sugar content analyses (ii) the preparation of the RNA samples and library construction for NGS sequencing, (iii) the pipeline used for transcript reconstruction and their annotation, (iiii) the application of VerySNP to accurately identify genetic variants in the transcriptomes of Gora and Sultanine. The final part of the chapter reports results and relative discussion.

Chapter 1

Background

Biotechnology and bioinformatics are often both involved in modern science breakthroughs. While biotechnologies are enhancing the speed in high-throughput results, bioinformatics is able to process the massive amount of data by both standardizing computational pipelines and developing data-specific tools. The advance of new technologies is pushing both genomics and transcriptomics further into the digital age.

1.1 Single Nucleotide Polymorphisms

Genetic information can be stored in a specific nucleic acid, the DNA, as a sort of hard copy of a code composed of four different nucleotide bases (A, T, C, G) in a linear, which makes it a long and stable molecule. The DNA molecule is the subject of many reaction into the cellular environment, and one of those is the DNA replication, occurring every time the cell duplicated itself. Errors happening in the DNA replication naturally increase the biodiversity and guarantee the species evolution process, by generating mutations that are either silent or favorable to the individual and his heir. When a single nucleotide changes with an allelic frequency bigger than 1% within a population, it is known as polymorphism. Single nucleotide polymorphisms (SNPs) are the most common polymorphisms in eukaryotic genomes and are more stably inherited than other molecular markers (Brookes, 1999). The polymorphism can be of two kind: transition, when the nucleotide changes in another of the same class (C to T and A to G); or transversion, when the base is substituted by one belonging to a different class (C to A, C to G, T to A and T to G).

As SNPs are highly conserved throughout evolution and within a population, the map of SNPs serves as an excellent genotypic marker for research. Indeed, SNPs have been used in genome-wide association studies (GWAS), e.g. as high-resolution markers in gene mapping related to diseases or normal traits. SNP application in crops range from linkage disequilibrium-based association mapping and genetic diagnostics, to genetic diversity analysis, cultivar identification, phylogenetic analysis and characterization of genetic resources (Rafalski, 2002). Anyway, the use of SNP will become more widespread with the increasing availability of crop genome sequence, the reduction in cost, and the increased throughput of SNP assay. In humans, the knowledge of SNPs will help in understanding how drugs act in individuals with different genetic variants, in identifying human diseases resulting from SNP mutation and as markers for genetic diseases that have complex traits. SNPs can also be used in cancer diagnostic, to study genetic abnormalities in cancer and to identify patterns of allelic imbalance (Mei *et al.*, 2000), which are all studies with potential prognostic and diagnostic uses. These studies may provide insights into how certain diseases develop, as well as information about how to create therapies for them.

1.2 SNP identification

A number of experimental methods for SNP discovery and genotyping have been developed since the early days, although all are not equally useful and it is unclear which are the most suitable and most efficient (Gupta *et al.*, 2001). Methods such as resequencing (Snager *et al.*, 1977), denaturing gradient gel electrophoresis (DGGE; Myers *et al.*, 1986), single strand conformational polymorphism analysis (SSCP; Orita *et al.*, 1989), minisequencing (Syvänen *et al.*, 1990), heteroduplex analysis (HA; White *et al.*, 1992), derived/cleaved amplified polymorphic sequences (dCAPs/CAPs; Konieczny and Ausubel,1993), dHPLC WAVE (Oefner and Underhill, 1995), pyrosequencing (Ronaghi *et al.*, 1998), TaqMan assay (Lee *et al.*, 1999), targeting induced local lesions in genomes (TILLING; McCallum *et al.*, 2000), and temperature gradient capillary electrophoresis (TGCE; Hsia *et al.*, 2005) have all been used with success. Significant efforts towards large-scale characterization of SNPs have been attempted with high-throughput techniques, such as DNA chips and microarrays (Gunderson *et al.*, 2005) and the SNPlex$^{\mathrm{TM}}$ genotyping system (Applied Biosystems; De la Vega *et al.*, 2005). However, these platforms are expensive and not flexible since in order to be economically efficient consider only a fixed pool of genetic loci. Moreover, they are not practical for small to medium size laboratories and thus alternative techniques must be employed.

Troggio *et al.* (2008) have compared three of the mentioned methods for SNP assay known to affordable, moderately high-throughput, and multi-purpose: SSCP on both non-denaturant gel electrophoresis and fluorescence-based capillary electrophoresis, and minisequencing. They concluded that results with SSCP fluorescence-based capillary electrophoresis were consistent with sequencing data and can be considered an efficient, accurate and reliable alternative to SSCP. However, SSCP analysis has the relevant drawback that it does not allow multiplexing, at least at the PCR level (Table 1.1).

Table 1.1: Features of SNP genotyping methods (Troggio *et al.*, 2008).

Methods	Most significant advantage	Disadvantage
SSCP-gel	Low-cost genotyping Inexpensive labelling method No expensive equipment required	Not suitable for high throughput Limited genotype discrimination
SSCP-capillary	Automated electrophoresis Accurate genotyping Reproducibility Rapid separation	Difficult to multiplex Expensive primer labelling
Minisequencing	Accurate genotyping Simplicity of assay Multiplexing capacity Easy data interpretation Mid-throughput	One SNP per reaction High cost Post-PCR purification Prior sequence information necessary

The evolution of SNP detection technology is characterized by the clever adoption of new biological methods, fluorescent and other reporters, computational algorithms, and highly sensitive analytical instruments. Although the ideal SNP detection method does not exist, the field has come a long way from the early days and the technologies are sufficiently robust that it is now possible to conduct large-scale genetic studies. As the cost of SNP detection continue to drop and throughput to increase, even the most ambitious studies will become economically feasible.

1.3 DNA sequencing technologies

1.3.1 Sanger Eve

The DNA sequencing through automated Sanger method is based on the selective incorporation of chain-terminating dideoxy-nucleotides by DNA polymerase during in vitro DNA replication (Sanger and Coulson, 1975). The four dideoxy-nucleotides (dATP, dGTP, dCTP and dTTP) are differently labeled, radioactively or fluorescently, enabling the identification of the unknown nucleotide sequence.

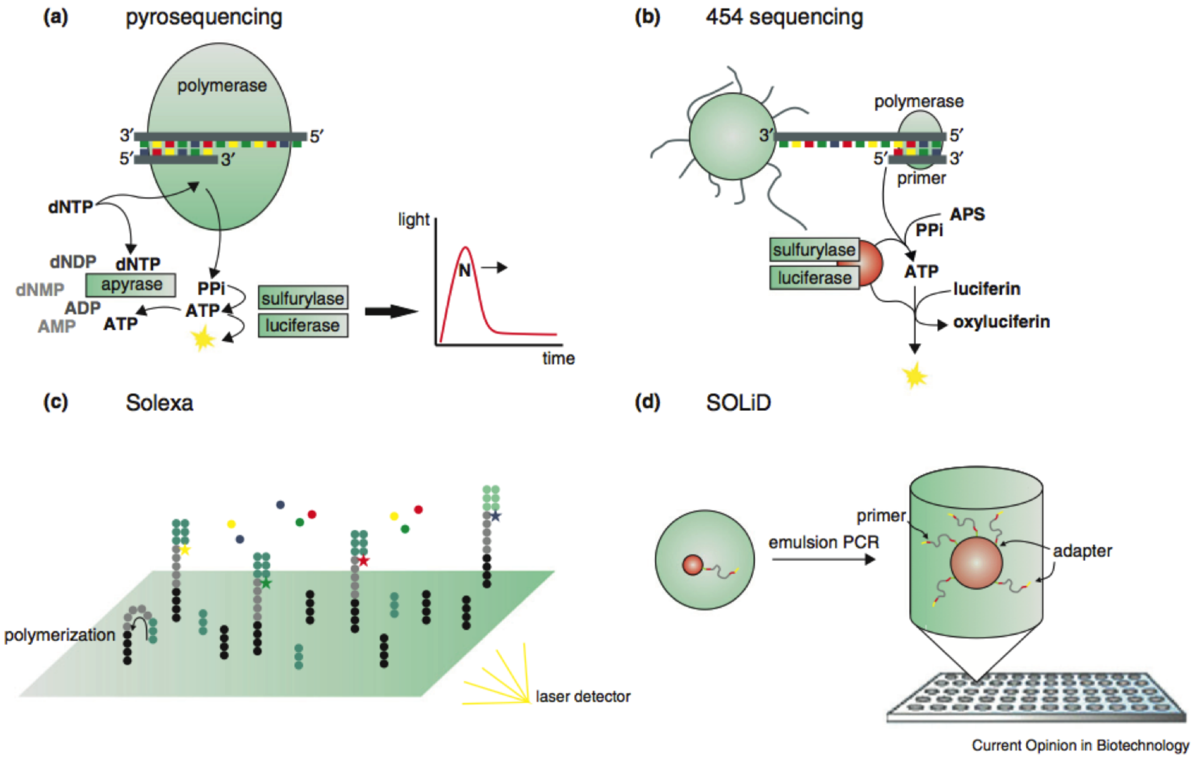
Developed by Frederick Sanger and colleagues in 1977 (Sanger *et al.*, 1977), it was the most widely used sequencing method for approximately 25 years and led to a number of monumental accomplishments, including the completion of the only finished-grade human genome sequence. Common challenges of DNA sequencing with the Sanger method include poor quality in the first 15-40 bases of the sequence due to primer binding and deteriorating quality of sequencing traces after 700-900 bases. Moreover, in cases where DNA fragments are cloned before sequencing, the resulting sequence may contain parts of the cloning vector. Such limitations showed a need for new and improved technologies, especially for large-scale, automated genome analyses.

Indeed, recently, Sanger sequencing has been supplanted by Next Generation Sequencing (NGS) methods, leaving to the automated Sanger method the title of ‘first-generation’ technology. However, the Sanger method remains in wide use, primarily for smaller-scale projects and for obtaining especially long contiguous DNA sequence reads (more than 500 nucleotides). The ultimate goal of high-throughput sequencing is to develop systems that are low-cost, and extremely efficient at obtaining extended read lengths. Longer read lengths of each single electrophoretic separation, substantially reduces the cost associated with de novo DNA sequencing and the number of templates needed to sequence DNA contigs at a given redundancy.

1.3.2 Next Generation Sequencing

Next Generation Sequencing (NGS) instruments, such as Illumina/Solexa, AB/SOLiD and Roche/454 (Mardis, 2008), have revolutionized genome analysis performing high-throughput sequencing able to produce thousands or millions of sequences in parallel (Figure 1.1). From gene discovery to regulatory elements associated with diseases or any other trait of interest, high-throughput sequencing rapidly increased the research pace. The fast and low-cost production of enormous volumes of data is the primary advantage over conventional methods, i.e automated Sanger sequencing, allowing an entire genome to be sequenced with a run time ranging from minutes to weeks.

Figure 1.1: Basic principles of NGS techniques. (a) pyrosequencing: the incorporation of a new nucleotide generates detectable light. (b) 454 sequencing: nucleotide incorporation is associated with the release of pyrophosphate resulting in a light signal. (c) Solexa: DNA fragments build double-stranded bridges and after the addition of the labeled terminators the sequencing cycle starts. (d) SOLiD: if the adapters are bound, emulsion PCR is carried out to generate so-called bead clones (Mutz *et al.*, 2013).



The innovation of NGS technologies is the sequencing by synthesis (SBS) technology, also called pyrosequencing. In contrast to the Sanger method, the incorporation of nucleotides during DNA sequencing is monitored by luminescence. Therefore, a multi-enzyme system composed of DNA polymerase, ATP sulfurylase, luciferase and apyrase is responsible for the amplification reaction and generates a lightning after nucleotide binding. The four different nucleotides are added sequentially and only incorporated nucleotides cause a signal.

The variety of NGS technology features supports the coexistence of multiple platforms in the marketplace, with some having clear advantages for particular applications over others. Six sequencing platforms are currently available (454, Illumina, SOLiD, Helicos, Ion Torrent, PacBio) and a couple (StarLight and Nanopore) are in advanced development.

Most platforms require short DNA templates (200-1000 bp), containing forward and reverse primer binding sites, the reason why a library of templates is needed. Libraries can be constructed in many different ways, which are related to the cost per sample. The most salient features of the platforms are described in the next section.

- **454** was the first commercial NGS platform. 454 was acquired by Roche, but is still known as by the 454. 454 uses beads that start with a single template molecule, which is amplified via emPCR (emulsion PCR). Millions of beads are loaded onto a picotitre plate designed so that each well can hold only a single bead. All beads are then sequenced in parallel by flowing pyrosequencing reagents across the plate (http://www.454.com).

- **Illumina** developed the second commercial NGS platform. Solexa was subsequently acquired by Illumina and is now known by the name Illumina. Illumina uses a solid glass surface to capture individual molecules and bridge PCR to amplify DNA into small clusters of identical molecules. These clusters are then sequenced with a strategy similar to Sanger sequencing, except only dye-labelled terminators are added, the sequence at that position is determined for all clusters, then the dye is cleaved and another round of dye-labelled terminators is added (http://www.illumina.com).

- **SOLiD** was the third commercial NGS platform. Invitrogen acquired Applied Biosystems, becoming Life Technologies, but the name SOLiD has been kept. SOLiD uses ligation to determine sequences and until the most recent of Illumina's software and reagents, SOLiD has always had more reads (at lower cost) than Illumina (http://www.appliedbiosystems.com).

- **Helicos** developed the HeliScope, which was the first commercial single-molecule sequencer. Unfortunately, the high cost of the instruments and short read lengths limited adoption of this platform. Helicos no longer sells instruments, but conducts sequencing via a service centre model (http://www.helicosbio.com).

- **Ion Torrent** uses a sequencing strategy similar to the 454, except that (i) instead of a pyrophosphatase cascade, hydrogen ions (H^+) are detected, which means no lasers, cameras or fluorescent dyes are needed. Furthermore, (ii) the sequencing chips used are conform to common design and manufacturing standards, reducing the manufacturing cost. In 2010, the first early access instruments were deployed and Ion Torrent was purchased by Life Technologies, but it is still known as Ion Torrent (http://iontorrent.com).

- **PacBio** has developed an instrument that sequences individual DNA molecules in real time. Individual DNA polymoerases are attached to the surface of microscope slides. The sequence of individual DNA strands can be determined because each dNTP has unique fluorescent label, immediately detected prior to being cleaved off during synthesis. Low cost per experiment, fast run times and cool factor generated much enthusiasm for this platform, which first early instruments were deployed in 2010 (http://www.pacificbioscience.com).

Next-generation sequencing technologies have broad applicability in many fields of research. They offer new high-throughput sequencing techniques that prove to be useful for many applications, including genomic (Zhou *et al.*, 2010), transcriptomic (Marguerat *et al.*, 2008), epigenomic (Cullum *et al.*, 2011), regulomic (Park, 2008), metagenomic (Voelkerding *et al.*, 2009), and diagnostic research (Jia and Zhao, 2012) at a a resolution that would have been inconceivable some years ago.

Although NGS technologies totally revolutionized the way to do genetics, some considerations need to be made. Depending on the sequencing technology used, the nucleotide reading includes mistakes at different frequencies. Incorrect base calling is commonly happening close to the 3'-end of the sequence as the raw data quality is lowering. The most common sequencing technologies provide short sequence reads (100 bp each), which need to be cleaned and trimmed for poor base quality, decreasing their already short length. The averagely short read length may decrease the accuracy of the mapping on the reference genome. Other errors rise in the generation of the reverse-DNA transcription and the following PCR steps required to build cDNA libraries (Reumers *et al.*, 2012), leading to discrepancies into the sample population.

1.3.3 Who's next?

StarLight and Nanopore are the upcoming sequencing technologies aiming to longer read length and reduced cost per sample. A brief anticipation about how those technologies would work has been described in the following section.

- **StarLight**, or more extensively Life Technologies Single Molecule Real-Team Sequencing Technology, uses quantum dots to achieve single-molecule sequencing. DNA is attached to the surface of a microscope slide where sequencing occurs in a manner similar to PacBio. A major advantage of StarLight relative to PacBio is that the DNA polymerase can be replaced after it has lost activity. Thus, sequencing can continue along the entire length of a template. The peculiar innovation is the ability to perform 3-Dimensional DNA sequencing of ultra-long DNA fragments, wherein DNA-sequence *vs* time *vs* imaging-reagent-space are simultaneously collected. This additional information provides the ability to simultaneously measure how sequencing correlates with any factor on DNA that can be spatially imaged (e.g., methylation, restriction sites, promoter sites, etc.). In addition, completely phased and ordered reads are simultaneously obtained, and the effective ”mate-pairs” for each DNA fragment increase combinatorially with the number of sequencers on each individual DNA fragment. This type of 3-D sequencing information is ideal for quantitating genomic structural variation and for generating *de novo* scaffolds for shorter read-length sequencing data. Many characteristics of the Starlight technology are known (Karrow, 2010), but timing of a commercial launch, target costs and other details are unknown (http://www.lifetechnologies.com).

- **Nanopore** is an under development method performing ‘strand sequencing’, a technique where intact DNA polymers pass through a nanopore, being sequenced in real time as the DNA translocates the pore. The theory behind nanopore sequencing is that when a nanopore is immersed in a conducting fluid and a potential (voltage) is applied across it, an electric current due to conduction of ions through the nanopore can be observed. The amount of current is very sensitive to the size and shape of the nanopore. If single nucleotides (bases), strands of DNA or other molecules pass through or near the nanopore, this can create a characteristic change in the magnitude of the current through the nanopore. DNA could be passed through the nanopore for various reasons. For example, electrophoresis might attract the DNA towards the nanopore, and it might eventually pass through it. Alternatively, enzymes attached to the nanopore might guide DNA towards the nanopore. The potential is that a single molecule of DNA can be sequenced directly using a nanopore, without the need for an intervening PCR amplification step or a chemical labelling step or the need for optical instrumentation to identify the chemical label. Nanopore technologies promise no read length associated limitation and the possibility to sequence at 25X depth of coverage the human genome in minutes at a cost of 100 dollars (https://nanoporetech.com/).

1.4 Gene expression by RNA-seq

The short-read massively parallel sequencing of RNA, better known as RNA-seq, is a technology that uses the capabilities of next-generation sequencing to reveal a snapshot of RNA presence and quantity from a genome at a given moment in time. In this direction,

Table 1.2: Utility of DNA sequencing platforms for RNA-seq experiment of different templates. The letters indicate the the review's (Glenn, 2011) opinion of the overall utility (grade) for a platform. Utility grades combine data characteristics (amount, quality, length), cost of data, and ease of assembling the data into the final desired product. Major considerations for utility grades are noted in the third column.

Platform	Opinion	Transcritpome
454 - GS Jr.	C	Need multiple runs, expensive
454 - FLX+	A/B	Good but expensive, not best for short RNAs
MiSeq	B/A	May need multiple runs, assembly more challenging than 454, longer reads may make it the best
HiSeq 2000	A/B	Good, assembly more challenging than 454 but much more data available for analyses
HiSeq 2500 - rapid run	A	Good, assembly more challenging than 454 but much more data available for analyses
Ion Torrent - 314	C	Good, but reads are shorter than Illumina, as expensive as 454
Ion Torrent - 318	B/C	Good, data more challenging to assemble than 454 to Illumina
Ion Torrent Proton	B/A	Assembly more challenging than 454, longer reads could make it best
SOLID - 5500	C/D	Short reads make assembly challenging or impossible
PacBio - RS	B	Expensive, short RNA will be challenging

NGS has been successfully applied to gene expression profiling, it has emerged as the major quantitative transcriptome profiling system and provides nearly unlimited possibilities in modern bioanalysis. For years mRNA expression has been measured by microarray techniques or real-time PCR techniques. However, microarray technology's sensitivity is limited towards the amount of RNA, the quantification of transcript levels and the sequence information; on the other hand, real-time PCR has high sensitivity but it is a quite expensive technique and not convenient for a genome-wide survey of gene expression (Mardis, 2008). RNA-seq has become a strong alternative to microarrays and real-time PCR, because it provides all the essential information about the transcriptome without requiring any previous knowledge about the genetic sequence of the organism under study, but the reference sequence. In Table 1.2 are summarized the principal advantages and disadvantages of NGS platforms to perform RNA-seq that a scientist needs to know in order to fairly consider which technology fits better with the purpose of his experiment.

Regardless from which technology was used to obtain the data, RNA-seq data could be used by means two main approaches: *ab-initio* or mapping strategy. *Ab-initio* or *de-novo* approach consists on the assembly of the individual raw data into the putative transcripts, which is mandatory when the reference genome sequence for the organism under study is not available. The technical limitations imposed by short-read sequencing lead to a number of computational challenges with the consequent explosion of the number of software trying to answer to that problem. Nevertheless, even the most recent automated methods failed to identify all constituent exons and, in cases in which all exons were reported, the protocols tested often failed to assemble the exons into complete isoforms (Steijger *et al.*, 2013). On the contrary, when a genome sequence is available, the transcript reconstruction and quantification can be performed exploiting the alignment. RNA-seq analysis by mapping implements a two-step approach in which initial read alignments are analyzed to discover exon junctions; these junctions are then used to guide the final alignment. Several programs can also use existing gene annotation to inform spliced-read placement (Engström *et al.*, 2013). Theoretically speaking *de novo* approaches should be preferred, because they would give the comprehensive picture of the transcriptome, but current performances of such methods impose severe limitations in their applications (Steijger *et al.*, 2013).

The RNA-seq analysis requires four main steps for each one several algorithms have been developed over the past years and afterwards adapted to specific applications. As a result, a variety of bioinformatic tools to obtain an appropriate analysis system optimized to fulfill any study requirements is currently available (Pagani *et al.*, 2011). The standard four steps to analyze RNA-seq can be summarized as following and schematically presented in Figure 1.2.

1. The raw image data are converted in short reads sequences. The conversion into base sequences is performed by platform specific base calling-algorithms provided by the manufacturer, along with a quality score calculated for each base, indicating the reliability of each base call. The nucleotide sequence and their quality score are compressed information stored into a format called FASTQ.

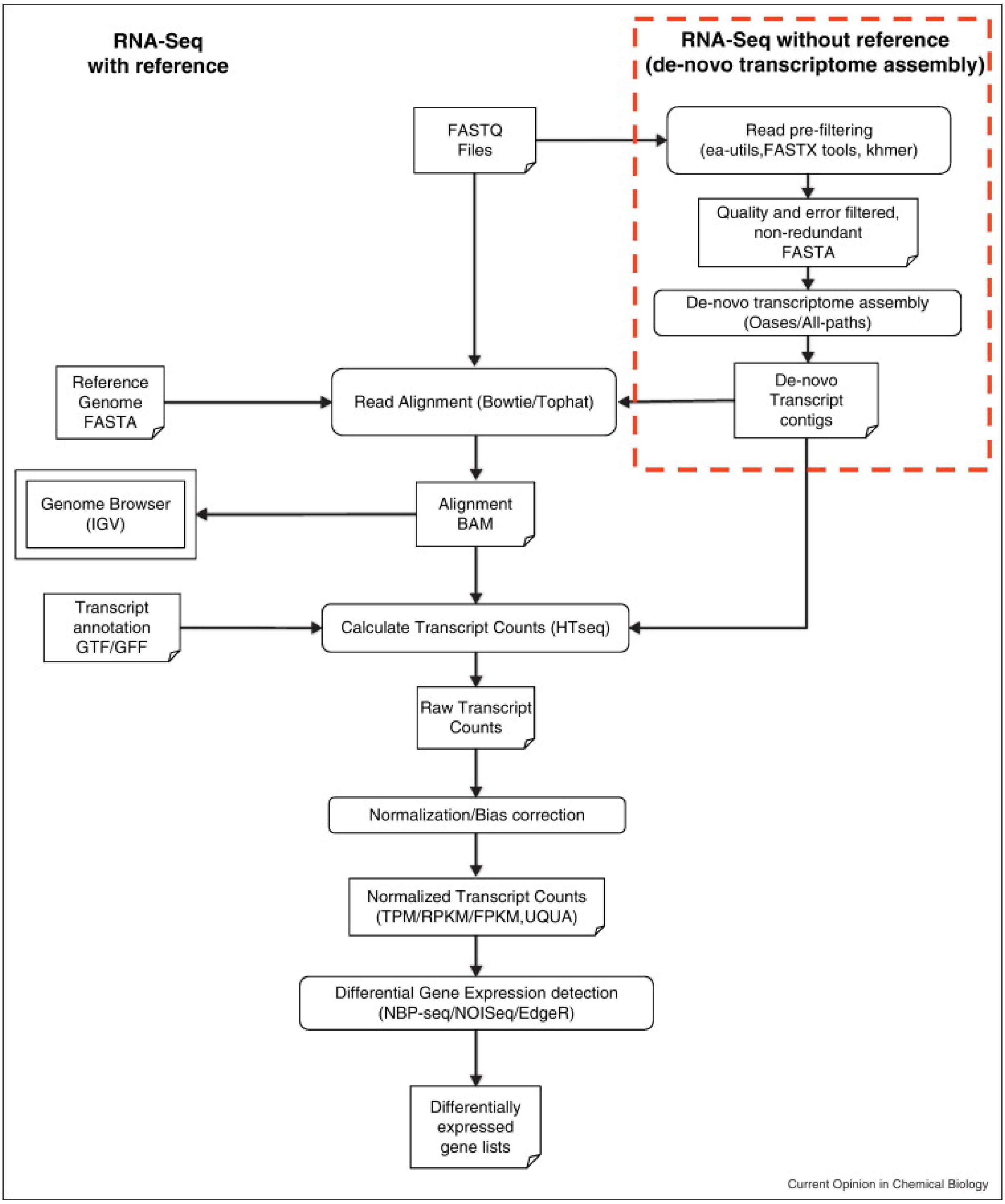
2. Align the short reads to a reference sequence, either genomic or transcriptomic. If available, the reference sequence in FASTA format can be downloaded by the ap-

propriated organism database. Mapping algorithms use indexing strategies, which enable them to align millions of short reads in a reasonable time, if compared to conventional alignment algorithms. Hash look up tables and Burrow Wheeler transformations are the two most popular indexing methods, of which the first shows high sensitivity while the second is much faster. Accordingly, read mapping tools have to balance between speed and sensitivity depending on the algorithm they are based on. The most fitting mapping software for each specific application case is determined by the reads length, which is sequencing technology related, and affects the calculation of allowed mismatches in the read alignment, which need to be tolerate because of the occurrence of sequencing errors, single nucleotide polymorphisms or mutations (Cullum *et al.*, 2011). RNA-seq reads are often aligned to the genomic reference sequence, instead of transcriptomic, because the latter is rarely available. This requires spliced read mapping software, which considers the genomic intron-exon structure by splitting unmapped reads and aligning the read fragments independently (Pagani *et al.*, 2011).

3. Calculate the expression level using peak calling algorithms. Aligned RNA-seq reads are quantified along the whole sequence generating en expression profile, delivered as a score, which needs to be normalized because of inherent bias in read quantification. Normalization of read counts enables the comparison of expression level between different genes as well as different experiments, which are affected by both the read sequencing depth and the number of reads mapped on genes of any length (Park, 2009).

4. Determine the differential gene expression. Genes, which are differentially expressed under different conditions, are detected by computational tools using normalized gene expression scores and statistical tests. These tools are classified as parametric or non-parametric algorithms. Parametric algorithms use common probability distributions such as Binomial or Poisson (Li and Tibshirani, 2011). Non-parametric ones model the noise distribution based on actual data. It was demonstrated that non-parametric algorithms show a lower dependency on sequencing depth and consequently achieve more robust results (Tarazona *et al.*, 2011).

Figure 1.2: Bioinformatics pipeline showing typical tasks involved in RNA-seq analysis. Additional steps required for de novo transcriptome assembly are shown in box at top right (McGettigan, 2013).


RNA-Seq
with reference
RNA-Seq without reference
(de-novo transcriptome assembly)
FASTQ
Files
Read pre-filtering
(ea-utils,FASTX tools, khmer)
Quality and error filtered,
non-redundant
FASTA
De-novo transcriptome assembly
(Oases/All-paths)
Reference
Genome
FASTA
Read Alignment (Bowtie/Tophat)
De-novo
Transcript
contigs
Genome Browser
(IGV)
Alignment
BAM
Transcript
annotation
GTF/GFF
Calculate Transcript Counts (HTseq)
Raw Transcript
Counts
Normalization/Bias correction
Normalized Transcript Counts
(TPM/RPKM/FPKM,UQUA)
Differential Gene Expression detection
(NBP-seq/NOISeq/EdgeR)
Differentially
expressed
gene lists
Current Opinion in Chemical Biology

Bibliography

[Brookes, (1999)] Brookes, A. J. (1999). The essence of SNPs. *Gene*, **234**(2), 177-186.

[Cullum *et al.*, (2011)] Cullum, R., Alder, O., Hoodless, P. A. (2011). The next generation: using new sequencing technologies to analyse gene regulation. *Respirology*, **16**, 210-222.

[De la Vega *et al.*, (2005)] De la Vega, F. M., Lazaruk, K. D., Rodhes, M. D.; Wenz, M. H. (2005). Assessment of two flexible and compatible SNP genotyping platforms: TaqMan® SNP genotyping assays and the SNPlex® genotyping system. *Mutation Research*, **573**, 111-135.

[Engström *et al.*, (2013)] Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Rätsch, G., *et al.*, Bertone, P. (2013). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods*, **10**(12), 1185-91.

[Glenn, (2011)] Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, *11*(5), 759-769.

[Gunderson *et al.*, (2005)] Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G., Chee, M. S. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics*, **37**(5), 549-554.

[Gupta *et al.*, (2001)] Gupta, P. K., Roy, J. K., Prasad, M. (2001). Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science*, **80**, 524-535.

[Hsia *et al.*, (2005)] Hsia, A.-P., Wen, T.-J., Chen, H. D., Liu, Z., Yandeau-Nelson, M. D., Wei, Y., *et al.*, Schnable, P. S. (2005). Temperature gradient capillary electrophoresis (TGCE)–a tool for the high-throughput discovery and mapping of SNPs and IDPs. *Theoretical and Applied Genetics*, **111**(2), 218-25.

[Karrow, (2010)] Karrow, J. (2010). Life Tech Details Real-Time Single-Molecule Tech at AGBT; Combines Qdots with FRET-Based Detection. *Genome Web*, March 2, 2010.

[Konieczny and Ausubel, (1993)] Konieczny, A., Ausubel, F. M. (1993). A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. *The Plant Journal*, **4**(2), 403-410.

[Jia and Zhao, (2012)] Jia, P., Zhao, Z. (2012). Personalized pathway enrichment map of putative cancer genes from next generation sequencing data. *PLoS ONE*, **7**, e37595.

[Lee *et al.*, (1999)] Lee, L. G., Livak, K. J., Mullah, B., Graham, R. J., Vinayak, R. S., Woudenberg, T. M. (1999). Seven-color, homogeneous detection of six PCR products. *Biotechniques*, **27**(2), 342-349.

[Li and Tibshirani, (2011)] Li, J., Tibshirani, R. (2011). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, **22**, 519-536.

[Mardis, 2008] Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, **9**, 387-402.

[Marguerat *et al.*, (2008)] Marguerat, S., Wilhelm, B. T., Bahler, J. (2008). Next-generation sequencing: applications beyond genomes. *Biochemical Society Transactions*, **36**, 1091-1096.

[McCallum *et al.*, (2000)] McCallum, C. M., Comai, L., Greene, E. A., Henikoff, S. (2000). Targeted screening for induced mutations. *Nature Biotechnology*, **18**(4), 455-457.

[McGettigan, 2013] McGettigan, P. A. (2013). Transcriptomics in the RNA-seq era. *Current Opinion in Chemical Biology*, **17**(1), 4-11.

[Mei *et al.*, (2000)] Mei, R., Galipeau, P. C., Prass, C., Berno, A., Ghandour, G., Patil, N., Wolff, R. K., Chee, M. S., Reid, B. J., and Lockhart, D. J. (2000). Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Research*, **10**,1126-1137.

[Mutz *et al.*, (2013)] Mutz, K.-O., Heilkenbrinker, A., Lönne, M., Walter, J.-G., Stahl, F. (2013). Transcriptome analysis using next-generation sequencing. *Current Opinion in Biotechnology*, **24**(1), 22-30.

[Myers *et al.*, (1986)] Myers, R. M., Maniatis, T., Lerman, L. S. (1987). Detection and localization of single base changes by denaturing gradient gel electrophoresis. *Methods Enzymology*, **155**, 501-527.

[Oefner and Underhill, (1995)] Oefner, P. J., Underhill, P. A. (1995). Comparative DNA sequencing by denaturing high-performance liquid chromatography (DHPLC). *American Journal of Human Genetics*, **57**, a266.

[Orita *et al.*, (1989)] Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K., Sekiya, T. (1989). Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, **86**(8), 2766-70.

[Pagani *et al.*, (2011)] Pagani, I., Liolios, K., Jansson, J., Chen, I.M., Smirnova, T., Nosrat, B., Markowitz, V.M., Kyrpides, N.C.. (2011). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, **40**, D571-D579.

[Park, (2008)] Park, P. J. (2008). Epigenetics meets next-generation sequencing. *Epigenetics*, **3**, 318-321.

[Park, (2009)] Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**, 669-680.

[Rafalski, (2002)] Rafalski, J. A. (2002a). Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Science.* **162**,329-333.

[Reumers *et al.*, (2012)] Reumers, J., De Rijk, P., Zhao, H., Liekens, A., Smeets, D., Cleary, J., *et al.*, Del-Favero, J. (2012). Optimized filtering reduces the error rate in detecting geno-mic variants by short-read sequencing. *Nature Biotechnology*, **30**(1), 61-8.

[Ronaghi *et al.*, (1998)] Ronaghi, M., Pettersson, B., Uhlén, M., Nyrén, P. (1998). PCR-Introduced Loop Structure as Primer in Research Reports. *Biotechniques*, **25**, 876-884.

[Sanger and Coulson, (1975)] Sanger, F., Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, **94**(3), 441-8.

[Sanger *et al.*, (1977)] Sanger, F., Nicklen, S., Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *PNAS*, **74**(12), 5463-7.

[Steijger *et al.*, (2013)] Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Abril, J. F., Akerman, M., *et al.*, Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, **10**(12), 1177-84.

[Syvänen *et al.*, (1990)] Syvänen, A.-C., Aalto-Setälä, K., Harju, L., Kontula, K., Söderlund, H. (1990). A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics*, **8**(4), 684-692.

[Tarazona *et al.*, (2011)] Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Research*, **21**, 2213-2223.

[Troggio *et al.*, (2008)] Troggio, M., Malacarne, G., Vezzulli, S., Faes, G., Salmaso, M., Velasco, R. (2008). Comparison of different methods for SNP detection in grapevine. *Vitis*, **47**(1), 21-30.

[Voelkerding *et al.*, (2009)] Voelkerding, K. V., Dames, S. A., Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clinical Chemistry*, **55**, 641-658.

[White *et al.*, (1992)] White, M. B., Carvalho, M., Derse, D., O'Brien, S. J., Dean, M. (1992). Detecting single base substitutions as heteroduplex polymorphisms. *Genomics*, **12**(2), 301-306.

[Zhou *et al.*, (2010)] Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., Yu, J. (2010). The next-generation sequencing technology and application. *Protein Cell*, **1**, 520-536.

Chapter 2

VerySNP

2.1 Abstract

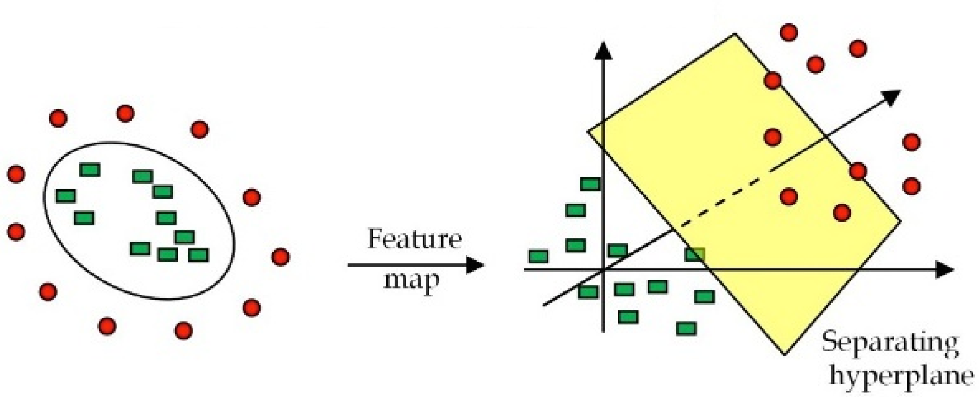
Several open-source tools have been recently developed to identify small variants in whole-genome data, the most popular being SAMtools and GATK. Commonly, variant calling provides a VCF file as output, which contains a list of candidates and additional information such as the variant call quality and its depth of coverage. Still the variant list presents an unsatisfactory accuracy due to high false positive calling rate. VCF parameters have been used to train a Support Vector Machine (SVM) that classifies the VCF records in true and false positive variants, cleaning the output from the most likely false positive results. We implemented the SVM approach in a new software, called VerySNP, and applied it to model and non-model organisms proving, in both cases, that this machine learning method efficiently recognizes true positive from false positive variants. The software is available at https://github.com/leonardelli/VerySNP.

2.2 Introduction

2.2.1 Support Vector Machines

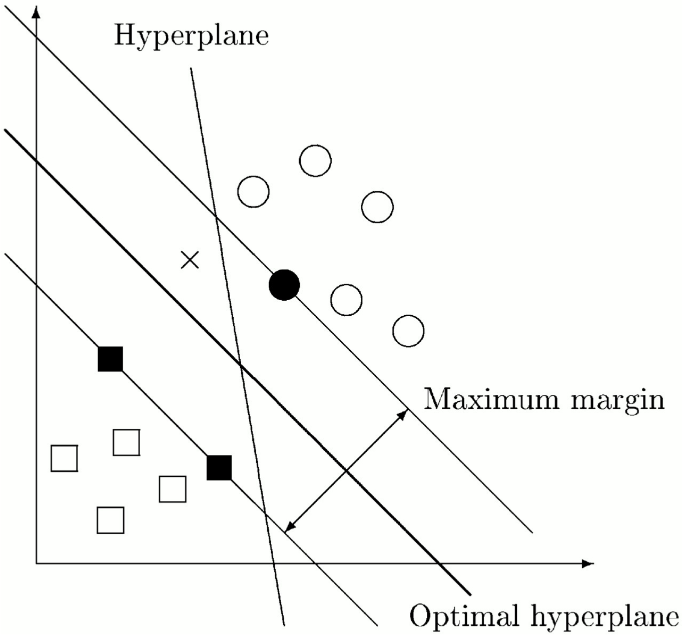
Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification or regression. Support vector machines represent an extension to non-linear models of the generalized portrait algorithm developed by Vladimir Vapnik. The SVM algorithm is based on the statistical learning theory and the Vapnik-Chervonenkis (VC) dimension introduced by Vladimir Vapnik and Alexey Chervonenkis (Vapnik *et al.*, 1998). The SVM is an efficient and reliable machine learning method to distinguish categorical data based on the contraction of a maximal margin hyperplane,

Figure 2.1: Calculating a list of feature for each point, the SVM spots the data in a higher space, called feature space, where the two clusters separation may be easier.



also referred to as the decision boundary, and the use of a kernel function to transform the data sets from the original input space into a high dimensional feature space (Figure 2.1). In the feature space, defined as a space for all possible combinations of predictive variables, highly non-linear relationships between the factors or attributes are qualified and examined using the margin maximization principle. The margin maximization principle has been proven mathematically to deliver robust and predictable performance on unseen data. The maximal margin hyperplane is defined as the hyperplane located at the largest distance to the nearest training data point of any class (Figure 2.2). In order to calculate it, the SVM selects two hyperplanes separating the two data sets with no points in between, and then tries to maximize their distance.

Figure 2.2: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. The points touching the lateral hyperplanes, called support vectors, confine the maximum margin, which is built drawing between the two clusters two parallel hyperplanes as far as possible from each other. The hyperplane drawn right in the middle of them represent the optimal separation between the two training sets (positive and negative) and it is the algorithm output to categorize new examples.



SVM represent the latest advancement in machine learning theory and delivers state of the art performance in numerous high value applications, which involved several types of biological data, including SNP identification in human sequencing data (Kong *et al.*, 2007). Kong *et al.* (2007) calculated the training features on the thermodynamic properties of nucleotides flanking the SNP site and they used the SVM model to recognize potential polymorphic sites, introducing a new feature, the SNP distribution score, which let them reach higher prediction rate (around 77%). While, another study exploiting flanking region thermodynamic properties to train RBF Networks, evaluated the SNP occurrence possibility in *Brassica napus*, as example of species lacking of a whole reference sequence (Hu *et al.*, 2011). SVM provided very efficient results in finding polymorphisms even when combined to other statistical approaches, such as the Fischer exact test in a hybrid method applied to Brassica oilseed rape genomic data (Xiong *et al.*, 2010).

Recently, the 1000 Genome Project Consortium built an integrated map of genetic variation from 1,092 human genomes and they generated the consensus exome SNP call set using an SVM approach (McVean *et al.*, 2012). For each candidate variant site they calculated features related to the aligned sequence reads, such as allele balance, strand bias, cycle bias, average depth and inbreeding coefficient statistics. Each feature was considered separately and classified as 'pass' or 'fail' based to a determined threshold. After multiple filter criteria they apply the SVM model assigning a score to each variant and variant with positive SVM score were considered as consensus SNP.

2.2.2 Variant calling methods

The variation of DNA sequence is at the same time the power and the subject of evolution; through genetic mutation organisms could gain new functions, after selection those functions could be inherited and the mutation would be fixed in the population. Once the DNA has changed, the genetic variation in the mutant organism can be identified in the DNA sequence if compared to a wild-type individual. When a single nucleotide changes with an allelic frequency bigger than 1% within a population, it is known as polymorphism.

Single Nucleotide Polymorphisms (SNPs) represent the most abundant type of genetic variations and are more stably inherited than other molecular markers (Brookes, 1999). They represent a valuable tool for several biological applications like linkage mapping, integration of genetic and physical maps, population genetics as well as evolutionary and protein structure-function relationship studies (Syvänen, 2001). The great interest in variant detection has been reflected in the development of a wide range of SNP genotyping methods (Mammadov *et al.*, 2012). Furthermore, the importance of finding only true variants is evident, considering the high cost of experimental validation through resequencing or SNP-chip (Ganal *et al.*, 2012), not only in terms of money, but also of time and samples.

The advent of next generation sequencing (NGS) technologies affects variant detection both directly and indirectly. Directly, because such techniques allow the production of a large amount of sequences cheaply and, indirectly, by increasing the number of available genome sequences. As a consequence, the most effective way to predict variants is based on mapping the DNA reads against a reference genome. Although NGS technologies are increasing the amount of genomic information at unprecedented pace, they are prone to an error rate of about one in one hundred base pairs (Loman *et al.*, 2012). These errors prevent reaching very high accuracy by means of *in silico* variant calling and, in general, by any data filtering procedure aimed at automatically identifying biologically relevant variants (Nakamura *et al.*, 2011; Taub *et al.*, 2010). Incorrect base calling is one of the most common sequencing errors especially near the 3' end of the sequence as the quality of raw data declines (Minoche *et al.*, 2011). Poor base quality along with short average read length may generate inaccurate data mapping on reference genomes. Further errors come from distortions with respect to the sample population, due to biases from the chosen sequence technology or from the reverse-DNA transcription and PCR steps required to generate cDNA libraries (Reumers *et al.*, 2012). Effective approaches are thus needed to distinguish real variants from the numerous sequencing artefacts.

Variant calling methods on data generated by Sanger sequencing were based on the analysis of trace files with a Bayesian statistics (Marth *et al.*, 1999). To handle NGS data such a kind of approach was, firstly, paired to the Artificial Neural Network (ANN) (Unne-berg *et al.*, 2005) and afterwards to other machine learning methods (Matukumalli *et al.*, 2006; Wegrzyn *et al.*, 2009) providing higher accuracy in variant identification. Until now several software have been developed, the most popular tools to process large-scale datasets are the functions *mpileup* in SAMtools package (Li *et al.*, 2009) and *UnifiedGenotyper* in GATK (Genome Analysis ToolKit; McKenna *et al.*, 2010), which are both binomial-based methods. GATK includes the Variant Quality Score Recalibration tool (*VQSR*; DePristo *et al.*, 2011), which identifies putative nucleotide variations using a multidimensional Gaussian distribution fitted to known true variant sites. Even though these tools accurately discover true variable sites, they still show high false positive rates, which is currently handled by using different empirically-derived filtering criteria (Koboldt *et al.*, 2012) on the several values showed in the VCF output.

Many factors contribute to defining a variant from mapped reads: the number of reads mapped on a region (read depth), the quality of the mapping, the distribution of nucleotides at the position, the distance of a potential polymorphic site from another, to cite some. Multiple factors may take part at the same time in defining a specific feature, for instance sequencing biases can affect both the read depth and the nucleotide frequencies (Nielsen *et al.*, 2011). Likewise, the genome nucleotide composition has effects on the overall nucleotide distribution, while other genetic parameters of the organism affect other features; e.g. the extent of Linkage Disequilibrium can affect the average distances among variants. Having such a complicate framework, the application of thresholds to rule out the most likely false positive predicitons is risky. Features are inter-dependent and should be considered together rather that one by one.

The Support Vector Machine (SVM) approach (Vapnik *et al.*, 1998) has gained increasing attention because of its successful application to many biological problems, including variant calling (Kong *et al.*, 2007, O'Fallon *et al.*, 2013). SVM-based methods are trained on a collection of known real and false variants, calculating some features for each of them. The software combines all features' values instead of using fixed thresholds. Here, we propose VerySNP, a SVM-based tool that classifies variant calling outputs to reduce false positive variants in downstream applications.

2.3 Methods

2.3.1 VerySNP: the tool

In this work, we used the freely available SVM package LIBSVM library (v 3.12; Chang and Lin, 2011). Once LIBSVM library is installed, a Support Vector Machine (Vapnik *et al.*, 1998) needs to be trained on true and false examples of what it is supposed to classify. Known true and false genetic variants have been used as positive and negative sets, respectively, to train VerySNP. The training dataset is balanced and contains the same number of positive and negative entries.

During the training, the SVM classifier learns how to discriminate true and false examples calculating for each one a specific list of features, which are known to affect the classification. The list of features involved in VerySNP training includes every parameter shown in 'Quality', 'Info' and 'Format' fields of a VCF file. A detailed description of VCF features can be found at SAMtools GitHub web page (http://samtools.github.io/hts-specs/VCFv4.2.pdf), while a complete summary is reported in Table 2.1. A binary SVM discriminates between two classes y_i, with $y_i \in \{+1, -1\}$. The discrimination between the two classes y_i can be made either through linear kernel function or Radial Basis Function (RBF) kernel. A 10-fold cross-validation evaluates the performance of the SVM on the training data. A grid search finds the best parameters (linear kernel: C; RBF-kernel: C and gamma) on the training folds. The parameter combination with the highest Matthews Correlation Coefficient (MCC) is finally chosen for testing. The test set is any candidate variant provided as output by GATK/SAMtools variant calling. Performing the test, VerySNP can predict if the candidate variant belongs either to the positive (+1) or the negative (-1) class.

In particular, VerySNP is composed of two main scripts written in Python (version 2.7.5) that implement the two steps of SVM approach: training and test. 'VerySNP_ training. py' needs the training set as input to build a prediction model, while 'VerySNP_test.py' classifies new unknown data either as positive or negative variants. More details about VerySNP usage are provided with the software package (readme. txt) available at https:// github.com/leonardell/VerySNP.

VerySNP classification performance was evaluated by calculating accuracy, specificity, sensitivity and precision (Loong *et al.*, 2003). Furthermore, Receiver Operating Char-

Table 2.1: Some of the parameters reported into VCF files by GATK and SAMtools, respectively, and used as VerySNP training features.

VCF name	Description	GATK	SAMtools
QUAL	SNP call quality	Yes	Yes
AC	Allele count in genotype, for each ALT allele	Yes	Yes
AF	Allele Frequency, for each ALT allele	Yes	Yes
GQ	Genotype Quality	Yes	Yes
PL	Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification	Yes	Yes
MQ	Mapping Quality	Yes	Yes
GT	Genotype	Yes	Yes
DP	Approximate read depth	Yes	Yes
FQ	Phred probability of all samples being the same	-	Yes
VDB	Variant Distance Bias	-	Yes
DP4	High-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases	-	Yes
PV4	P-value for strand bias, baseQ bias, mapQ bias and tail distance bias	-	Yes
AN	Total number of alleles in called genotypes	Yes	-
BaseQRankSum	Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities	Yes	-
Dels	Fraction of Reads Containing Spanning Deletions	Yes	-
FS	Phred-scaled p-value using Fisher's exact test to detect strand bias	Yes	-
HaplotypeScore	Consistency of the site with at most two segregating haplotypes	Yes	-
MLEAC	Maximum likelihood expectation (MLE) for the allele counts, for each ALT allele	Yes	-
MLEAF	Maximum likelihood expectation (MLE) for the allele frequency, for each ALT allele	Yes	-
MQ0	Total Mapping Quality Zero Reads	Yes	-
MQRankSum	Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities	Yes	-
QD	Variant Confidence/Quality by Depth	Yes	-
ReadPosRankSum	Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias	Yes	-
AD	Allelic depths for the ref and alt alleles	Yes	-

acteristic (ROC) curve and precision-recall curve have been drawn in order to compare VerySNP performance with the state of the art (SNPSVM and VQSR). The ROC curve is generated by plotting the fraction of false positives out of the total actual negatives (FPR, False Positive Rate) as a function of the fraction of true positives out of the total actual positives (TPR, True Positive Rate), while varying the discrimination threshold between the two classes y_i of a binary classifier.

2.3.2 Benchmarking

VerySNP was tested on three datasets originated from the genome sequencing of different organisms. Firstly, on yeast (*Saccharomyces cerevisiae*, strain EM93), a model organism with a rather small genome (12.1 Mb; Mewes *et al.*, 1997), easily manageable for bioinformatic analysis. Subsequently, the method was applied on two cultivated grapevine (*Vitis vinifera* L.) varieties, Pinot Noir (clone ENTAV 115) and Gewürztraminer (clone SMA 918), to test the tool's performance on a non-model organism with a rather large (504.6 Mb) and highly heterozygous genome (Velasco *et al.*, 2007). Pinot Noir is a black-berried internationally-grown variety and is a parent of the PN40024 near-homozygous line chosen as reference genome (Jaillon *et al.*, 2007). Gewürztraminer, belonging to the Savagnin or Traminer family, is a white-berried variety genetically distant from Pinot (Bowers *et al.*, 1996; Lacombe *et al.*, 2012). All the samples were sequenced through Illumina technology, but with different depth of coverage (yeast 125X, Pinot Noir 107X and Gewürztraminer 20X).

Yeast and Gewürztraminer reads have been publicly released (for yeast EM93 at DDBJ database, http://trace.ddbj.nig.ac .jp, with accession number ERP002541; for Gewürztraminer, EBI-ENA database project ID: PRJEB6378), while for Pinot Noir we exploited in-house data (Fondazione Edmund Mach). All data were aligned against the proper reference genome using Bowtie2 software (version 2.1.0; Langmead *et al.*, 2012) with standard options and VCF files were predicted by applying *mpileup* of SAMtools (version 0.1.18) and *UnifiedGenotyper* of GATK (version 2.3-9 with java version 1.7.0_17) with default options too. Among the predictions made by the two softwares we selected positions of true and false variants in different ways for the three organisms.

The sequence of 2,965 probes identifying validated variants in yeast EM93 were taken from Esberg *et al.* (2011) and were used to select true variants from SAMtools prediction, where 2,989 variants were called in correspondence of the 25 bp probes, and from GATK, where 3,419 variants were found into the probes regions (Gresham *et al.*, 2006). Original data of Esberg *et al.* are available at the EBI-EMBL database with the ArrayExpress accession number E-MEXP-3246. Since no monomorphic sites were available for yeast from public sources, we decided to collect false variants realigning simulated reads from the reference genome against itself. An appropriate tool for this scope is *ArtificialFastqGenerator* (Frampton *et al.*, 2012), which generates artificial paired-end reads, randomly derived from the reference genome sequence, to provide a gold-standard for reads alignment and variant calling. Artificial reads were generated by the software complying with the nucleotide quality scores of the original reads and including the error model of Illumina technology. We run the simulation several times generating an average depth of coverage of 120X for each simulation. The variant calling provided 2,366 false variants with SAMtools and 2,161 with GATK. These sets were highly redundant in genomic positions and, counting only unique coordinates, we got 433 unique false variant positions for SAMtools and 412 for GATK.

True and false variants for Pinot and Gewürztraminer were obtained from the analysis of SNP-chip array experiments. Grapevine genomic DNA samples were hybridized on Vitis17KSNP chip (GrapeReSeq Consortium https://urgi.versailles.inra.fr/Projects/Grape ReSeq) and data were analyzed with GenomeStudio Data Analysis Software. Based on signal clustering, Genome Studio identified the high quality hybridization sites either as heterozygous or homozygous giving a score for the cluster called *GenTrain* value. Clusters automatic evaluation for *GenTrain* values lower than 0.7 might be incorrect and so we performed a manual evaluation for all the ambiguous cases. In total we got 5,161 heterozygous sites in Pinot Noir and 4,958 in Gewürztraminer, while the other 12,495 homozygous sites in Pinot Noir and 12,640 in Gewürztraminer. When confirmed heterozygous sites were called as homozygous by the variant caller were considered those sites as false variants, on the contrary the true variants included all the confirmed correct calls.

We compared VerySNP performance against SNPSVM (O'Fallon *et al.*, 2013), another software exploiting SVM approach, and VQSR, which includes the complete pipeline of GATK best practice guidelines to predict polymorphisms (https://www.broadinstitute.org /gatk/guide/best-practices) learning from true variant examples only.

All tested softwares require a training step using a VCF file. VCF files were next produced out of each sample by applying SAMtools and GATK, which predicted the largest number of variants presented in our benchmarking sets (Tables 2.2 and 2.3). The number of true and false sites used as training for each sample is summarized in Table 2.2. The overlap between the predicted variants and the benchmarking sets represents all known variants available for training the models, from which we built the actual training sets balancing the number of known true variants to the number of known false ones. The variant fraction not considered for training was exploited to evaluate the tools' performance.

Table 2.2: The whole variant call provided by *mpileup* of SAMtools and *UnifiedGenotyper* of GATK, respectively (Tot predicted variants) and the number of known true and false variants in three different samples (Yeast EM93, Pinot Noir ENTAV 115 and Gewürztraminer SMA 918). The predicted true/false variants come from overlapping the total amount of predicted variants and the known true/false variants.

	Tot Predicted variants	Known True set	Predicted True set	Known False set	Predicted False set
Yeast SAMtools	42,766	2,965	2,989	2,339	11
Yeast GATK	48,122	2,965	3,419	2,114	24
Pinot Noir SAMtools	3,097,569	5,161	4,617	12,495	1,541
Pinot Noir GATK	4,597,394	5,161	4,948	12,495	1,651
Gewürztraminer SAM	2,696,200	4,958	4,043	12,640	2,177
Gewürztraminer GATK	3,036,621	4,958	4,435	12,640	2,228

Table 2.3: Performance comparison among three variant calling (VerySNP, SNPSVM and VQSR) in three different samples: Yeast EM93; Pinot Noir ENTAV 115 and Gewürztraminer SMA 918. The predicted true/false sets come from overlapping the total amount of predicted variants and the evaluation sets (known true and known false variants left out of the training on purpose to evaluate the tool performance).

Yeast EM93			Tot Predicted variants	Evaluation True set	Predicted True set	Evaluation False set	Predicted False set
VQSR			47,451	824	820	412	19
SNPSVM	SAM		49,381	2,567	2,514	11	10
	GATK		50,081	3,007	2,960	24	13
VerySNP	SAM	Linear	42,714	2,567	2,560	11	5
		RBF	42,714	2,567	2,562	11	6
	GATK	Linear	48,036	3,007	2,974	24	17
		RBF	48,036	3,007	2,983	24	14

Pinot Noir ENTAV 115			Tot Predicted variants	Evaluation True set	Predicted True set	Evaluation False set	Predicted False set
VQSR			3,792,196	3,429	3,358	132	129
SNPSVM	SAM		5,239,730	3,200	3,173	124	15
	GATK		14,832,057	3,429	3,388	132	13
VerySNP	SAM	Linear	3,082,834	3,200	3,171	124	110
		RBF	3,082,834	3,200	3,161	124	111
	GATK	Linear	4,588,257	3,429	3,366	132	120
		RBF	4,588,257	3,429	3,306	132	123

Gewürztraminer SMA 918			Tot Predicted variants	Evaluation True set	Predicted True set	Evaluation False set	Predicted False set
VQSR			2,564,249	2,385	2,086	178	175
SNPSVM	SAM		14,111,458	2,040	2,036	174	174
	GATK		10,526,342	2,385	2,346	178	8
VerySNP	SAM	Linear	2,683,903	2,040	1,999	174	167
		RBF	2,683,903	2,040	2,000	174	166
	GATK	Linear	3,032,207	2,385	2,350	178	170
		RBF	3,032,207	2,385	2,350	178	170

Equation 2.1 shows how to calculate VerySNP sensitivity or its True Positive Rate (TPR) knowing the number of Positive (P), True Positive (TP) and False Negative (FN). Equation 2.2 executes the specificity or True Negative Rate, knowing the number of Negative (N), True Negative (TN) and False Positive (FP). Equation 2.3 is the precision or Positive Predictive Value of the software. Equation 2.4 provides the fall-out or False Positive Rate and it represents the specificity complementary. Equation 2.5 calculates the accuracy of a binary classifier performance. Equation 2.6 is the Matthews Correlation Coefficient.

$$TPR = TP/P = TP/(TP + FN) \tag{2.1}$$

$$SPC = TN/N = TN/(FP + TN) \tag{2.2}$$

$$PPV = TP/(TP + FP) \tag{2.3}$$

$$FPR = FP/N = FP/(FP + TN) = 1 - SPC \tag{2.4}$$

$$ACC = (TP + TN)/(P + N) \tag{2.5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.6}$$

2.3.3 Performance evaluation

To estimate VerySNP performances, we calculated ROC and precision-recall curves on the evaluation sets composed of known true and false variants intentionally left out from the training set. Specifically, each evaluation set was composed of the 8% of the initial negative variants and the unused true variants remaining after balancing the true and false entries of the training sets (for detailed description see Table 2.4). The varied value used to draw the curves was different for each software: the probability to be a true variant from the output of VerySNP, the quality value from the output of SNPSVM and the VQSLOD parameter reported by VQSR VCF output file.

2.3.4 Feature selection

To better understand the role of each VCF value into the classification process we exploited the Recursive Feature Elimination (RFE) and cross-validation, as described in Abeel *et*

Table 2.4: Number of known true and false variants included in the training sets and in the evaluation sets. For each sample, SAM means the variant positions were retrieved applying SAMtools *mpileup*, while GATK denotes variant calls performed by *UnifiedGenotyper*.

Tool	Sample	Training set		Evaluation set	
		True	False	True	False
SAM	Yeast	422	422	2567	11
	Pinot Noir	1,417	1,417	3,200	124
	Gewürztraminer	2,003	2,003	2,040	174
GATK	Yeast	388	388	3,031	24
	Pinot Noir	1,519	1,519	3,429	132
	Gewürztraminer	2,050	2,050	2,385	178

al. (2010) already used. The feature set under analysis includes 23 features adopted when the variant calling was made through *UnifiedGenotyper* (GATK) and 20 features when performed by *mpileup* (SAMtools) (Table 2.1).

Starting with the whole feature set, a linear SVM classifies the training sets and the RFE iteratively removes the least important feature in terms of weight in the SVM hyperplane. At each step, a linear SVM is re-estimated on the same training sets calculating the remaining features only, until all features are eliminated. We used the scikit-learn package for the RFE (Guyon *et al.*, 2002).

2.4 Results

A very early result of our work was the production of true and false variant sets useful to train or to validate new methods for variant calling. For each sample (yeast, Pinot Noir and Gewürztraminer) we provide the variant coordinates on the available genome sequence (see the VCF files provided in the GitHub repository along with the software).

Figure 2.3: The ROC curves obtained plotting the True Positive Rate (TPR or sensitivity) vs. the False Positive Rate (FPR or fall-out) for three variant callers (VerySNP, SNPSVM and VQSR) in three different samples: a) yeast; b) Pinot Noir and c) Gewürztraminer.

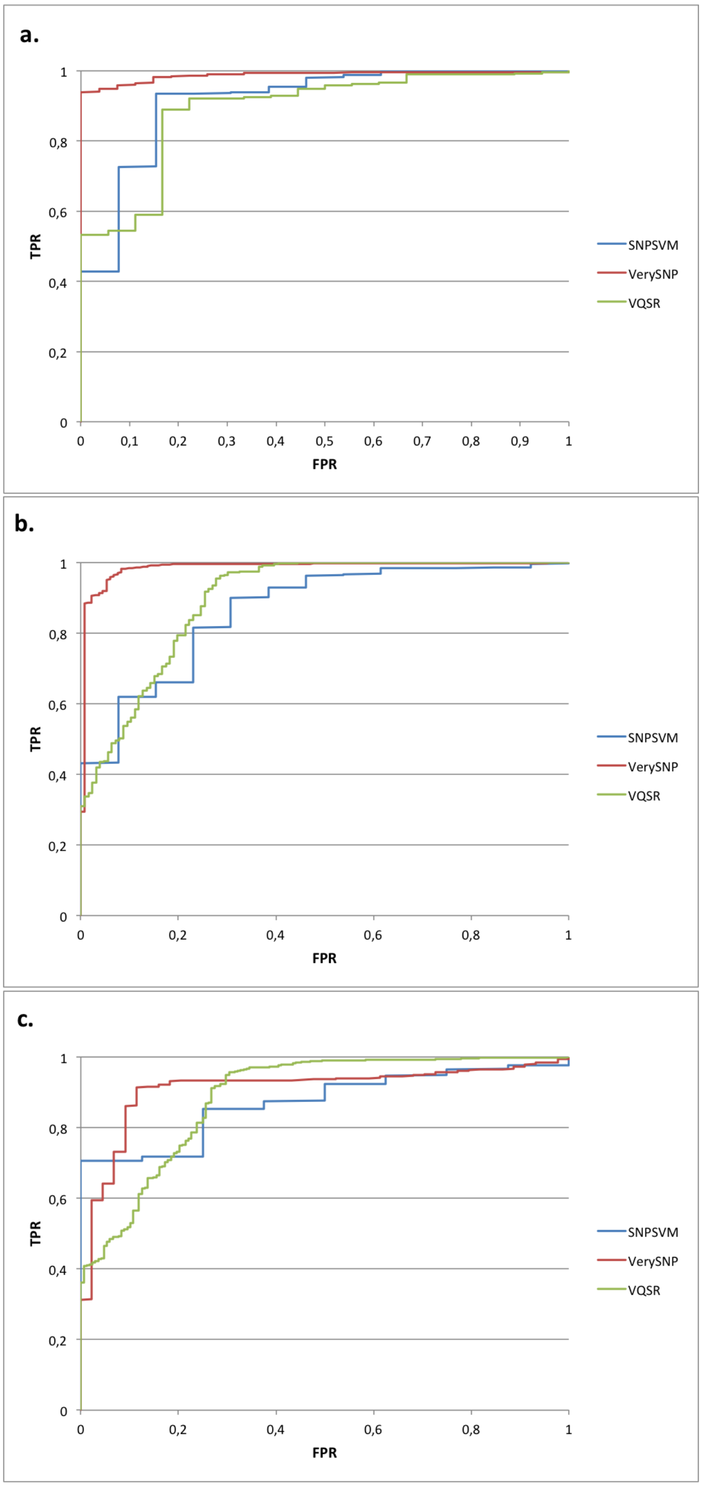


Table 2.5: Area under the ROC curves resulted applying VerySNP, SNPSVM and VQSR to yeast, Pinot Noir and Gewürztraminer.

Tools	Yeast	Pinot Noir	Gewürztraminer
VerySNP	0.988	0.986	0.910
SNPSVM	0.914	0.864	0.871
VQSR	0.891	0.896	0.887

Figure 2.4: Venn diagram showing the number of true (T) and false (F) variants in the evaluation set and in the prediction of VerySNP, SNPSVM and VQSR applied to yeast dataset (TPs = True Positives; FPs = False Positives): a) Number of true positives (Ps) in the evaluation set (green circle) overlapped to all VerySNP (blue circle) and SNPSVM (purple circle) predictions (TOT variants); b) Number of true variants of the evaluation set called by VQSR (green circle), VerySNP (blue circle) and SNPSVM (purple circle) (TPs); c) Number of false variants in the evaluation set predicted by VQSR (green circle), VerySNP (orange circle) and SNPSVM (red circle) (FPs).

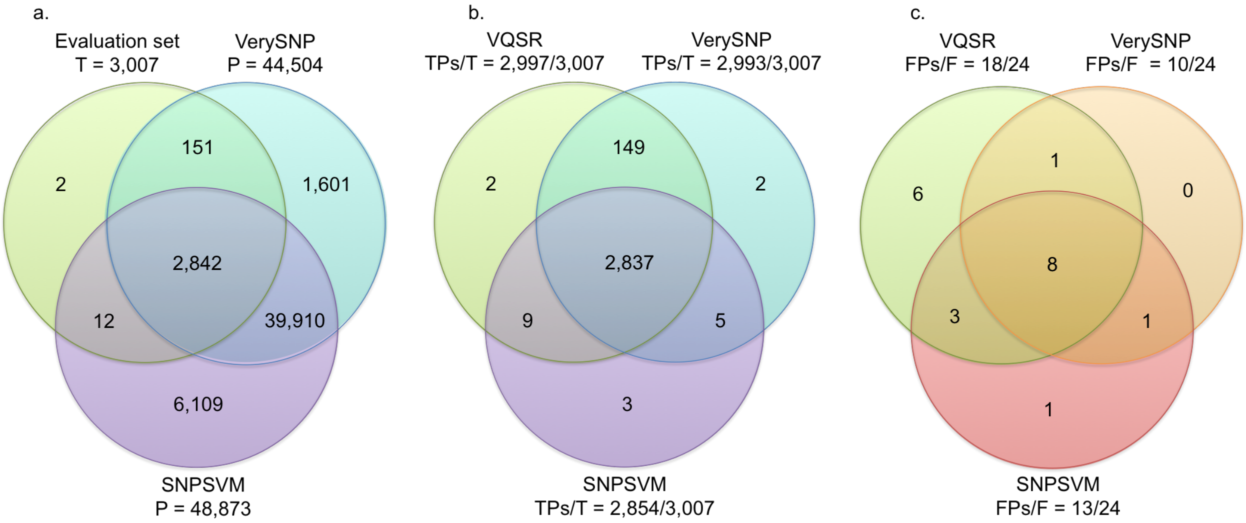


Figure 2.3 shows the performance of VerySNP when trained on GATK variant calling and RBF kernel. The best model is finally used for the evaluation set. Very similar results were obtained using GATK and linear kernel, while using SAMtools variant calling along with its training sets has given slightly worse performances (see Table 2.2). Considering the area under the ROC curves (Table 2.5) we can conclude that VerySNP showed the largest areas with regard to SNPSVM and VQSR for all tested samples (average value

0.961) and it proved to be a very good tool to accurately identify positive variants and correctly recognize false ones.

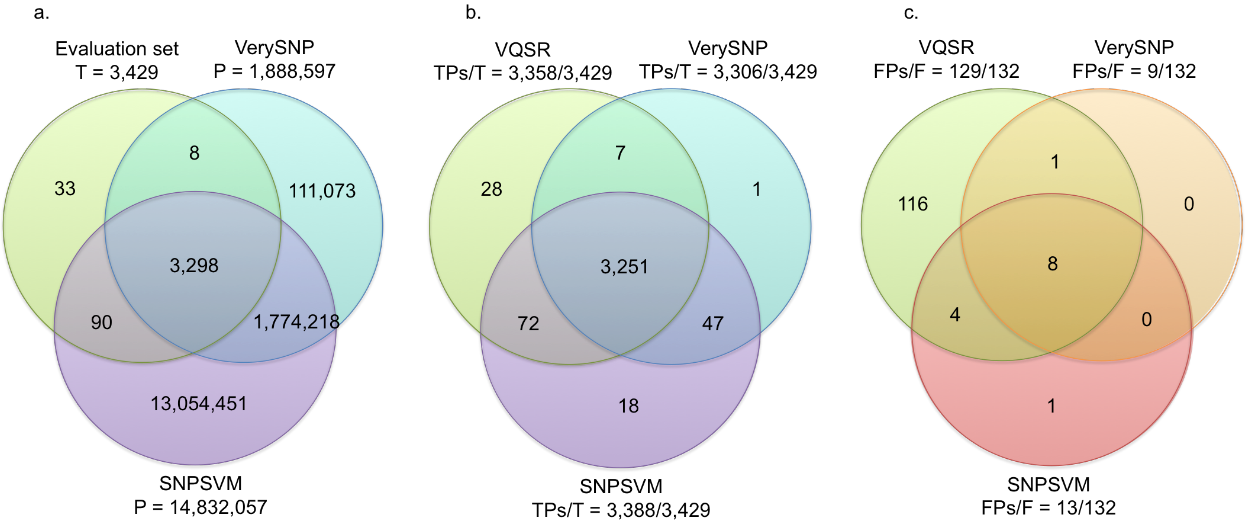
In our study SNPSVM and VQSR showed fairly overlapping ROC curves and the average values of the area under the curve were 0.883 and 0.858, respectively. In yeast (Figure 2.3 [a]), SNPSVM outperforms VQSR after a level of specificity equal to 0.08. Looking at grapevine samples, although Pinot Noir curves (Figure 2.3 [b]) were rather similar, the SNPSVM curve was lower than VQSR for specificity higher than 0.12. Similarly, in Gewürztraminer (Figure 2.3 [c]) SNPSVM showed lower sensitivity than VQSR whereas specificity was higher than 0.26.

The effectiveness of VerySNP and SNPSVM as binary classifiers in reducing the false positives rate can be easily pointed out considering the results presented in Figures 2.4 (yeast), 2.5 (Pinot Noir) and 2.6 (Gewürztraminer). In particular, figure 2.6 [a] and 2.6 [b] show that out of 178 false positive variants available as evaluation set, VQSR wrongly predicted 175 as true variants, while the two SVM methods failed in 8 cases only (Figure 2.6 [c]). Similar results were obtained for yeast (Figure 2.4) and Pinot Noir (Figure 2.5). It is worth mentioning that, although SNPSVM and VerySNP are both based on SVM, they are trained on different features, directly calculated from the aligned reads by SNPSVM, while already calculated through Bayesian methods from the aligned reads in VerySNP.

The variant calling goal is to detect true variants avoiding false positives and not missing any true variant. Tool's performance can be measured in those terms by using the precision-recall curve, where precision (or positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Drawing the precision-recall curves for each software applied to one sample at the time (Figure 2.7) reveals three similar behaviors, showing just a small range of differences in the AUC. While the areas under the precision-recall curves in yeast showed no differences (AUC was 1.000 for all three software), in Pinot Noir the AUC went from 1.000 of VerySNP to 0.996 of VQSR and in Gewürztraminer the range was slightly larger, going from 1.000 of VerySNP to 0.991 of VQSR.

The comparison between precision-recall and ROC curves highlights that the main difference among the three tools is the ability to distinguish false positives rather than recover the whole amount of true positives. Indeed, the larger difference between the areas under the ROC curves of the three software applied to Gewürztraminer is 0.039, almost ten times bigger than between the areas under the precision-recall curves (0.004).

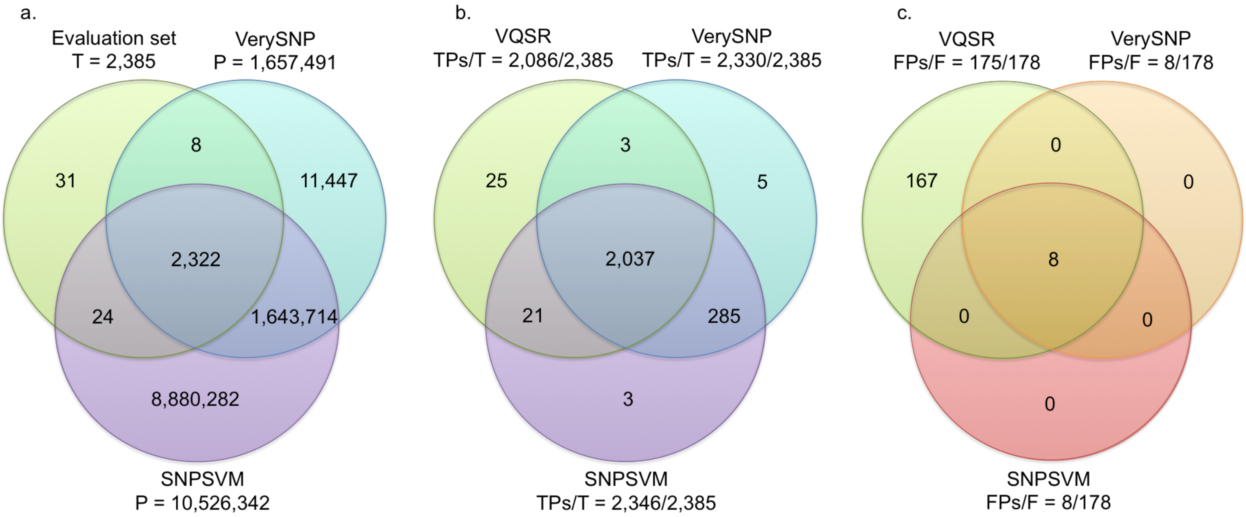
Figure 2.5: Venn diagram showing the number of true (T) and false (F) variants in the evaluation set and in the prediction of VerySNP, SNPSVM and VQSR applied to Pinot Noir dataset (TPs = True Positives; FPs = False Positives): a) Number of true positives (Ps) in the evaluation set (green circle) overlapped to all VerySNP (blue circle) and SNPSVM (purple circle) predictions (TOT variants); b) Number of true variants of the evaluation set called by VQSR (green circle), VerySNP (blue circle) and SNPSVM (purple circle) (TPs); c) Number of false variants in the evaluation set predicted by VQSR (green circle), VerySNP (orange circle) and SNPSVM (red circle) (FPs).



The whole variants profile of the three studied samples is unknown, making hard to estimate the missing true variants, except for the evaluation set. The high number of variants predicted by SNPSVM (10,526,342) in Gewürztraminer is quite impressive when compared to VQSR and VerySNP (1,657,491), raising the question of which tool is the closest to the real picture.

RFE and cross-validation analysis shown a quite high optimal number of VCF features required for best performances. Among the 23 values of GATK VCF outputs, the best classifications are reached with 12 for Pinot Noir, 13 and 21 for yeast and Gewürtztraminer (Table 2.6), respectively. Among the top ranking features of all samples there are features linked to the alignment quality, as the combine depth of aligned reads (DP) and the Mapping Quality (MQ), and values referred to features of the hypothetical variant, like the number of alleles (AC) and their frequency (AF).

Figure 2.6: Venn diagram showing the number of true (T) and false (F) variants in the evaluation set and in the prediction of VerySNP, SNPSVM and VQSR applied to Gewürztraminer dataset (TPs = True Positives; FPs = False Positives). a) Number of true positives (T) in the evaluation set (green circle) overlapped to all VerySNP (blue circle) and SNPSVM (purple circle) predictions (P); b) Number of true variants (TPs) of the evaluation set called by VQSR (green circle), VerySNP (blue circle) and SNPSVM (purple circle); c) Number of false variants (FPs) in the evaluation set predicted by VQSR (green circle), VerySNP (orange circle) and SNPSVM (red circle) (FPs).



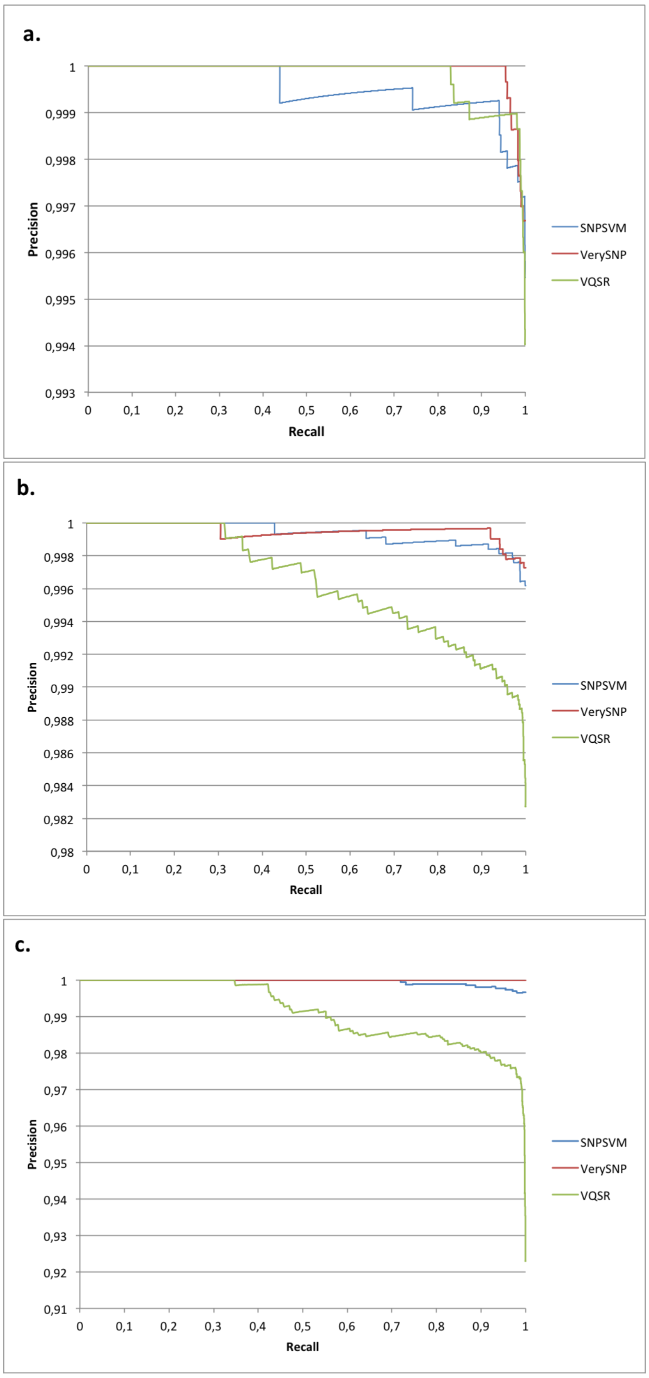
2.5 Discussion

Machine Learning techniques, and in particular SVMs, have often been applied to solve biological problems because of their high accuracy and efficiency, which are indispensable properties to detect variant as well. Since the most popular variant calling, GATK and SAMtools, usually call a set of variants large enough to include almost all possible true variants, here we propose to enhance the accuracy by reducing the false positive variant prediction rate with a SVM-based approach. VerySNP was designed to classify GATK and SAMtools calls in true and false variants taking into consideration all VCF features concerning reads alignment and nucleotide quality at the variant site.

Table 2.6: Cross validation of feature subsets and Recursive Feature Elimination (RFE) analysis. The cross validation of feature subsets evaluates the optimal number of informative features and we observed a minimum of 13, 12 and 21 features to reach the best classification, respectively for yeast, Pinot Noir and Gewürztraminer. The RFE ranks the features as reported in the table. Some features have been repeated multiple times depending on how many values they include.

	GATK Features	Yeast	Pinot Noir	Gewürztraminer
1	Quality	9	8	1
2	AC	1	1	1
3	AF	1	1	1
4	AN	11	13	3
5	BaseQRankSum	1	3	1
6	DP	2	1	1
7	FS	5	2	1
8	HaplotypeScore	7	4	1
9	MLEAC	1	1	1
10	MLEAF	1	1	1
11	MQ	1	1	1
12	MQ0	1	5	1
13	MQRankSum	4	1	1
14	QD	1	1	1
15	ReadPosRankSum	1	1	1
16	AD	3	10	1
17	AD	1	6	1
18	DP	6	1	1
19	GQ	1	11	1
20	GF	1	1	1
21	PL	8	9	1
22	PL	1	7	1
23	PL	10	12	2

Figure 2.7: Precision-recall curves of SNPSVM, VerySNP and VQSR applied to yeast (a), Pinot Noir (b) and Gewürztraminer (c). While the areas under the curves (AUCs) was 1.000 for all three software in yeast, in Pinot Noir the AUC measured 1.000 using VerySNP, 0.999 using SNPSVM and 0.996 using VQSR. In Gewürztraminer the range was slightly larger, going from 1.000 using either VerySNP or SNPSVM to 0.991 using VQSR.



VerySNP was tested on a model organism (yeast) and two cultivated varieties of a non-model plant organism (grapevine). Grapevine represents a typical non-model organism for its size, its long life cycle and its difficult genetic manipulation. Nonetheless, due to its economic importance there is much interest in studying its genetic variations and associate them to the plant phenotype. For these reasons we have developed a tool flexible enough to work both on model and non-model organisms.

Having Pinot and Traminer different geographical origins and being genetically well distinct (Lacombe *et al.*, 2012), we expected to observe significant differences in reads alignment against the reference genome and consequently in variant calling. This was not the case when looking at the ROC curves. Indeed, VerySNP learnt quite well from the VCF features and in both cases it very accurately recognized false variants from true ones. The choice of testing VerySNP on yeast also came from the need to apply the tool on publicly available data, in order to let the scientific community test our results. Moreover, yeast is commonly used to test bioinformatic tools because of its relatively small genome. We used in-house produced sequencing reads in order to be sure that the strain was exactly the same as the one described in the literature reporting the variant validation.

While VQSR requires tens of thousands of true examples to precisely fit its Gaussian distributions, SVM-based approaches, like SNPSVM and VerySNP, can make accurate calls, by learning from few hundreds of true and false variants, allowing precise variant calling in non-model organisms, such as grapevine, where a limited set of validated variants is available (e.g. Lijavetzky *et al.*, 2007; Pindo *et al.*, 2008; Vezzulli *et al.*, 2008). Finally, it can be foreseen that as more accurate training sets are developed, the prediction faithfulness of these tools will significantly increase.

2.6 Conclusion

Variant calling is a challenging process especially in non-model organisms due to the lack of largely validated variant sets and the high complexity of their genome sequence. The SVM tools have been proved to outperform other approaches in reducing false positive rate. Therefore, we provide a software that helps to tackle this problem exploiting the SVM ability to learn which variant has features closely related to known true, rather than false, variants. Valuable information is taken from the VCF files and used to detect the most likely candidate variants by applying the SVM model on the variant calling outputs.

Bibliography

[McVean *et al.*, (2012)] McVean, G. A. *et al.*, 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature, 491, 56-65.

[Abeel *et al.*, (2010)] Abeel, T., Helleputte, T., Van de Peer, Y., *et al.*, Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* (Oxford, England), **26**(3), 392-8.

[Bowers and Meredith, (1996)] Bowers, J. E., Meredith, C. P. (1996). Genetic Similarities among Wine Grape Cultivars Revealed by Restriction Fragment-length Polymorphism (RFLP) Analysis, *Journal of the American Society for Horticultural Science*, **121**(4), 620-624.

[Brookes, (1999)] Brookes, A. J. (1999). The essence of SNPs. *Gene*, **234**(2), 177-186.

[Chang and Lin, (2011)] Chang, C.C., Lin, C.J. (2011). LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology (TIST)*, **2**(3),1-27. Software available at http://www.csie.ntu.edu.tw/,cjlin/libsvm

[DePristo *et al.*, (2011)] DePristo, M., Banks, E., Poplin, R. E., *et al.*, Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nature Genetics*, **43**(5), 491-498.

[Esberg *et al.*, (2011)] Esberg, A., Muller, L. a H., McCusker, J. H. (2011). Genomic structure of and genome-wide recombination in the Saccharomyces cerevisiae S288C progenitor isolate EM93, *PloS One*, **6**(9), e25211.

[Frampton *et al.*, (2012)] Frampton, M., Houlston, R. (2012). Generation of artificial FASTQ files to evaluate the performance of next-generation sequencing pipelines, *PloS One*, **7**(11), e49110.

[Ganal *et al.*, (2012)] Ganal, M. W., Polley, A., Graner, E.-M., *et al.*, Durstewitz, G. (2012). Large SNP arrays for genotyping in crop plants, *Journal of Biosciences*, **37**(5), 821-828.

[Gresham *et al.*, (2006)] Gresham, D., Ruderfer, D. M., Pratt, S. C., *et al.*, Kruglyak, L. (2006). Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray, *Science*, **311**(5769), 1932-6.

[Guyon *et al.*, (2002)] Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002). Gene selection for cancer classification using support vector machines, *Machine Learning*, **46**(1-3), 389-422.

[Hu *et al.*, (2011)] Hu, X., Li, R., Meng, J., Xiong, H., Xia, J., Li, Z. (2011). Evaluation of the occurrence possibility of SNP in Brassica napus with sliding window features by using RBF networks. *Wuhan University Journal of Natural Sciences*, **16**(1), 73-78.

[Koboldt *et al.*, (2012)] Koboldt, D. C., Zhang, Q., Larson, D. E., *et al.*, Lin, L., *et al.*, Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing, *Genome Research*, **22**(3), 568-76.

[Kong *et al.*, (2007)] Kong, W., Choo, K. W. (2007). Predicting single nucleotide polymorphisms (SNP) from DNA sequence by support vector machine, *Frontiers in Bioscience*, **12**, 1610-1614.

[Jaillon *et al.*, (2007)] Jaillon, O., Aury, J.-M., Noel, B., *et al.*, Choisne, N., *et al.*, Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, *Nature*, **449**(7161), 463-7.

[Loman *et al.*, (2012)] Loman, N. J., Misra, R. V, Dallman, T. J., *et al.*, Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms, *Nature Biotechnology*, **30**(5), 434-9.

[Lacombe *et al.*, (2013)] Lacombe, T., Boursiquot, J.-M., Laucou, V., *et al.*, This, P. (2013). Large-scale parentage analysis in an extended set of grapevine cultivars (*Vitis vinifera* L.), *Theoretical and Applied Genetics*, **126**(2), 401-14.

[Langmead *et al.*, (2012)] Langmead, B., Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2, *Nature Methods*, **9**(4), 357-359.

[Li *et al.*, (2009)] Li, H., Handsaker, B., Wysoker, A., *et al.*, Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, **25**(16), 2078-9.

[Lijavetzky *et al.*, (2007)] Lijavetzky, D., Cabezas, J. A., Ibez, A., *et al.*, Martinez-zapater, J. M. (2007). High throughput SNP discovery and genotyping in grapevine

(*Vitis vinifera* L .) by combining a re-sequencing approach and SNPlex technology, *BMC Genomics*, **11**, 1-11.

[Loong *et al.*, (2003)] Loong, T. (2003). Clinical review Understanding sensitivity and specificity with the right, *British Medical Journal*, **327**, 716-9.

[Mammadov *et al.*, (2012)] Mammadov, J., Aggarwal, R., Buyyarapu, R., Kumpatla, S. (2012). SNP markers and their impact on plant breeding, *International Journal of Plant Genomics*, **2012**, 1-11.

[Marth *et al.*, (1999)] Marth, G. T., Korf, I., Yandell, M. D., *et al.*, Gish, W. R. (1999). A general approach to single-nucleotide polymorphism discovery, *Nature Genetics*, **23**(4), 452-6.

[Matukumalli *et al.*, (2006)] Matukumalli, L. K., Grefenstette, J. J., Hyten, D. L., *et al.*, Van Tassell, C. P. (2006). Application of machine learning in SNP discovery, *BMC Bioinformatics*, **7**, 4.

[Mckenna *et al.*, (2010)] Mckenna, A., Hanna, M., Banks, E., *et al.*, Depristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Research*, **20**, 1297-1303.

[Mewes *et al.*, (2011)] Mewes, H. W., Albermann, K., Bahr, M., *et al.*, Zollner, A. (1997). Erratum: Overview of the yeast genome, *Nature*, **387**(6634), 737.

[Minoche *et al.*, (2011)] Minoche, A. E., Dohm, J. C., Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems, *Genome Biology*, **12**(11), R112.

[Nakamura *et al.*, (2011)] Nakamura, K., Oshima, T., Morimoto, T., *et al.*, Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers, *Nucleic Acids Research*, **39**(13), e90.

[Nielsen *et al.*, (2011)] Nielsen, R., Paul, J. S., Albrechtsen, A., Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data, *Nature Reviews Genetics*, **12**(6), 443-51.

[O'Fallon *et al.*, (2013)] O'Fallon, B. D., Wooderchak-Donahue, W., Crockett, D. K. (2013). A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data, *Bioinformatics*, **29**(11):1361-6.

[Pindo *et al.*, (2008)] Pindo, M., Vezzulli, S., Coppola, G., *et al.*, Troggio, M. (2008). SNP high-throughput screening in grapevine using the SNPlex$^{\mathrm{TM}}$ genotyping system, *BMC Plant Biology*, **8**, 12.

[Reumers *et al.*, (2012)] Reumers, J., De Rijk, P., Zhao, H., *et al.*, Del-Favero, J. (2012). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing, *Nature Biotechnology*, **30**(1), 61-8.

[Syvänen *et al.*, (2001)] Syvänen, A. C. (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms, *Nature reviews Genetics*, **2**(12):930-42.

[Taub *et al.*, (2010)] Taub, M. A, Corrada Bravo, H., Irizarry, R. A. (2010). Overcoming bias and systematic errors in next generation sequencing data, *Genome Medicine*, **2**(12), 87.

[Unneberg *et al.*, (2005)] Unneberg, P., Strömberg, M., Sterky, F. (2005). SNP discovery using advanced algorithms and neural networks, *Bioinformatics*, **21**(10), 2528-30.

[Vapnik *et al.*, (1998)] Vapnik, V.N. (1995). The Nature of Statistical Learning Theory, *Springer-Velag.*

[Velasco *et al.*, (2007)] Velasco, R., Zharkikh, A., Troggio, M., *et al.*, Viola, R. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety, *PloS one*, **2**(12), e1326.

[Vezzulli *et al.*, (2008)] Vezzulli, S., Micheletti, D., Riaz, S., *et al.*, Velasco, R. (2008). A SNP transferability survey within the genus *Vitis*, *BMC Plant Biology*, **8**, 128.

[Wegrzyn *et al.*, (2009)] Wegrzyn, J. L., Lee, J. M., Liechty, J., Neale, D. B. (2009). PineSAP–sequence alignment and SNP identification pipeline, *Bioinformatics*, **25**(19), 2609-10.

[Xiong *et al.*, (2010)] Xiong, H., Hu, X., Xia, J., Li, R., Shi, F., Meng, J., Li, Z. (2010). A hybrid Fisher / SVM method for SNP discovery in Brassica oilseed rape. *Journal of Food, Agriculture & Environment*, **8**(3&4), 705-708.

Chapter 3

Grapevine acidity

3.1 Abstract

Grapevine berry acidity at harvest plays a key role in wine fermentation, affecting the final quality of the product and, thus, its economical return. For this reason, the knowledge of the molecular bases of berry acidity is of great importance. The discovery of a stable natural grapevine lacking in acid content, Gora Chirine, gives the perfect opportunity to study a highly complex trait like acidity. In particular, Gora berry transcriptome has been compared with the one of Sultanine, which is a normal acidity variety, genetically close to Gora. Their transcripts have been reconstructed using the Pinot Noir genome sequence as reference, identifying 29,903 and 31,503 transcripts in Gora and in Sultanine, respectively. Those transcripts that found a correlation in the Cribi V2 reference annotation have been classified with gene ontology terms, allowing to detect the most likely involved into the acid transport and compartmentalization. The RNAs alignments to the reference genome also enabled the call of genetic variants, of which 225,864 SNPs in Gora and 188,781 SNPs in Sultanine were assessed as true variants by VerySNP. The overlap of those two SNP groups highlighted the differences between Gora and Sultanine: 84,359 in Gora and 47,276 in Sultanine were recognized as unique SNPs of each particular genotype. Considering the transcripts found in GO categories of interest for the acidity trait and the transcripts showing at least one SNP specific of Gora genotype, we calculated the number of those SNPs generating a non-synonymous mutation as 81. This set represents a valuable list of candidate transcritps potentially related to grapevine berry acidity.

3.2 Introduction

3.2.1 Organic acids in fruit-crops

Organic acids support numerous and varied aspects of cellular metabolism in all plants. They are among the main determinants of the organoleptic quality of fleshy fruits and their products, but the type of organic acid found, and the levels to which they accumulate are extremely variable between species, developmental stages and tissue types. Currently, an insufficient understanding of the heterogeneous and complex pathways through which the principal organic acids are synthesized, degraded and regulated, prevents targeted genetic manipulation aimed at modifying fruit acid metabolism in response to environmental conditions (Boudehri *et al.*, 2009). Acidity is of great interest in agriculture due to its strong influence to the harvested date in crops, primarily in those fruits requiring further processing, like wine grapes fermentation. The balance of acids in wine grape must (juice) is central for supporting desirable growth (and preventing undesirable growth) of microorganisms responsible for wine fermentation. Acids concentration can also affect final wine characteristics through involvement in secondary processes such as carbonic maceration and malolactic fermentation, and can even alter the growth capabilities of malolactic bacteria (Kunkee, 1991).

In both climacteric and a non-climateric fruits malate is one of the most prevalent acids, followed by citric and tartaric acid, which contribute to the total cell acidity. Malate is an important participant in numerous cellular functions, from controlling stomatal aperture, improving plant nutrition, and increasing resistance to heavy metal toxicity (Fernie and Martinoia, 2009; Schulze *et al.*, 2002), to other processes more intricately linked with metabolic pathways. The non-climacteric fruits of *Vitis vinifera* (grape) do not contain large amounts of citrate, and the large quantity of tartrate presents in the fruit is not used in primary metabolic pathways. Therefore malate is the only high-proportion organic acid that is actively metabolized throughout ripening of grapes (Sweetman *et al.*, 2009).

Malate is thought to be synthesized during the green stage of fruit growth, through the metabolism of assimilates translocated from leaf tissues, as well as photosynthetic activity within the fruit itself. Just before veraison, or at the inception of fruit ripening, malate accumulation switches to malate degradation and the sugar synthesis begins (Ruffner and Hawker, 1977). In post-veraison fruit, malate is liberated from the vacuole and becomes available for catabolism through various avenues, including the TCA cycle and respiration, gluconeogenesis, amino acid interconversions, ethanol fermentation, and the production of complex secondary compounds such as anthocyanins and flavonols (Famiani *et al.*, 2000; Farineau and Laval-Martin, 1977; Ruffner, 1982; Ruffner and Kliewer, 1975). With the accumulation of sugars and inhibition of glycolysis in ripening grapes (Ruffner and Hawker, 1977), malate is likely a vital source of carbon for these pathways. Once grapes reach veraison, sugar metabolism begins to support hexose accumulation and synthesis rather than catabolism, through regulation of key enzymes of the glycolytic and gluconeogenic pathways (Ruffner and Hawker, 1977). Therefore, at this stage, sugars relinquish the role of major carbon source for energy metabolism and biosynthesis. Malate released from the vacuole during ripening has the potential to fulfill this function, and can do so through involvement in gluconeogenesis, respiration (aerobic and anaerobic), and biosynthesis of secondary compounds.

Vacuolar transporters play a critical role in the switch from malate accumulation to degradation in grape berries, as the acid must be released from the vacuole before it can be metabolized. This involves activities of anion transporters that allow passage of malate through the tonoplast, as well as proton pumps that use the hydrolysis of high energy molecules (ATP and PPi) to drive the import of protons into the vacuole. The latter create an proton gradient that enables malate to be transported into the vacuole against its own concentration (Luttge and Ratajczak, 1997). Several vacuolar dicarboxylate channels have been identified in plants (Emmerlich *et al.*, 2003; Hafke *et al.*, 2003; Kovermann *et al.*, 2007). However the regulation of vacuolar pH does not only relies on primary pumps and anion transporters. By example, H^+/K^+ exchangers convert the proton gradient in a potassium gradient. A complete description of membrane transport is largely beyond the scope of this thesis, but its must be understood that all process affecting the energy balance of the cell, the primary pumps, the secondary transport of most solutes at the tonoplast, will finally affect the vacuolar pH and the concentration of all solutes inside the vacuole.

More recently, Aprile *et al.* (2011), pointed out the knock-out of the *Arabidopsis* H^+-ATPase proton pump *AHA10* citrus homolog and the *Petunia* H^+-ATPase proton pump *Ph*PH5 citrus homolog, both targeted to the vacuolar membrane (Verweij *et al.*, 2008), as responsible for the sweet mutation in Faris lemon variety.

In some fruits, particularly grape, the exposure of the ripening fruit to warmer climatic conditions leads to lower levels of malate at harvest (Lakso and Kliewer, 1978; Ruffner *et al.*, 1976). The temperature-sensitivity of fruit malate degradation may be influenced by activities of enzymes involved in pathways such as the TCA cycle and respiration, ethanol fermentation and gluconeogenesis (Hawker, 1969; Lakso and Kliewer, 1975; Romieu *et al.*, 1992; Taureilles-Saurel *et al.*, 1995).

Most data available on the inheritance of organic acids on fruit trees deals with ripe stage as a major target for breeding. Like grapevine, Apple acidity is at first determined by malic acid. Kenis *et al.* (2008), detected a major year-stable QTL accounting for 20-34% of the total variance on LG16, on Telamon and Braeburn, consistent with that observed on Fiesta (Leibhard *et al.*, 2003).

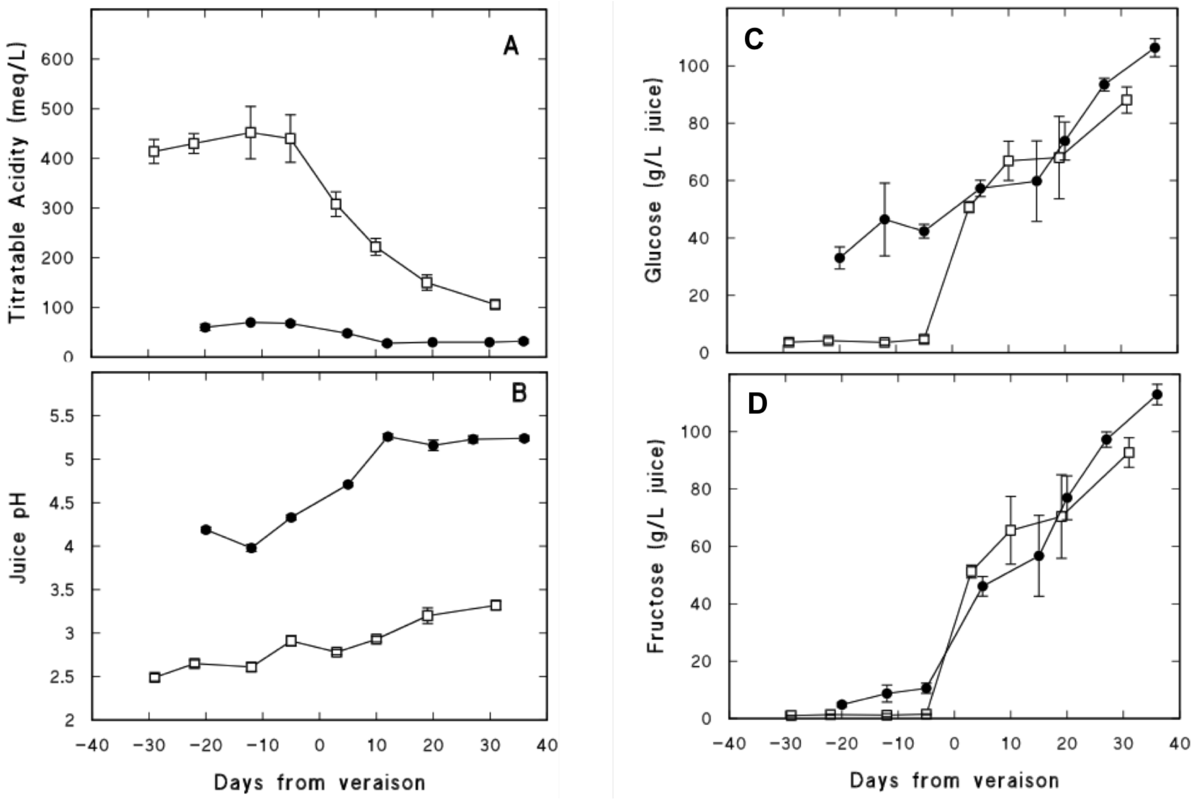
In peaches (*Prunus persica*), the low acidity phenotype depends on a dominant D allele, localized at the proximal end of LG 5, based on segregation studies of 1,718 individuals resulting from a F2 progeny (Boudehri *et al.*, 2009). Putative transcripts in this small region can also be browsed at http://www.rosaceae.org/gb/gbrowse/malus_x_domestica/ and would noticeably include a putative K^+/H^+ symporter as the most probable target for mutation.

3.2.2 Gora and Sultanine

We compared two grapevine varieties, Gora Chirine and Sultanine. These two varieties present a really similar genetic background, but Gora shows an exceptionally low acidity content, while Sultanine has standard acidity. They have also other differences, such as berry size (larger berries in Gora), presence of seeds (two-three seeds per berry in Gora and no seed in Sultanine) and flowering activity. Gora and Sultanine are genetically very close (as demonstrated by AFLP analysis and SSR profiles), which should exclude sexual recombination events between their two genotypes.

Goras phenotype has been studied by Diakou *et al.* (1997 and 2000) and they proved Gora has lower levels of all organic acids (malic, tartaric and citric acid), as shown in Figure 3.1. They also investigate the ability to synthesize and degrade malic acid in Gora, discovering that malic acid is synthesized in Gora cytosol and quickly degraded. Furthermore, the enzyme PEPC, a key enzyme in malate synthesis seems not to be responsible for the low acidity level in Gora since its activity was found to be higher than in normal acidity berries (Diakou *et al.*, 2000). Interestingly, the same authors were able to show that the vacuolar pH of the low acidity and normal acidity varieties was similar (between 2.7 to 3.0) unlike the juice pH (vacuolar + cytosolic) which was much higher in Gora (around 4.3 instead of 3.0). Gora presents an higher glucose level already at berry green stage (Figure 3.1), suggesting some aspects of ripening stage are already active in green stage for sugar-acid metabolism, but not for cell wall and other metabolisms.

Figure 3.1: Titratable acidity (A) and pH (B) of grape berries of cvs Cabernet Sauvignon (white squares) and Gora Chirine (black circles) sampled from before veraison to harvest. Concentration of glucose (C) and fructose (D) in the juice of grape berries of cvs Cabernet Sauvignon (white squares) and Gora Chirine (black circles) sampled from before veraison to harvest (Diakou *et al.*, 1997).



All these observations move the attention to the vacuolar storage of protons and acids. Hence, the best candidate genes for the acidless mutation are transporter proteins located into the tonoplast membrane, such as H^+-ATPase (both vacuolar and plasmic forms), malate transporters (ALMT9, TDT), sugar transporters (Glucose/H^+ antiporter) and H^+/K^+ antiporter. As a matter of fact, all transport impacting the osmotic or electric components of the pmf (proton motive force) would possibly impact vacuolar pH. As reported by Aprile *et al.* (2011) an H^+-ATPase proton pump is responsible for acidless mutant in lemon. On the other hand, Bai *et al.* (2012) found an aluminium-activated malate transporter-like that determine low acidity trait in apple. Shimada *et al.* (2006) investigated the induction of a citrate/H^+ symporter expression when oranges loose acidity along ripening. In sweet melon, Cohen *et al.* (2014) discovered a 12-bp insertion into the *PH* gene sequence, coding for a H^+ transporter of the endoplasmic reticulum, which affects the protein structure by extending or shifting one of the transmembrane domains to include the duplicated amino acids. The malfunctioning of the mutated transporter is the reason of low-acid melon (*Cucumis melo*) phenotype.

3.3 Methods

3.3.1 Samples and acid/sugar content analyses

Gora and Sultanine triplicates were sampled on 4th July 2012 in the experimental vineyard of INRA-Supagro in Montpellier at 0h, 6h, 12h, and 18h, in order to address all genes expressed within one nycthemeral sample (Rienth *et al.*, 2014). Berries were separated from the cluster, cutted in halves with a scalpel and eventually deseeded (Gora) before freezing in liquid nitrogen. The process was conducted sequentially on individual fruits in order that fixation occurred less than one minute following separation from the cluster. Berries were then reduced to a fine powder in liquid nitrogen. Aliquots of the powder were either analyzed for sugar and acids, or mixed before RNA extraction. The similar berry weight between Gora berries and Sultanine berries warrants that RNA-seq results will not be affected by different skin to flesh ratios (Table 3.1).

Table 3.1: Gora and Sultanine berries have been sampled and their acid/sugar content analyzed. The following table report differences and analogies in berry composition calculated on 12 samples per cultivar. In the last raw we refer to the total content of all the previous mentioned compounds: glucose (G), fructose (F), malate (M) and tartrate (T).

	Gora	Sultanine
Flesh and skin FW (g)	0.50 ± 0.12	0.44 ± 0.07
pH	4.17	2.58
Malate (mM)	7 ± 2	148 ± 10
Tartrate (mM)	31 ± 2	109 ± 05
Glucose (mM)	233 ± 20	69 ± 8
Fructose (mM)	47 ± 08	23 ± 2
G+F+M+T (mM)	317 ± 27	348 ± 20

3.3.2 Preparation of RNA samples and extraction

RNA-seq data were produced via an external service (Genopole, Toulouse) following the the protocols reported below. Samples were grinded in liquid nitrogen and the total cellular RNA was extracted using a Spectrum Plant Total RNA kit (Sigma, Inc., USA) with a DNAse treatment. RNA concentration was first measured using a NanoDrop ND-1000 Spectrophotometer then with the Quant-iT$^{\text{TM}}$RiboGreen®(Invitrogen, USA) protocol on a Tecan Genius spectrofluorimeter. RNA quality was assessed by running 1 μL of each RNA sample on RNA 6000 Pico chip on a Bioanalyzer 2100 (Agilent Technologies, Inc., USA). Samples with an RNA Integrity Number (RIN) value greater than eight were deemed acceptable according to the Illumina TruSeq RNA protocol.

The TruSeq RNA sample Preparation v2 kit (Illumina Inc., USA) was used according to the manufacturer's protocol with the following modifications. In brief, poly-A containing mRNA molecules were purified from 2 μg total RNA using poly-T oligo attached magnetic beads. The purified mRNA was fragmented by addition of the fragmentation buffer and was heated at 94°C in a thermocycler for 4 min. The fragmentation time of 4 min was used to yield library fragments of 250-500 bp. First strand cDNA was synthesized using random primers to eliminate the general bias towards 3' end of the transcript. Second strand cDNA synthesis, end repair, A-tailing, and adapter ligation was done in accordance with the manufacturer supplied protocols. Purified cDNA templates were enriched by 15 cycles of PCR for 10 s at 98°C, 30 s at 65°C, and 30 s at 72°C using PE1.0 and PE2.0 primers and with Phusion DNA polymerase (NEB, USA). Each indexed cDNA library was verified and quantified using a DNA 100 Chip on a Bioanalyzer 2100 then equally mixed by ten (from different samples). The final library was then quantified by real time PCR with the KAPA Library Quantification Kit for Illumina Sequencing Platforms (Kapa Biosystems Ltd, SA) adjusted to 10 nM in water and provided to the Get-PlaGe core facility (GenoToul platform, INRA Toulouse, France http://www.genotoul.fr) for sequencing.

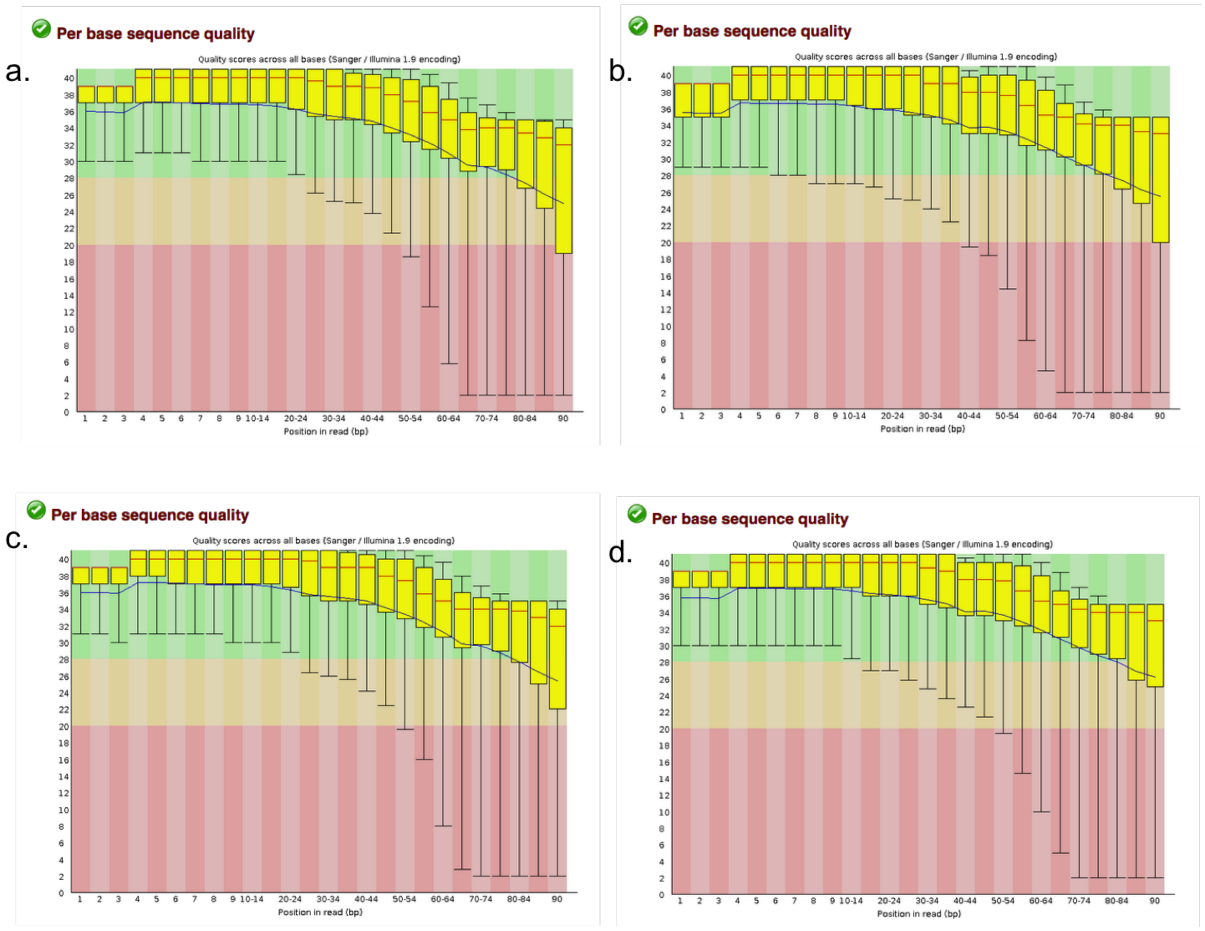
Final mixed cDNA library was sequenced using the Illumina mRNA-Seq, paired-end protocol on a HiSeq2000 sequencer, for 2 × 100 cycles. Library was diluted to 2 nM with NaOH and 2.5 μL transferred into 497.5 μL HT1 to give a final concentration of 10 pM. 120 μL was then transferred into a 200 μL strip tube and placed on ice before loading onto the cBot, mixed library, from 10 individual indexed libraries, being run on a single lane. Flow cell was clustered using TruSeq PE Cluster Kit v3, following the Illumina PE_Amp_Lin_Block_V8.0 recipe. Following the clustering procedure, the flow cell was loaded onto the Illumina HiSeq 2000 instrument following the manufacturer's instructions. The sequencing chemistry used was v3 (FC-401-3001, TruSeq SBS Kit) with the 2 × 100 cycles, paired-end, indexed protocol. Image analyses and base calling were performed using the HiSeq Control Software (HCS 1.5.15) and Real-Time Analysis component (RTA 1.13.48). Demultiplexing was performed using CASAVA 1.8.1 (Illumina) to produce paired sequence files containing reads for each sample in Illumina FASTQ format.

3.3.3 Transcript reconstruction and GO annotation

Gora and Sultanine have been sampled just before the veraison, the stage at which green berries start to become colored. The RNA as been extracted from a pool of berries sampled over a week before veraison and then sequenced. Transcriptomic reads quality was manually inspected using FastQC software (v0.11.2; written in java and available at http://www.bioinformatics.babraham.ac.uk/ projects/fastqc), which showed a quite high level of duplication and a bias at the 5'-end, the latter due to the preferential amplification of GC-rich regions, normally happening when the random priming technique has been used on the samples to reverse transcribe the RNA samples (Benjamini and Speed, 2012). The GC-bias mostly affects the gene expression calculation, which was not the objective of our experiment, and thereby there was no need to fix it. Afterwards, the reads have been cleaned and trimmed using Prinseqlite (http://prinseq.sourceforge.net) as proven in Figure 3.2, which shows the per base quality of both Gora and Sultanine right and left reads after the cleaning and trimming. Gora and Sultanine good quality reads have been aligned to the Pinot Noir reference genome (Jaillon *et al.*, 2007) with TopHat (v2.0.11; Kim *et al.*, 2013) setting the standard deviation for the distribution on inner distances between mate pairs to 100 bp and decreasing the default minimum intron length to 25 bp. The alignment successfully showed a percentage of reads mapped in proper pairs equal to 78% for Gora and 79.86% for Sultanine.

The transcript reconstruction into Gora and Sultanine RNA-seq alignments have been performed by Cufflinks (v2.2.1; Trapnell *et al.*, 2012) and their predicted transcripts have been labeled either as overlapping between the two grape varieties or as uniquely present in one of them; even though, in this preliminary study, we will always consider all Gora transcripts. In order to avoid false positive transcripts and to properly consider only the regions that are actually transcribed, we analyzed the total depth of the RNA reads coverage distribution along the reference and established a minimum value above which a certain transcript can be considered as expressed. The exact value was calculated with a parametric method, the negative binomial distribution, which provides the maximum number of reads that do not make a certain region transcribed by measuring the number of negative attempts to have a transcript expressed while adding reads. The negative binomial distribution is often used when the variance is much higher than the average, which is a common case in RNA-seq reads alignments. We applied the probability to have reads by chance aligned under a certain transcript as less than or equal to 5%. Given this assumption we took into account only the putative transcripts with average depth of coverage higher than 27 for Gora and higher than 31 for Sultanine (Anders and Huber, 2010).

Figure 3.2: The FastQC software per base sequence quality: the already cleaned and trimmed Gora left (a) and right (b) reads of the RNA-seq paired ends; the already cleaned and trimmed Sultanine left (c) and right (d) reads of RNA-seq paired ends.



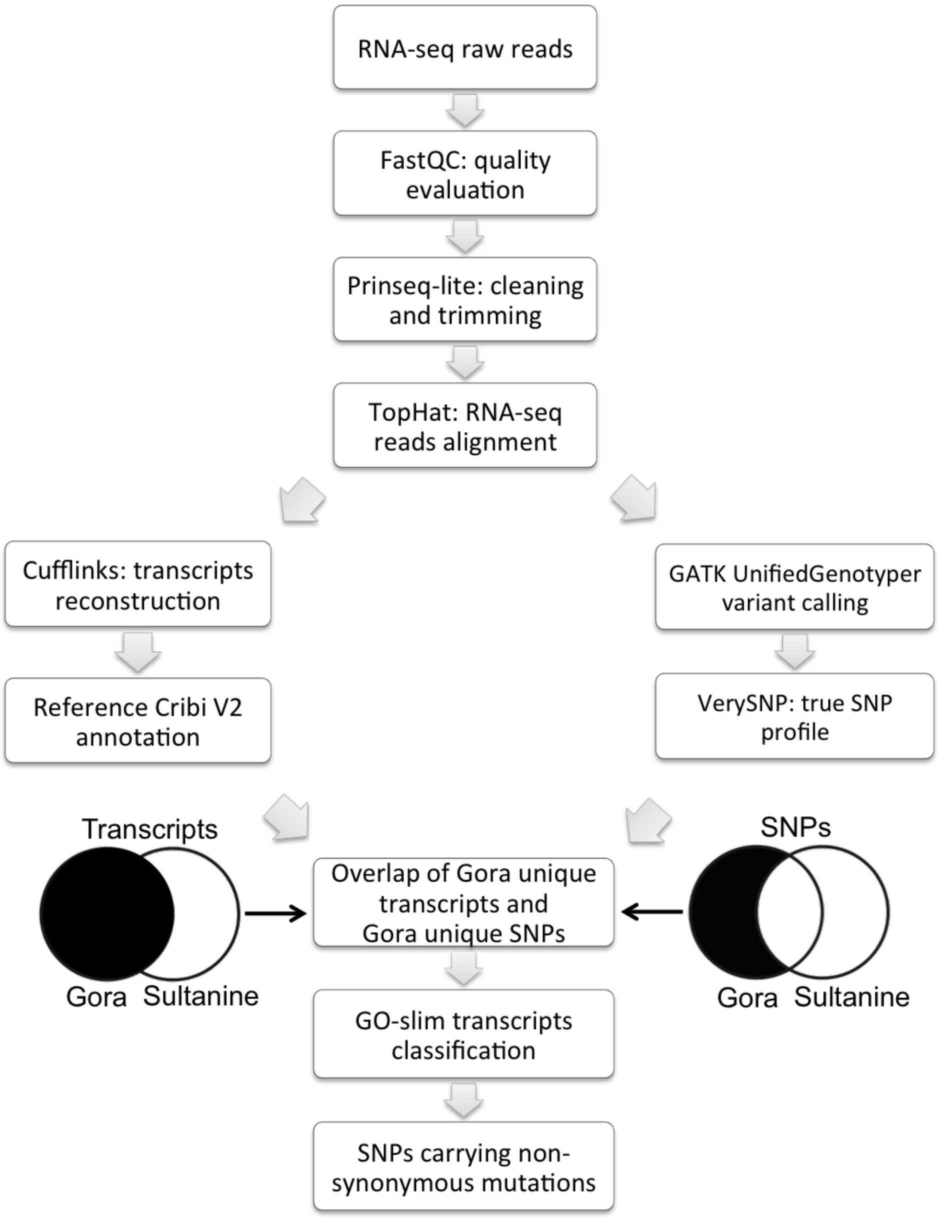
In addition, Gora putative transcripts profiles have been compared to the Pinot Noir reference Cribi V2 gene prediction (http://genomes.cribi.unipd.it/DATA/V2) to better identify which V2 genes belonged to Gora and Sultanine transcriptome and to compare them in terms of functional annotation. Indeed, Gora's veraison transcripts which have found a correspondence in the V2 reference gene prediction, have been classified in the general categories provided by the Gene Ontology (GO), which help to understand the biological role played by the expressed transcripts (Table 3.5, 3.6 and 3.7). We used the Plant GO slim; GO slims are shorter versions of GO ontologies, containing only a particular subset of the terms in the whole GO and they may be build around particular areas of ontologies or specific to species, like plants.

3.3.4 VerySNP application

While the GATK variant calling (*UnifiedGenotyper*) was performed on the berry RNA-seq reads alignment of both Gora and Sultanine, the DNA of the two grape varieties were extracted from their leaf, respectively, and both hybridized on Vitis17KSNP chip (GrapeReSeq Consortium https://urgi.versailles.inra.fr/Projects/Grape ReSeq). The Vitis17KSNP chip data analyzed with GenomeStudio Data Analysis Software (Illumina Corp.) have been used as true and false variants to train VerySNP. Out of the 1,994 SNPs called by GATK and confirmed by the SNP-chip in Gora, we could count 1,877 as true and 117 as false SNPs. At the same way, in Sultanine we got 2,136 SNPs called by GATK and addressed by the SNP-chip, of which 1,999 resulted as true and 137 as false calls. Known true and false SNPs were balanced to the least numerous class (always the false set in our case) and used to train VerySNP (see Chapter 2 for more details). After the training, the 10-fold validation showed an accuracy average equal to 81.9% in Gora and to 71.3% in Sultanine, while the average precision was 81.8% and 78.4%, respectively. Among the 10 models proposed by the 10-fold validation, VerySNP defined the best one to set apart true and false SNPs in each cultivar by calculating the Matthew Correlation Coefficient (MCC), which resulted of 0.91 in Gora's and 0.70 in Sultanine's training set validation. That SVM model was applied to the whole GATK variant call to recognize which SNPs were actually true in Gora as well as in Sultanine, respectively.

Finally, we used the Bedtools function *intersect* (Quinlan and Hall, 2010) to cross the SNP profiles with the transcripts prediction and be able to see the SNP distribution in the transcriptome. Thanks to a short python script we edited, each SNP caught in the cross was characterized either as missense, nonsense or synonymous, knowing the reference sequence and the alternative allele showed in Gora/Sultanine reads.

Figure 3.3: The comparative analysis between Gora and Sultanine started with the RNA-seq raw reads and followed all the steps to the alignment for both cultivars. Afterwords, the alignment was the input for two different analysis: the transcript reconstruction through Cufflinks, which was then crossed with the reference Cribi V2 annotation, and the variant calling through GATK UnifiedGenotyper, which produced a list of putative SNPs classified in true and false calls by VerySNP. At this point, all Gora transcript and Gora specific SNPs (black sections of the two overlapped circles) were considered in the transcripts GO classification and in the SNP characterization.


RNA-seq raw reads
FastQC: quality evaluation
Prinseq-lite: cleaning and trimming
TopHat: RNA-seq reads alignment
Cufflinks: transcripts reconstruction
Reference Cribi V2 annotation
GATK UnifiedGenotyper variant calling
VerySNP: true SNP profile
Transcripts
Gora
Sultanine
SNPs
Gora
Sultanine
Overlap of Gora unique transcripts and Gora unique SNPs
GO-slim transcripts classification
SNPs carrying non-synonymous mutations

The whole procedure applied to the comparative analysis of Gora and Sultanine, including transcriptome, SNP data and gene ontology, is summarized in Figure 3.3.

3.4 Results

A preliminary observation of acid and glucose contents analyses suggests that Gora is primary unable to withstand a proton gradient and its pH is not below the pK of malate and tartrate. Gora's acids, which can obviously be synthesized (presence of tartrate, Diakou *et al.* (2000) labelled malate and PEPC kinase activity with $^{14}CO_2$), can not be accumulated by protonation (see Martinoia *et al.* (2007) for better explanation of organic acid trapping mechanism). The storage of both acids is affected, not only the one of malate. The loss in osmotic pressure is compensated mostly by glucose. Very rapidly, when the fruit starts to ripen, the glucose to fructose ratio reaches 1 in Gora and Sultanine as well (not shown). Then, the strong glucose to fructose ratio in Gora confirms the fruit is at green stage, together with the green color and the hard texture of the fruit, in spite of an acid composition even lower than never observed in ripe berries (0.5 M glucose, 0.5 M fructose, more than 50 mM tartrate and malate possibly lower, depending on environmental conditions). By many aspects, Gora looks like a ripe vacuole, in a green cytoplasm.

The whole set of Gora and Sultanine RNA-seq reads have been separately aligned to the Pinot Noir reference genome and their total percentage of alignment was equal to 78.00% in Gora and to 84.04% in Sultanine. Paired-end reads alignment is performed by placing the right read at a known distance from its left read, but reasons like the large genetic distance between the reference sequence and the mapped individual, or the advent of large insertions and deletion, may prevent the read mapping in proper pair. Although Pinot Noir is not genetically very close to Gora and Sultanine, the percentage of reads mapped in proper pair was calculated to be 72.59% in Gora and 79.86% in Sultanine; in both cases not far from the whole amount of mapped reads. Around 5% of the mapped reads do not match in proper pairs in both Gora and Sultanine alignments. This may be indicative of noticeable structural variations (i.e. large INDELs) between reference genome Pinot Noir 40024 and Sultanine related cultivars (Di Genova *et al.*, 2014). The unmapped reads, likely matching regions specific of Gora and Sultanine genomes, concerned the 15.96% and the 22% of Gora and Sultanine total raw reads set, respectively (Table 3.2).

Table 3.2: Of the whole set of RNA-seq reads (1) the number of total mapped reads was found to be the 78% in Gora and the 84.04% in Sultanine (2), while counting the reads mapped in a proper pair only, the rates go to 72.59% in Gora and to 79.86% in Sultanine (3). The remaining reads fraction amounted to 15.96% in Gora and 22% in Sultanine and is considered as unmapped (4). The Cufflinks software predicted a total number of transcripts (5) that was then analyzed. The minimum value in depth of reads coverage for which a certain transcript was considered as expressed was calculated (6).

	Number of	Gora	Sultanine
1	Total raw paired-ends	40,339,336	43,733,962
2	Total mapped reads	31,677,180	36,752,428
3	Properly mapped reads	29,284,248	34,925,020
4	Unmapped reads	8,662,156	6,981,534
5	Cufflinks transcripts	29,903	31,503
6	Minimum coverage considered	27	31

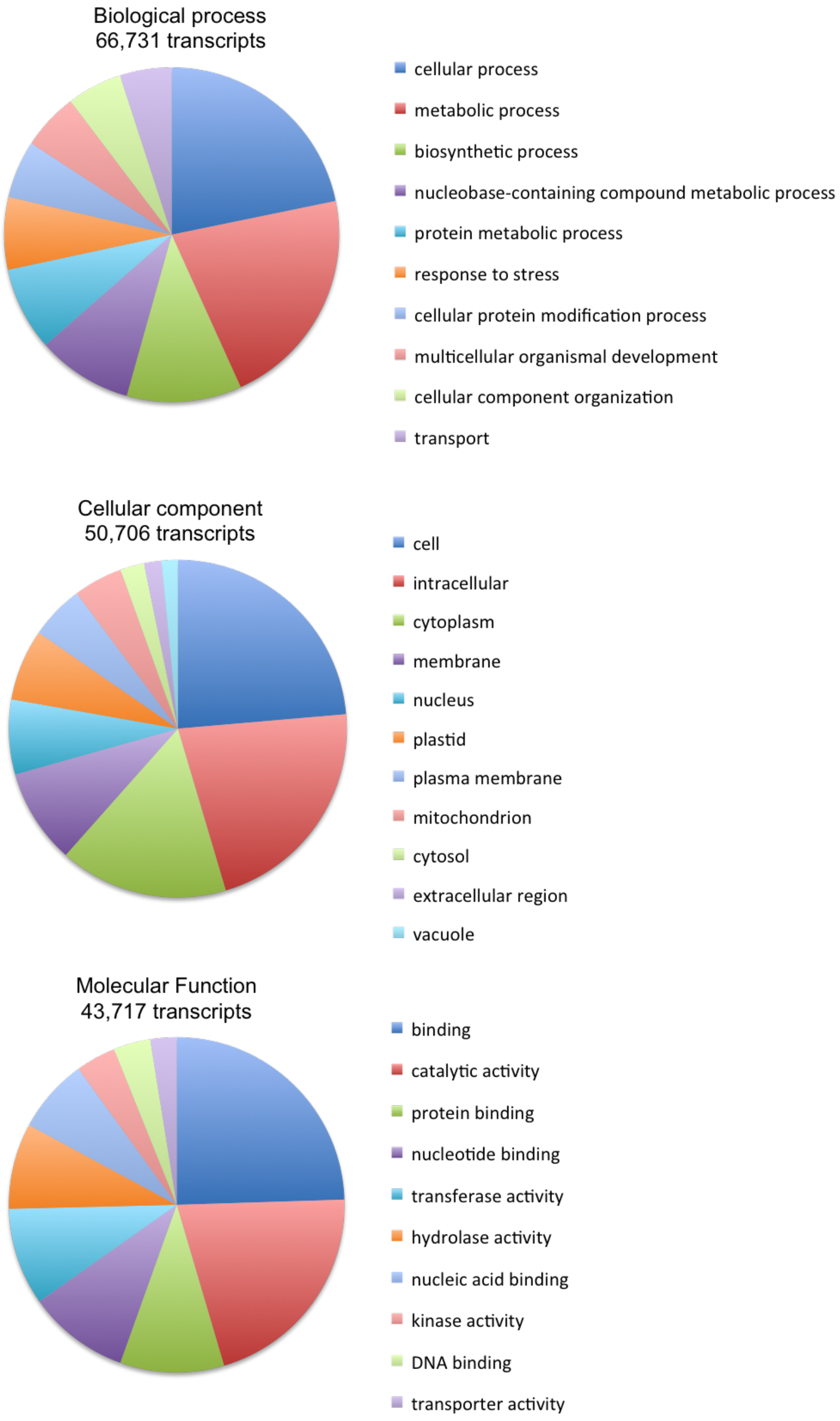
The transcripts reconstructed by Cufflinks from the RNA-seq reads alignment on Pinot Noir were about 29,903 in Gora and 31,503 in Sultanine, of which we considered only the ones with a average depth of coverage of 27 in Gora and 31 in Sultanine (Table 3.2). We mapped those transcripts on the Cribi V2 annotation of Pinot Noir reference and detected which Pinot Noir mRNA sequences corresponded to Gora and Sultanine transcripts. We have found a significant match of the Gora transcripts with 12,811 Cribi V2 coding sequences. The considerable differences between the transcripts showing a correspondence in the Cribi V2 annotation and the original number of transcripts reconstructed by Cufflinks can be explained mainly as inaccuracy of Cufflinks working without the reference annotation and partially as due to the genetic distance between Gora/Sultanine and Pinot Noir. In the V2 mapped Gora transcripts we could find 7,238 true and unique SNPs. Indeed, the alignment of Gora and Sultanine RNA-seq reads to the reference genome was also required to call the genetic variations occurring in such sequences. At this aim we used *UnifiedGenotyper*, the variant calling function of the Genome Analysis Toolkit (GATK), which is a software package developed at the Broad Institute to analyze high-throughput sequencing data. This software provided 317,990 variants in Gora and 357,186 in Sultanine transcriptome, but only 225,864 and 188,781 were classified as true variants by VerySNP in Gora and Sultanine, respectively. Comparing Gora and Sultanine true variant sets we gathered the SNPs shown uniquely by one variety or the other, which amounted to 84,359 in Gora and 47,276 in Sultanine (Table 3.3). The number of polymorphisms which are divergent between Gora and Sultanine is lower than the number of SNPs retrieved as different from Pinot Noir, but obviously excludes that Gora and Sultanine can be considered as clones. This result does not meet our expectations.

Table 3.3: Gora and Sultanine RNA-seq reads have been aligned to Pinot Noir 40024 reference genome to generate the variant calling through GATK *UnifiedGenotyper*, which provided hundreds of thousands of variants (1). Only a fraction of variants have been classified by VerySNP as true variants (2). Looking for differences between the two cultivars, the study was focused on the investigation of the variants uniquely showed either by Gora or Sultanine (3). After mapping the transcripts into the reference gene prediction, the coding sequences (CDSs) where the variants fell were retrieved (4) and characterized by the kind of mutation they may originate: missense (5), synonymous (6) and nonsense (7).

	Number of	Gora	Sultanine
1	GATK variant calls	317,990	357,186
2	True variants by VerySNP	225,864	188,781
3	Unique variants	84,359	47,276
4	Unique Variants in CDSs	12,811	6,082
5	Missense variants in CDSs	6,701	3,295
6	Synonymous variants in CDSs	6,034	2,736
7	Nonsense variants in CDSs	75	50

Out of the 84,359 Gora true SNPs solely present in all Gora's transcripts, 41,440 were found in introns, 18,126 in UTRs, 17,555 fell outside gene predictions and 7,238 in CDSs (coding sequences). The SNPs retrieved in CDSs are 7,238 and they are found in 12,811 Gora transcripts, since more than one transcripts can cover the same SNP position. Each transcript carries at least one SNP and each SNP can cause different mutation depending on the transcript frame which is considered. In total, we identified 12,811 putative single nucleotide mutations and, tracking their position into the codons, we characterized the kind of mutation they may originate at the protein level. Equally, the number of Sultanine true SNPs uniquely present in all Sultanine transcripts were 47,276; of which the majority was found in introns, amounting to 28,556 SNPs, while 7,378 were mapped in UTRs, 7,902 fell outside gene predictions and only 3,440 in CDSs.

Figure 3.4: All Gora transcripts crossed with Gora unique SNPs occurring in CDSs have been classified in the three main GO-slim categories: biological process (66,731 transcripts), cellular component (50,706 transcripts) and molecule function (43,717 transcripts). Each category includes other more specific classifications, of which we graphically illustrated the widest top-ten.



As already mentioned for Gora, the number of SNPs retrieved in Sultanine is repeated for each transcript covering such position in the genome and for this reason 3,440 SNPs found in CDSs have been counted as 6,082 (Table 3.3).

We compared the amino acid coded by the Pinot Noir reference with the alternative amino acid coded by Gora and Sultanine and, whether the amino acid was found identical, we defined the SNP mutation as synonymous, otherwise as missense mutation, when the reference codon and the alternative codon generated two different amino acids. SNPs resulting in a premature stop codon into the protein sequence, by replacing the original amino acid with a "stop" codon, also referred as nonsense mutation, occur at extremely low frequency. In particular, out of 12,811 Gora's variants in CDSs: 6,701 would cause a missense mutation; 6,034 a synonymous one and only 75 a nonsense mutation. Likewise, out of 6,082 Sultanine's variants in CDSs: 3,295 would cause a missense mutation; 2,736 a synonymous one and only 50 a nonsense mutation. The observed SNP frequency and the type of caused mutation appears in line with several previous reports, where synonymous mutations are quite ordinary and nonsense mutation are rare. The number of missense mutation is, however, elevated but we can not be able to tell whether they change the cell biology or not, until we verify each of them with biological experiments. A further approach would be to check which SNP changes the general properties of amino acids, i.e. the presence of proline stiffens the protein backbone; in other cases it could switch the amino acid polarity or modify their charge, from neutral to positive and negative and viceversa, causing a different protein 3D folding.

Thanks to the match with the Pinot Noir reference, we were able to classify Gora transcripts according to the Gene Ontology vocabulary, which highlighted many significant information. The GO is the only annotation system able to cluster gene functions into more general categories and manage to keep the classification objective, independent from the gene names or the pathways they are involved into. Studying an unknown phenomenon, such as the identification of what caused the low acidity content in Gora, the GO classification is an appropriate way to have a straight view of the results. On the GO slim we re-mapped the Cribi V2 transcripts ID names corresponding to all Gora transcripts predicted by Cufflinks. Figure 3.4 graphically illustrates the widest GO slim categories in which Gora transcripts have been classified, while a full list of GO slim classes ranked by the number of transcripts per class is reported in Tables 3.5, 3.6 and 3.7. Among the several classes found, as suggested by previous literature on acidless mutants, we focussed our attention on the transport class within the biological process category (Table 3.5), which counts up to 1,819 Gora transcripts. Likewise, considering the cellular component category (Table 3.6), the vacuole term seems to be the most interesting to examine, also because it includes a more limited number of Gora transcripts (592). Similarly, the transporter activity mentioned in the molecular function category (Table 3.7) is definitively worth to look into, representing 783 Gora transcripts. Focusing on these selected transcripts we mapped which SNPs were present in those Gora's sequences only and characterized the kind of mutation they may originate. Table 3.4 shows the amount of synonymous, missense and nonsense SNPs found into the transcript sequences classified as involved in the transport process, in the vacuole compartment and in transporter activities. The SNPs causing missense and nonsense mutations into transcripts encoding for transporter proteins as well as located into the vacuole, are the most promising candidates as determinants of the berry acidity content (last row of Table 3.4). The functional validation of the role played by these proteins will need, however, an experimental validation.

Table 3.4: Gora transcripts corresponding to the GO terms most likely involved into the crop acidless mutation have been selected and the Gora specific SNPs shown into those sequences were characterized either as synonymmous, missense or nonsense, depending on which kind of mutation they originate.

GO category	GO term	Go ID	Synonymous	Missense	Nonsense
Biological process	Transport	GO:0006810	879	928	12
Cellular component	Vacuole	GO:0005773	304	280	6
Molecular function	Transporter activity	GO:0005215	391	386	6
SNPs represented by all classes			77	79	2

To complete Gora and Sultanine comparison we characterized the SNPs commonly showed by both cultivars, which means that observing both alignments of Gora and Sultanine to the reference Pinot Noir, there are polymorphism showed in the same chromosome and relative position in both cultivars. Common SNPs amounted to 141,505 and only 91 of those showed a different alternative allele between Gora and Sultanine. When the alternative allele is identical between two varieties, we suppose the event of differentiation of the varieties occurred after those mutation events happened and, thus, they have inherited exactly the same mutations. On the contrary, when mutations happen independently in two separate organisms, it is quite rare that the two individual present the same identical event.

3.5 Discussion

The use of genetic mutants is a very valuable methodology to dissect the genetic determinants of a specific phenotype. The scope of this approach is to associate the mutant phenotype to a definite genotype. The outcome of this genetic analysis is the identification of one or more chromosomal regions responsible for the trait of interest. Further experiments are then needed to pinpoint the single gene, within that region, associated to the mutant phenotype.

In this study we took advantage of the availability of a grapevine cultivar named Gora Chirine showing a mutated phenotype for the pH, sugar, malate and tartrate concentrations in the berry, when compared to the very close relative cultivar Sultanine. Aim of the project was the identification of a small number of single nucleotide polymorphisms in the transcriptome of the two cultivars, potentially linked to the difference in berry acidity. The approach has been a combination of RNA-seq data analysis, gene ontology annotation and SNP detection with the objective to reduce the number of gene candidates from thousands to few dozens. A number amenable to experimental validation.

The first step was the reconstruction of the berry transcripts of Gora and Sultanine starting from RNA-seq data originated from a pool of berries harvested at the peak of acid content, few days before veraison. In order to avoid possible interference with the nycthemeral cycle (Rienth *et al.*, 2014), four triplicate samples were harvested at six hours interval and pooled, before RNA extraction. It was confirmed that both tartaric and malic acids were dramatically reduced in Gora, in green and hard berries, before the onset of ripening. Moreover, the loss of 0.2 M tartrate plus malate was compensated by the accumulation of 0.2 M glucose, thus the osmotic pressure was kept constant.

Around 78% and 84% of Gora and Sultanine reads, respectively, were mapped to the Pinot Noir reference genome. This value is a little bit smaller than the 89% reported previously for Corvina (Venturini *et al.*, 2013) and it might be due to technical reasons as well as to the genetic distance to the reference genome. By using the RNA-seq analysis software Cufflinks, the aligned reads were assembled into 29,904 and 31,503 transcripts in Gora and Sultanine, respectively. A similar transcripts reconstruction based on RNA-seq data was already performed in two other *V. vinifera* cultivars: the Uruguayan Tannat and the Italian Corvina. In the first case a total of 34,680 genes were predicted in RNA isolated from berry skin and seeds harvested at 3 pre-veraison stages (Da Silva *et al.*, 2013), while in the second case a much larger number of genes (40,610) was identified, probably since 45 different samples from different organs and tissues were considered (Venturini *et al.*, 2013).

The observation that a fraction of reads did not map to the reference genome was not unexpected: this accounted for 16% of total reads in Gora and 22% in Sultanine. Unmapped reads correspond to sequences with a number of mismatches to the reference above the fixed threshold calculated by the aligner and, apart from cases of sequencing errors, likely represent transcripts not shared with Pinot Noir nuclear genome, but they could be either part of the chloroplast and mitochondrial genome, or viral RNA, either potentially variety-specific regions of the nuclear DNA. In the case of Tannat 1,873 genes fell in this class (Da Silva *et al.*, 2013), while a smaller number was found in the transcriptome of Corvina (180 private genes) (Venturini *et al.*, 2013). In this study we concentrated our attention on the reads mapping on the Pinot Noir reference genome because this would have helped the comparative analysis of Gora versus Sultanine and would have made gene prediction and annotation more straightforward. Moreover, acidity is a common trait all over *Vitis* cultivars and more likely controlled by conservative genes than by variety-specific sequences.

Indeed, we took advantage of a very recent annotation with Gene Ontology terms of the Pinot Noir gene predictions and based on a massive sequencing of RNAs derived from many different tissues, stress conditions and *Vitis* genotypes (CRIBI V2, Vitulo *et al.*, 2014). In this particular case of interest, the classification of the transcripts into GO categories guided the research towards the putative responsible of the acidless phenotype in Gora, rather than to highlight differences or enrichments of GO classes of the Gora berry transcriptome when compared to normal acidity *Vitis* varieties.

In parallel, the comparative study between Gora and Sultanine focused on the characterization of the single nucleotide polymorphisms in their transcriptome. The working hypothesis was, indeed, that a point mutation in the coding sequence of a gene involved in berry acidity was possibly the cause of the acidless phenotype.

SNP calling performed with the variant calling function of the Genome Analysis Toolkit (GATK) recognized more than 300,000 SNPs in both cultivars, when compared to the Pinot Noir reference genome. These figures were largely diminished following the application of VerySNP to select true SNPs: 225,864 (71% of the total GATK calling) and 188,781 SNPs (53% of the total GATK calling) were considered as true SNPs in Gora and Sultanine, respectively. These results have been a convincing prove of the efficacy of applying VerySNP to transcriptomic data to classify as true and false the SNPs outputted by variant caller algorithms.

Since the comparison of these numbers with those reported in other studies is very difficult, being affected by several parameters (e.g. reads coverage, gene predictions, SNP calling, SNP selection, etc.) and requiring the different samples to be processed and analyzed in parallel within the same experiment, we can only discuss that a high number of single nucleotide polymorphisms was expected due to the high genetic variability within the *Vitis vinifera* species and to the known genetic distance between Pinot Noir (West European group) and Sultanine or Gora (East group) (Bacilieri *et al.*, 2013).

Of the 225,864 and 188,781 SNPs, a fraction of 37% and 25% corresponded to single nucleotide variants was found exclusively in Gora and Sultanine and therefore named *unique variants*. Such an elevated proportion of specific SNPs in Gora and Sultanine was a complete surprise, as it exceeds the expectations for clonal variation (Cabezas *et al.*, 2011). The unique variants have been, then, classified according to the gene location they were found (UTRs, introns and CDSs). Among them, the most relevant are likely those positioned in the CDSs because they can directly affect the final protein product. The analysis showed that 12,811 transcripts in Gora and 6,082 transcripts in Sultanine presented unique variants in the coding regions. Not all the unique mutations might have, however, similar effect on the encoded product.

Missense and nonsense mutations lead either to a change of amino acid or to a premature stop in translation, and those can be very interesting candidates for causing the acidless phenotype. This case is supported by the recent finding in a low acid apple, showing a recessive gene with a premature stop codon in an aluminum activated malate transporter as responsible for the low acidic content (Bai *et al.*, 2012; Khan *et al.*, 2013). Nonsense mutations are clearly the most severe, causing the formation of a truncated product that, in most cases, will not be functional. Instead, the effect on the function of the protein product of missense mutations will be case-specific, being linked to the physico-chemical properties of the changed amino acid and to its role and position into the protein chain, which might affect the enzymatic activity or the three-dimensional structure. Since the analysis of this group of mutations would have required very long time due to its particularities and to the high number (6,701 in Gora and 3,295 in Sultanine), we focused our attention on the transcripts containing premature stops (nonsense mutations), namely 75 in Gora and 50 in Sultanine. To further narrow down the number of candidates, the analysis of the premature stops has been combined to the information arising from the GO annotation of the transcripts where they occurred. The knowledge available for Gora suggested, indeed, that a defect in the vacuolar transport might be the reason of the low acidity of the berry juice. By crossing the transcripts of Gora belonging to the three selected GO classes (GO:0006810, GO:0005773 and GO:0005215) with the presence of at least one nonsense mutation in the CDS, we counted a total of 22 transcripts.

As previously said, the power of this kind of analysis is also the possibility to reduce thousands transcripts involved in the Gora berry maturation, to few tens transcripts with a putative role in the acidless phenotype. Another example, beyond the vacuolar transporters, could be the restriction to candidate genes with a role in 'regulation of gene expression, epigenetic (GO:0040029 in Table 3.5) starting from the 29,903 transcripts found in Gora by Cufflinks, of which 2,409 are Gora transcripts mapping in Cribi V2 annotation, showing only 441 transcripts annotated with the 'regulation of gene expression, epigenetic function and including in their sequence at least one SNP not shared with Sultanine; among them only 204 transcripts carried a SNP originating a non-synonymous mutation.

However, our analysis has some limits. The study we carried out to identify candidate mutations responsible for the acidless phenotype makes sense only if we assume that a point mutation is causing the loss of acidity. Although SNPs are very common genetic variants, they are not the only ones, and it might be a small or large insertion/deletion (INDELs) responsible for the observed phenotype in Gora, instead. In the case of sweet melon it has been discovered, indeed, that a small 12-bp insertion in the *PH* gene, coding for a H^+ transporter of the endoplasmic reticulum, was responsible for the acidless phenotype. Another limitation of our analysis resides in the fact that it was restricted to the encoding fraction of the genome (transcriptome). There are several examples in the literature showing that mutations in intergenic regions, including promoters and other regulatory regions, often have strong phenotypic effects. Interestingly, one of the most well known examples is the grapes color, determined by two *Myb* transcription factor genes, *VvMybA1* and *VvMybA2* regulating anthocyanin biosynthesis. Inactivation of these two functional genes, through the insertion of the *Gret1* retrotransposon in the *VvMybA1* promoter and through a non-synonymous single nucleotide polymorphism present in the *VvMybA2* coding region, gives rise to a white berry phenotype (Kobayashi *et al.*, 2004, 2005; Walker *et al.*, 2007).

Having the genome sequences of Gora and Sultanine will permit to widen the examination for candidate mutations not only because it will largely increase the sequence range of the analysis, but also because the detection of INDELs in transcripts is very difficult due to splicing and alternative splicing events. A de novo assembly of the Sultanine draft genome sequence has been recently published (Di Genova *et al.*, 2014), however at this stage, its quality is far from being comparable to the one of Pinot Noir and its gene annotation is still missing. For these reasons in this thesis we decided to use the Pinot Noir genome as reference sequence. Undoubtedly, it would be worth to align the fraction of unmapped reads we obtained, against the Sultanine genome, to confirm if such reads belong to variety-specific sequences or are still unmapped, maybe because falling into hardly accessible part of the chromosome (centromeres and telomeres). Anyway, if the high fragmentation of the Sultanine *de novo* assembly genome in several contigs prevents the possibility of a successful alignment of genomic reads, the alignment of RNA-seq reads and, the subsequent splicing prediction, are even more sensible to such fragmentation, dropping any chance of success.

3.6 Conclusion

The procedure of combining transcriptome analysis and annotation together with single nucleotide polymorphisms in related grapevine genotypes, as described here, has shown to be quite effective in reducing the number of potential candidates for the trait of interest. In this study starting from more than 80,000 unique single nucleotide polymorphisms of Gora, we have found 75 located in coding regions and causing non-sense mutations. Gene ontology annotation of the transcripts carrying these mutations, has allowed to identify in this group, those most likely linked to the acid metabolism. Of particular interest, appear 22 transcripts located in the vacuole or assigned a transport activity. Although biological validation of the results will confirm the role of these transcripts in determining berry acidity, we can foresee that the approach we have used can be successfully applied to several other studies.

Table 3.5: All Gora transcripts crossed with Gora unique SNPs occurring in CDSs have been classified in GO-slim categories, which labeled 66,731 transcripts as part of the biological process. The following table reports the GO ID and the number of transcripts classified in each GO term involved into the biological process category.

GO ID	GO term name	Count
GO:0009987	cellular process	7,935
GO:0008152	metabolic process	7,827
GO:0009058	biosynthetic process	4,041
GO:0006139	nucleobase-containing compound metabolic process	3,365
GO:0019538	protein metabolic process	2,941
GO:0006950	response to stress	2,553
GO:0006464	cellular protein modification process	2,025
GO:0007275	multicellular organismal development	2,006
GO:0016043	cellular component organization	1,939
GO:0006810	transport	1,819
GO:0009628	response to abiotic stimulus	1,537
GO:0009056	catabolic process	1,501
GO:0007154	cell communication	1,358
GO:0009791	post-embryonic development	1,333
GO:0005975	carbohydrate metabolic process	1,304
GO:0009605	response to external stimulus	1,161
GO:0006629	lipid metabolic process	1,087
GO:0007165	signal transduction	1,062
GO:0006259	DNA metabolic process	996
GO:0009653	anatomical structure morphogenesis	970
GO:0009719	response to endogenous stimulus	879
GO:0009607	response to biotic stimulus	876
GO:0007049	cell cycle	736
GO:0030154	cell differentiation	590
GO:0009908	flower development	581
GO:0040007	growth	462
GO:0040029	regulation of gene expression, epigenetic	441
GO:0000003	reproduction	431
GO:0009790	embryo development	401

GO:0006091	generation of precursor metabolites and energy	359
GO:0016049	cell growth	340
GO:0006412	translation	340
GO:0019748	secondary metabolic process	299
GO:0009856	pollination	250
GO:0008219	cell death	239
GO:0016265	death	239
GO:0009991	response to extracellular stimulus	172
GO:0019725	cellular homeostasis	169
GO:0015979	photosynthesis	162
GO:0009606	tropism	121
GO:0009875	pollen-pistil interaction	119
GO:0007267	cell-cell signaling	87
GO:0009838	abscission	14
GO:0009835	fruit ripening	9
GO:0007610	behavior	8

Table 3.6: All Gora transcripts crossed with Gora unique SNPs occurring in CDSs have been classified in GO-slim categories, which labeled 50,706 transcripts as involved in some cellular components. The following table reports the GO ID and the number of transcripts classified in each GO term of the cellular component category.

GO ID	GO term name	Count
GO:0005623	cell	8,983
GO:0005622	intracellular	8,282
GO:0005737	cytoplasm	6,128
GO:0016020	membrane	3,452
GO:0005634	nucleus	2,724
GO:0009536	plastid	2,598
GO:0005886	plasma membrane	1,966
GO:0005739	mitochondrion	1,783
GO:0005829	cytosol	874
GO:0005576	extracellular region	630
GO:0005773	vacuole	592

GO:0005794	Golgi apparatus	490
GO:0005783	endoplasmic reticulum	438
GO:0030312	external encapsulating structure	374
GO:0005618	cell wall	365
GO:0005856	cytoskeleton	270
GO:0005840	ribosome	261
GO:0009579	thylakoid	249
GO:0005730	nucleolus	233
GO:0005768	endosome	197
GO:0005777	peroxisome	163
GO:0005654	nucleoplasm	142
GO:0005635	nuclear envelope	64
GO:0005615	extracellular space	13
GO:0005578	proteinaceous extracellular matrix	7
GO:0005764	lysosome	2

Table 3.7: All Gora transcripts crossed with Gora unique SNPs occurring in CDSs have been classified in GO-slim categories, which labeled 43,717 transcripts as having a specific molecular function. The following table reports the GO ID and the number of transcripts classified in each GO term called into the molecular function category.

GO ID	GO term name	Count
GO:0005488	binding	7,582
GO:0003824	catalytic activity	6,483
GO:0005515	protein binding	3,098
GO:0000166	nucleotide binding	3,027
GO:0016740	transferase activity	2,905
GO:0016787	hydrolase activity	2,557
GO:0003676	nucleic acid binding	2,219
GO:0016301	kinase activity	1,190
GO:0003677	DNA binding	1,104
GO:0005215	transporter activity	783
GO:0003723	RNA binding	763
GO:0003700	sequence-specific DNA binding transcription factor activity	329

GO:0004871	signal transducer activity	280
GO:0004518	nuclease activity	248
GO:0003682	chromatin binding	202
GO:0030246	carbohydrate binding	197
GO:0005198	structural molecule activity	154
GO:0008289	lipid binding	124
GO:0008135	translation factor activity, nucleic acid binding	107
GO:0030234	enzyme regulator activity	99
GO:0003774	motor activity	98
GO:0004872	receptor activity	86
GO:0019825	oxygen binding	48
GO:0005102	receptor binding	14
GO:0030528	transcription regulator activity	0
GO:0045182	translation regulator activity	0

Bibliography

[Anders and Huber, (2010)] Anders, S., Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**(10), R106.

[Aprile *et al.*, (2011)] Aprile, A., Federici, C., Close, T. J., Bellis, L. De, Cattivelli, L., Roose, M. L. (2011). Expression of the H + -ATPase AHA10 proton pump is associated with citric acid accumulation in lemon juice sac cells. *Functional & Integrative Genomics*, **11**(4), 551-563.

[Bacilieri *et al.*, (2013)] Bacilieri, R., Lacombe, T., Le Cunff, L., Di Vecchi-Staraz, M., Laucou, V., Genna, B., *et al.*, Boursiquot, J.-M. (2013). Genetic structure in cultivated grapevines is linked to geography and human selection. *BMC Plant Biology*, **13**, 25.

[Bai *et al.*, (2012)] Bai, Y., Dougherty, L., Li, M., Fazio, G., Cheng, L., Xu, K. (2012). A natural mutation-led truncation in one of the two aluminum-activated malate transporter-like genes at the Ma locus is associated with low fruit acidity in apple. *Molecular genetics and genomics*, **287**(8), 663-78.

[Benjamini and Speed, (2012)] Benjamini, Y., and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, **40**(10), e72.

[Boudehri *et al.*, (2009)] Boudehri, K., Bendahmane, A., Cardinet, G., Troadec, C., Moing, A., Dirlewanger, E. (2009). Phenotypic and fine genetic characterization of the D locus controlling fruit acidity in peach. *BMC Plant Biology*, **9**, 59.

[Cabezas *et al.*, (2011)] Cabezas, J. A., Ibáñez, J., Lijavetzky, D., Vélez, D., Bravo, G., Rodríguez, V., *et al.*, Martinez-Zapater, J. M. (2011). A 48 SNP set for grapevine cultivar identification. *BMC Plant Biology*, **11**(1), 153.

[Cohen *et al.*, (2014)] Cohen, S., Itkin, M., Yeselson, Y., Tzuri, G., Portnoy, V., Harel-Baja, R., *et al.*, Schaffer, A. A. (2014). The PH gene determines fruit acidity and contributes to the evolution of sweet melons. *Nature Communications*, **5**(4026), 1-9.

[Da Silva *et al.*, 2013] Da Silva, C., Zamperin, G., Ferrarini, A., Minio, A., Dal Molin, A., Venturini, L., *et al.*, Delledonne, M. (2013). The high polyphenol content of grapevine cultivar tannat berries is conferred primarily by genes that are not shared with the reference genome. *Plant Cell*, **25**, 4777-4788.

[Di Genova *et al.*, (2014)] Di Genova, A., Almeida, A. M., Muñoz-Espinoza, C., Vizoso, P., Travisany, D., Moraga, C., *et al*, Maass, A. (2014). Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biology*, **14**, 7.

[Diakou *et al.*, (1997)] Diakou, P., Moing, A., Svanella, L., Ollat, N., Rolin, D. B., Gaudillere, M., Gaudillere, J. P. (1997). Biochemical comparison of two grape varieties differing in juice acidity. *Australian Journal of Grape and Wine Research*, **3**, 1-10.

[Diakou *et al.*, (2000)] Diakou, P., Svanella, L., Raymond, P., Gaudillre, J.-P., Moing, A. (2000). Phosphoenolpyruvate carboxylase during grape berry development: protein level, enzyme activity and regulation. *Australian Journal of Plant Physiology*, **27**, 221-229.

[Emmerlich *et al.*, 2003] Emmerlich, V., Linka, N., Reinhold, T., Hurth, M. A., Traub, M., Martinoia, E., Neuhaus, H. E. (2003). The plant homolog to the human sodium/dicarboxylic contransporter is the vacuolar malate carrier. *Proceedings of the National Academy of Sciences, USA*, **100**, 11122-11126.

[Famiani *et al.*, (2000)] Famiani, F., Walker, R. P., Técsi, L., Chen, Z., Proietti, P., Leegood, R. C. (2000). An immunohistochemical study of the compartmentation of metabolism during the development of grape (*Vitis vinifera* L .) berries. *Journal of Experimental Botany*, **51**(345), 675-683.

[Farineau and Laval-Martin, 1977] Farineau, J., Laval-Martin, D. (1977). Light versus dark carbon metabolism in cherry tomato fruits. II. Relationship between malate metabolism and photosynthetic activity. *Plant Physiology*, **60**, 877-880.

[Fernie and Martinoia, (2009)] Fernie, A. R., Martinoia, E. (2009). Malate. Jack of all trades or master of a few? *Phytochemistry*, **70**(7), 828-832.

[Hafke *et al.*, 2003] Hafke, J. B., Hafke, Y., Smith, J. A. C., Lüttge, U., Thiel, D. (2003). Vacuolar malate uptake is mediated by an anion-selective inward rectifier. *The Plant Journal*, **35**, 116-128.

[Hawker, 1969] Hawker, J. S. (1969). Changes in the activities of malic enzyme, malic dehydrogenase, phosphopyruvate carboxylase and pyruvate decarboxylase during the development of a non-climacteric fruit (the grape). *Phytochemistry*, **8**, 19-23.

[Khan *et al.*, 2013] Khan, S., Beekwilder, J., Schaart, J., Mumm, R., Soriano, J., Jacobsen, E., Schouten, H. (2013). Differences in acidity of apples are probably mainly caused by a malic acid transporter gene on LG16. *Tree Genetics & Genomes*, **9**(2), 475-487.

[Kenis *et al.*, (2008)] Kenis, K., Keulemans, J., Davey, M. W. (2008). Identification and stability of QTLs for fruit quality traits in apple. *Tree Genetics & Genomes*, **4**(4), 647-661.

[Kim *et al.*, (2013)] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, **14**(4), R36.

[Kovermann *et al.*, 2007] Kovermann, P., Meyer, S., Hortensteiner, S., Picco, C., Scholz-Starke, J., Ravera, S., Lee, Y., Martinoia, E. (2007). The Arabidopsis vacuolar malate channel is a member of the ALMT family. *Plant Journal*, **52**, 1169-1180.

[Kunkee, (1991)] Kunkee, R. E. (1991). Some roles of malic acid in the malolactic fermentation in wine making. *FEMS Microbiology Letters*, **88**, 55-72.

[Kobayashi *et al.*, (2004)] Kobayashi, S., Goto-yamamoto, N., Hirochika, H. (2004). Retrotransposon-Induced Mutations in Grape Skin Color. *Science*, **304**, 982.

[Kobayashi *et al.*, (2005)] Kobayashi, S., Yamamoto, N.G., Hirochika, H. (2005). Association of *VvmybA1* gene expression with anthocyanin production in grape (*Vitis vinifera*) skin-color mutants. *Journal of the Japanese Society for Horticultural Science*, **74**, 196-203.

[Jaillon *et al.*, (2007)] Jaillon, O., Aury, J.-M., Noel, B., *et al.*, Choisne, N., *et al.*, Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, *Nature*, **449**(7161), 463-7.

[Lakso and Kliewer, 1975] Lakso, A. N., Kliewer, W. M. (1975). The Influence of Temperature on Malic Acid Metabolism in Grape Berries. *Plant Physiology*, **56**, 370-372.

[Lakso and Kliewer, 1978] Lakso, A. N., Kliewer, W. M. (1978). The Influence of Temperature on Malic Acid Metabolism in Grape Berries. II. Temperature Responses of Net

Dark CO_2 Fixation and Malic Acid Pools. *American Journal of Enology and Viticulture*, **29**,145-149.

[Leibhard *et al.*, (2003)] Liebhard, R., Koller, B., Gianfranceschi, L., Gessler, C. (2003). Creating a saturated reference map for the apple (Malus x domestica Borkh.) genome. *Theoretical and Applied Genetics*, **106**(8), 1497-508.

[Lüttge and Ratajczak, 1997] Lüttge, U., Ratajczak, R. (1997). The physiology, biochemistry and molecular biology of the plant vacuolar ATPase. *Advances in Botanical Research*, **25**, 253-296.

[Martinoia *et al.*, (2007)] Martinoia, E., Maeshima, M., Neuhaus, H. E. (2007). Vacuolar transporters and their essential role in plant metabolism. *Journal of Experimental Botany*, **58**(1), 83-102.

[Quinlan and Hall, 2010] Quinlan, A. R., Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

[Rienth *et al.*, (2014)] Rienth, M., Torregrosa, L., Kelly, M. T., Luchaire, N., Pellegrino, A., Grimplet, J., Romieu, C. (2014). Is transcriptomic regulation of berry development more important at night than during the day? *PloS One*, **9**(2), e88844.

[Romieu *et al.*, (1992)] Romieu, C., Tesniere, C., Than-Ham. L., Flanzy, C., Robin, J. P. (1992). An examination of the importance of anaerobiosis and ethanol in causing injury to grape mitochondria. *American Journal of Enology and Viticulture*, **43**, 129-133.

[Ruffner *et al.*, (1976)] Ruffner, H. P., Hawker, J. S., Hale, C. R. (1976). Temperature and enzymic control of malate metabolism in berries of *Vitis vinifera. Phytochemistry* **15**, 1877-1880.

[Ruffner, (1982)] Ruffner, H. P. (1982). Metabolism of tartaric and malic acids in *Vitis*: a review - Part B. *Vitis*, **21**, 346-358.

[Ruffner and Hawker, (1977)] Ruffner, H. P., Hawker, J. S. (1977). Control of glycolysis in ripening berries of *Vitis vinifera. Phytochemistry* **16**, 1171-1175.

[Ruffner and Kliewer, (1975)] Ruffner, H. P., Kliewer, W. M. (1975). Phosphoenolpyruvate carboxykinase activity in grape berries. *Plant Physiology*, **56**, 67-71.

[Schulze *et al.*, (2002)] Schulze, J., Tesfaye, M., Litjens, R. H. M. G., Bucciarelli, B., Trepp, G., Miller, S., *et al.* Vance, C. P. (2002). Malate plays a central role in plant nutrition. *Plant and Soil*, **247**, 133-139.

[Shimada *et al.*, (2006)] Shimada, T., Nakano, R., Shulaev, V., Sadka, A., Blumwald, E. (2006). Vacuolar citrate/H+ symporter of citrus juice cells. *Planta*, **224**(2), 472-80.

[Sweetman *et al.*, (2009)] Sweetman, C., Deluc, L. G., Cramer, G. R., Ford, C. M., Soole, K. L. (2009). Regulation of malate metabolism in grape berry and other developing fruits. *Phytochemistry*, **70**(11-12), 1329-44.

[Taureilles-Saurel *et al.*, 1995] Taureilles-Saurel, C., Romieu, C. G., Robin, J. P., Flanzy, C. (1995). Grape (*Vitis vinifera* L.) malate dehydrogenase. II. Characterization of the major mitochondrial and cytosolic isoforms. *American Journal of Enology and Viticulture*, **46**, 29-36.

[Trapnell *et al.*, (2012)] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, **7**(3), 562578.

[Venturini *et al.*, (2013)] Venturini, L., Ferrarini, A., Zenoni, S., Tornielli, G. B., Fasoli, M., Dal Santo, S., *et al.*, Delledonne, M. (2013). De novo transcriptome characterization of *Vitis vinifera* cv. Corvina unveils varietal diversity. *BMC Genomics*, **14**(1), 41.

[Verweij *et al.*, (2008)] Verweij, W., Spelt, C., Di Sansebastiano, G.-P., Vermeer, J., Reale, L., Ferranti, F., *et al.*, Quattrocchio, F. (2008). An H+ P-ATPase on the tonoplast determines vacuolar pH and flower colour. *Nature Cell Biology*, **10**(12), 1456-1462.

[Vitulo *et al.*, (2014)] Vitulo, N., Forcato, C., Carpinelli, E. C., Telatin, A., Campagna, D., DAngelo, M., *et al.*, Valle, G. (2014). A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. *BMC Plant Biology*, **14**(1), 99.

[Walker *et al.*, (2007)] 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. *The Plant Journal*, **49**, 772-785.

Acknowledgements

I would like to thank the ICT Doctoral School and the University of Trento, which co-supervisioned my PhD with the École Doctorale SIBAGHE and the INRA-Supagro in Montpellier for the resources they provided. I am extremely thankful to the French institutions for the hospitality they showed to me in my abroad period and to the Fondazione Edmund Mach, for the economical support and the availability of computer capacity while I was working in Italy.

Regarding the results accomplished in Chapter 2, I would like to thank Elisa Banchi, Silvia Vezzulli and Riccardo Velasco for the Vitis17KSNP chip hybridization. I am also grateful to Duccio Cavalieri, Irene Stefanini, Valentina Cappelletti and Matteo Ramazzotti for the yeast benchmark set and Federico Vaggi for stimulating conversation.

I thank Charles Romieu, who is part of the INRA-Supagro institution, and provided the RNA-seq data analyzed in Chapter 3, while for the SNP validation through the Vitis17KSNP chip hybridization I am thankful once again to Elisa Banchi, Silvia Vezzulli and Riccardo Velasco.

I am thankful to all my current and former advisors for their support: Charles Romieu, Patrice This, Claudio Moser, and Enrico Blanzieri. They had to make the effort to be a team whose members were far away from each other but, however, always ready to promptly comunicate and solve the bureaucracy mysteries.

A key figure in this PhD project development is Alessandro Cestaro, who has been a highly skilled mentor with a profound knowledge in bioinformatics and biology, and who had the strength to address this intricate PhD journey in the right direction since his involvement. I would like to sincerely thank him for taking me under his wings and proving how many things I was able to achieve in the last years. Alessandro is an exquisite person to relate with and I strongly wish to keep working together on future projects.

I deeply thank my parents to be tirelessly close to me in every instant of my life and my sisters, who have the incredibly capacity to always make me smile, no matter how dark the moment could be.

I definitively need to thank my boyfriend, Luca, with whom the burden felt less heavy. He is the one person who emotionally walked into this PhD journey with me day by day and, if I didn't exhausted him already, I truly wish to be with him also on the next adventure.

The path to achieve the PhD is not supposed to be easy, and I can now say this one was particularly troubled, but isn't the toughest training making the better soldier? Let's hope so!