Scuola di Dottorato Ubaldo Montelatici
In Scienze e Tecnologie Vegetali Microbiche e Genetiche
Dottorato di ricerca in Scienze Genetiche, Microbiologiche e Bioinformatica
Ciclo XXV

AND

Faculty of Sciences, Department Plant Biotechnology and Bioinformatics
Doctoral Study Programme in Science: Biochemistry and Biotechnology

Genomewide analysis of gene expression in *Vitis Vinifera ssp.*

Università Degli Studi di Firenze
Ghent University

Università degli studi di Firenze

Scuola di Dottorato Ubaldo Montelatici

Dottorato di ricerca in Scienze Genetiche, Microbiologiche e Bioinformatica

Ciclo XXV

Settore scientifico disciplinare di riferimento: AGR07/BIO18

Coordinatore del dottorato

Prof. Milvia Luisa Racchi

University of Ghent

Faculty of Sciences, Department Plant Biotechnology and Bioinformatics

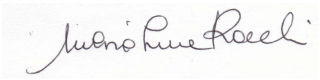
Doctoral Study Programme in Science: Biochemistry and Biotechnology

Genomewide analysis of gene expression in *Vitis Vinifera ssp.*

Emilio Potenza

21/12/2012

Supervisor:
prof. Milvia Luisa Racchi



Supervisor:
prof. Yves Van de Peer

Author: Emilio Potenza

b

Partnership agreement governing the joint supervision and awarding of a doctorate diploma between ***University of Florence*** *and* ***Ghent University***

Ghent University, hereinafter referred to as 'UGent',
hereby duly represented by Prof. dr. P. Van Cauwenberge, Rector,
domiciled : Sint-Pietersnieuwstraat 25, 9000 Ghent (Belgium)

University of Florence, hereinafter referred to as 'UNIFI',
hereby duly represented by Prof. dr. A. Tesi Rector,
domiciled: Florence, Piazza S. Marco 4, (Italy)

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Dichiarazione

Con la presente affermo che questa tesi è frutto del mio lavoro e che, per quanto io ne sia a conoscenza, non contiene materiale precedentemente pubblicato o scritto da un'altra persona né materiale che è stato utilizzato per l'ottenimento di qualunque altro titolo o diploma dell'Università o altro istituto di apprendimento, a eccezione del caso in cui ciò venga riconosciuto nel testo.

Firenze, **21/12/2012**
Emilio Potenza

d

Preface

This PhD thesis was prepared at the department of Agriculture Biotechnology, Genetic division, at the University of Florence, Faculty of Agriculture in collaboration with the Department of Plant System Biology (PSB[1], a VIB - Ghent University research department, Bioinformatics and systems biology division), at the Faculty of Sciences, Department Plant Biotechnology and Bioinformatics at the University of Ghent with the supervision of the Department of Comparative Genomics at the 'Fondazione Edmund Mach'.

This PhD thesis was prepared with a a partnership agreement between **University of Florence** and **Ghent University** for the the joint supervision and awarding of a doctorate diploma. Thanks to the *'Scuola di Dottorato Ubaldo Montelatici In Scienze e Tecnologie Vegetali Microbiche e Genetiche, Dottorato di ricerca in Scienze Genetiche microbiologiche e Bioinformatica'*, Ciclo XXV at UNIFI, and to the *Doctoral Study Programme in Science: Biochemistry and Biotechnology* at UGent.

[1] http://www.psb.ugent.be/

This PhD thesis was fully funded by the *GMPF PhD Programme*[2] and by the **'Fondazaione Edmund Mach'**[3].

GMPF is an international PhD Programme in Fruit Plants Genomics and Molecular Physiology that represents 17 research institutions from 11 different countries. The programme was launched in 2009 and it is based at San Michele all'Adige (TN), Italy. GMPF offers once a year PhD scholarships for high level international research projects and collaborations in fruit biology through a network involving prestigious institutions.

Firenze, **21/12/2012**
Emilio Potenza

[2]http://www.gmpf.eu
[3]http://www.fmach.it/

Mr. Asimov, tell us something about
the thermodynamic properties of the compound
thiotimoline...[4]

Acknowledgements

This PhD thesis would not have been possible without the support of many people. The author wishes to express his gratitude to his supervisors, Prof. Milvia Luisa Racchi, Prof. Yves Van de Peer and Dr. Alessandro Cestaro who were abundantly helpful and offered invaluable assistance, support and guidance. Deepest gratitude are also due to the GMPF PhD programme, for providing me with a good environment and facilities to complete this project. Special thanks also to all his graduate friends for sharing the literature and invaluable assistance. Not forgetting to his best friends who always been there. The author would also like to convey thanks to the Edmund Mach Foundation for providing the financial means and laboratory facilities. The author wishes to express his love and gratitude to his beloved families; for their understanding and endless love, through the duration of his studies.

[4]Professor Ralph S. Halford, during Isaac's PhD defence

Summary (Italian)

Le analisi di RNA-seq rappresentano oggi il principale metodo per studiare nel suo insieme il trascrittoma di un organismo. La tecnologia RNA-seq è potenzialmente capace di superare le limitazioni dei precedenti sistemi di indagine, principalmente perché l'enorme mole di dati prodotta consente di giungere ad identificare mRNA non ancora annotati. La capacitá di generare un elevatissima copertura 'coverage' del trascrittoma ci ha permesso di scoprire che perfino nell'uomo una parte rilevante del trascrittoma è ancora poco conosciuta.

Nonostante i considerevoli sforzi fatti per studiare questi meccanismi, non siamo ancora in grado di associare una funzione chiara a tutti i trascritti identificabili per mezzo dei sequenziamenti massivi. Non di meno, le nuove technologie di sequenziamento hanno incrementato le possibilità di identficare il 'rumore' del processo di trascrizione, fornendoci anche la possibilitá di investigare aspetti poco conosciuti come lo 'splicing' alternativo e altri eventi rari.

Con l'avvento delle nuove tecnologie di sequenziamento, il numero di geni coinvolti in eventi di splicing alternativo sta crescendo esponenzialmente,

anche se resta ancora poco chiaro quale sia la funzionalità delle isoforme rilevate dalle analisi. Purtroppo nelle piante gli studi sono pochi e condotti principalmente su organismi modello.

La prima parte della tesi si focalizza sull'analisi dello splicing alternativo nella bacca di vite (*Vitis spp.*). Utilizzando la tecnologia RNA-seq Illumina, il trascrittoma di 10 cultivar di vite è stato sequenziato ed analizzato poco dopo l'invaiatura, ad una maturità tecnica comune per tutti i campioni. Le 10 varietà sono state selezionate in base ai loro differenti profili metabolici, specialmente per quanto riguarda la loro capacità di produrre e accumulare flavonoidi.

I dati presentati in questo studio sono ad oggi la piú completa analisi di sequenziamento effettuata in vite, e permettono per la prima volta di analizzare lo splicing alternativo con una altissima risoluzione. Per analizzare e confrontare i livelli di splicing alternativo abbiamo sviluppato un nuovo software, findAS, che è in grado di identificare tutti i principali eventi alternativi al modello, utilizzando le informazioni contenute nei files di allineamento e i corrispettivi modelli genici.

La nostra analisi in vite suggerisce che almeno il 40% dei geni multiesonici hanno eventi di splicing alternativo, abbiamo inoltre dimostrato l'esistenza di una ampia classe di isoforme poco abbondanti. In media sono stati predette ~110,000 giunzioni di splicing per ogni cultivar oggetto di analisi, la maggioranza delle quali presenta la sequenza consenso tipica.

Abbiamo osservato che la maggior parte degli eventi alternativi ha un basso 'coverage', infatti più del 70% degli eventi di splicing alternativo ha un rapporto tra l'isoforma alternativa e l'isoforma costitutiva inferiore a 0,1. Inoltre la maggior parte dei siti di splicing alternativi e rari si trovano molto vicini ai siti costitutivi. Questo potrebbe essere dovuto sia alla presenza di errori che per effetto di una qualche forma di 'rumore' del processo di trascrizione. In real-

tà quello che potrebbe sembrare un semplice rumore stocastico del processo di trascrizione, risulta essere spesso ampiamente conservato nelle 10 cultivar analizzate.

Questo comportamento complesso potrebbe dimostrare l'esistenza di una funzione importante anche per gli eventi a basso 'coverage', probabilmente legata a meccanismismi di regolazione quali NMD e RUST. I nostri dati illustrano le principali caratteristiche dello splicing alternativo nella bacca di vite e forniscono alcune indicazioni sul ruolo che potrebbe svolgere l'efficienza nel processo di splicing per la regolazione dell'espressione genica.

La seconda parte di questa tesi si concentra sullo studio dei geni differenzialmente espressi nelle 10 cultivar di vite che costituiscono il materiale sperimentale della tesi. Le analisi RNA-seq sono utilizzate sempre di più anche per studi di espressione differenziale, utilizzando il numero delle 'reads' che si allineano in corrispondenza di un gene come una stima del livello di espressione.

Con una serie di analisi statistiche, il conteggio delle 'reads' per ogni gene è stato utilizzato per stimare il livello di espressione di un singolo gene e per capire se il gene risultasse differenzialmente espresso nel confronto tra le 10 cultivar. Il risultato finale di questo studio è stata l'identificazione di un insieme di geni caratteristici per ogni cultivar, cioè un insieme di geni che risultano essere differenzialmente espressi per una cultivar nel confronto con tutte le altre 9 varietà.

Sebbene un'analisi globale dell'espressione genica basata su una singola replica non consente una solida interpretazione biologica, questa analisi RNA-Seq ci fornische una chiara visione della partecipazione di più famiglie multigeniche durante lo sviluppo della bacca, identificando quali membri sono espressi e caratterizzandone i loro profili di espressione nel dettaglio,

mostrando che possono partecipare alla sintesi e all'accumulo di metaboliti secondari in 10 diverse cultivar. Rispetto agli studi precedenti sono stati identificati molti nuovi trascritti coinvolti nella maturazione della bacca, aprendo la strada ad una descrizione più accurata e più dettagliata dei processi molecolari coinvolti nello sviluppo degli acini che sono alla base delle loro proprietà organolettiche.

Nell'insieme la tesi ha dimostrato che grazie alle enormi potenzialità fornite dalle nuove tecnologie di sequenziamento, è possibile analizzare approfonditamente i meccanismi di espressione e regolazione che avvengono nella bacca di vite. Sperando in una prospettiva a medio termine, che cio possa contribuire a migliorare ulteriormente la qualità di vini. La procedura bioinformatica, gli strumenti di calcolo sviluppati e più in generale, l'approccio utilizzato in questa tesi sono applicabili, non solo alla vite ma anche ad altre specie.

Summary (English)

RNA-seq analysis represents nowadays the emerging method for the genome-wide transcriptome analysis. Due to the large amounts of reads produced, this technology is able to overcome most of the limitations of previous methods, especially in finding new mRNAs. This unprecedented depth of sequence coverage, which now we are able to produce, has shown that still a relevant part of the transcriptome is not well characterized even in human.

Although considerable efforts have been recently made to analyze these mechanisms, a portion of the transcripts so far identified have no clear function. Nevertheless, the new sequencing technology have increased the chance to identify the transcriptional noise, providing the opportunity to investigate some unknown aspects of Alternative Splicing (AS) such as low-abundance AS events.

Here, we performed by means of RNA-seq a comparative genome-wide analysis of berry transcriptome in 10 grapevine cultivars selected on the base of different metabolic profiles. The data presented in this study provides, up to date, the most comprehensive set of RNA-seq gene expression variants in

grape, and is been expected to facilitate detection of alternative splicing events with high resolution. The first part of the thesis is focused on the analysis of the Alternative Splicing for the berries transcriptome. With the incoming NGS data the ratio of genes for which alternative splicing could be detected increased exponentially. Nevertheless it is still unclear how functionally relevant these splice forms are. Moreover, only few comprehensive studies of plant transcriptomes are available, and mainly in model organisms.

Our analysis suggests that in *Vitis Vinifera*, at least 40% of intron-containing genes are alternatively spliced. We demonstrated the existence of a large class of low abundance isoforms, encompassing approximately ~110,000 splice junctions for each cultivar, that mostly derived from junctions with typical consensus sequence. We have found that the majority of the mRNA diversity observed derives from low-abundance events. More than 70% of total events showed a read coverage ratio between the alternative and the constitutive form lower than 0,1.

In addiction rarely used splice sites shown an enrichment near often-used splice site of the constitutive form, suggesting that transcription is affected by a kind of 'noise'. However this putative noise is extensively conserved between the 10 cultivars analyzed.

This complex behaviour could hint to a relevant functionality even for low-coverage splicing events which are probably related to regulatory mechanisms. Our data provide a comprehensive analysis of alternative splicing in Vitis vinifera and giving some lights and proposes hypothesis about the roles of spliceosome efficiency in regulating gene expression.

RNA-seq is also increasingly being used to quantify gene expression, as the number of mapped reads to a given gene or transcript is an estimation of the level of expression of that feature. The second part of the thesis is focused on

a digital expression analysis (DGE) between the same 10 grapevine cultivars. Using NOISeq (R package) to the reads count for each gene we were able to detect all the genes differentially expressed in a pairwise comparison, 45 different tables were obtained with the genes that are likely to be differentially expressed in the pair of cultivars. The results was the identification of the most peculiar genes for each cultivar and their putative function.

Although a global analysis of gene expression based on a single replication does not allow a solid biological interpretation, this RNA-Seq analysis clearly provided a comprehensive view of the participation of several multigene families in berry development and ripening, identifying which members are expressed and characterizing their expression profiles in detail, showing which are likely to participate in the synthesis and accumulation of secondary metabolites in 10 different cultivars. In comparison to previous studies, the RNA-Seq method identified many additional transcripts, paving the way for a more accurate and more detailed description of the molecular processes involved in the development of grape berries and the basis of their organoleptic properties

The overall thesis shows that with the incoming of the omics technologies is finally possible explore the gene expression and its regulation in grape, with a long term goal of such a search to provide an improvement to the wine quality. The bioinformatic pipeline, the computational tools developed and in general the used approach in this thesis are applicable to other species with only some minor edits.

When people thought the Earth was flat, they were wrong.
When people thought the Earth was spherical they were wrong.
But if you think that thinking the Earth is spherical
is just as wrong as thinking the Earth is flat, then
your view is wronger than both of them put together.

Isaac Asimov

1 Introduction

1.1 The central Dogma

1.1.1 The history of the dogma

Figure 1.1: The Central Dogma: The simplest concept of biology

I would like to begin this dissertation with one of the most famous concept in Molecular Biology, the Central Dogma, perhaps a very basic concept and yet one of the most profound for a biologist.

The purest and original form of the central dogma states that the genetic information flow starts from DNA and terms with a protein, passing trough the RNA step, but most important, the genetic information cannot be passed from protein to any other form. The Figure 1.1 shows the very basic assumption of the central dogma. Using the metaphor of the dictionary, basically a dictionaries exist to convert DNA to RNA and RNA to Protein, but there isn't a dictionary to convert a protein to anything else. Once a genetic message is translated into a proteins, the flow of genetic information is stuck!

The central dogma was originally proposed in 1958 by Francis Crick [1], the same Francis Crick that only few year before has discovered the double-helix-structure of DNA in collaboration with J.D Watson [2]. In the glorious publication, *On Protein Synthesis* [1], Crick discusses two majors topics that he calls 'the central dogma' and 'the sequence Hypothesis'.

It may be easiest to first explain a more common definition of the Central Dogma, even if it is not the original definition of 1953, which describes the direction of information flux in molecular biology. According to a definition that Crick himself revisited in 1970 [3] after several misunderstanding about the real meaning of the famous dogma, the genetic information can be transmitted in one way from DNA to RNA to protein, more precisely the central dogma is a negative statement ('transfers from protein to nucleic acid do not exist' [3]). The dotted gray arrows of Fig.1.2 underscore the original conception of the central dogma, the information is not allowed to be transferred to DNA, RNA, nor more protein, if it is already stored as protein.

On the other side, the sequence hypothesis complements the basic state-

ment of the central dogma, 'transfers from nucleic acid to protein exist' [3]. Basically, the sequence hypothesis states that the genetic information can be stored in a specific nucleic acid (DNA), as a sort of hard copy of the code, using 4 different nucleotide bases (A, T, G, C) in a linear, long and stable molecule. Then from DNA the information can be transcribed in a more manageable version,the RNA, as a soft copy of the original code. In the final end the sequence can be transcribed in a simple chain of amino acids, the building blocks of every protein.

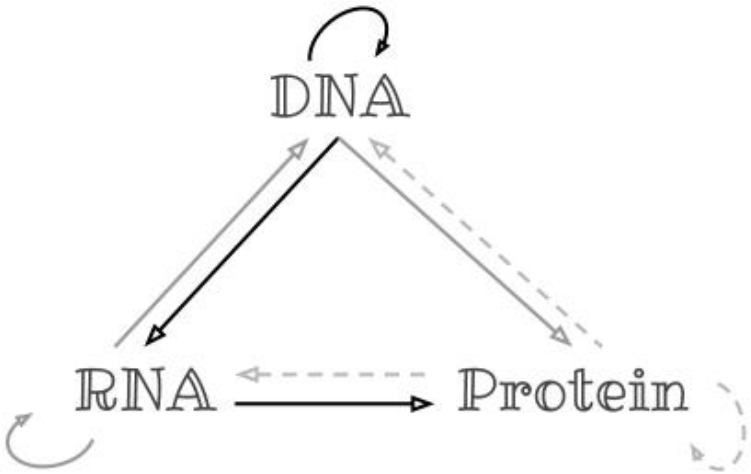


Figure 1.2: The Central Dogma: with dark line we sign the general transferts, with light-gray the special transfers and with a dotted line the unknown trasfers, following the Crick definition of 1970 [3]

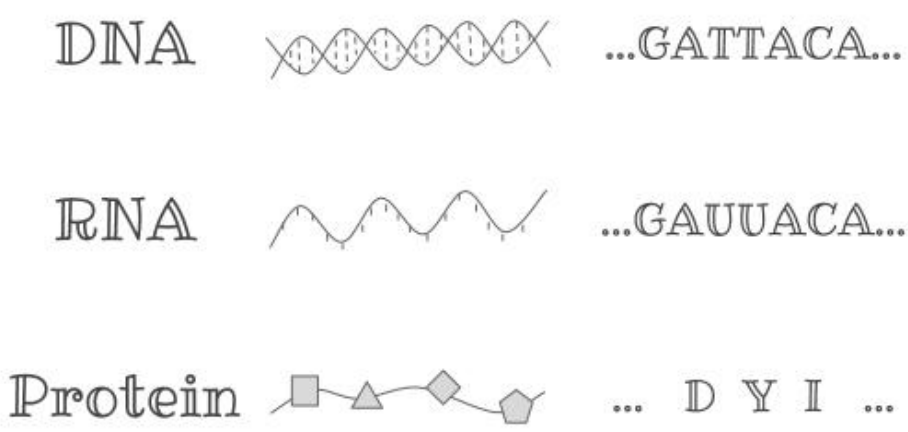


Figure 1.3: This flow of information is dependent on the genetic code, which defines the relation between the sequence of bases in DNA (or its mRNA transcript) and the sequence of amino acids in a protein.

1.1.2 What exactly do I mean by genetic information?

DNA and RNA are long linear biopolymers, with an overall term of nucleic acid. A nucleic acid is composed by a long chain of nucleotide, each of which contains a pentose sugar (ribose for RNA or deoxyribose for DNA), a phosphate group and a base. The genetic information is indeed stored in the sequence of bases along a nucleic acid chain.

Furthermore, the nucleic acid forms a double helix structure with a double stranded biopolymers. The base pairs generated within the double helix are almost fixed (A-T, C-G). The hydrogen bounds, generated within a double stranded molecule, provide an high stability and long life for the DNA

molecule but also a suitable structure for reproduce the genetic information by copying chain partially or entirely [4].

As showed in Fig.1.3, the specific language of DNA is written using only four letter. The nucleotide bases adenine (A), thymine (T), cytosine (C), and guanine (G). RNA is indeed very similar to a DNA molecule but mostly can be found a single-stranded molecular and also there is one letter substitution, uracil (U) for thymine. Last but not the least even the protein is a linear molecule, which is able to folds into complex 3D structures, with several manner depending on the physical an chemical properties of its primary sequence of amino acids, such as Methionine (M), Glycine (G), and Proline (P).

This flow of information is directly dependent on the genetic code. The genetic code basically describe the relation between the DNA sequence (or its mRNA transcript) and the amino acids sequence. The code is nothing more than a sequence of three bases, codon, that identifies an amino acid. Codons are almost the same in every organisms, each codons in a mRNA molecule are read using another type of RNA, the tRNA molecules, which carries the amino acid during the protein synthesis.

Coming back to the basic statement of the central dogma, 'transfers from protein to nucleic acid do not exist' [3], the transfer of RNA's four letter sequences to protein's twenty letters of amino acids is an irreversible process. The main reason for that irreversibility is because multiple codons can code for a single molecule. In the glycine's case GGA, GGU, GGC, and GGG all encode the amino acid, but in order to 'untranslate' a glycine we have to know from which combination of codons the glycine has been generated. In a more scientific way this is known as codon degeneracy.

1.1.3 What is a gene?

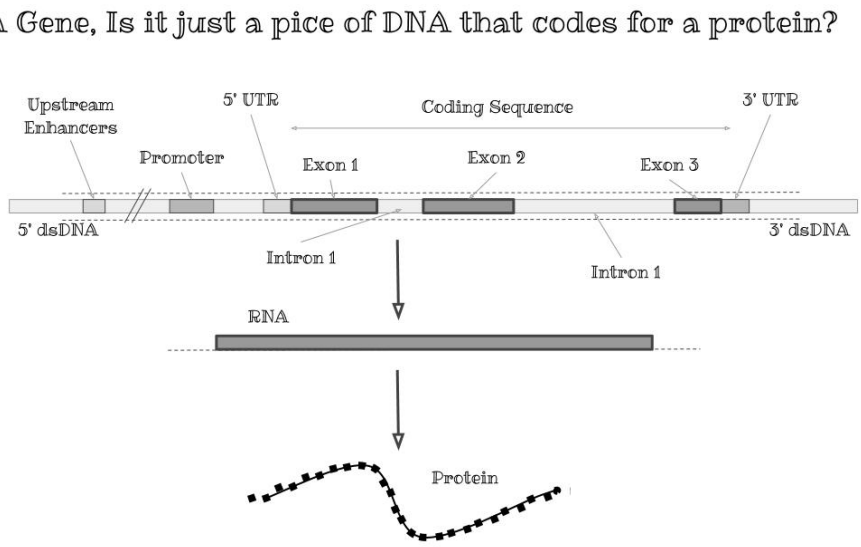


Figure 1.4: Is it just a protein coding region or somthing relatad to any kind of functional activity?

Starting from the first definition of gene, commonly used in classical genetics, a gene was an abstract unit of inheritance, that in some way ferried its properties to the child from the parents. As the biochemistry came into its own, a gene became something directly linked with enzymes or proteins, only one for each gene. Finally with the advent of molecular biology, genes became a real stuff, a long piece of DNA which could be used to build their associated protein, after a conversion to the messenger RNA [4].

Anyway this picture is still the working model for many scientists. As long as we keep going deeply into the genomes analysis, it seems, every day more clear, that the information is extracted from the chromosomes in a complex way, even more than was supposed at the beginning of this kind genetic studies.

For example, it is now clear to every scientist, that an RNA molecules can have an active role in many cellular processes, not only as passive state through which each gene send its message into the cell. In particular, several RNA molecules such as transfer RNA (tRNA) and ribosomal RNA (rRNA), microRNA or RNAi are part of the protein-synthesizing and transcripts regulation [5] [6].

A lot of observations suggest that we need to define a new paradigm [7] in order to redefine the meaning of 'gene' ad 'genetic information' taking into account the large variety of regulatory RNAs.

The classical definition of a gene should be changed not only because from a single gene many different isoforms can be generated, in several post-transcriptional or post-translational changes, but also because it is now clear that, in a genome, there are several different kind of 'genes', with fundamental function that by the way do not encode proteins [7].

It has been suggested to change the protein-centric view of the gene definition many different time and with several different definition such as 'fuzzy transcription clusters with multiple products' [8] or 'union of genomic sequences encoding a coherent set of potentially overlapping functional products' [9], surprising both very close to the earliest concept of of genetic locus.

1.1.4 The classic violations

If we purely consider this matter from the information theory point of view, given 3 different stationary states (DNA, RNA and protein) there are nine existing transfer possibility as showed in Fig.1.2. In this theoretical architecture, RNA can be transferred into DNA, protein into other protein, DNA directly into protein, and so on and so fort. Since the elaboration of the cen-

tral dogma more than 50 years ago, many discovery contradict that idea of genome [10]. Moreover, to describe the real nature of genetic information should be changed also the idea of a discrete 'gene' as the fundamental unit. It was already demonstrated that a genome is transcribe extensively but also has been found that from most of the loci are expressed 'a complex mix of overlapping, interleaved and bidirectional coding and non-coding, sense and 'antisense' transcripts' [7], [11][12][13].

Probably the fist amazing discovery that had challenged molecular biology's paradigm was alternative splicing. The first description for such unusual mechanism was detected in viruses around 1977 [14][15]. With alternative splicing, from a single gene the cell can produce different isoforms an so proteins removing introns and sewing together the exons in several different orders. However, alternative splicing did not in itself change massively the gene statement; no more one gene one protein, but basically one gene many protein.

Another early violation involves RNA-to-DNA information transfer and involves a specific enzyme family called reverse transcriptases. Starting from an RNA molecule a reverse transcriptase can reproduce a complementary DNA molecule. This mechanism was first discovered in some retroviruses[1], but than we also discovers the telomerase, that is a human reverse transcriptase responsible for protecting from shortening the ending parts of the human chromosomes, one of the causes for the cell ageing [16].

Without any doubt the most consistent today's assault on the gene concept comes from the news unimaginable scopes of RNA. For example the role of distal 5' exons, that are only used in specific conditions and can be far away from the genetic locus [17][18], the possibility to have alternative initiation sites [19], and also wide variety of short and long RNAs [12][13].

[1] A class of virus that includes HIV and Hepatitis B Virus

Thus, the real boundaries of a classic gene cannot be defined unequivocally, furthermore it has been demonstrated the existence of the so called chimeric transcripts [20]. If we consider all these evidence we can assume a higher order network organization built around the action of many different transcription factories [21].

In conclusion we should consider the idea to refresh the way we look at the genome. We should definitely change the way we parse the biological information content. One of the possible way is to start considering RNA, and not any more the genes, as the central entities, annotated with a genomic origin, their function, if present also their open reading frame and moreover the interaction with specific molecules.

1.1.5 Transcription

Every single cell in every single complex organism share the same 'building' information within their DNA, but how is it possible to generate that large magnitude of different kind of cell? The simple answer is that the genome are conditionally used. The first step that allow the complex regulation of the genome usage is the transcription of the DNA in RNA [22].

First of all I would like to briefly clarify the basic step that occur during the RNA transcription, as aforementioned, the genetic information is stored into the DNA of chromosomes, and the process that 'transcribe' that information into an RNA soft copy, single stranded, is called transcription. The enzyme responsible to made a complementary single strand of DNA is called RNA polymerase (RNA pol). The classic role of an RNA molecule is pass from the nucleus into the cell cytoplasm where the translation into protein occur. This type of RNA is called mRNA , i.e messenger RNA [23].

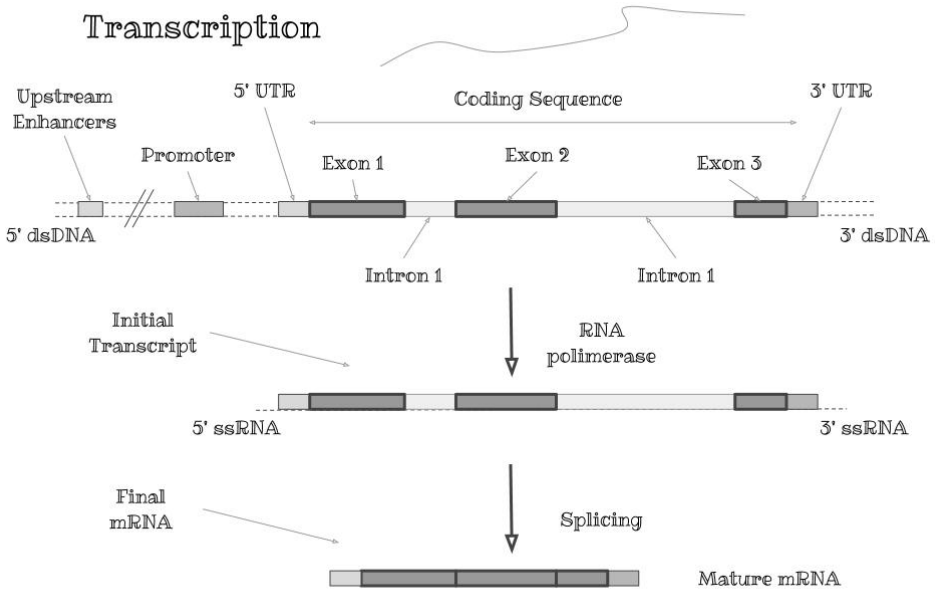


Figure 1.5: The transcription process, where DNA is converted to RNA, a more portable set of instructions for the cell.

The RNA polymerase is a protein complex that work under the regulatory effect of several transcription factors, co-factor, activator or repressors. It have been found in every species, one type in Bacteria but three different types in Eukaryotes [24].

The initiation is the first step of the transcription when the RNA pol bind the promoter region of one gene. It is important to note that during this step, RNA pol can be specifically driven to one gene by the so called sigma subunit. It have found that exist several sigma subunit specific to different promoters, moreover the sigma subunit actively cooperate in regulating the gene expression giving the unwinding capability to the holoenzyme. The promoter itself is a key part for the regulatory process and probably that the reason why the

complexity of the promoter increase a lot from Bacteria to Eukaryotes [25].

In the Eukaryotes, RNA pol use many cofactors, i.e. transcription factors, in order to tune the gene expression, Eukaryotic genes are also regulated by enhancers and/or silencers: sequence elements located some kilo bases from the regulated gene. The DNA looping facilitates the interaction between enhancer, transcription factor, promoter and RNA pol. Proteins that facilitate this looping are called activators, while those that inhibit it are called repressors [26].

Once transcription is initiated, the holoenzyme unwinds DNA double helix and RNA polymerase start adding nucleotides to the 3' end of the growing RNA chain, while reading the DNA strand as template.

Finally the process need to be terminated. Terminator sequences are found close to the ends of coding sequences. This final step is quite different in Bacteria and Eukaryotes, in bacteria we found two ways, the rho-dependent (the transcription is blocked by RNA hairpin loop) and rho-independent (rho factor directly unwinds the DNA-RNA hybrid). In Eukaryotes, RNA pol I can be blocked by some cofactor and RNA pol III if transcribe a polyuracil stretch, but RNA pol II can continue for thousands of bases after the coding sequence. This tail seems to be cleaved by a complex associate with the polymerase just before the polyadenilation at the 3' end [27].

The RNA molecule produced at the end of the process can be itself considered as a 'finished product' with a proper functional activity within the cell, but of course sometimes, the mRNA could be translated, in order to generate a protein molecule [28] [29].

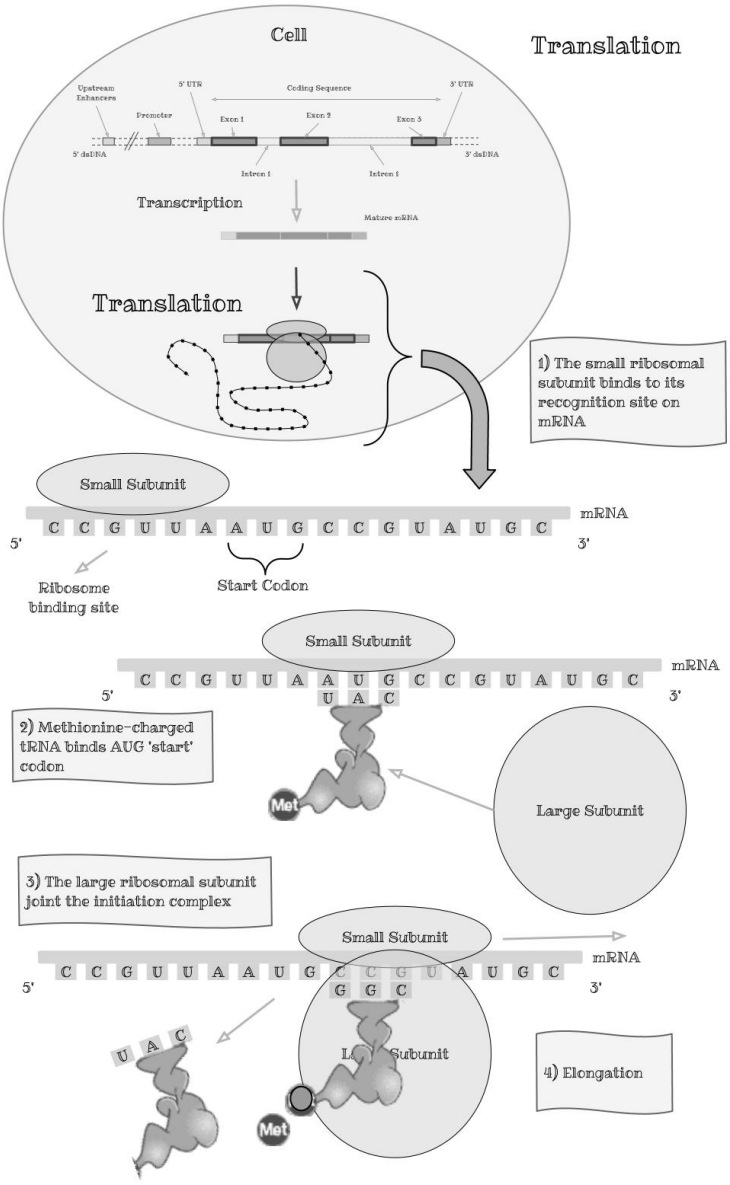


Figure 1.6: Translation. Diagram showing how the translation of the mRNA and the synthesis of proteins is made.

1.1.6 Translation

Just after this long introduction on the genetic information flow, I cat forgot to mention the final step for many generated transcript, the translation into protein.

During the translation the polypeptide, i.e. the amino acid chain, is assembled by ribosomes by using tRNA as adaptor molecule between an amino acid and the RNA molecule [30]. Basically each tRNA carried one amino acid and an 'anticodon' which can base pair with a codon on the RNA molecule. A codon is nothing more than a triplet of bases specifies a given amino acid but one amino acids can be specified by more than one codon [31].

In Eukaryotes, translation occurs across the membrane of the endoplasmic reticulum, the ribosome is composed of more than fifty different proteins plus two structural rRNAs, each part of the 30s subunit or the 70s[2] subunit [32].

Translation [3] proceeds in four phases (amino acids are brought to ribosomes and assembled into proteins as summarized in Fig. 1.6):

1 **Initiation**. The 30s subunit bind the mRNA. The ribosome then moves until it finds the first start codon, AUG, the mRNA, where it binds a Methionine by the Met-tRNA within the P site.

2 **Elongation**. The 50s subunit now binds (A site) and each new aminoacyl-tRNA enters in the A site.

3 **Translocation**. The entire ribosome now 'translates' over one codon position, so that the nascent chain is now bound to the P site. While

[2]The 's' is a unit of sedimentation, referring to how fast a particle settles out during centrifugation.

[3]http://en.wikipedia.org

reading the mRNA as template the amino acid chain is generated one step at the time and later folded into an active protein

4 **Termination**. The first time that a STOP codon enter the A site. Some factors actually facilitate the polypeptide release because they bind an A site with the termination codon.

1.2 Gene regulation

Since the beginning of genomics was discovered that the number of protein coding genes doesn't change with the increase of the organism complexity [33]. Today it is largely accepted that the complexity is generated by the regulatory process that manage the gene expression in a very sophisticated way [34].

Regulation of gene expression includes a wide range of mechanisms, for example to trigger developmental pathways, respond to environmental stimuli, or adapt to new food sources. These mechanisms modulate the transcription of a specific gene, increasing or decreasing the level of specific gene products (protein or RNA). Virtually any step of gene expression can be modulated, from transcriptional initiation, to RNA processing, and to the post-translational modification of a protein [4].

Gene expression regulation involve many different factors. From regulatory proteins, many different kind of transcription factors and non-coding RNAs. Furthermore many studies supposed that the control of epigenetic pathways could be driven by a large class of non-protein-coding RNAs [35] [36].

[4] http://en.wikipedia.org

In this section I would like to introduce the recent understanding regarding the gene regulation, particularly for the complex network activity of transcription factors, the regulation carried out by the hidden layer of RNA molecules and by all the other kind of *cis*- and *trans*-acting regulatory factors [7].

1.2.1 Transcription factor and regulatory networks

In molecular biology and genetics, the term 'transcription factor' (TF) is used to define in general a protein that binds to specific DNA sequences, thereby controlling the transcription process in many different ways, as enhancer or silencer of the RNA production [37] [38].

Alone or interacting with other molecules, TFs perform this function as an activator or as a repressor of the RNA ploymerase activity. The action of a single TF is often general, i.e. the same TF is used to act in the some way in different cell types by recognizing binding sites that occur in many places around the gene and around the genome. Anyway the role of differences in TFs versus coding changes during the development of an organism is a topic of debates [39].

A Transcription factor can use a variety o different ways to influence gene expression, basically a TF bind the DNA either enhancer or promoter regions of DNA adjacent to the genes. Transcription initiation, as already explained, requires the binding of RNA pol II, TFs, and cofactors to cis-action regulatory sequences.

In summary we can define two types of binding sites for a TF, enhancer or silencer. The binding site is actually a short degenerated sequence from 6 to 20 bp, that can be located around or within a gene. From its position a TF : (i) can influence the stability of the RNA pol and the DNA bond, (ii) can alter

histone conformation by catalyzing the acetylation or deacetylation, (iii) can recruit directly other proteins or cofactors. It is important to note that histone acetylation weakens the association of DNA with histones, and of course the deacetylation strengthens the association of DNA with histones. If the histones are strongly bound, the DNA result less accessible to the transcription, as so it is possible to down-regulate the transcription [40].

The complexity arising from enhancer or silencer is increased by the fact that often multiple TFs, multiple enhancer and many other types of molecules, combine their effects on gene expression.

1.2.2 Hidden layer of RNA

It is now clear also that not only TFs, promoter, enhancer or silencer are the players of the gene regulation game. Some evidence shown that the mammals genome are pervasive transcribed, apparently in a developmentally specific manner [41] [42]. Initially, the first suspect of this kind of behaviour was the expression noise, but there are now substantial genome-wide evidence pointing the intrinsic functionality of these transcript [43] [44], suggesting that the expression 'noise' is more probably an expression 'choice'.

For example there are growing evidence that both coding and non coding gene produce various RNA isoforms, shorter then the hypothetical full-length [45]. These truncated isoforms can be produced as short RNA (sRNA) or by the cleavage of longer RNAs. For example miRNA are thought to be produced by introns or from degradeted ncRNAs [46]. Just to be clear the study of those complex interaction is at the real beginning, it has been difficult to distinguish between artefacts, natural products of the RNA's degradation and a sRNA with a functional role. The use of the high-throughput sequencing is expected to resolve as much as possible the issue.

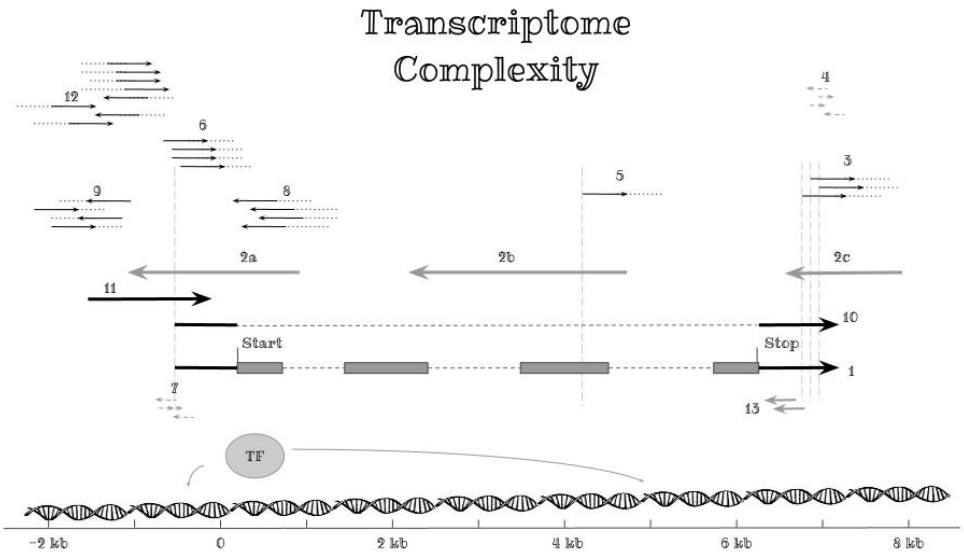


Figure 1.7: Complexity of transcriptome around a hypothetical gene. Not in scale. CAGE tags (dots indicate unknown ends) identify TSSs or other capped molecules; (1) protein coding mRNA; (2, a-c), antisense RNAs (3'-3' overlap, full overlap, 5'-5' overlap); (3) CAGE tags, likely polyadenylated; (4) termination-associated sRNAs (TASRs); (5) exonic long-capped transcripts; (6) CAGE tags identifying TSS; (7) PASRs and tiny 18 nt long RNAs (tiRNAs); (8) antisense transcription events; (9) bidirectionally RNAs from core promoters; (10) ncRNA splicing isoforms only partially overlapping to coding mRNA sequences; (11) PALRs; (12) PROMPTs; (13) miRNAs and endogenous siRNAs; (14) other sRNAs. The list of different types of RNAs is continuously growing ...[45].

Recent research indicates that most of the eukaryotic genome is transcribed in both directions: sense and antisense [47]. Which strand should transcribed and how much are made in some extent by transcription factors activity, local chromatin modifications, boundary elements and other regulatory RNA species. Long ncRNAs can regulate the chromatin structure [48] or recruit

other DNA modifying complexes to regulate TSS (Transcription start sites) accessibility to transcription factors and RNA polymerase II. The supposed complexity around a gene is shown in fig. 1.7.

One of the main actors of the transcriptome complexity is the antisense transcript. Natural antisense transcript are RNA molecules transcribed in the opposite direction of the coding region, with a partial overlap or complete with the sense mRNA. Today 'the presence of antisense transcript is no longer a curiosity but rather a pervasive feature of mammalian genomes' [49]. Both ends of the protein-coding gene have the propensity to produce natural antisense transcript but the regulation effect is not always linked to the expression of the sense gene. In term of alternative splicing they tend to undergo less frequently than the sense transcripts.

Many differ functional activity have been discovered knocking down the endogenous antisense transcript, such as showing either an increase or a decrease of the sense transcript, suggesting that the antisense-mediated regulation must operate through a variety of mechanisms [50]. For example the transcription-related modulation, where the act of collision in the antisense direction modulate transcription in the sense RNA, like when the polymerase is blocked by another polymerase running in the opposite direction [51] but also when the the antisense RNA is involved in remodelling the chromatin structure [52] or opening the transcription bubble [53].

Natural antisense RNA may also be in related to the DNA methylation, within some epigenetic regulation process. This gene regulatory model is based on the direct or indirect interaction between RNA-DNA or RNA-chromatin [54]. The third type of mechanism by which antisense transcript can regulate the gene expression is based on the formation of a RNA duplex sense-antisense. The nuclear RNA duplex can produce alternative splicing effect masking the constitutive splice site [55], and also the duplex is the target for

many enzymes that can alter the localization, the stability and the transport of the sense mRNA transcript. In the last mechanism that I would like to explore, a duplex RNA-DNA is formed in the cytoplasm. Nuclear RNA hairpins can affect sense RNA stability, cover the microRNA binding sites or even serve as template for generating endogenous small interfering RNAs (siRNA).

ncRNA classification

One of the easiest criteria that is used in order to classify non-coding RNAs (ncRNAs) is their size. Therefore, we have divided the ncRNAs into two major groups small (17-200 nt) and large (more than 200 nt) and both of them can be split into different sub-groups depending on their function, origin, mode of action or particular feature.

Short ncRNAs range from 17 nucleotides (nt) to approximately 200 nt in length. The short ncRNAs can be subdivided into:

- **siRNA**. Small-interfering RNAs
- **sRNA**. Small non-coding RNAs
- **piRNA**. Piwi-interacting RNAs
- **miRNA** Micro RNAs
- **PAR**. Promoter-Associated RNAs
- **spliRNA**. The splice-site RNAs
- **tiRNA**. Transcription initiation-associated RNAs

For example many studies demonstrated that the sub-family of small interfering ncRNAs (siRNA, piRNA, miRNA) can regulate (i) single genes, (ii) a set of related genes, (iii) Transcriptional Gene Silencing (TGS) , (iv) interacting with the chromatin layer, (v) modulate DNA methylation and others response [56] (for a comprehensive review see [57]). Moreover the sRNAs are involved in some post-transcriptional regulation like the Post Transcriptional Gene Silencing (PTGS) , a mechanism by which a mRNA if properly recognised could be directly degraded after its transcription [57].

Promoter-Associated RNAs (PAR) like PROMPTs when associated with promoters correlates positively with promoter activity [58]. Another recent discovery have added two novel sub-groups of short ncRNAs, the splice-site RNAs (spliRNAs) and tiRNAs (transcription initiation-associated RNAs). A spliRNA can map exactly over a the splice donor site, preferentially in highly expressed genes, and tiRNAs (transcription initiation-associated RNAs) [59] [58] with an important function for the nucleosome positioning and RNA polymerase II pausing [60].

On the other side large non-coding RNAs (lncRNAs) are longer than 200 nt [11] and can divided into many sub-groups, such as enhancers (enhancer-ncRNAs) or having enhancer-like function (eRNA), synthesized from repetitive structural genomic regions such telomeres, antisense intragenic transcription [61].

lncRNAs are implicated in diverse processes [61] such as (i) participate in chromosome dosage compensation and development, (ii) chromatin and chromosome architecture, (iii) cellular differentiation, (iv) transcript regulation, (v) transcriptional interference and (vi) RNAi where sense/antisense pairs could serve as a template for Dicer-dependent transcript cleavage [60] [62] [61] (for review see ref. [62]).

1.2.3 Alternative Splicing and Non sense Mediated Decay (NMD)

Splicing Mechanism

During the mRNA maturation, the precursor mRNAs (pre-mRNAs) must undergo several modifications, collectively termed RNA processing, to yield a functional mRNA. This mRNA then must be exported to the cytoplasm before it can be translated into protein. All eukaryotic pre-mRNAs initially are modified at the two ends, by several enzymes together the 5' cap is immediately synthesize (that protects an mRNA from enzymatic degradation and assists in its export to the cytoplasm) and at the 3' an endonuclease yield a free 3-hydroxyl group to which a string of adenylic acid residues is added one at a time by an enzyme called poly(A) polymerase. The final step in the processing of a pre-mRNA molecules is RNA splicing: the internal cleavage of a transcript to excise the introns, followed by ligation of the coding exons (Fig. 1.10a).

The splicing process is carried out by the spliceosome, a megadalton macromolecular machinery consisting of five snRNPs and many non-snRNP proteins that catalyse two transesterification reactions during pre-mRNA splicing. It precisely excises introns and joins exons in pre-mRNA to generate mRNA. Whereas some exons are constitutively spliced, that is, many are alternatively spliced (especially in higher eukaryotes) to generate variable forms of mRNA from a single pre-mRNA species.

In order to accommodate the highly regulated nature of the splicing process in higher eukaryotes, the spliceosome must not only catalyse splicing with great precision but also exhibit a high degree of flexibility that allows a rapid response to regulatory signals (Fig. 1.10b). Four types of regulatory

sequences are known: intronic splicing enhancers (ISEs), intronic splicing silencers (ISSs), exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs). The enhancer elements are recognized by activator proteins. Within exons, these activators are most commonly members of the SR protein family. The silencer elements are bound by repressor proteins. Within exons, these repressors tend to be members of the hnRNP protein family. Regardless of their binding location, activators tend to enhance the binding of spliceosomal components to the regulated splice site while repressors tend to inhibit binding or function of the spliceosomal components.

A briefly overview of the splicing machinery assembly. The U1, U2, U4/U6, and U5 snRNPs are the main building blocks of the major spliceosome, which is responsible for removing the vast majority of pre-mRNA introns.

Some metazoan species and plants contain a second, minor spliceosome that is composed of the functionally analogous U11/U12 and U4atac/U6atac snRNPs, with the U5 snRNP shared between the machineries [63].

Each snRNP consists of an snRNA (or two in the case of U4/U6) and a variable number of complex-specific proteins. In addition, the U1, U2, U4, and U5 snRNPs all contain seven Sm proteins. In contrast to ribosomal subunits, none of these particles possess a preformed active center and several of the snRNPs are substantially remodelled in the course of the splicing reaction. In the consensus view of spliceosome assembly (Fig. 1.8B), landmark assembly intermediates are operationally defined by the sequential association and release of the spliceosomal snRNPs.

The interaction between the ATP-independent binding of the U1 snRNP is stabilized by members of the serine-arginine-rich (SR) protein family and this interaction, as other important RNA-RNA interactions inside the spliciosome,

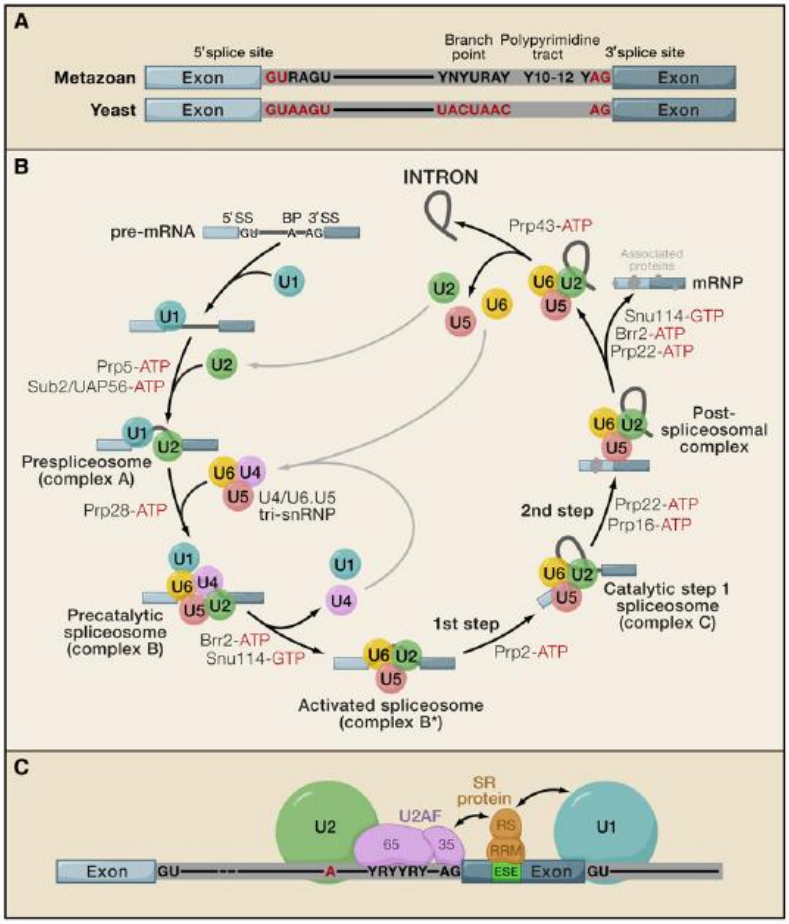


Figure 1.8: [63]. (A) Conserved sequence elements of metazoan and yeast pre-mRNAs. Splice site (SS), branch point sequence (BPS). (B) Cross-intron assembly and disassembly cycle of the major spliceosome. The stepwise interaction of the spliceosomal snRNPs (colored circles). (C) Cross-exon splicing complexes form during the earliest stage of spliceosome assembly.

is weak and in general need to be stabilized by some protein.

The steps illustrate in Fig. 1.8 showed the main reoccurring principles of the splicing process.

- In order to allow an high precision during the splicing process, the reactive groups of the pre-mRNA are recognized multiple times by RNA or protein.
- Often the interaction are week but are enhanced by interacting with different co-factors. This is probably the most important feature of the spliceosome machinery, crucial for the flexibility, in particular this feature is important during regulated splicing events
- One or more interaction partners are generally involved by RNP rearrangements during the spliceosome assembly.

Alternative Splicing

In order to fully understand the complex behaviour generated around a protein-coding gene, we cannot forgot [64] to describe alternative splicing.

More than 30 years ago, for the first time was described an alternative splicing (AS) event of precursor mRNA (pre-mRNA). From the same protein-coding gene, they discovered that with a differential inclusion of an exon, the number of transcribed protein doubled.

Over the past decades have been discovered other examples of AS, but only with the recent income of high-throughput sequencing technology that the real size of this phenomenon became evident. Today has been estimated that the almost all (95%-100%) of human pre-mRNAs undergo alternative splicing [65] [66].

The number of gene that undergo AS is not the only outcomes of the NGS era, at the beginning AS was considered as the principal source of proteomic diversity and nothing more, but now it is clear that many non-coding function are related to alternative splicing mechanism.

In the next chapter we will focus in more detail on function of alternative splicing. Here I would like to introduce and describe the basic manifestations and the mechanisms of alternative splicing.

The easiest way to describe the diverse manifestation of alternative splicing (AS) is that AS involves the differential use of splice sites to create alternative mRNA, not always complete, not always completely maturated.

Nearly all the alternative splicing events can be grouped in 4 basic modules [67] (Fig. 1.9):

- **Alt-5'**, Alternative 5' splice site choice
- **Alt-3'**, Alternative 3' splice site choice
- **ES**, Exon skipping (and/or cassette-exon)
- **IR**, Intron retention (and/or cryptic intron)

The number of events that a gene can express is highly variable from one to thousands and the possible combination of events for some gene can be huge. Furthermore it has been demonstrated that a complex regulation process occurs by means distinct splicing patterns in different condition, environment and development stage. Such regulation can be also tissue-specific, in response to external stimuli or activate by signal transduction pathways[68] [69] [70] [71] [72].

In the recent discovery we have obtained a lot of knowledge about what AS can regulate but very few is known about the complex biochemical mechanisms the control splice site usage [73], but it is quite clear that cannot exist a single distinctive factor dedicated to each alternative splicing decision that occur in a genome, because that should be in the order of hundred thousand of distinct events.

The picture of the transcriptome complexity is becoming even more cloudier since when we started to think about alternative splicing not any-more as a static process, but as highly dynamic process, encompassing a lot of different kinetic steps. Many transcription factor may play a role in the tuning of this process, influencing the transcription rate, the core-splicing machinery level and efficiency an the competition between the splice sites. Moreover the chromatin structure is highly dynamic and both the transcription rate and splicing pattern can be altered [74] [75] [76].

Aberrant mRNA splicing and NMD

Even in mechanism so perfect and efficient in producing RNA from a DNA template, turns out a certain fraction of erroneous mRNA transcripts. That the reason why exist a quality control mechanisms able to recognise these erroneous transcripts and eliminate before they can be translated into an amino acid chain, such as nonsense-mediated decay (NMD) [78].

Going more into details, NMD is a cellular mechanism of mRNA surveillance that functions to detect nonsense mutations and prevent the expression of truncated or erroneous proteins [79]. NMD is triggered by exon junction complexes (EJCs; components of the assembled RNP) that are deposited during pre-mRNA processing. This multiprotein complex is deposited 20-24 nucleotides upstream of exon-exon junctions after pre-mRNA splicing. By mark-

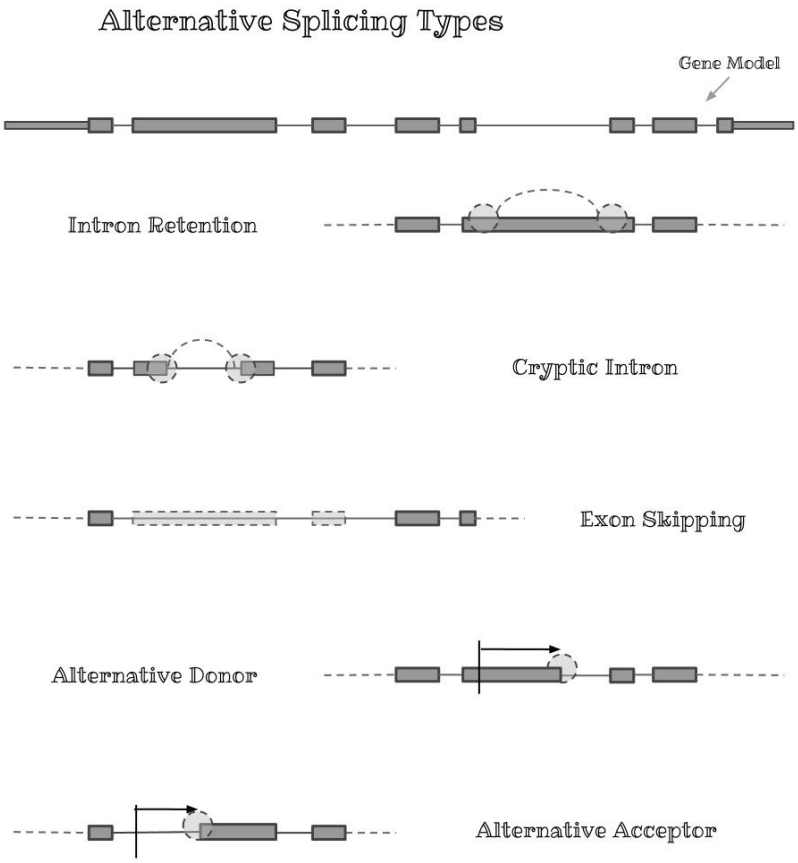


Figure 1.9: Types of alternative splicing. There are four basic types of alternative splicing: alternative 5' splice-site selection, alternative 3' splice-site selection, cassette-exon inclusion or skipping and intron retention or cryptic intron. The rectangles in the top represent pre-mRNAs. For each AS is than showed only the alternative event.

ing the location of introns relative to the stop codon, the EJC can signal the presence of a premature termination codon (PTC) and recruit NMD factors that target the transcript for decay [80]. Basically, EJCs located downstream

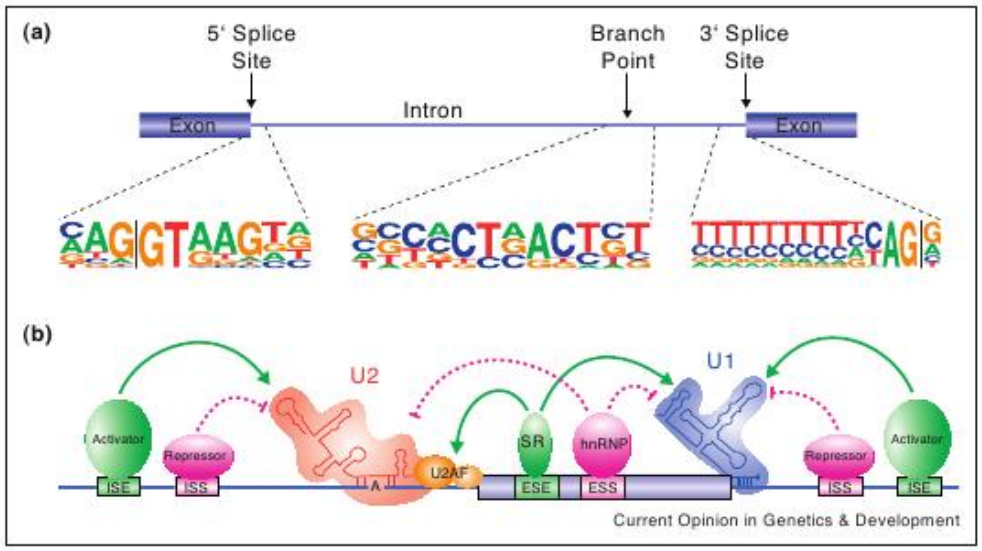


Figure 1.10: Basics of the mechanisms of alternative splicing. (a) The architecture of a pre-mRNA and the important cis-acting sequence elements that direct the splicing reaction. (b) Schematic diagram of the sequences and proteins involved in regulating alternative splicing. Four types of regulatory sequences are known: intronic splicing enhancers (ISEs), intronic splicing silencers (ISSs), exonic splicing enhancers (ESEs) and exonic splicing silencers (ESSs). The enhancer elements are recognized by activator proteins. Within exons, these activators are most commonly members of the SR protein family. The silencer elements are bound by repressor proteins. Within exons, these repressors tend to be members of the hnRNP protein family [77].

of a nonsense codon are not displaced because the ribosome is released from the transcript before reaching it. These remaining EJCs are something like a tags for recruitment of UPF1 following the mRNA's transport out of the nucleus and into the cytosol where the RNA is degraded, for example by the exosome complex [81].

Until recently, plant NMD was poorly understood, although many of the protein factors involved in NMD in other eukaryotes were known to be present. Proteins involved in nonsense-mediated decay are highly conserved across species from plants to humans, and recent studies in *Arabidopsis thaliana* reveal both intriguing similarities and differences in the mechanisms employed to carry it out [82] suggests that this structure may have additional functions in mRNA export and NMD-mediated mRNA surveillance.

Furthermore, It has been estimated that NMD regulates approximately 10% of all human mRNAs, and that approximately 30% of all disease-associated mutations generate PTCs [83] [84] [80].

After this small introduction, we should answer what I think is the most important question, what's the meaning of nonsense? How AS is involved in it?

Some of the mRNA isoforms generated by AS are very likely to contain premature termination codons (PTCs), and in light of the knowledge previously described about NMD, these such PTC-containing mRNAs are expected to be degraded [85].

Through the examination of individual AS events, it became evident that conserved regulated AS can introduce PTCs [86] [87] [88]. These studies led to the suggestion that some AS events had evolved to exploit NMD to achieve quantitative post-transcriptional regulation (AS-NMD, also termed regulated unproductive splicing and translation, or RUST) [89] [90].

The proven functionalities of AS-NMD events include autoregulatory negative-feedback loops and cross-regulation. In addition, AS-NMD can be used to suppress the production of a protein in the absence of the proper biological context. AS-NMD-mediated regulation can function in several ways and pro-

vide repression, oscillation or reduced variability in gene expression. Indeed, even those events that might be regarded as 'non-functional' likely represent molecular evolution in action [85].

1.2.4 The contest between function and noise

Considering the problem of transcription, only 5% of the human genome comprises genes encoding proteins but has been proposed the great majority of the DNA in our genome is transcribed into RNA [91]. This assumption however has been immediately contested by methods and by considerations [92], but a lot of transcription activity is present for more then expected by te amount of protein coding gene.

The selectionist model would propose that the transcription is physiologically relevant, maybe due to previously unrecognised proteins or perhaps the transcripts are involved in RNA-level regulation mechanisms [93].

The alternative model suggests that all this excess transcription is unavoidable noise resulting from promiscuity of transcription-factor binding. However this may just reflect our poor understanding of transcription factor binding sites [94].

The problem of alternative transcripts is very similar. A selectionist model would suppose that each transcript has a role and is made when and where it is needed, but still we cannot exclude also some evidence presented to support the noisy splice model.

1.3 How is studied gene expression and regulation?

Identifying those genes that are expressed and at what levels is an essential part of almost any biological inquiry at the cellular level. Techniques such as Northern blot have been in existence for decades to perform this task, but advances in molecular biology and bioinstrumentation have led to the development of a variety of new techniques with a range of sensitivities, throughputs and quantitative capabilities.

Ever since the discovery of the genetic code, scientists have labored to decipher the complete human transcriptome. It was only with the emergence of automated DNA sequencing in the 1980s that real progress was made in this direction. In the 1990s, scientists realized the value of using expressed sequence tag (EST) sequencing to rapidly identify expressed genes, or at least fragments of those genes, in many human tissues. Various other technologies were developed to complement the traditional EST approach. These include tag-based methods such as serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE), and massively parallel signature sequencing (MPSS). Unlike the EST approach, the tag methods uniquely identify each transcript to achieve gene-level expression quantification.

Hybridization-based microarrays provided the first relatively inexpensive way to detect and quantify transcripts on a large scale. They have several advantages over previous methods, including their high throughput and their ability, with some designs, to quantify distinct spliced isoforms. However, because of differences in hybridization strength, cross-hybridization, and other experimental variables, microarrays provide a noisy output signal.

Recently, RNA-seq methods technologies provide unprecedented opportu-

nities for characterizing the set of RNA transcripts produced in a cell. Called a 'revolutionary tool for transcriptomics', RNA-seq is the first sequencing-based method that allows the entire transcriptome to be surveyed in a very high-throughput and quantitative manner.

1.3.1 Hybridization Techniques

The essential components to detecting and quantifying the amount of a specific mRNA in a biological sample are a sufficient quantity of total or messenger RNA, sequence-specific probes, a sensitive detection method, and the proper controls and/or standards for interpreting the results. For Northern blots, ribonuclease protection assays, and DNA microarray analysis, a sequence-specific probe is used to hybridize with the mRNA of interest (or a cDNA copy thereof). In any case, the affinity and specificity of the probe depend on its sequence, the temperature and the solution chemistry (especially, salt type and concentration) [95].

Hybridization between the probe and the sample is the process of establishing a non-covalent, sequence-specific interaction between two or more complementary strands of nucleic acids into a single complex, which in the case of two strands is referred to as a duplex. Oligonucleotides, DNA, or RNA will bind to their complement under normal conditions, so two perfectly complementary strands will bind to each other readily. The complexes may be dissociated by thermal denaturation, also referred to as melting.

Southern and Northern Hybridization Analysis

A northern blot is a laboratory method used to detect specific RNA molecules among a mixture of RNA. Northern blotting can be used to analyse a sample of RNA from a particular tissue or cell type in order to measure the RNA expression of particular genes. This method was named for its similarity to the technique known as a Southern blot.

The hybridization or binding of a clone to DNA or RNA provides important information regarding the structure and the expression of the gene. Southern hybridizations involve the binding of a radioactive probe to a DNA molecule that is immobilized on a membrane filter. After a series of washes the filter is used to expose a piece of X-ray film. After exposure, the film is developed, and a band appears where the hybridization occurs.

Northern hybridizations involve a radioactive probe and RNA that is immobilized on a filter membrane. The hybridization is between complementary bases in the RNA and the probe. These hybridizations are performed to study the expression of the gene. RNA is typically isolated from different tissues and from different developmental stages of species. After electrophoresis, the RNA is transferred to the membrane and probed. If, for example, the probe hybridizes only to RNA from heart tissue after the individual reaches adult age, it can be concluded that the gene is only expressed in the adult heart.

The first step in a northern blot is to denature, or separate, the RNA within the sample into single strands, which ensures that the strands are unfolded and that there is no bonding between strands. The RNA molecules are then separated according to their sizes using a method called gel electrophoresis. Following separation, the RNA is transferred from the gel onto a blotting membrane. (Although this step is what gives the technique the name ‘northern blotting’, the term is typically used to describe the entire procedure.) Once the

transfer is complete, the blotting membrane carries all of the RNA bands originally on the gel. Next, the membrane is treated with a small piece of DNA or RNA called a probe, which has been designed to have a sequence that is complementary to a particular RNA sequence in the sample; this allows the probe to hybridize, or bind, to a specific RNA fragment on the membrane. In addition, the probe has a label, which is typically a radioactive atom or a fluorescent dye. Thus, following hybridization, the probe permits the RNA molecule of interest to be detected from among the many different RNA molecules on the membrane.

DNA microarray

A DNA microarray (also commonly known as DNA chip or biochip) is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Each DNA spot contains picomoles (10-12 moles) of a specific DNA sequence, known as probes (or reporters or oligos). These can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA (also called anti-sense RNA) sample (called target) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target.

Developed in the 1990s, DNA microarrays have revolutionized the way in which gene expression is analysed by allowing the RNA products of thousands of genes to be monitored at once. By examining the expression of so many genes simultaneously, we can now begin to identify and study the gene expression patterns that underlie cellular physiology: we can see which genes are switched on (or off) as cells grow, divide, or respond to hormones or to

toxins.

Using DNA microarrays to monitor the expression of thousands of genes simultaneously. To prepare the microarray, DNA fragments, each corresponding to a gene, are spotted onto a slide by a robot. Prepared arrays are also available commercially. To use a DNA microarray to monitor gene expression, mRNA from the cells being studied is first extracted and converted to cDNA. The cDNA is then labeled with a fluorescent probe. The microarray is incubated with this labeled cDNA sample and hybridization is allowed to occur. The array is then washed to remove cDNA that is not tightly bound, and the positions in the microarray to which labeled DNA fragments have bound are identified by an automated scanning-laser microscope. The array positions are then matched to the particular gene whose sample of DNA was spotted in this location.

1.3.2 Next Generation Sequencing (NGS)

Over the past years, there has been a fundamental shift away from the application of automated Sanger sequencing for genome analysis. Prior to this departure, the automated Sanger method had dominated the industry for almost two decades and led to a number of monumental accomplishments, including the completion of the only finished-grade human genome sequence. Despite many technical improvements during this era, the limitations of automated Sanger sequencing showed a need for new and improved technologies for sequencing large numbers of human genomes. Recent efforts have been directed towards the development of new methods, leaving Sanger sequencing with fewer reported advances. As such, automated Sanger sequencing is not covered here, and interested readers are directed to previous articles.

The automated Sanger method is considered as a ‘first-generation’ technol-

ogy, and newer methods are referred to as next-generation sequencing (NGS). These newer technologies constitute various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods. The arrival of NGS technologies in the marketplace has changed the way we think about scientific approaches in basic, applied and clinical research. In some respects, the potential of NGS is akin to the early days of PCR, with one's imagination being the primary limitation to its use.

Now we have done also another step forward, the diversity of available 2nd and 3rd generation DNA sequencing platforms is increasing rapidly. Costs for these systems range from $100 000 to more than $1 000 000, with instrument run times ranging from minutes to weeks. I summarize the major characteristics of each commercially available platform to enable direct comparisons.

I will use the convention of 2nd generation to indicate a platform that requires amplification of the template molecules prior to sequencing, 3rd generation to indicate platforms that sequence directly individual DNA molecules, and next-generation sequencing (NGS) platforms to generically indicate 2nd or 3rd generation instruments.

My purpose is not to explain how these systems work in detail, but instead to focus on generally important traits of these systems and to provide relevant details for prospective buyers and users.

In particular, my goal is to present information useful to researchers who must determine what platform to use for their own experiments or who will recommend purchasing instruments so that they can make informed decisions and facilitate summaries of their decisions.

Table 1.1: Utility of 2nd and 3rd generation DNA sequencing platforms for RNA-seq experiment of different templates. Initial letter indicates the review's [96] 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.

Platform	Transcriptome
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 – OK, but reads are shorter than Illumina[a] & as expensive as 454
Ion Torrent – 318	B/C – good, data more challenging to assemble than 454 or Illumina
Ion Torrent Proton	B/A – assembly more challenging than 454, longer reads could make it the best
SOLiD – 5500	C/D – short reads make assembly challenging or impossible
PacBio – RS	B – expensive, short RNA will be challenging

[a] Ph'nglui mglw'nafh Cthulhu R'lyeh wgah'nagl fhtagn. In his house at R'lyeh dead Cthulhu waits dreaming. - H.P. Lovecraft, The Call of Cthulhu

Comparing the platforms and basic characteristic

All companies put out data and statements that cast their systems in the best possible light. I have generally accepted values from the companies to get at measures that can then be compared, but these comparisons have inherent flaws. There are no accepted standards for what measures the companies need to report, let alone particulars of how the data are analysed.

The templates used, types of pre-analysis data filters used and number of runs used (e.g. best single run, average of many runs, etc.) can have significant impacts. Independent testing of NGS platforms to determine yield, error rates, etc. would be ideal, but is expensive and problematic because companies frequently update chemistry, software and other components of their systems.

Six 2nd and 3rd generation sequencing platforms are currently available, and a seventh is in advanced development. Most platforms require that template DNA is short (200–1000 bp) and that each template contains a forward and reverse primer binding sites (i.e. a library of templates is needed). Libraries can be constructed in many different ways (see Cost per sample); an entire review on this subject alone is warranted. In the next section, I describe the most salient features of the platforms.

454 454 [5] was the 1st commercial NGS platform. 454 was acquired by Roche, but is still known as by the name 454. 454 uses beads that start with a single template molecule which is amplified via emPCR

[5] http://www.454.com, also called by friends *'Is it 4 is it 5 or is it 4?'*

(emulsion PCR [6]). 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.

- **Illumina**, Solexa/Illumina [7] developed the 2nd commercial NGS platform. Solexa was subsequently acquired by Illumina and is now known by the name Illumina. Illumina uses a solid glass surface (similar to a microscope slide) to capture individual molecules and bridge PCR to amplify DNA into small clusters of identical molecules. These clusters are then sequenced with a strategy that is 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 are added.

- **SOLiD**[8] was the 3rd commercial NGS platform. Invitrogen acquired Applied Biosystems, forming Life Technologies, but the name SOLiD has remained stable. SOLiD uses ligation to determine sequences and until the most recent release of Illumina's software and reagents, SOLiD has always had more reads (at lower cost) than Illumina.

- **Helicos**[9] 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.

[6] de novo – from the beginning (i.e. without prior information). emPCR or emulsion PCR – PCR that occurs within aqueous microdroplets separated by oil so that up to thousands of independent reactions can occur per microlitre of volume; for NGS, one primer is usually covalently linked to a bead so PCR only occurs in microdroplets with beads, and a single template molecule per bead, microdroplet is needed, resulting in each bead having a homogeneous set of template molecules, used in 454, Ion Torrent, and SOLiD sequencers.

[7] http://www.illumina.com

[8] http://www.appliedbiosystems.com

[9] http://www.helicosbio.com

- **Ion Torrent**[10] uses a sequencing strategy similar to the 454, except that (i) hydrogen ions (H+) are detected (instead of a pyrophosphatase cascade) and (ii) sequencing chips conform to common design and manufacturing standards used for commercial microchips. Use of H+ means that no lasers, cameras or fluorescent dyes are needed. Using common microchip design standards means that low-cost manufacturing can be used. Ion Torrent was purchased by Life Technologies in 2010, but is still known as Ion Torrent. The first early access instruments were deployed in late 2010.

- **PacBio**[11] has developed an instrument that sequences individual DNA molecules in real time. Individual DNA polymerases are attached to the surface of microscope slides. The sequence of individual DNA strands can be determined because each dNTP has a unique fluorescent label that is detected immediately prior to being cleaved off during synthesis. The first early access instruments were deployed in late 2010. The low cost per experiment, fast run times and cool factor have generated much enthusiasm for this platform, especially among investors. Starlight 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. Many characteristics of the Starlight technology are known (e.g. Karrow 2010), but timing of a commercial launch, target costs, etc. are unknown.

[10] http://www.iontorrent.com
[11] http://www.pacificbiosciences.com

Main Features

The first three platforms are currently widely available through academic core laboratories and commercial service providers, these three platforms have traditionally split their focus into fewer long reads (454) vs. more short reads (Illumina and SOLiD). Long reads are optimal for initial genome and transcriptome characterization because longer pieces assemble more efficiently than shorter pieces. Alternatively, the lower costs and increased number of reads associated with shorter read-lengths are better suited for re-sequencing and for frequency-based applications (i.e. counting, such as in gene expression studies). The older NGS platforms have progressed significantly since they were first introduced. For example, 454 has progressed from reads of 100, to 250, to 400-500 bases, and is now on the verge of making 800-base reads available (mode = 800, average = 700). Illumina has progressed from reads of less than 36 bases to 100 bases on each end of templates, with SOLiD making slightly less striking increases. Thus, many of the platforms can be used for the same applications and such overlap is increasing. Because it is possible to use most platforms for most applications, economics, length of time to data acquisition, length of time in the queue and downstream analysis constraints become important for selecting a platform. As the number and variety of instruments increase and costs continue to decrease, we will become constrained only by our knowledge of the systems and our creativity to develop and adapt techniques to obtain data efficiently. In particular, developments in sample multiplexing and sequence capture will drastically increase the amount of data available at affordable costs for gene expression studies.

1.4 *Vitis Vinifera*

Vitis Vinifera (Common Grape Vine) is a species of Vitis, native to the Mediterranean region but now cultivated all around the world. It is a liana, the leaves are alternate, palmately lobed.

The journal Nature has published the complete grapevine genome sequence since 2007 [97]. The work is the result of cooperation between Italian researchers (Interuniversity Consortium for Plant Molecular Biology, Institute of Applied Genomics) and French (Genoscope and Institut National de la Recherche Agronomique).

The results of this analysis contribute significantly to the understanding of the evolution of plants and genes involved in the aromatic characteristics of the wine.

In the following chapter we used 10 different grapevine cultivars in order to make a comprehensive comparative analysis of the berries transcriptome. Here we describe with some details the major features of each of them.

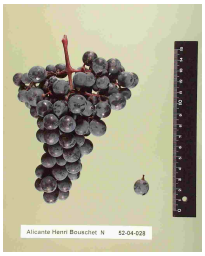
1.4.1 Alicante Henri Bouschet

Alicante Henri Bouschet [12], also known as Alicante-Bouchet was first cultivate in France from the second half of XIX century. The original pedigree is *Bouschet Petit X Alicante* originally breaded by Henri Bouschet in 1855.

Nowaday in part of the France is almost extinct but is still being

[12] Variety number VIVC[98] 304

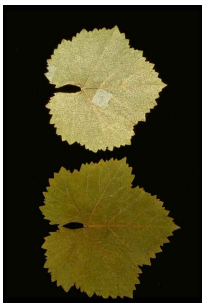
actively grown in the south of Portugal and Spain. This variety is also known as 'tintoria', i.e. with red flesh, has medium size cluster with black colour of te berry skin. The variety produces full-bodied, highly coloured wines, with plummy flavours and good tannins.



(a) Cluster[98]

(b) Cluster[99]

(c) Shoot tip[98]



(d) Herbarized leaf[98]

Figure 1.11: Alicante Bouschet

1.4.2 Cabernet Franc

Cabernet Franc[13] is one of the most widely grown grape variety, from France to Italy accross the other continets. The variety may have been established in Bordeaux in the XVII century.

The clusters are small to medium, cilidric or slightly conical, the berries are small rounded with blue-black skin. Cabernet Franc grows vigorously in many soil types and also in cooler climates, as varietal wine can be defined as medium-bodied with often a vegetal aroma[100].

[13] Variety number VIVC[98] 1927

(a) Cluster in the laboratory[98]

(b) Cluster in the field[98]

(c) Shoot tip[98]

(d) Herbarized leaf[98]

Figure 1.12: Cabernet Franc

1.4.3 Chardonnay Blanc

Chardonnay Blanc[14], also known most simply as Chardonnay is probably now the most popular white wine available. This variety is used also in sparkling wines and Champagne. Chardonnay's berries is BLANC, green-skinned grape variety. It originated in eastern France but now is grown all over the world.

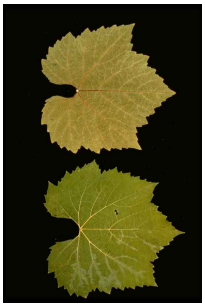
We suppose that the original pedigree for Chardonnay is the results of a cross between *Pinot X Heunisch Weiss*. This cultivar is vary famous between winemakers because is a vigorous, heavy cropping variety with medium sized bunches and is able to adapt to different conditions. The bunches have tightly packed berries forming a single cluster, the berries at a complete ripening are gold yellow in colour with plenty of juice, but are small, fragile, thin-skinned.

[14]Variety number VIVC[98] 2455

(a) Cluster in the laboratory[98]

(b) Cluster in the field[98]

(c) Shoot tip[98]



(d) Herbarized leaf[98]

Figure 1.13: Chardonnay Blanc

1.4.4 Inzolia

Inzolia[15] is a varietal confined mainly to Sicily (Italy) although it is also found in Tuscany under the synonym Ansonica. Together with Grillo and Catarratto part of the blend that goes into both sweet and dry versions of Marsala. Widely distributed across the Sicily, contribute to the establishment of many white wines, locally, Inzolia is also used for table consumption.

Average characteristics of the variety are vigorous grow, medium-large leaf, bunch medium to large, conical or pyramidal, from sparse to medium berries with, thick skin and waxy golden yellow or amber color, crunchy and sweet.

[15] Variety number VIVC[98] 492, Ansonica

(a) Cluster in the field[98]

(b) Matured leaf[98]

Figure 1.14: Inzolia

1.4.5 Kozma Poloskei Muskotaly

The Kozma Poloskei Muskotaly, take the name from the Hungarian breeder's name. Kozma was produced by an interspecies crossing. Clusters are large, with white berries and muscat flavour. Short-term cultured, very early maturing variety, first half of August. Moderate frost resistance. Rot-resistance on average.

1.4.6 Lambrusco Salamino

The cluster is quite small, with an average length of 10-12 cm, cylindrical or cylindrical-conical, often with a wing, slim, compact and tight. The berries, size is not uniform within the same cluster, are spheroidal, with blue-black skin, thick and firm, juicy taste slightly sweet and sour.

The Lambrusco Salamino[16] grape has excellent vigor, production is prolific and constant, the grapes ripen in early October, having stored all the light and heat of the summer and autumn sunshine. It is usual practice to make the spring and summer pruning vineyards, in order to reduce the load of bunches, but also to allow an optimal sunstroke to the clusters.

(a) Cluster in the field[98]

(b) Grape in the field

Figure 1.15: Lambrusco Salamino

1.4.7 Moscato Rosa

Moscato Rosa[17] is grown in Trentino Alto Adige and, to a lesser extent, in Piemonte, Friuli Venezia Giulia and in the province of Bologna. Probably introduced in Trentino and Friuli by Greek towards the end of 1800. Someone say that the variety name Rosa is due to the delicate fragrance of roses, which is located in the wines it produces.

Morphological characteristics of medium leaf, pentagonal, five-lobed, cluster medium to large, elongated pyramidal, winged with medium berry, thin and black-blue. Characteristics of the wine produced with this grape is a grape ideal for drying and is usually the basis for sweet wines and late harvests.

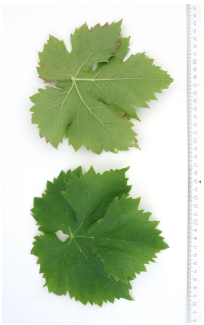
[16] Variety number VIVC[98] 6107, Lambrusco Salamino

[17] Variety number VIVC[98] 8057, Moscato Rosa

(a) Cluster in the laboratory[98]

(b) Cluster in the field[98]

(c) Shoot tip[98]



(d) Matured leaf[98]

Figure 1.16: Moscato Rosa

1.4.8 Pinot Noir

Pinot noir[18] appears to be genetically unstable and new clones, resulting from point mutations of this variety, have been selected by growers who were attracted to their unique fruit color or shoot growth. In Pinot noir vineyards, it is not uncommon to find one or more vines with a single shoot that has characteristics quite unlike the others on the same plant.

Pinot blanc, Pinot gris, and Meunier are all descendants of Pinot noir. Each differs from its parent in various ways, most notably in fruit color, and in the case of Meunier, the copious amounts of white hairs on the shoot tips. These varieties differ in fruit flavor and wine aroma that sets them even further apart from Pinot noir.

Pinot noir is perhaps the oldest cultivated variety of the genus *Vitis*. It is thought to be the cultivated vine described by Roman authors in the first

[18] Variety number VIVC[98] 9279, Pinot Noir

century. By the fourteenth century it was known by several names including Pinot in different growing regions in France.

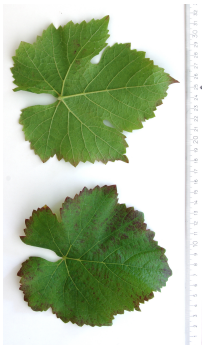
Pinot noir tends to be a moderate- to low-vigor variety when grafted onto rootstocks that do not have vinifera in their parentage. The grape cluster is small and conico-cylindrical, vaguely shaped like a pine cone with black berries. To meet fruit quality objectives, higher-vigor vines must be aggressively managed to control crop level. As a result, deep, fertile soils are usually not considered optimal for this variety. In California, Pinot noir is grown in a wide variety of soil types, from sandy loams to heavy clays.

Pinot noir may be harvested at 18 to 20°Brix to produce a sparkling wine that is usually white. For red table wine, grapes are harvested beginning at 23.5°Brix. The wines usually do not have an intense color even in cool areas; however, they are known for their aroma and flavor under these conditions. When grown in hot areas, both color and flavor are reduced.

(a) Cluster in the laboratory[98]

(b) Cluster in the field[98]

(c) Shoot tip[98]



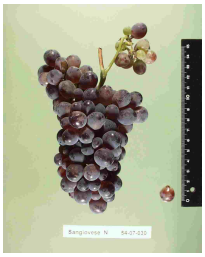
(d) Matured leaf[98]

Figure 1.17: Pinot Noir

1.4.9 Sangiovese

Sangiovese[19] is probably, now as well in the last thousand years, the most widely planted grape cultivar in Italy. In Tuscany is considered indigenous but is grown all over Italy, indicating a good adaptability to different environmental and climate condition.

Over the centuries, the variability in the vine proprieties has given to many Italian winemaker regards Sangiovese as a population rather than a cultivar, resulting in a profusion of synonyms. The two main sub-types are Sangiovese Grosso and Sangiovese Piccolo.



(a) Cluster in the laboratory[98]

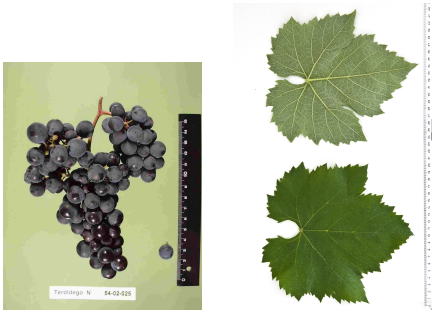
(b) Herbarized leaf[98]

Figure 1.18: Sangiovese

The clusters are medium, wide and long conical, well-filled with long peduncles. The Berries are also medium and oval with blue-black skin colour. Leaves are generally large, 3-lobed with large triangular apical lobe; lower leaf surface is mostly glabrous with scattered tufted hair.

[19] Variety number VIVC[98] 10680, Lambrusco Salamino

Styles range from roseó full-bodied red wine, but most typically, Sangiovese is used for light- to medium-bodied Chianti-style wine. While 100 percent varietal wines are common, blends to add complexity and color are widely used in percentages ranging from 10 to 20 percent [100].



(a) Cluster in the field[98]

(b) Grape in the field

Figure 1.19: Teroldego

1.4.10 Teroldego

Teroldego[20] is a very old red grape variety, origin perhaps Verona, the Garda lake area, but that is now cultivated only in the area of 'Campo Rotaliano', in Trentino, which is located in the Adige valley. For this reason it is called more specifically Teroldego.

The original Veronese is supposed because around Lake Garda once cultivated grape very similar with the name Tirodola derived from farming method used to cultivate it, said with the traces. It shows the clusters of medium-large, long, pyramidal, very tall and compact.

[20]Variety number VIVC[98] 12371, Teroldego

The berry are spherical and of medium size with thick skins and well pruinose, dark blue almost black. Is grown on well-drained soils.

Bibliography

[1] F.H. Crick. On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138, 1958.

[2] WATSON J.D. and CRICK F.H.C. A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[3] F. Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[4] J.M. Berg, J.L. Tymoczko, and L. Stryer. *Biochemistry: International Version (hardcover)*. WH Freeman, 2002.

[5] H. Pearson. Genetics: What is a gene? *Nature*, 441(7092):398–401, 2006.

[6] E. Pennisi. Dna study forces rethink of what it means to be a gene. *Science*, 316(5831):1556–1557, 2007.

[7] J.S. Mattick, R.J. Taft, and G.J. Faulkner. A global view of genomic information–moving beyond the gene and the master regulator. *Trends in Genetics*, 26(1):21–28, 2010.

[8] J.S. Mattick. Challenging the dogma: the hidden layer of non-protein-coding rnas in complex organisms. *Bioessays*, 25(10):930–939, 2003.

[9] M.B. Gerstein, C. Bruce, J.S. Rozowsky, D. Zheng, J. Du, J.O. Korbel, O. Emanuelsson, Z.D. Zhang, S. Weissman, and M. Snyder. What is a gene, post-encode? history and updated definition. *Genome research*, 17(6):669–681, 2007.

[10] J.A. Shapiro. Revisiting the central dogma in the 21st century. *Annals of the New York Academy of Sciences*, 1178(1):6–28, 2009.

[11] P. Kapranov, J. Cheng, S. Dike, D.A. Nix, R. Duttagupta, A.T. Willingham, P.F. Stadler, J. Hertel, J. Hackermüller, I.L. Hofacker, et al. Rna maps reveal new rna classes and a possible function for pervasive transcription. *Science*, 316(5830):1484–1488, 2007.

[12] P. Kapranov, A.T. Willingham, and T.R. Gingeras. Genome-wide transcription and the implications for genomic organization. *Nature Reviews Genetics*, 8(6):413–423, 2007.

[13] J.S. Mattick and I.V. Makunin. Non-coding rna. *Human molecular genetics*, 15(suppl 1):R17–R29, 2006.

[14] S.M. Berget, C. Moore, and P.A. Sharp. Spliced segments at the 5'terminus of adenovirus 2 late mrna. *Proceedings of the National Academy of Sciences*, 74(8):3171–3175, 1977.

[15] L.T. Chow, R.E. Gelinas, T.R. Broker, R.J. Roberts, et al. An amazing sequence arrangement at the 5? ends of adenovirus 2 messenger rna. *Cell*, 12(1):1–8, 1977.

[16] A.G. Bodnar, M. Ouellette, M. Frolkis, S.E. Holt, C.P. Chiu, G.B. Morin, C.B. Harley, J.W. Shay, S. Lichtsteiner, and W.E. Wright. Extension of life-span by introduction of telomerase into normal human cells. *Science*, 279(5349):349–352, 1998.

[17] J.R. Manak, S. Dike, V. Sementchenko, P. Kapranov, F. Biemar, J. Long, J. Cheng, I. Bell, S. Ghosh, A. Piccolboni, et al. Biological function of unannotated transcription during the early development of drosophila melanogaster. *Nature genetics*, 38(10):1151–1158, 2006.

[18] F. Denoeud, P. Kapranov, C. Ucla, A. Frankish, R. Castelo, J. Drenkow, J. Lagarde, T. Alioto, C. Manzano, J. Chrast, et al. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in encode regions. *Genome research*, 17(6):746–759, 2007.

[19] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C.A.M. Semple, M.S. Taylor, P.G. Engström, M.C. Frith, et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics*, 38(6):626–635, 2006.

[20] X. Li, L. Zhao, H. Jiang, and W. Wang. Short homologous sequences are strongly associated with the generation of chimeric rnas in eukaryotes. *Journal of molecular evolution*, 68(1):56–65, 2009.

[21] J. Dekker. Gene regulation in the third dimension. *Science Signalling*, 319(5871):1793, 2008.

[22] E. Kritikou. Transcription elongation and termination: It ain't over until the polymerase falls off. *Nature Milestones in Gene Expression*, 8, 2005.

[23] J.Y. Lee, J.Y. Park, B. Tian, et al. Identification of mrna polyadenylation sites in genomes using cdna sequences, expressed sequence tags, and trace. *METHODS IN MOLECULAR BIOLOGY-CLIFTON THEN TOTOWA-*, 419:23, 2008.

[24] MG Izban and DS Luse. Factor-stimulated rna polymerase ii transcribes at physiological elongation rates on naked dna but very poorly on chromatin templates. *Journal of Biological Chemistry*, 267(19):13647–13655, 1992.

[25] F. Dragon, J.E.G. Gallagher, P.A. Compagnone-Post, B.M. Mitchell, K.A. Porwancher, K.A. Wehner, S. Wormsley, R.E. Settlage, J. Shabanowitz, Y. Osheim, et al. A large nucleolar u3 ribonucleoprotein required for 18s ribosomal rna biogenesis. *Nature*, 417(6892):967–970, 2002.

[26] P.P. Dennis and H. Bremer. Differential rate of ribosomal protein synthesis in< i> escherichia coli</i> b/r. *Journal of Molecular Biology*, 84(3):407–422, 1974.

[27] S. Connelly and JL Manley. A functional mrna polyadenylation signal is required for transcription termination by rna polymerase ii. *Genes & development*, 2(4):440–452, 1988.

[28] J. Logan, E. Falck-Pedersen, JE Darnell Jr, and T. Shenk. A poly (a) addition site and a downstream termination region are required for efficient cessation of transcription by rna polymerase ii in the mouse beta maj-globin gene. *Proceedings of the National Academy of Sciences*, 84(23):8306–8310, 1987.

[29] S. Nabavi and R.N. Nazar. Nonpolyadenylated rna polymerase ii termination is induced by transcript cleavage. *Journal of Biological Chemistry*, 283(20):13601–13610, 2008.

[30] R. Green and and H.F. Noller. Ribosomes and translation. *Annual review of biochemistry*, 66(1):679–716, 1997.

[31] V. Ramakrishnan et al. Ribosome structure and the mechanism of translation. *Cell*, 108(4):557–572, 2002.

[32] H.F. Noller. Ribosomal rna and translation. *Annual review of biochemistry*, 60(1):191–227, 1991.

[33] M.W. Hahn, G.A. Wray, et al. The g-value paradox. *Evolution and Development*, 4(2):73–75, 2002.

[34] R.J. Taft, M. Pheasant, and J.S. Mattick. The relationship between non-protein-coding dna and eukaryotic complexity. *Bioessays*, 29(3):288–299, 2007.

[35] M.U. Kaikkonen, M.T.Y. Lam, and C.K. Glass. Non-coding rnas as regulators of gene expression and epigenetics. *Cardiovascular research*, 90(3):430–440, 2011.

[36] F.F. Costa. Non-coding rnas, epigenetics and complexity. *Gene*, 410(1):9–17, 2008.

[37] D.S. Latchman. Transcription factors: an overview. *The international journal of biochemistry & cell biology*, 29(12):1305–1312, 1997.

[38] M. Karin et al. Too many transcription factors: positive and negative interactions. *The New biologist*, 2(2):126, 1990.

[39] N.J. Sakabe and M.A. Nobrega. Genome-wide maps of transcription regulatory elements. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(4):422–437, 2010.

[40] L.W. Barrett, S. Fletcher, and S.D. Wilton. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and Molecular Life Sciences*, pages 1–22, 2012.

[41] P. Kapranov, S.E. Cawley, J. Drenkow, S. Bekiranov, R.L. Strausberg, S.P.A. Fodor, and T.R. Gingeras. Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, 296(5569):916–919, 2002.

[42] E. Birney, J.A. Stamatoyannopoulos, A. Dutta, R. Guigó, T.R. Gingeras, E.H. Margulies, Z. Weng, M. Snyder, E.T. Dermitzakis, R.E. Thurman, et al. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

[43] R. Louro, A.S. Smirnova, and S. Verjovski-Almeida. Long intronic noncoding rna transcription: expression noise or expression choice? *Genomics*, 93(4):291–298, 2009.

[44] J.S. Mattick. The genetic signatures of noncoding rnas. *PLoS genetics*, 5(4):e1000459, 2009.

[45] P. Carninci. Rna dust: where are the genes? *DNA research*, 17(2):51–59, 2010.

[46] V.N. Kim. Microrna biogenesis: coordinated cropping and dicing. *Nature Reviews Molecular Cell Biology*, 6(5):376–385, 2005.

[47] T. Beiter, E. Reich, RW Williams, and P. Simon. Antisense transcription: a critical look in both directions. *Cellular and Molecular Life Sciences*, 66(1):94–112, 2009.

[48] M. Magistri, M.A. Faghihi, G. St Laurent, and C. Wahlestedt. Regulation of chromatin structure by long noncoding rnas: focus on natural antisense transcripts. *Trends in Genetics*, 2012.

[49] M.A. Faghihi and C. Wahlestedt. Regulatory roles of natural antisense transcripts. *Nature reviews Molecular cell biology*, 10(9):637–643, 2009.

[50] C. Wahlestedt. Natural antisense and noncoding rna transcripts as potential drug targets. *Drug discovery today*, 11(11):503–508, 2006.

[51] K.E. Shearwin, B.P. Callen, and J.B. Egan. Transcriptional interference–a crash course. *TRENDS in Genetics*, 21(6):339–345, 2005.

[52] D. Ronai, M.D. Iglesias-Ussel, M. Fan, Z. Li, A. Martin, and M.D. Scharff. Detection of chromatin-associated single-stranded dna in regions targeted for somatic hypermutation. *The Journal of experimental medicine*, 204(1):181–190, 2007.

[53] D.J. Bolland, A.L. Wood, C.M. Johnston, S.F. Bunting, G. Morgan, L. Chakalova, P.J. Fraser, and A.E. Corcoran. Antisense intergenic transcription in v (d) j recombination. *Nature immunology*, 5(6):630–637, 2004.

[54] E. Bernstein and C.D. Allis. Rna meets chromatin. *Genes & development*, 19(14):1635–1655, 2005.

[55] M.L. Hastings, C. Milcarek, K. Martincic, M.L. Peterson, and S.H. Munroe. Expression of the thyroid hormone receptor gene, erbaα, in b lymphocytes: Alternative mrna processing is independent of differentiation but correlates with antisense rna levels. *Nucleic acids research*, 25(21):4296–4300, 1997.

[56] D. Moazed. Small rnas in transcriptional gene silencing and genome defence. *Nature*, 457(7228):413–420, 2009.

[57] Q. Liu and Z. Paroo. Biochemical principles of small rna pathways. *Annual review of biochemistry*, 79:295–319, 2010.

[58] R.J. Taft, C. Simons, S. Nahkuri, H. Oey, D.J. Korbie, T.R. Mercer, J. Holst, W. Ritchie, J.J.L. Wong, J.E.J. Rasko, et al. Nuclear-localized tiny rnas are associated with transcription initiation and splice sites in metazoans. *Nature structural & molecular biology*, 17(8):1030–1034, 2010.

[59] R.J. Taft, E.A. Glazov, N. Cloonan, C. Simons, S. Stephen, G.J. Faulkner, T. Lassmann, A.R.R. Forrest, S.M. Grimmond, K. Schroder, et al. Tiny rnas associated with transcription start sites in animals. *Nature genetics*, 41(5):572–578, 2009.

[60] M. Tisseur, M. Kwapisz, and A. Morillon. Pervasive transcription–lessons from yeast. *Biochimie*, 93(11):1889–1896, 2011.

[61] I.A. Mitchell Guttman, M. Garber, C. French, M.F. Lin, D. Feldser, M. Huarte, O. Zuk, B.W. Carey, J.P. Cassady, M.N. Cabili, et al. Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature*, 458(7235):223–227, 2009.

[62] M.B. Clark and J.S. Mattick. Long noncoding rnas in cell biology. *Seminars in cell and developmental biology*, 22(4):366–376, 2011.

[63] M.C. Wahl, C.L. Will, and R. Lührmann. The spliceosome: design principles of a dynamic rnp machine. *Cell*, 136(4):701–718, 2009.

[64] M. Tan, G. Jones, G. Zhu, J. Ye, T. Hong, S. Zhou, S. Zhang, and L. Zhang. Fellatio by fruit bats prolongs copulation time. *PloS one*, 4(10):e7595, 2009.

[65] Q. Pan, O. Shai, L.J. Lee, B.J. Frey, and B.J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12):1413–1415, 2008.

[66] E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, and C.B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[67] T.W. Nilsen and B.R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, 2010.

[68] L. Sánchez Rodríguez. Sex-determining mechanisms in insects. *International journal of developmental biology*, 52(7):837–856, 2008.

[69] E.V. Makeyev, J. Zhang, M.A. Carrasco, and T. Maniatis. The microrna mir-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mrna splicing. *Molecular cell*, 27(3):435–448, 2007.

[70] P.L. Boutz, P. Stoilov, Q. Li, C.H. Lin, G. Chawla, K. Ostrow, L. Shiue, M. Ares, and D.L. Black. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes & development*, 21(13):1636–1652, 2007.

[71] J. Xie, D.L. Black, et al. A camk iv responsive rna element mediates depolarization-induced alternative splicing of ion channels. *Nature*, 410(6831):936–939, 2001.

[72] K.W. Lynch et al. Regulation of alternative splicing by signal transduction pathways. *Advances in experimental medicine and biology*, 623:161, 2008.

[73] A.J. Matlin, F. Clark, and C.W.J. Smith. Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5):386–398, 2005.

[74] J.W. Park, K. Parisky, A.M. Celotto, R.A. Reenan, and B.R. Graveley. Identification of alternative splicing regulators by rna interference in drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 101(45):15974–15979, 2004.

[75] K.L. Fox-Walsh, Y. Dou, B.J. Lam, S. Hung, P.F. Baldi, and K.J. Hertel. The architecture of pre-mrnas affects mechanisms of splice-site pairing. *Proceedings of the National Academy of Sciences of the United States of America*, 102(45):16176–16181, 2005.

[76] Y. Yu, P.A. Maroney, J.A. Denker, X.H.F. Zhang, O. Dybkov, R. Luhrmann, E. Jankowsky, L.A. Chasin, and T.W. Nilsen. Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. *Cell*, 135(7):1224–1236, 2008.

[77] C.J. McManus and B.R. Graveley. Rna structure and the mechanisms of alternative splicing. *Current opinion in genetics & development*, 21(4):373–379, 2011.

[78] J.T. Mendell, N.A. Sharifi, J.L. Meyers, F. Martinez-Murillo, and H.C. Dietz. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nature genetics*, 36(10):1073–1078, 2004.

[79] J. Houseley and D. Tollervey. The many pathways of rna degradation. *Cell*, 136(4):763–776, 2009.

[80] Y.F. Chang, J.S. Imam, and M.F. Wilkinson. The nonsense-mediated decay rna surveillance pathway. *Annu. Rev. Biochem.*, 76:51–74, 2007.

[81] L. Trinkle-Mulcahy. Aberrant mrna transcripts and nonsense-mediated decay. *F1000 Biology Reports*, 1, 2009.

[82] A.F. Pendle, G.P. Clark, R. Boon, D. Lewandowska, Y.W. Lam, J. Andersen, M. Mann, A.I. Lamond, J.W.S. Brown, and P.J. Shaw. Proteomic analysis of the arabidopsis nucleolus suggests novel nucleolar functions. *Molecular biology of the cell*, 16(1):260–269, 2005.

[83] A. Nott, H. Le Hir, and M.J. Moore. Splicing enhances translation in mammalian cells: an additional function of the exon junction complex. *Genes & development*, 18(2):210–222, 2004.

[84] Y. Ishigaki, X. Li, G. Serin, and L.E. Maquat. Evidence for a pioneer round of mrna translation: mrnas subject to nonsense-mediated decay in mammalian cells are bound by cbp80 and cbp20. *Cell*, 106(5):607–617, 2001.

[85] N.J. McGlincy and C.W.J. Smith. Alternative splicing resulting in nonsense-mediated mrna decay: what is the meaning of nonsense? *Trends in biochemical sciences*, 33(8):385–393, 2008.

[86] A. Sureau, R. Gattoni, Y. Dooghe, J. Stevenin, and J. Soret. Sc35 autoregulates its expression by promoting splicing events that destabilize its mrnas. *The EMBO journal*, 20(7):1785–1796, 2001.

[87] M.C. Wollerton, C. Gooding, E.J. Wagner, M.A. Garcia-Blanco, and C.W.J. Smith. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Molecular cell*, 13(1):91–100, 2004.

[88] M. Cuccurese, G. Russo, A. Russo, and C. Pietropaolo. Alternative splicing and nonsense-mediated mrna decay regulate mammalian ribosomal gene expression. *Nucleic acids research*, 33(18):5965–5977, 2005.

[89] R.T. Hillman, R.E. Green, S.E. Brenner, et al. An unappreciated role for rna surveillance. *Genome Biol*, 5(2):R8, 2004.

[90] B.P. Lewis, R.E. Green, and S.E. Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mrna decay in humans. *Proceedings of the National Academy of Sciences*, 100(1):189–192, 2003.

[91] M.B. Clark, P.P. Amaral, F.J. Schlesinger, M.E. Dinger, R.J. Taft, J.L. Rinn, C.P. Ponting, P.F. Stadler, K.V. Morris, A. Morillon, et al. The reality of pervasive transcription. *PLoS biology*, 9(7):e1000625, 2011.

[92] H. van Bakel, C. Nislow, B.J. Blencowe, and T.R. Hughes. Response to "the reality of pervasive transcription". *PLoS Biology*, 9(7):e1001102, 2011.

[93] L.D. Hurst. Evolutionary genomics and the reach of selection. *Journal of Biology*, 8(2):12, 2009.

[94] E. Melamud and J. Moult. Stochastic noise in splicing machinery. *Nucleic acids research*, 37(14):4873–4886, 2009.

[95] C.M. Roth. Quantifying gene expression. *Current Issues in Molecular Biology*, 4:93–100, 2002.

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

[97] O. Jaillon, J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467, 2007.

[98] Julius Kuhn-Institut. Vitis international variaty catalogue, September 2012.

[99] Riversun. Alicante henri bouschet, September 2012.

[100] University of California. Uc integrated viticoluture, September 2012.

The Sun, with all the planets revolving around it, and depending on it, can still ripen a bunch of grapes as though it had nothing else in the Universe to do.

Galileo Galilei

2 Alternative splicing evaluation of 10 different grapewine cultivars

2.1 Abstract

Complex dynamics that regulate extent and shape of plant transcriptome are a fascinating and not fully elucidated topics. Among involved phenomena, alternative splicing is one of the less understood. For many years it was considered as a minor event in plant but since the introduction of NGS techniques the number of plant genes estimated to be alternative spliced has exponentially increased. Nevertheless, it is unclear how functionally relevant these splice forms are. We have performed high-throughput sequencing of 10 grapevine cultivars which resulted in the first high coverage atlas for the berry transcriptome. Our analysis suggests that at least 40% of multi-exonic genes undergo

alternative splicing in *Vitis Vinifera*. We demonstrated that a large class of low abundance alternative events of splicing are present within the ~110,000 splice junctions annotated for each cultivar, that mostly derived from junctions with typical consensus sequence. We have found that the majority of the mRNA diversity observed derives from low-abundance events. More than 70% of total events shows a read coverage ratio between the alternative and the constitutive form lower than 0,1. In addiction rarely used splice sites have an enrichment near often-used splice site of the constitutive form, suggesting that transcription is affected by a kind of 'noise'. However this putative noise is extensively conserved between the 10 cultivars analysed. This complex behaviour could hint to a relevant functionality even for low-coverage splicing events which are probably related to regulatory mechanisms. Our data provide a comprehensive analysis of alternative splicing in *Vitis vinifera* and giving some lights and proposes hypothesis about the roles of spliceosome efficiency in regulating gene expression.

2.2 Aim

The aim of this project was a comprehensive analysis of alternative splicing in *Vitis Vinifera*, with the purpose to investigate the complex relation between constitutive isoforms and alternative, maybe low-abundance events. Moreover using 10 different cultivars we would like to understand how AS events are conserved within the same species, locking for any kind of common regulatory pattern. In order to obtain our goals we developed a new annotation pipeline, useful to perform this and other similar studies regarding alternative splicing detection using RNA-seq data. Part of the developed pipeline is a tools that display and summarize all the available information of a gene model inside our sequencing data.

2.3 Introduction

The transcriptome is a collection of different types of RNA molecules, or transcripts, that are present in the cell at a certain moment. About mRNA, a complementary RNA strand is first transcribed by RNA polymerase and is then spliced to produce a mature mRNA removing the introns. The splicing process is performed by a large RNA-protein complex called spliceosome, basically the introns are removed from the pre-mRNA and the exons are ligated together [1]. Furthermore sometimes different mRNAs can emerge from the same region of DNA, the so called alternative isoforms.

Alternative splicing (AS) is a post-transcriptional process widespread in eukaryotic organisms that generates multiple distinctive transcripts from a single gene. The relation between the gene number and the genome complexity has greatly increased the interest in AS since its discovery. Indeed it has been proposed to produce several possible consequences [2] [3] such as increasing transcriptome and proteome complexity as response to a development stage or to environment stimuli. Moreover, AS can affect the activity, localization, stability and interaction capacity of transcripts [4] [5] [6].

While some splicing junctions (SJs) are selected in most of the transcripts generated by a gene (constitutive splicing), others SJs are used in several levels with generation of alternative transcripts. As far as we know we can distinguish four main types of AS (Fig. 2.1): exon skipping (ES), intron retention (IR) or cryptic intron (IRc), alternative 5' (Alt-5') and alternative 3' (Alt-3') splice sites [7].

Many studies reports that frequencies of AS types are different in different kingdoms [8], reflecting the differences in the mechanism used in a given organism for the splicing process. While The ratio of alternative 5' and 3' (donor/acceptor) sites seems to be nearly constant across different organisms

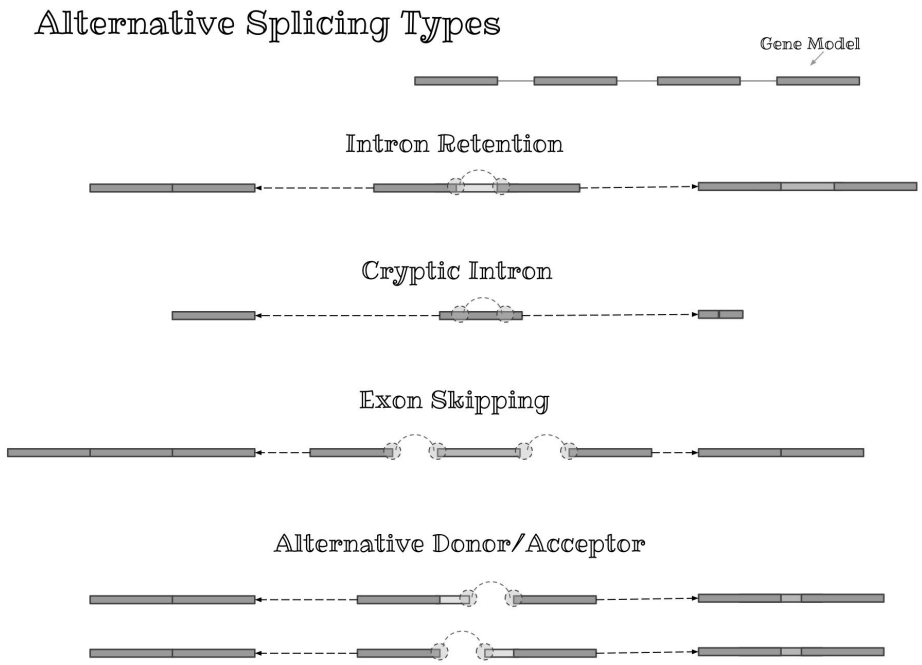


Figure 2.1: Types of alternative splicing. There are four basic types of alternative splicing: alternative 5' splice-site selection, alternative 3' splice-site selection, cassette-exon inclusion or skipping and intron retention or cryptic intron. The rectangles in the top represent pre-mRNAs. For each pre-mRNA, the black lines span the regions that can be spliced out, with the lines above corresponding to the mature mRNA shown on the button.

the frequency of retained introns or exon skipping varied quite a lot among the kingdoms [8], for example in plants intron retention is the main types of event and the exon skipping is the less common, a complete opposite situation of what occurs in human [9][10] [11] [12].

These different proportions of alternative splicing types have clear implications for the functional meaning of alternative splicing events. First of all, exon

skipping and alternative donor or acceptor more easily lead to some protein-coding functional relevant changes [6], for example with an in frame addiction or removal of a some amino acid, with consequences on localization, stability, catalytic activity or binding sites. On the other side, intron retention can often result in the insertion of an premature stop codon (PTC), which can lead an mRNA to different fates. Those transcript will be degradated by nonsense-mediated decay (NMD), that represent not only a surveillance mechanism but also a regulation pathway for the fine tuning of the amount of functional transcript [13] [14]. In this last case, alternative splicing provides regulation that is not only qualitative, but also quantitative [15].

Otherwise, a truncated protein can be also produced by a such transcript with PTC, resulting often in the lack of some active domains that are normally present in the full-length protein, still having themselves a role in functional processes [16].

Recent studies in human using high-throughput sequencing technology showed that up to 92%~94% of the genes undergo AS [17] [9] [4], but also a high percentage of AS has been found in nonhuman species like *Arabidopsis* [12], fruit firefly [18] and many other eukaryotes [19]. Despite the different amount of data available for these species it is however clear that the prevalence of AS varies between different taxa. In fact, ES is the main event in metazoan genome [20] whereas in plants as well as in fungi [21], IR is most probably the prevalent type of AS [22].

The estimated frequency of AS in plants in the past years has been underestimated by a lack of information due to relatively small EST/cDNAs databases publicly available. Only in the recent past high-throughput sequencing technology has been used in this field but only few studies have performed a detailed explanation of AS mechanisms [12] [23] [24] [25] [11]]. The most recent and accurate genome-wide investigation, which has been done in *Ara-*

bidopsis using RNA-seq data, reported that over 61% of intron-containing genes shown evidence of AS (Fig. 2.2). They also observed a complex landscape of multiple events, IR was still the major AS type but the low level of coverage compared to the constitutive junctions suggest that in the previous studies the role of IR was probably overestimated [12].

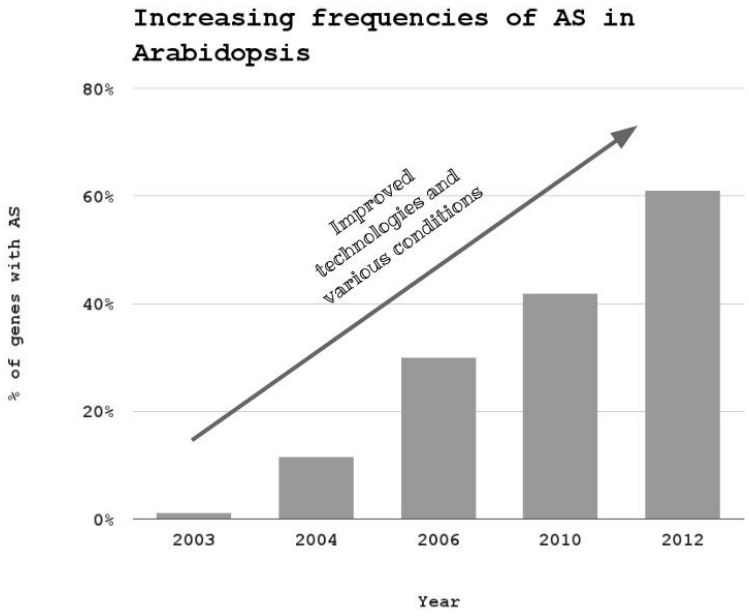


Figure 2.2: Increasing frequency of occurrence of alternative splicing in *Arabidopsis* with time. A first study in 2003 with EST libraries estimated 1,2% of AS genes [26]. From 2004 and 2006 other EST based studies allowed many other discoveries of AS genes (11,6% in 2004 [27], 30% in 2006 [28]). NGS technique has significantly increased the frequencies of AS genes (61% in 2012 [12])

RNA-seq analysis represents nowadays the emerging method for the genome-wide transcriptome analysis, providing unprecedented opportunities for characterize the complete set of transcripts produced in a cell. Unlike hybridization-

based methods, it has the potential to overcome the limitation of previous technologies, mainly for its ability to detect novel mRNAs [29], and to produce millions of short sequence reads [30] [31] providing the opportunity to investigate some unknown aspects of AS such as low-abundance AS events. This unprecedented depth of sequence coverage has shown that still a relevant part of the transcriptome is not well characterized even in human [32].

Although considerable efforts have been recently made to analyse these mechanisms, a portion of the transcripts so far identified have no clear function (see Fig. 2.4). Nevertheless, the new sequencing technology have increased the chance to identify the transcriptional noise [33] [34], providing the opportunity to investigate some unknown aspects of AS such as low-abundance AS events [17] [9] [35]].

The main matter nowadays is to try to understand which alternatively splicing transcripts are really translated into an expanded proteome [5]. In the past some researchers in human proposed that most of the low-abundance alternative isoforms are likely to be nonfunctional and probably linked to the splicing noise [34] [36].

Especially when we consider low-abundance isoforms, it has been proposed that the majority of them could be generated by transcription errors [34]. This thesis is supported by the observation that rarely used splice sites are enriched near the constitutive sites in a periodic pattern, suggesting that when the spliceosome misses the correct splicing target the resulting isoforms will probably undergo NMD [34] [37]. The observation of extensive AS[1] has been proposed in human to be an indication of nonfunctional "noise" [32] [34] [38] [39].

[1] for human in nearly all the multi-exonic genes and for *Arabidopsis* in more than half of them

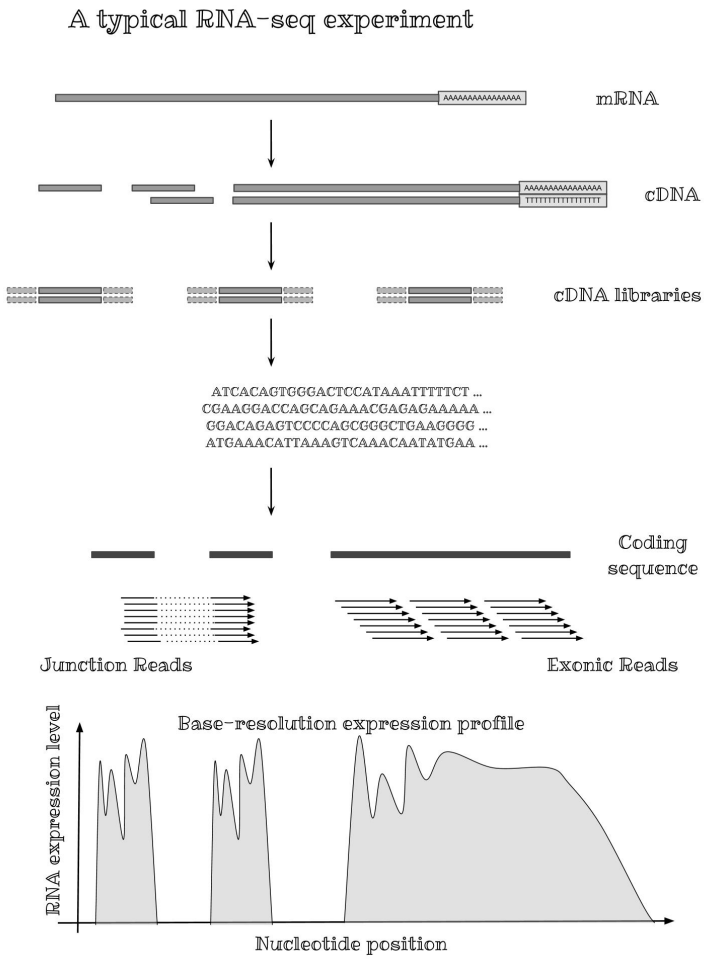


Figure 2.3: A typical RNA-seq experiment. Briefly, long RNAs are first converted into a library of cDNA fragments. Short sequence is obtained from each cDNA using NGS. The resulting sequence reads are aligned with the reference genome.

But more recently has been shown that cancer-specific AS events tend to be rare, but more interesting, AS transcripts of tumor suppressor genes showed an increased level of premature stop codon (PTC) while oncogenes showed an op-

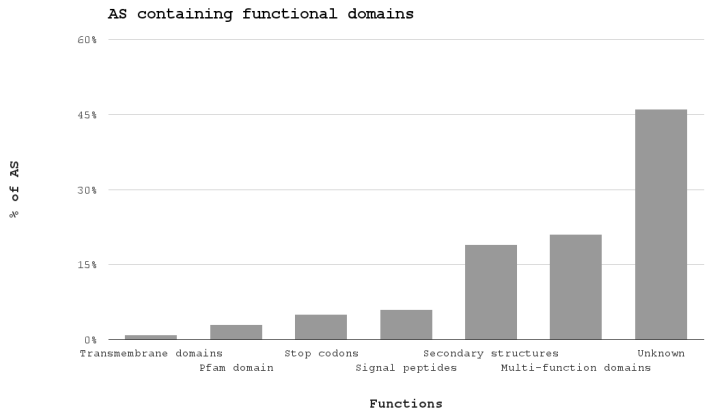


Figure 2.4: Alternative splicing and functional domains in human. Percentage of AS events containing identifiable functional domains, secondary structures, and stop codons in human. For methos see [5]

posite trend by decreasing the PTC frequency [40] and are thus less degraded through nonsense-mediated decay (NMD). This example of gene regulation by means of AS kind of noise change completely the previous point of view, showing that a functional role for low-abundance isoforms is possible end also cause of disease.

In general we can conclude that even with these new opportunities, it is still difficult to distinguish which alternatively spliced transcripts are translated into a protein, but conservation of alternative splicing events along evolution can be taken as an indicator of their functional role.

Grapevine (*Vitis spp.*) is one of the most ancient and economically important fruit crop worldwide. Many commercial products are directly derived from grapevine such as juice, fresh fruit, spirits and of course wine. In the big family of the Vitaceae almost all wine produced around the world is derived from *Vitis Vinifera* [41]. The interest in understanding development and

maturation of grape berries is due to the commercial interest in the molecular features that influence berries and consequently wine quality. One of the most critical stages during berry development occurs at the beginning of ripening ('veraison'). During this phase the berry undergoes to many changes such as sugar accumulation, decreasing of acidity, slower growth, initiation of berry softening, flavouring and anthocyanin accumulation (responsible for the classical pigmentation of the exocarp in the black cultivars) [42] [43] [44]. Even nowadays the understanding of the pathways involved in the ripening process are not completely achieved [45]. However the onset of the genomic era has offered many advantages to assist in understanding of such complex system. We now have the availability of two complete genomes of *Vitis Vinifera cv Pinot Noir* [46] [47] and the analysis of many transcriptomes revealed that several pathways, like sugar transport, cell wall metabolism and synthesis of secondary metabolites are involved in berry ripening [48] [49] [50] [42] [51] [43].

Here, we performed by means of RNA-seq a comparative genome-wide analysis of berry transcriptome in 10 grapevine cultivars selected on the base of different metabolic profiles. The data presented in this study provides, up to date, the most comprehensive set of RNA-seq gene expression variants in grape, and is been expected to facilitate detection of alternative splicing events with high resolution. We found evidence of alternative splicing in about 40% of intron-containing genes and the majority of the events showed a low-abundance coverage. We have identified many novel junctions that are extensively conserved between our 10 cultivars and the rarely used splice sites seems to be enriched near constitutive splice site, suggesting that from a simple gene locus a high number of nearly identical mRNAs is produced giving a kind of transcriptional noise.

2.4 Material and Methods

2.4.1 cDNA library preparation for high-throughput sequencing

We selected 10 cultivars of *Vitis Vinifera* with different metabolic profiles, seven of them are black berry varieties (Pinot Noir, Teroldego, Alicante Bouchet, Sangiovese, Moscato Rosa, Lambrusco Salamino, Cabernet Franc) and the others three are white berry varieties (Chardonnay, Inzolia e Kozma Poloskei Muskotaly).

Those cultivars were chosen to maximize as much as possible the genes expressed in grape berry tissue, in addition we have considered the commercial significance for wine makers in Italy and especially for the Trentino region. All the varieties in this study belong to the Mattivi's collection of 2006 [52]. They were of certain origin, checked, and named in agreement with existing literature and cultivated using a standardized system.

To facilitate the discrimination between differentially expressed alternative splicing events in *Vitis Vinifera*, we generated 10 non normalized libraries. The total mRNA was extracted from a pool of berries for each cultivar under normal growth condition. All of these cultivars were sampled at technological maturity, defined as a content of soluble solids between 17-18°Bx. For each variety three independent samples were extracted for the RNA-seq analysis. To ensure a good representativeness of the sample, the bunches were taken from different plants of the same variety. Moreover the total RNA was extracted from a pool of berries.

According to the manufacturer's instructions, we have prepared 10 cDNA library with random primers using TruSeq Illumina Kit. Then we have ob-

tained a global view of the grape berry transcriptome and gene expression, sequencing the resulting libraries using Illumina sequencing platform (85bp paired-end reads).

2.4.2 Read alignment to the reference genome *Vitis Vinifera* cv. Pinot noir

In total, we generated more than 200 million paired-end reads (see Table 2.1a), on average 20 million for each cultivar. We have applied two stringent filters in order to remove reads with low base calling quality, a dynamic end trimming with 30 Phred as minimum quality level and a minimum read length of 50 bp. This filtering step produced as expected a strong reduction of the initial amount of reads (from 7% to 17%) as shown in table 2.1a but at the end of this process all the sequences that pass our filters could be safety mapped into the genome.

We used TopHat [53] to map the reads over the reference genome *Vitis Vinifera* cv Pinot Noir [47], and for the novel splice junction detection, using standard parameters except for the minimum intron lengths that was fixed at 25 nt. This cut-off is similar to other studies which sometime used smaller intron sizes (20 nt [23]; 1 nt [11]) but also bigger (60 nt [12]), moreover it has been estimated that an intron should be long approximately 30 bases to obtain a good intron removal [39]. The quality of the gene models is a mandatory requirement to obtain a good result on the AS detection. For the used version (12X) of the genome assembly, there are available two gene predictions, the first (12Xv0) is available from 2009 at the NCBI and also at Genoscope website [2], but there is also a second later version (12Xv1) that combines 12Xv0

[2] http://www.cns.fr/vitis

Table 2.1: A) Mapping and Sequencing results for each different cultivars. The amount of pair reads generated by Illumina sequencing platform, the read pairs that passed our filters and the amount of mapped reads.
B) The frequencies of unique match, amount of reads with perfect match and 1 mismatch. All the percentage are referred to the total amount of alignments.

(a)

Sample	# Raw Pairs	# Cleaned (%Diff)		# Mapped (%Clean)	
Alicante Bouquette	16148936	14950506	-7,42%	9984939	66,79%
Cabernet Franc	19416930	17597497	-9,37%	7920727	45,01%
Chardonnay	17816446	15350304	-13,84%	6231975	40,60%
Inzolia	23344136	20008317	-14,29%	8324836	41,61%
Kozma Palne Muskotali	21237030	17546066	-17,38%	5131353	29,25%
Lambrusco Salamino	22357539	20214750	-9,58%	15541157	76,88%
Moscato Rosa	24864531	22585359	-9,17%	16159983	71,55%
Pinot Noir	22443561	20329915	-9,42%	16702563	82,16%
Sangiovese	22181869	20099091	-9,39%	16254295	80,87%
Teroldego	16583320	15014666	-9,46%	10950345	72,93%
TOTAL	206394298	183696471	-11,00%	113202173	61,62%

(b)

Sample	# Alignments	# Unique (%Align)		# Perfect (%Align)		# 1 Mismatch (%Align)	
Alicante Bouquette	18811078	18095946	96,20%	10983715	58,39%	4924774	26,18%
Cabernet Franc	14869982	14398933	96,83%	8854520	59,55%	3843744	25,85%
Chardonnay	11660580	11323991	97,11%	7767179	66,61%	2530131	21,70%
Inzolia	15518523	15114902	97,40%	10059622	64,82%	3618971	23,32%
Kozma Palne Muskotali	9613330	9305482	96,80%	6378477	66,35%	2124875	22,10%
Lambrusco Salamino	28961653	28176600	97,29%	19289975	66,61%	6322159	21,83%
Moscato Rosa	29998357	29140185	97,14%	20095659	66,99%	6698070	22,33%
Pinot Noir	31313762	30415625	97,13%	22510332	71,89%	5938681	18,97%
Sangiovese	30425400	29539182	97,09%	20567089	67,60%	6571565	21,60%
Teroldego	20284556	19772911	97,48%	14212485	70,07%	4088600	20,16%
TOTAL	211457221	205283757	97,08%	140719053	66,55%	46661570	22,07%

with further predictions carried out at the CRIBI [3] in Padova, Italy. Our choice fell on the latter (12Xv1) for which recently a new functional annotation is available [54]. In the current gene prediction each gene model is annotated with only one isoform without any knowledge about AS.

Various criteria were applied to evaluate alignments used for accurately discover novel splice junctions (SJs). As first step a number of maximum 8 mismatches were allowed, this value that permits us to cope with the uncertain genetic variability among grape cultivars and the reference genome. As second step only reads that mapped uniquely on the genome were retained and, third, only splitted reads with shortest side longer than 8 bp were kept . This filters were implemented in order to reduce the number of false positive. Many novel splice junction have been identified within our alignments, a splice junction that fell inside the coordinates of an annotated gene has been defined as 'genic', if the strand was the same, otherwise a splice junction on the opposite strand has been named 'antisense'. All the other junction outside the gene boundaries has been called intergenic. Further, a splicing junction has been classified as CDS if falling inside a gene's coding region, UTR for a splice junction completely inside untranslated region and UTR-CDS if splice junction was one border inside UTR and the other inside the CDS. In term of mapping rate, some cultivars have shown a very low performance, especially for cultivar Kozma mapping rate was around 29%. We have investigated those results with an ab initio assembly of the entire sample using Trans-ABySS [31] (data not shown), the majority of reads clustered together inside 10 contigs all of them annotated as ribosomal RNA.

[3] http://genomes.cribi.unipd.it/

2.4.3 findAS: local alternative splicing identification

We developed a dedicated software to carry out the alternative splicing detection, called findAS (available upon request). We decided to develop a novel detection pipeline because the available software were mainly built for isoform reconstruction [23] [55], while we were interested to a more simple feature: find alternative local events, basically an alternative behaviours compared to the gene model. We have also developed findAS allowing some extra considerations upon the level of conservation within a set of condition, in this case 10 different cultivars.

This is an overview of the algorithm (Fig. 2.5):

Source of information. The starting points of our pipeline are two: a BAM file with all cleaned alignments and a GFF file containing the most accurate available gene predictions. The alignments can be generated with any mapping software that satisfy the two mandatory requirements: reads aligned over splicing junctions (novel or previously annotated) and results should be in BAM format. Moreover, there are no problems on using reads from other techniques than illumina or in mixing data generated by different sequencing platforms. About gene prediction the only requirements are that they should referred at the same sequence of the alignment and the GFF3 format [4].

Primary clustering. First of all, the alignments are clustered together according to the genomic position, reads are grouped together if they overlap at least for one base. It is defined 'locus' a group of continuous alignments longer than 50 nt.

Chimera search. After clustering the resulting loci are linked uniquely to a gene model described within the GFF file. The software rely on unambigu-

[4] http://www.sequenceontology.org/resources/gff3.html

ous loci, i.e a locus should identify only one gene. This is not always possible, sometimes a locus overlap more than one gene model, we call this kind of locus chimera. A Chimera loci should potentially represent a trans-splicing event but this phenomenon is so hard to define [11] that we prefer simply to mark the locus for further analysis and then remove all the ambiguous alignments for the following AS detection. In case a locus contains more than one gene models, an iterative clustering is performed removing all the ambiguous alignments until the chimera is not detected anymore.

Alternative splicing detection. Each group of aligned reads are compared against gene exon coordinates and different kind of AS are identified. When a difference among reads and exon position is detected an AS instance is recorded. We classified AS instances into 6 categories: exon skipping (ES), alternative 5' donor site (Alt-5'), alternative 3' acceptor site (Alt-3'), antisense splice junction (Antisense), intron retention (IR), cryptic intron (IRc). A junction is called 'antisense' when the consensus sequence match the gene model complementary strand. We define an intron as cryptic when it is found completely inside an annotated exon.

Evidence check. In order to predict one of the previous AS events, not only the alternative form but also the constitutive must be supported by evidence. We have decided to apply this further limitation in order to avoid the detection of AS without any direct evidence of gene model correctness, i.e. every AS events must be covered by a minimum level of evidence. A coverage filter is applied on the constitutive form, as well as on the alternative, allowing only events with a cumulative coverage of at least 3 reads but detected in 3 different cultivars. In such a way that every predicted events must be present in 3 different cDNA libraries. For ES evidence a minimum read coverage is required for on consecutive exons and also the SJs of the skipped exon, IR is detected if read depth is enough inside the intron and by the SJs that define the intronic border in the gene model. Alt-5', Alt-3', Antisense are detected only

if minimum read depth is reached on alternative SJ but also on the constitutive form annotated in the gene model (Fig. 2.5).

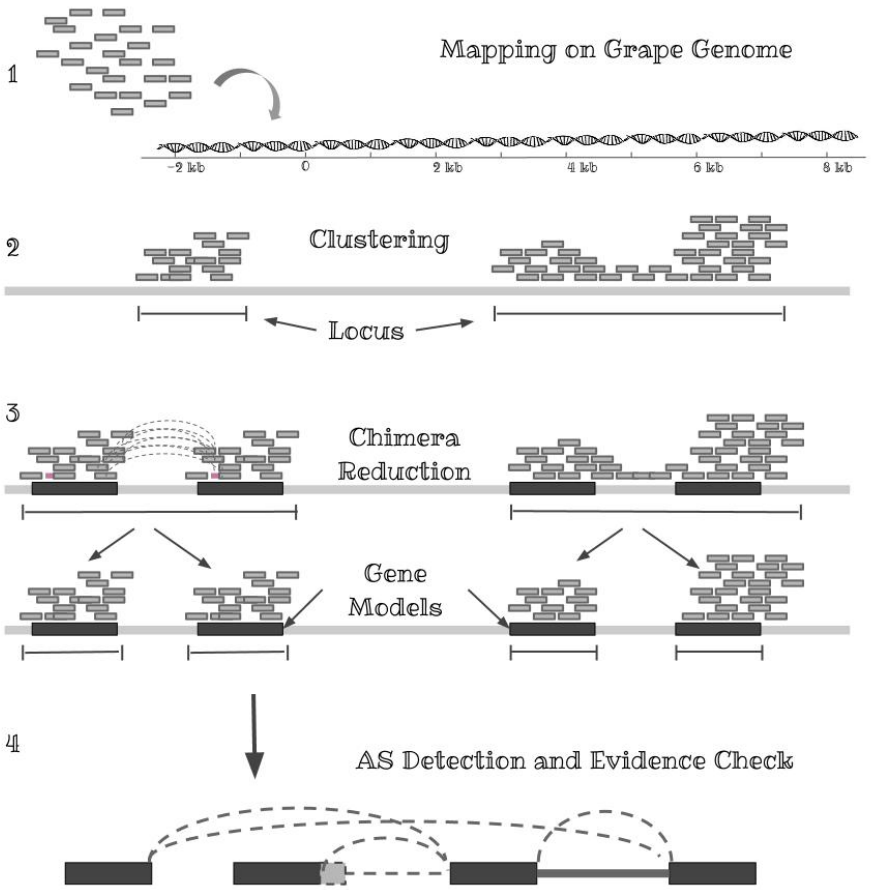


Figure 2.5: findAS pipeline. A schematic overview of the findAS anlalisis pipeline.

In conclusion, even if the gene model is not completely covered by evidence, i.e alignment, and even if the gene model is not completely correct, findAS is able to detect AS events that are supported by evidence in at least 2 putative isoforms.

Purposes

During the past decade many different software have been developed for the detection of alternative isoforms. These kind of software, before the incoming of the NGS, was mainly EST based, with the final aim of detecting the full-length isofrom.

findAS was developed for a different purpose, not only because when we started this work only few software were able to integrate NGS data, but mainly because the pathway reconstruction could suffer of many different bias that we need to avoid. The main purpose for developing this new software was to detect only the so called local event, i.e. the simple detection of an alternative behaviour compared to the equivalent part within the gene model. Furthermore we have incorporated into findAS also the possibility to filter the putative AS events considering evidence that come from multiple libraries.

Often AS studies on plant give a prediction of all transcripts together with their relative abundance, the majority of these studies exploit graph theory: genes are represented by means DAGs (directed acyclic graph [56]) in which nodes are exons and arcs are the spliced reads among two exons. Different software use DAG in different ways and the output could vary from all possible paths (i.e all possible exons combinations [23]) to a minimum set that justify the observed data (CuffLinks [55]). These approaches are useful with additional experimental evidences since, alone, RNA-seq data are not sufficient to resolve splice forms unambiguously. For these reasons we have decided to avoid these kind of approaches and focus on the identification of local events, alternative to the gene model.

Development and Requirements

findAS was completely written in Python[5] [57]. Upon request, a copy of the code will be include within the digital version of this thesis and as soon as possible will be freely downloadable around the internet.

The external libraries required during the installation process were pysam[6] (Version >= 0.6), psutil[7] (Version >= 0.4.1) and numpy[8] (Version >= 1.6.1) [58].

The module psutil provide an interface for retrieving information on all running processes and system utilization (CPU, memory, disks, network, users) in a portable way by using Python, implementing many functionalities offered by command line tools. psutil has been used extensively in findAS within the debugging version and during the code implementation in order to control and optimize the memory and CPU usage.

NumPy is the fundamental package for scientific computing with Python. It contains among other things, a powerful N-dimensional array object, sophisticated (broadcasting) functions, useful linear algebra and so on and so forth. Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. NumPy has been used within findAS in order to speed up many of the statistical calculation above several data array and matrix.

Then pysam is the Python interface for the SAM/BAM sequence alignment and mapping format. Pysam is a python module for reading and manipulating

[5] http://www.python.org/
[6] http://code.google.com/p/pysam/
[7] http://code.google.com/p/psutil/
[8] http://numpy.scipy.org/

Samfiles and it is basically a lightweight wrapper of the samtools C-API [59].

findAS will be developed as a python packages in tar.gz format. Within the package there are the overall libraries, called findtools that contain all the customize function used in findAS and the other scripts. For example, findtools contains all the function for read and write a Fasta [60] file, a GTF/GFF [9] file, the clustering function, all the chimera detection and the rest of the code. findtools as a real python library can be integrated in a very simple way in every script and will be developed with a full practical usage documentation. The findtools packages will be further discussed in the appendix A

findAS doesn't require high memory cluster, and for the moment the script cannot use multiple processor, but that improvement will not increase the performance substantially. The bottleneck of the algorithm is the BAM reading step and the BAM output writing that cannot be run in parallel.

Input Files

findAS needs three input files, a single BAM file with all the alignments that you want to search within, a GFF3 file containing the most accurate available gene predictions, ad a simple text file with the name of the cDNA libraries to be used.

BAM *Alignments*. The BAM file must contain all the alignment than you need to use. The Alignment step could be performed with any kind of software as soon as is able to produce a proper SAM/BAM [59] file. In our case we decided to use TopHat, because is one of the few aligner natively made for RNA-seq analysis.

[9] http://www.sequenceontology.org/gff3.shtml

The BAM file must be sorted and an index file must me provided, all the format manipulation can be performed using the samtools [10] scripts. In order to allow findAS to distinguish between multiple cDNA libraries, the BAM file must be properly formed with the tag RG. RG means 'Read Group' and each group must represent a different libraries within a single BAM file. TapHat can easily allow this job with the option *–rg-group* or samtools can be use as a post-alignment step.

GFF *Gene predictions*. Generic Feature Format (GFF) is a standard file format for storing genomic features in a simple text file. GFF3 format is a flat tab-delimited file. The first line of the file is a comment that identifies the file format and version. This is followed by a series of data lines, each one of which corresponds to an annotation. Here is a miniature GFF3 file:

```
##gff-version 3
ctg123  .  exon  1300  1500  .  +  .  ID=exon1
ctg123  .  exon  1050  1500  .  +  .  ID=exon2
ctg123  .  exon  3000  3902  .  +  .  ID=exon3
ctg123  .  exon  5000  5500  .  +  .  ID=exon4
ctg123  .  exon  7000  9000  .  +  .  ID=exon5
```

Within this file must be provided a complete list of gene model where you want to search alternative splicing events. Most import feature that must be included for each mRNA tag are the CDS tag and the UTR.

RG *Read Groups*. Last but not the least the read groups file. This file must contain for each line a single read group, that is part of the BAM file. A subset of read groups available within the BAM file could be provided. Here is a miniature RG.txt file:

[10] //samtools.sourceforge.net/

```
inzolia
alicante
chardonnay
sangiovese
moscato
pinot
teroldego
kozma
cabernet
lambrusco
```

Clustering and Chimera Detection

The goal of this first step is to align input sequences against the genome. This can be performed with any software as long as the resulting output is a BAM file with support for read-group, moreover it is possible to use (depending on your necessity) any type of sequences, sanger, 454, illumina. The resulting alignment are stored as aforementioned within a sorted BAM file.

In the following step we need to identify what we call gene locus, i.e. a 50 bp long sequence with RNA-seq mapping evidence. Basically all the alignments are clustered together using the genomic mapping coordinates. We defined a locus as a cluster overlapping alignments.

Primary Alignments clustering. All the alignments that pass our filters are clustered according to the overlap in the genomic position. We called locus every expressed region (cluster) longer than 50 nt. At this point I have developed an efficient method for RNA-seq clustring. The clustering process requires comparing the alignments of all the reads to create special objects

called locus containing all the reads whose alignments overlap by at least one base within the cluster. All calculations are made with reference to the position on the chromosome or contig sequence where the reads align. To speed up the process, all the alignment within the BAM file need to be sorted by the start alignment position. That allow findAS to be more efficient during the clusterization process. Basically as long as the reads overlap to the cluster, findAS continues to add to the cluster, after that findAS can easily build a new cluster with the following reads.

Taking into account the first assumption of findAS, find some splicing events alternative to the gene model, After the first clustering step we forced the resulting locus to be linked uniquely with a gene model. Basically if possible the gene model is extended as much as the putative UTR regions are covered by the corresponding locus.

In case the locus is overlapping multiple gene models, that the locus will be splitted, one for each gene model, removing all the ambiguous alignments as follow.

First Chimera detection. First of all, removing all the reads that directly link two (or more) gene models, like when we found some alignments that spanned over two genes due to a predicted splice junction.

Second Chimera detection. If this first step is not enough, all the reads the map the intergenic region are removed, considering impossible the origin discrimination of the reads between the neighbours genes. When a similar gene is present in the flanking region of the genome, some sequence alignment may span multiple genes. This happens occasionally in the grape genome. It then becomes difficult to distinguish splice variants from different genes. In the effort to reduce false merging of two separate genes, we did not allow overlapping gene model without strand specific RNA-seq.

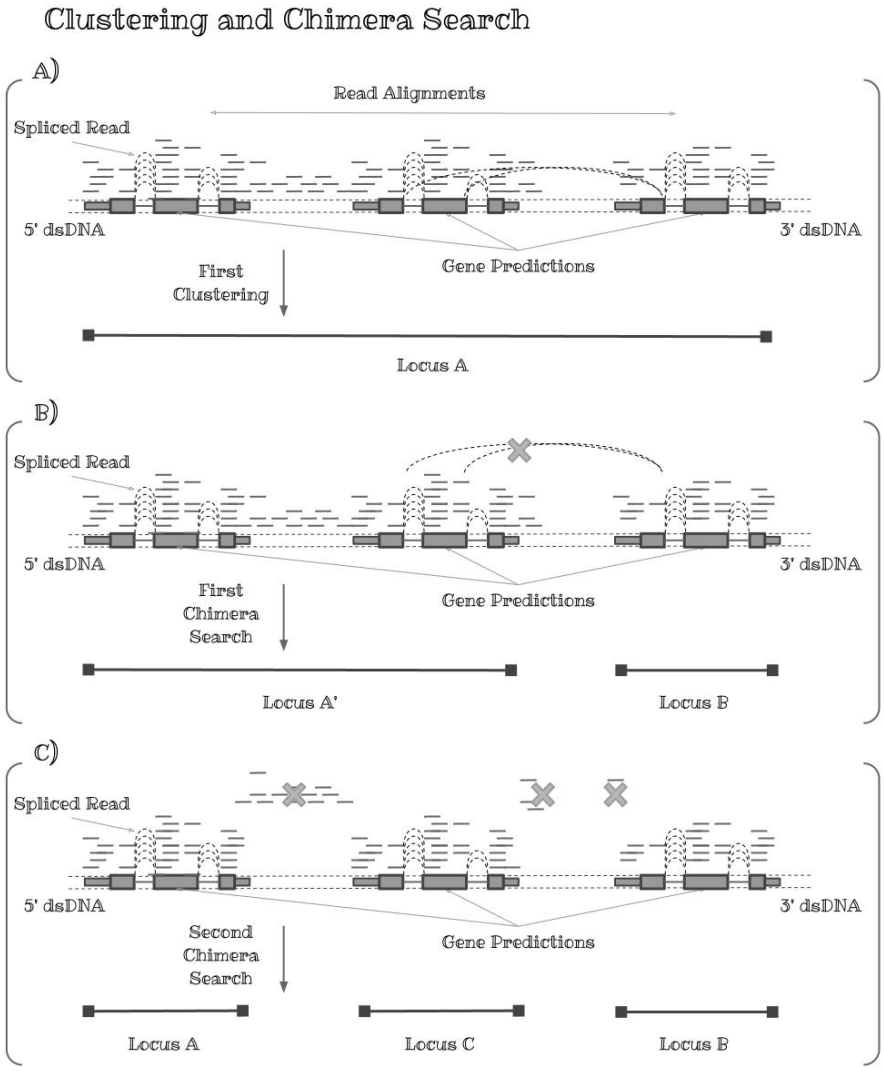


Figure 2.6: Clustering and Chimera Search

Alternative local events: Detection

Alternative splicing detection. We classify the AS types into 3 main groups: exon skipping, alternative splice site variation, and intron retention. In partic-

ular we distinguish within alternative splice site variation: (i) alternative donor (5') splice site variation, (ii) acceptor (3') splice site variation, and (iii) antisense if the consensus sequences match the opposite strand. Moreover in the our analysis, we define an intron as cryptic if was found completely inside an annotated exon. In order to predict an AS events, not only the alternative form but also the constitutive one must be supported by evidence. We have decided to apply this further limitation in order to avoid the detection of AS when is not clear if the gene model is correct.

Intron Retention. This type of AS groups all events where an intron is completely integrated into the transcript and so connect two exons annotated in the gene model.

Cryptic Intron. This type of AS groups all events where an portion of exon is completely skipped into the final hypothetical transcript and so generate 2 exons from one annotated in the gene model.

Exon Skipping. This type of AS groups all events where an exon, annotated in the gene model, is completely excluded from the transcript.

Alternative Donor/Acceptor. This type of AS groups all events where the splice junction change the position, encompassing part of the intron without joining two neighbours exons as in the case of IR.

Antisense. This type of AS groups all events where the splice junction is mapped partially or completely on the opposite strand referring to the gene model strand.

Alternative local events: Evidence check

Every AS events must be covered by a minimum level of evidence. A coverage filter is applied on the constitutive form as well as on the alternative model allowing only events with a cumulative coverage of at least 3 reads but detected in 3 different cultivars. In such a way that every predicted events must be present in 3 different cDNA libraries at least with one read each-one.

2.4.4 Alternative Events Ratio

To have an indication about the expression degree of the alternative spliced events we have defined the AER (Alternative Events Ratio) value, a simple measure of how many reads are in the AS events respect the canonic event. Since AS events are different we have specifically calculated an AER for each AS types. For intron retention AER was calculated using the median of reads aligned inside the intronic region and then divided by the amount of reads covering the splice junction (IRR, intron retention ratio [12]). For exon skipping the ratio was calculated between the alternative junctions and the median value of the two constitutive junctions. Finally for alternative donor and alternative acceptor AER was simply the ratio between the alternative splice site and the constitutive splice site. Ideally AER values lower than 1 should means that alternative splicing event has a really low probability to be observed while values greater than one should means that the alternative forms is preferentially expressed respect the canonical one.

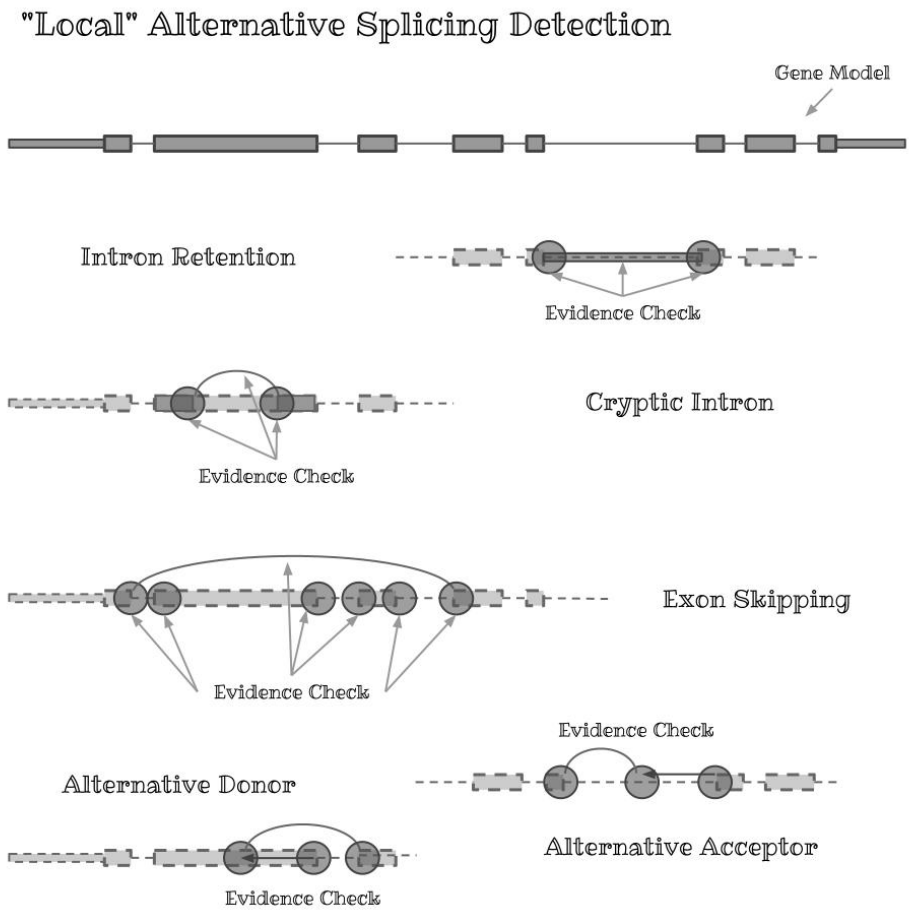


Figure 2.7: Alternative local events: Detection and Evidence Check. Picture showed positions in which read coverage is evaluated to design an AS event. Intron retention and cryptic intron are kept only if read depth on dark grey regions is above the cut-off of 3, same value is applied to all the other checks. Exons skipping are allowed after evidence checks on the alternative junction and the exon border skipped by the junction. Same checks are then applied on the alternative splice site, the alternative junction and the constitutive splice site.

2.4.5 AS bias for CDS exons

V1 gene prediction encompass 29,971 gene models in which CDS exons are 142,632 while UTR exons account for 42,320. In the light of 'stochastic noise' theory all exons have the same probabilities to undergo an AS event, no matter if they are coding or non-coding. From this assumption the number of AS events should be proportional to the number of different exons while from our data only 6% of AS events are UTR exons instead of the expected 22.9%.

2.5 Results

2.5.1 Extensive coverage for *Vitis Vinifera* transcriptome

Alternative splicing events predicted in this work result from an alignment of an amount of reads that gives, as far as we know, the most comprehensive picture of grape berry transcriptome to date. Almost all aligned reads (~97%) uniquely map onto the reference genome and ~66% of which with perfect match (see Table 2.1b). Additionally, the alignments shown an extensive coverage for the whole grape genome (Fig. 2.8).

The amount of mapped reads ranges from the minimum of 29% for Kozma to the 82% for Pinot N, the observed differences of the overall mapping level are quite expected, red wine berry cultivars better maps on reference genome of Pinot while white berry cultivars show a lower level of mapped reads (see Table 2.1a); the low amount of reads aligned from Kozma has a different explanation since it is due to a strong contamination of ribosomal RNA filtered out during the cleaning step (see material and methods).

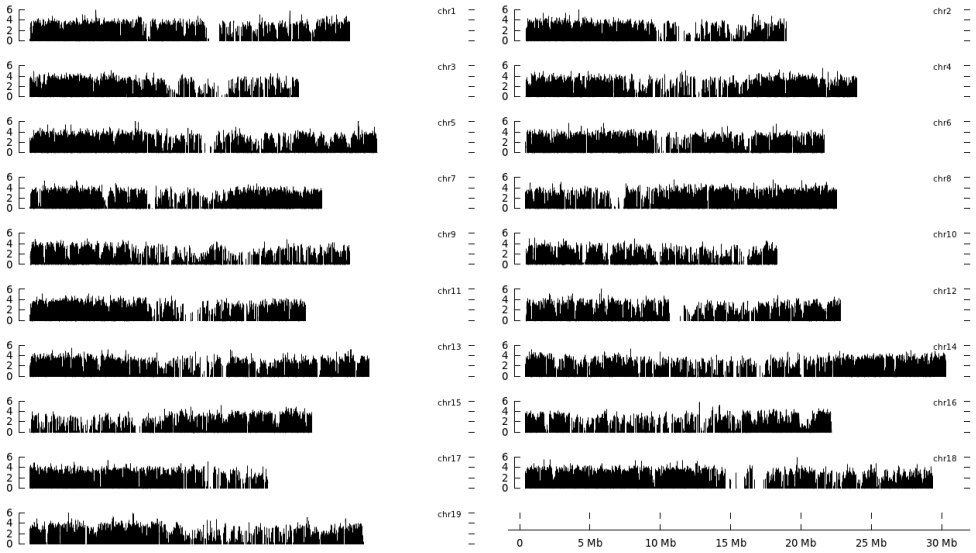


Figure 2.8: Read coverage per chromosome. Log2 scale of read coverage in windows of 1 kb per chromosome. Distribution of RNA-seq coverage as the sum of each cDNA libraries is shown along chromosomes of reference genome Pinot Noir. A vertical black line is referred to a log2 of the number of reads in a windows of 1 kb plotted against chromosome coordinates.

2.5.2 Level of detection of splice junctions with multiple cultivars

We have defined on average 107330 of SJs for each cultivars. The majority of SJs in our results resides in annotated or predicted genes (~95%) but anyway several novel SJs have been found (see Table 2.2 for detailed statistics in each cultivar). Moreover most of the overall predicted junctions are located in the coding sequence. Considering a mean value, ~92% of SJs are located inside the CDS, ~6% inside the UTR regions and ~2% span over CDS and UTR.

In average ~31% (from ~24% to ~37%) of the total SJs have been detected

as novel junctions and also it is important to note that the amount of novel SJs correlates with the amount of data available for each cultivar. As shown in figure 2.9 the fraction of novel junctions identified, follows the same pattern of the histogram representing the amount of alignment (green bars). According to this observation We suppose that the different level of detection for novel junctions should be referred mainly to the different amount of data instead to the genetic variability between our cultivars.

Inspection of dinucleotides at the intron borders indicates that the majority of SJs uses the canonical splice sites of plant introns. We have identified 95.0% GT-AG SJs and GC-AG, AT-AC respectively 2.5% and 1.9%.

Table 2.2: Splicing junction discovery rate. In the following table are showed the total amount of splicing junctions annotated in each cultivar, the relative fraction that overlap a gene prediction and the relative fraction of novel junctions. In the right part are showed the amount of junction annotated inside UTR regions, CDS or UTR-CDS if the junction overlap both coding and noncoding regions.

Sample	**n°SJs**	**SJs in GP (UTR; UTR-CDS; CDS)**		**SJs new**
Alicante Bouquette	103060	95,7%	(5,50%, 2,48% , 92,2%)	29,7%
Cabernet Franc	100838	95,8%	(5,35%, 1,97%, 92,67%)	29,5%
Chardonnay	92375	95,9%	(4,99%, 1,7%, 93,32%)	24,2%
Inzolia	100461	95,5%	(5,45%, 1,82% 92,73%	27,4%
Kozma Palne Muskotali	88659	96,1%	(5,05%, 1,67%, 93,28%)	24,5%
Lambrusco Salamino	116750	95%	(5,92%, 2,49%, 91,59%)	34,6%
Moscato Rosa	122479	94,8%	(6,24%,2,4%, 91,36%)	37,4%
Pinot Noir	118070	94,9%	(6,08%, 2,65%, 91,27%)	35%
Sangiovese	120250	94,9%	(5,97%, 2,48%, 91,55%)	36,2%
Teroldego	110365	95,1%	(5,84%, 2,25%, 91,91%	31,7%
Average	107330,7	95,37%	(5,68%, 2,22%, 92,10%)	31,02%

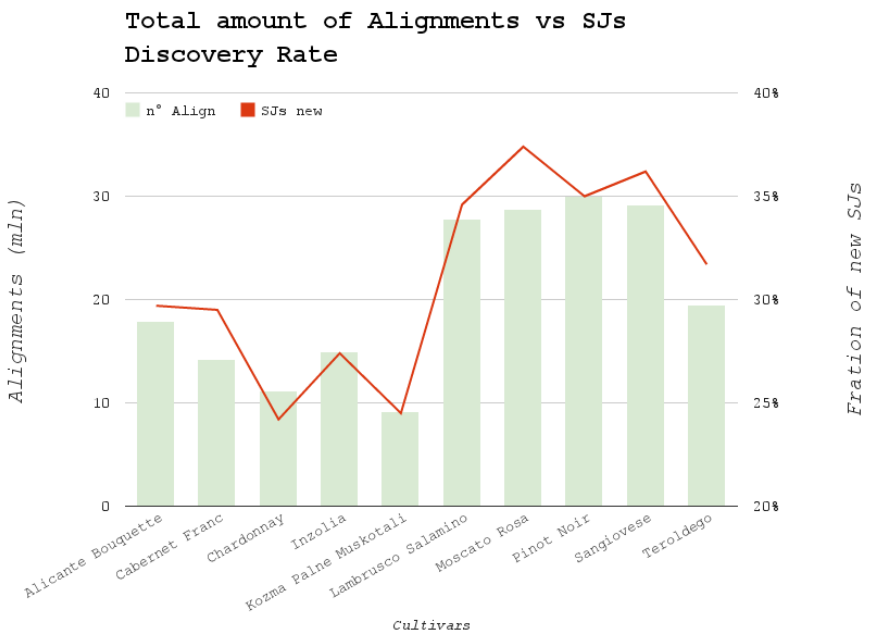


Figure 2.9: Splicing junction discovery rate. Splice junction discovery rate compared with the total amount of alignments obtained from each cultivars. The red line on the right axis shows the fraction of novel SJs annotated within each cultivar.

2.5.3 Abundance of alternative splicing classes

Often AS studies on plant give a prediction of all transcripts together with their relative abundance, the majority of these studies exploit graph theory: genes are represented by means DAGs (directed acyclic graph [56]) in which nodes are exons and arcs are the spliced reads among two exons. Different software use DAG in different ways and the output could vary from all possible paths (i.e all possible exons combinations [23]) to a minimum set that justify the observed data (CuffLinks [55]). These approaches are useful with additional experimental evidences since, alone, RNA-seq data are not sufficient to resolve

splice forms unambiguously. For these reasons we have decided to avoid these kind of approaches and focus on the identification of local events, alternative to the gene model.

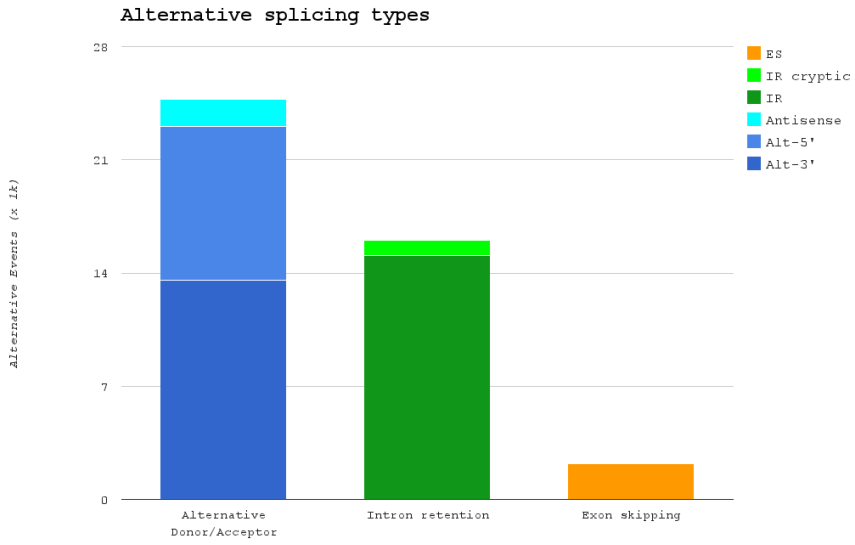


Figure 2.10: Alternative splicing types. The frequencies of the major categories of alternative splicing are shown overall the total amount of unique alternative splicing events identified within 10 *Vitis Vinifera* cultivars.

Our detection of AS relies, as explained in the methods section, only on cluster of reads that can be uniquely associated to a gene prediction. The final results of this method are 43,775 loci that correspond to 24,415 gene predictions covered by 196 million of aligned reads. Inside each locus we have looked for AS events in into 6 main groups: exon skipping (ES), alternative 5' donor site (Alt-5'), alternative 3' acceptor site (Alt-3'), antisense splice junction (Antisense), intron retention (IR), cryptic intron (IRc). In order to

predict an AS event, not only the alternative form but also the constitutive one must be supported by evidence. A coverage filter was applied on the constitutive form as well as on the alternative allowing only events with a cumulative coverage of at least 3 reads but detected in 3 different cultivars, the aim was to support our prediction with evidence in 3 different cDNA libraries reducing the influence of sequencing and mapping errors for low-coverage events.

Table 2.3: Alternative splicing detection result. The frequencies and the raw count of the major categories of alternative splicing are shown overall the total amount of unique alternative splicing events identified within 10 *Vitis Vinifera* cultivars, in the below part of the table. In the table above the alternative splicing events count was distinguished for the evidence within each cultivar.

Sample	**Alt-3'**	**Alt-5'**	**Antisense**	**IR**	**IRc**	**ES**
Alicante Bouquette	6795	4583	834	6412	410	913
Cabernet Franc	6424	4494	743	6827	358	1035
Chardonnay	5088	3532	671	5091	286	627
Inzolia	6176	4354	816	6726	379	785
Kozma Palne Muskotali	4759	3451	592	5136	257	747
Lambrusco Salamino	8616	5912	1071	10327	582	1163
Moscato Rosa	9331	6494	1133	10544	599	1287
Pinot Noir	8816	6144	1173	10368	574	1201
Sangiovese	9040	6298	1052	10578	530	1285
Teroldego	7676	5392	958	8570	488	1049
Total Events		**57,6%**		**37,32%**		**5,08%**
%	31,62%	22,03%	3,95%	35,14%	2,18%	5,08%
#	13578	9462	1695	15091	937	2182

Table 2.3 shows how many unique events were identified for each of the

major categories of alternative splicing. 40.4% of intron-containing genes had at least 1 alternative event. The most common event was intron retention with about 37% and the less common event, exon skipping (~5%). These estimations agree with previous studies on other plants [7][22][20][61], nevertheless is worthing to note that are exon junctions that especially undergo AS events, Alt-5' and Alt-3' together account for ~57% of the total (Fig. 2.10). These relative ratios among different AS events are conserved in all cultivars (Tab. 2.3).

Only genes with reads from at least 3 cultivars were retained in this analysis and what we observed is that the number of conserved AS change a lot among cultivars. 11,367 AS (the 26% of the total) are shared among only 3 cultivars and this value decrease till 3,735 (8.7%) for AS events conserved in all the cultivars (suppl. fig. 2). Moreover AS events conserved among all cultivars seems to have different features respect to AS events not conserved, for example the number of predicted AS events for a single gene. The majority of genes predict to be alternative spliced has just one AS event (suppl fig 6) but looking at the relative ratio the percentage change from 23 for all alternative spliced genes to 50% for AS genes conserved in more than 7 cultivars.

2.5.4 Relative low abundance of alternative events

We have obtained an indication of the expression degree for each AS events (ES, IR, IRc, Alt-5', Alt'-3, Antisense) by calculating the read coverage of the alternative event divided by the coverage of the consensus form. We called this relationship Alternative Events Ratio (AER).

We observe that on average 74% of all types of AS events (where the fraction for each types IR, ES, Alt-3'/Alt-5' are respectively 72%, 89%, 74%.) has an AER value lower than 0.1. These values reflect in some way the occurrence

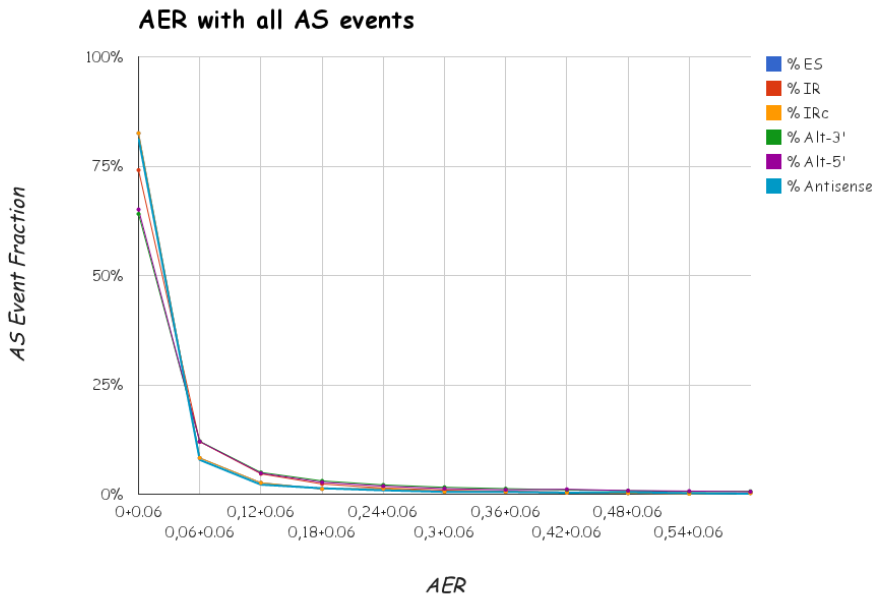


Figure 2.11: Alternative Events Ratio (AER). A) The relative coverage abundance between the putative alternative events and the related gene model. Each dot represents the AS events count within an ASR window of 0.06.

between the AS events and the constitutive form, suggesting that most of these events are considerably rare (Fig. 2.11).

The distance of Alt-3' and Alt-5' junction respect the constitutive exon borders (see Fig. 2.13). Almost all the events are in a range of less than 10 nt from the canonical exon/intron border.

In order to discover if there was some evidence of periodicity showing for example an over-representation for the position in frame, we have divided these AS events in two subcategories (see Figure 2.14), AER >=1 and AER<1, we have choose these two categories with the hypothesis that AS with

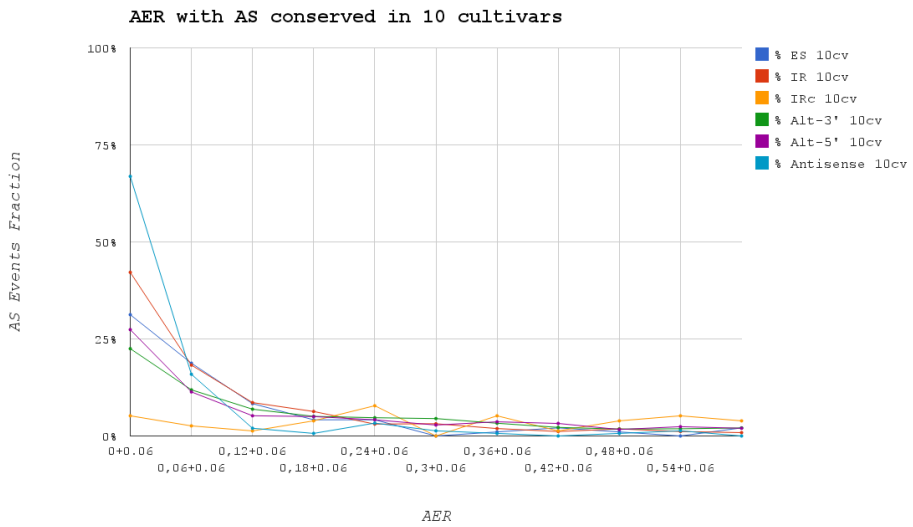


Figure 2.12: Alternative Events Ratio (AER). B) The relative coverage abundance between the putative alternative events conserved in all the cultivars and the related gene model. Each dot represents the AS events fraction within an ASR window of 0.06. The overall counts of AS events conserved in all the cultivars are the following: ES 96, IR 1046, IRc 77, Alt-3' 1405, Alt-5' 960, Antisense 151.

AER>=1 has more chance to have a functional role respect those with AER < 1. Performing a binomial test with expected frequency of 33% as a random choice for the position in frame, we have found that in Alt-3' AER>=1 is the only case where the in frame position is prevalent (44,27% P-value = 6,2e-10), instead Alt-5' AER<1 showed a prevalence for the positions not in frame (27,7% P-value = 2,2e-16). In the other two we have found that most probably there wasn't any prevalence (Alt-3' AER<1, Alt-5' AER>=1 respectively 32,18% P-value 0,05; 30,99% P-value = 0,365).

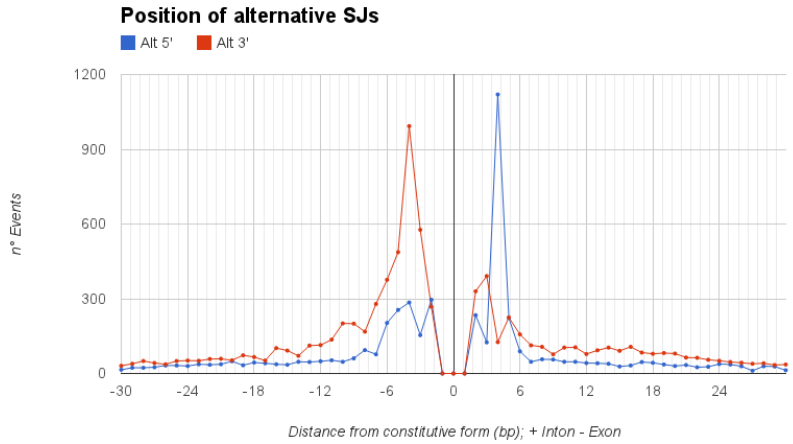
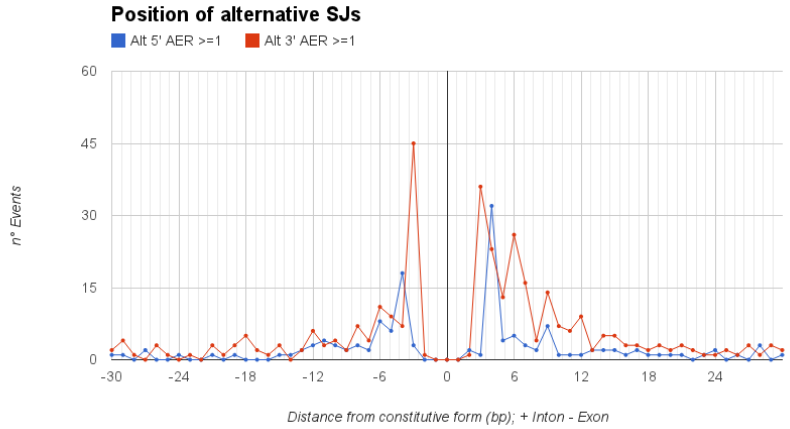


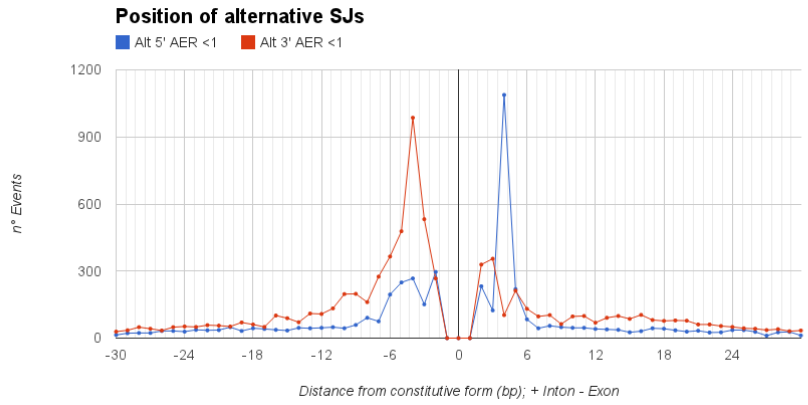
Figure 2.13: Distance of alternative SJs to the constitutive form. Alternative SJs location compared to the constitutive SJ, i.e. annotated within the gene model. The positive values on the x-axis represent the intronic region and the negative value represent the exonic region.

2.5.5 Functional annotation

We used the gene ontology to analyse the most frequent function for the 100 most alternatively spliced genes. The GO term were assigned with Argot2 [62] performing Blast and HMM searches against Uniprot and P-fam databases, respectively (see Fig. 2.16). The most common class in the Biological Process category were cellular process (27%), metabolic process (25%), response to stimulus (7%), biological regulation (7%). In the Cellular Component, the main locations are the organelle (29%) and the membrane (18%). The largest classes in the Molecular Function are catalytic activity (45%), binding (43%) and transporter activity (8%).



(a) $AER >= 1$



(b) $AER < 1$

Figure 2.14: Distance of alternative SJs to the constitutive form in two AER sub-categories. Alternative SJs positioning to the relative position of the SJ annotated within the gene model, within two coverage categories, higher (A) and lower (B) than AER 1.

2.6 Discussion

Alternative splicing is the most prominent mechanism that generates structural transcriptome complexity with many different outcomes: i) proteome expansion, ii) introduction of PTC which causes down-regulation by NMD; iii) UTRs variability that affects mRNA translation probability, localization and stability.

Despite recent advances in sequencing technologies, the study of the plant transcriptome is still in its early stages, and even those organism largely studied, as human, alternative splicing remain a complex arguments. In human, for example, recent evidence suggests that more than 90% of genes undergo to alternative splicing [17] [9] [4], however the functional role of such high frequency of AS transcripts is quite controversial and other studies suggesting that the majority of these alternative events are simply due to noise introduced by the splicing process [32] [34] [38] [39]].

On this study we have decided to analyse only single AS events without any attempt to produce the entire transcriptome. Do not predict the whole hypothetical transcripts could be sound as a limitation but, on the other hand, this decision allows us to use directly our observations without making any additional a priori assumptions [63]. Nevertheless even if our data does not allow us to predict mRNAs is clear that the generating transcript variants is not the only function of alternative splicing.

We have found evidence of alternative splicing in about 40% of intron-containing genes and for each event we have found evidence in at least 3 cultivars. We have identified many novel SJs and the majority of them showed a low level of expression, in general much of the observed mRNA diversity includes low-abundance events. We have identified many novel junctions that are extensively conserved in the analysed cultivars (in average an alternative

SJs is conserved in 5 cultivars) (Fig. 2.15) and also rarely used splice sites showed an enrichment close to the often-used splice site, i.e, the constitutive form. Moreover cultivars that have alternative events for a gene they also present the constitutive form in the 90% of cases, suggesting that the alternative form is preferentially expressed in combination with the constitutive one.

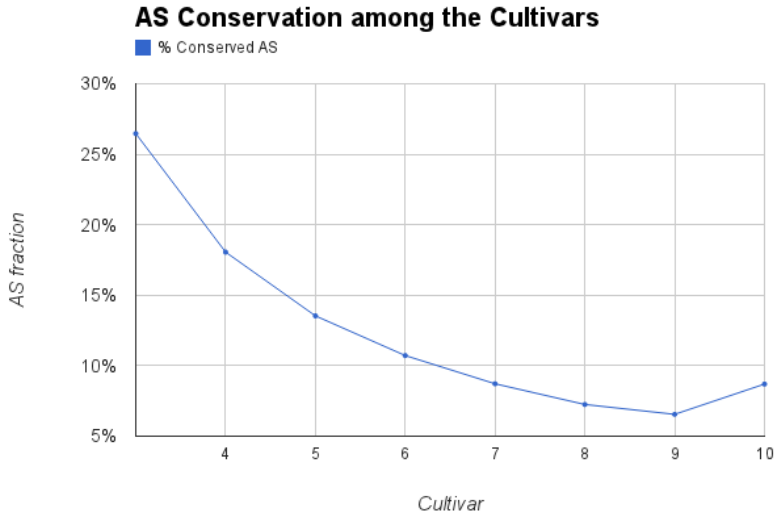


Figure 2.15: Relative abundance of AS events shared among cultivars. AS events are differentially conserved among cultivars, AS events shared by at least 3 cultivars are 11,367 and the number steadily decrease till 28,13 for the AS events shared among 9 cultivars. The number of AS events common to all cultivars is still considerable: 3,735 that's represent more or less the 8.7% of all predicted AS.

All these features fit the hypothesis that explain most alternative splicing as a consequence of stochastic noise in the splicing machinery. Agreeing with this hypothesis fluctuations of cellular environment determine an imperfect selection of splice sites that finally produces many, low-level, different alternative transcripts [38]. Since these fluctuations randomly affect the spliceosome mis-position, the expected number of AS events, for a single gene, should be

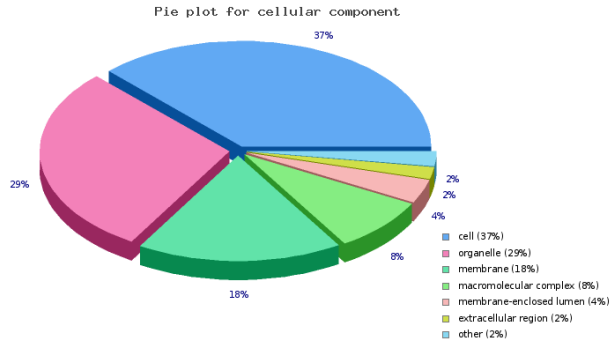
proportional to the number of exons and to how much the gene is expressed. These predictions seem confirmed from our data in which the number of predicted AS events per gene visually correlates with the gene exons number and with his expression level (calculated either as raw number of reads or rpkm value) (supplementary figure 3-5).

However, even though that a loss of efficiency of the spliceosome can caused a lot of putative alternative splicing events in grape, not all the characteristics of observed AS events can be explained without thinking to a functional role. Notably examples are: the high number of not in frame alternative SJs conserved among 5 or more cultivars, and the bias of AS events for CDS exon over UTR ones. Though we have not the final answer to this question and further analysis will be necessary nevertheless we can already suppose that, caused or not from the stochastic noise, low abundance AS events are playing a role in regulating gene expression.

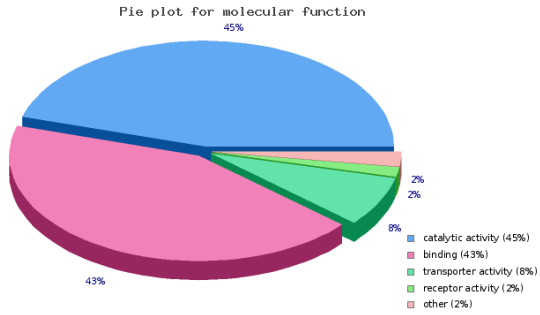
2.7 Author Contribution

Being first author I played the lead role in designing and implementing findAS and the other statistical analysis and comparisons. cDNA library preparation and sequencing was done by Elisa Asquini. I wrote the manuscript, though considerable contributions were made by dr. Alessandro Cestaro.

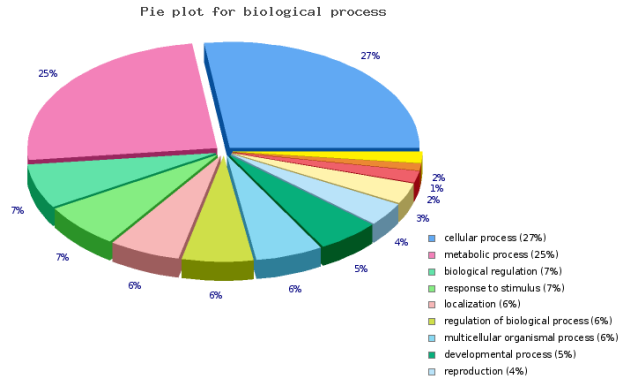
Prof. M.L. Racchi, Dr. Alessandro Cestaro and prof. Y. Van de Peer have supervised the project.



(a) Cellular Component



(b) Molecular Function



(c) Biological Process

Figure 2.16: Functional Annotation. The most frequent function for the 100 most alternatively spliced genes

Bibliography

[1] M.C. Wahl, C.L. Will, and R. Lührmann. The spliceosome: design principles of a dynamic rnp machine. *Cell*, 136(4):701–718, 2009.

[2] B.R. Graveley. Alternative splicing: increasing diversity in the proteomic world. *TRENDS in Genetics*, 17(2):100–107, 2001.

[3] T.W. Nilsen and B.R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, 2010.

[4] S. Stamm, S. Ben-Ari, I. Rafalska, Y. Tang, Z. Zhang, D. Toiber, TA Thanaraj, and H. Soreq. Function of alternative splicing. *Gene*, 344:1–20, 2005.

[5] L. Chen, J.M. Tovar-Corona, and A.O. Urrutia. Alternative splicing: A potential source of functional innovation in the eukaryotic genome. *International Journal of Evolutionary Biology*, 2012, 2012.

[6] A.M. Mastrangelo, D. Marone, G. Laidò, A.M. De Leonardis, and P. De Vita. Alternative splicing: Enhancing ability to cope with stress via transcriptome plasticity. *Plant Science*, 185:40–49, 2012.

[7] A.S.N. Reddy. Alternative splicing of pre-messenger rnas in plants in the genomic era. *Annu. Rev. Plant Biol.*, 58:267–294, 2007.

[8] A.M. McGuire, M.D. Pearson, D.E. Neafsey, J.E. Galagan, et al. Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol*, 9(3):R50, 2008.

[9] Q. Pan, O. Shai, L.J. Lee, B.J. Frey, and B.J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12):1413–1415, 2008.

[10] J.M. Baek, P. Han, A. Iandolino, and D.R. Cook. Characterization and comparison of intron structure and alternative splicing between medicago truncatula, populus trichocarpa, arabidopsis and rice. *Plant molecular biology*, 67(5):499–510, 2008.

[11] G. Zhang, G. Guo, X. Hu, Y. Zhang, Q. Li, R. Li, R. Zhuang, Z. Lu, Z. He, X. Fang, et al. Deep rna sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome research*, 20(5):646–654, 2010.

[12] Y. Marquez, J.W.S. Brown, C. Simpson, A. Barta, and M. Kalyna. Transcriptome survey reveals increased complexity of the alternative splicing landscape in arabidopsis. *Genome research*, 22(6):1184–1195, 2012.

[13] B.P. Lewis, R.E. Green, and S.E. Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mrna decay in humans. *Proceedings of the National Academy of Sciences*, 100(1):189–192, 2003.

[14] D.A.W. Soergel, L.F. Lareau, and S.E. Brenner. Regulation of gene expression by coupling of alternative splicing and nmd. *Madame Curie Bioscience Database*, 2000.

[15] W.B. Barbazuk, Y. Fu, and K.M. McGinnis. Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome research*, 18(9):1381–1392, 2008.

[16] X.C. Zhang and W. Gassmann. Rps4-mediated disease resistance requires the combined presence of rps4 transcripts with full-length and truncated open reading frames. *The Plant Cell Online*, 15(10):2333–2342, 2003.

[17] E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, and C.B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[18] B.R. Graveley, A.N. Brooks, J.W. Carlson, M.O. Duff, J.M. Landolin, L. Yang, C.G. Artieri, M.J. van Baren, N. Boley, B.W. Booth, et al. The developmental transcriptome of drosophila melanogaster. *Nature*, 471(7339):473–479, 2010.

[19] N. Kim, A.V. Alekseyenko, M. Roy, and C. Lee. The asap ii database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic acids research*, 35(suppl 1):D93–D98, 2007.

[20] E. Kim, A. Magen, and G. Ast. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, 35(1):125–131, 2007.

[21] E. Kim, A. Goren, and G. Ast. Alternative splicing: current perspectives. *Bioessays*, 30(1):38–47, 2008.

[22] B.B. Wang and V. Brendel. Genomewide comparative analysis of alternative splicing in plants. *Proceedings of the National Academy of Sciences*, 103(18):7175–7180, 2006.

[23] S.A. Filichkin, H.D. Priest, S.A. Givan, R. Shen, D.W. Bryant, S.E. Fox, W.K. Wong, and T.C. Mockler. Genome-wide mapping of alternative splicing in arabidopsis thaliana. *Genome research*, 20(1):45–58, 2010.

[24] T. Lu, G. Lu, D. Fan, C. Zhu, W. Li, Q. Zhao, Q. Feng, Y. Zhao, Y. Guo, W. Li, et al. Function annotation of the rice transcriptome at single-nucleotide resolution by rna-seq. *Genome research*, 20(9):1238–1249, 2010.

[25] A.P.M. Weber, K.L. Weber, K. Carr, C. Wilkerson, and J.B. Ohlrogge. Sampling the arabidopsis transcriptome with massively parallel pyrosequencing. *Plant physiology*, 144(1):32–42, 2007.

[26] W. Zhu, S.D. Schlueter, and V. Brendel. Refined annotation of the arabidopsis genome by complete expressed sequence tag mapping. *Plant Physiology*, 132(2):469–484, 2003.

[27] K. Iida, M. Seki, T. Sakurai, M. Satou, K. Akiyama, T. Toyoda, A. Konagaya, and K. Shinozaki. Genome-wide analysis of alternative pre-mrna splicing in arabidopsis thaliana based on full-length cdna sequences. *Nucleic acids research*, 32(17):5096–5103, 2004.

[28] M.A. Campbell, B.J. Haas, J.P. Hamilton, S.M. Mount, and C.R. Buell. Comprehensive analysis of alternative splicing in rice and comparative analyses with arabidopsis. *BMC genomics*, 7(1):327, 2006.

[29] A. Sánchez-Pla, F. Reverter, M.C. Ruíz de Villa, and M. Comabella. Transcriptomics: mrna and alternative splicing. *Journal of Neuroimmunology*, 2012.

[30] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[31] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S.D. Jackman, K. Mungall, S. Lee, H.M. Okada, J.Q. Qian, et al. De novo assembly and analysis of rna-seq data. *Nature methods*, 7(11):909–912, 2010.

[32] M. Pertea. The human transcriptome: An unfinished story. *Genes*, 3(3):344–360, 2012.

[33] M. Ebisuya, T. Yamamoto, M. Nakajima, and E. Nishida. Ripples from neighbouring transcription. *Nature cell biology*, 10(9):1106–1113, 2008.

[34] J.K. Pickrell, A.A. Pai, Y. Gilad, and J.K. Pritchard. Noisy splicing drives mrna isoform diversity in human cells. *PLoS genetics*, 6(12):e1001236, 2010.

[35] F. Ozsolak and P.M. Milos. Rna sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2):87–98, 2010.

[36] Z. Su, J. Wang, J. Yu, X. Huang, and X. Gu. Evolution of alternative splicing after gene duplication. *Genome research*, 16(2):182–189, 2006.

[37] Y. Dou, K.L. Fox-Walsh, P.F. Baldi, and K.J. Hertel. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *Rna*, 12(12):2047–2056, 2006.

[38] E. Melamud and J. Moult. Stochastic noise in splicing machinery. *Nucleic acids research*, 37(14):4873–4886, 2009.

[39] O. Jaillon, K. Bouhouche, J.F. Gout, J.M. Aury, B. Noel, B. Saudemont, M. Nowacki, V. Serrano, B.M. Porcel, B. Ségurens, et al. Translational control of intron splicing in eukaryotes. *Nature*, 451(7176):359–362, 2008.

[40] L. Chen, J.M. Tovar-Corona, and A.O. Urrutia. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. *Human molecular genetics*, 20(22):4422–4429, 2011.

[41] J.F. Terral, E. Tabard, L. Bouby, S. Ivorra, T. Pastor, I. Figueiral, S. Picq, J.B. Chevance, C. Jung, L. Fabre, et al. Evolution and history of grapevine (vitis vinifera) under domestication: new morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient european cultivars. *Annals of botany*, 105(3):443–455, 2010.

[42] S.T. Lund, F.Y. Peng, T. Nayar, K.E. Reid, and J. Schlosser. Gene expression analyses in individual grape (vitis vinifera l.) berries during ripening initiation reveal that pigmentation intensity is a valid indicator of developmental staging within the cluster. *Plant molecular biology*, 68(3):301–315, 2008.

[43] S. Zenoni, A. Ferrarini, E. Giacomelli, L. Xumerle, M. Fasoli, G. Malerba, D. Bellin, M. Pezzotti, and M. Delledonne. Characterization of transcriptional complexity during berry development in vitis vinifera using rna-seq. *Plant physiology*, 152(4):1787–1795, 2010.

[44] S.T. Lund and J. Bohlmann. The molecular basis for wine grape quality-a volatile subject. *Science Signalling*, 311(5762):804, 2006.

[45] S.D. Castellarin, G.A. Gambetta, H. Wada, K.A. Shackel, and M.A. Matthews. Fruit ripening in vitis vinifera: spatiotemporal relationships among turgor, sugar accumulation, and anthocyanin biosynthesis. *Journal of experimental botany*, 62(12):4345–4354, 2011.

[46] R. Velasco, A. Zharkikh, M. Troggio, D.A. Cartwright, A. Cestaro, D. Pruss, M. Pindo, L.M. FitzGerald, S. Vezzulli, J. Reid, et al. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PloS one*, 2(12):e1326, 2007.

[47] O. Jaillon, J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467, 2007.

[48] F.G. da Silva, A. Iandolino, F. Al-Kayal, M.C. Bohlmann, M.A. Cushman, H. Lim, A. Ergul, R. Figueroa, E.K. Kabuloglu, C. Osborne, et al. Characterizing the grape transcriptome. analysis of expressed sequence tags from multiple vitis species and development of a compendium of gene expression during berry development. *Plant physiology*, 139(2):574–597, 2005.

[49] N. Terrier, D. Glissant, J. Grimplet, F. Barrieu, P. Abbal, C. Couture, A. Ageorges, R. Atanassova, C. Leon, J.P. Renaudin, et al. Isogene specific oligo arrays reveal multifaceted changes in gene expression during grape berry (vitis vinifera l.) development. *Planta*, 222(5):832–847, 2005.

[50] S. Pilati, M. Perazzolli, A. Malossini, A. Cestaro, L. Demattè, P. Fontana, A. Dal Ri, R. Viola, R. Velasco, and C. Moser. Genome-wide transcriptional analysis of grapevine

berry ripening reveals a set of genes similarly modulated during three seasons and the occurrence of an oxidative burst at veraison. *BMC genomics*, 8(1):428, 2007.

[51] L. Deluc, J. Grimplet, M. Wheatley, R. Tillett, D. Quilici, C. Osborne, D. Schooley, K. Schlauch, J. Cushman, and G. Cramer. Transcriptomic and metabolite analyses of cabernet sauvignon grape berry development. *BMC genomics*, 8(1):429, 2007.

[52] F. Mattivi, R. Guzzon, U. Vrhovsek, M. Stefanini, and R. Velasco. Metabolite profiling of grape: flavonols and anthocyanins. *Journal of agricultural and food chemistry*, 54(20):7692–7702, 2006.

[53] C. Trapnell, L. Pachter, and S.L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[54] J. Grimplet, J. Van Hemert, P. Carbonell-Bejerano, J. Díaz-Riquelme, J. Dickerson, A. Fennell, M. Pezzotti, and J.M. Martínez-Zapater. Comparative analysis of grapevine whole-genome gene predictions, functional annotation, categorization and integration of the predicted gene sequences. *BMC Research Notes*, 5(1):213, 2012.

[55] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. Van Baren, S.L. Salzberg, B.J. Wold, and L. Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[56] A.S.N. Reddy, M.F. Rogers, D.N. Richardson, M. Hamilton, and A. Ben-Hur. Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements. *Frontiers in plant science*, 3, 2012.

[57] G. Van Rossum and F.L. Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica, 1995.

[58] D. Ascher, P.F. Dubois, K. Hinsen, J. Hugunin, T. Oliphant, et al. Numerical python, 2001.

[59] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[60] W.J. Wilbur and D.J. Lipman. Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences*, 80(3):726–730, 1983.

[61] H. Ner-Gaon, R. Halachmi, S. Savaldi-Goldstein, E. Rubin, R. Ophir, and R. Fluhr. Intron retention is a major phenomenon in alternative splicing in arabidopsis. *The Plant Journal*, 39(6):877–885, 2004.

[62] M. Falda, S. Toppo, A. Pescarolo, E. Lavezzo, B. Di Camillo, A. Facchinetti, E. Cilia, R. Velasco, and P. Fontana. Argot2: a large scale function prediction tool relying on semantic similarity of weighted gene ontology terms. *BMC bioinformatics*, 13(Suppl 4):S14, 2012.

[63] M.F. Rogers, J. Thomas, A.S.N. Reddy, A. Ben-Hur, et al. Splicegrapher: detecting patterns of alternative splicing from rna-seq data in the context of gene models and est data. *Genome Biol*, 13:R4, 2012.

A computer would deserve to be called intelligent
if it could deceive a human into believing that
it was human.

Alan Turing

3 Grapewine Digital Gene Expression

3.1 Abstract

10 *Vitis Vinifera* cultivars were selected with different metabolic profiles [1] in order to perform a comparative analysis of berry transcriptome at single base resolution. This article focuses on digital expression analysis (DGE) from a biological point of view as much as it focuses on tools for studying a cell's transcriptome and about tools available to correlate gene expression changes to functional changes.

While the microarray-based (analog) gene-expression profiling technology has dominated the 'omics' era, Next-Generation Sequencing (NGS) based gene-expression profiling (RNA-Seq) is likely to replace this analog technology in the present [2]. RNA-Seq shows much promise for transcriptomic stud-

ies as the genes of interest do not have to be known a priori, new classes of RNA, SNPs and alternative splice variants can be detected.

However, new technology also brings with it new issues to resolve: the specific technical properties of RNA-Seq data differ to those of analog data, leading to novel systematic biases which must be accounted for the analysis of this type of data. Additionally, multireads and splice junctions can cause problems when mapping the sequences back to a genome.

The data presented in this study provide the most comprehensive set of RNA-seq gene expression variants in *Vitis Vinifera* to date, and therefore we have detected and annotated differentially expressed genes and those candidate that are likely to be specific for each single cultivar.

3.2 Introduction

The *Vitis Vinifera* (common grapevine) belongs to the family Vitaceae, which comprises about 60 inter-fertile wild *Vitis* species. Grapes can be grown in Asia, North America and Europe at latitudes from 50°N to 40°S and up to 3,000 meters above sea level and so under subtropical, Mediterranean and continental-temperate climatic conditions [3] [4]. Grapes and their derivatives have a large and expanding worldwide market with almost 98% of grape vineyards planted with *Vitis Vinifera L. ssp. vinifera (or sativa)* cultivars of Eurasian origin[4], that derive from wild forms *Vitis vinifera L. ssp. sylvestris* [3].

Vitis Vinifera is the single *Vitis* species that acquired significant economic interest over time; there are also some other important species that are used as breeding rootstock due to their resistance against grapevine pathogens (Phyl-

loxera, Oidium and mildew), for example the North American *V. rupestris*, *V. riparia* or *V. berlandieri*. *V. vinifera* is a liana growing to several meters tall, with flaky bark. The leaves are alternate, palmately lobed, 5-20 cm long and broad. The fruit is a berry, known as a grape; in the wild species it is 6 mm diameter and ripens dark purple to blackish with a pale wax bloom; in cultivated plants it is usually much larger, up to 3 cm long, and can be green, red, or purple [1]. Then it is relevant to mention also that *Vitis vinifera* bears hermaphroditic self-fertilizing flowers but outbreeding by means of wind and insect pollination is the norm; as a result, cultivars are highly heterozygous and carry many deleterious recessive mutations. [4].

There are two available genome resource, both of them are *V. Vinifera cv. Pinot Noir* but the first one refers to an homozygous plant [5] and the second to an heterozygous plant [4]. Compared to other perennials, the genome size is relatively small, 475 Mb, similar to rice (Oryza sativa, 430 Mb;), barrel medic (Medicago truncatula, 500 Mb;) and black cottonwood poplar (Populus trichocarpa, 465 Mb;)[4]. For the purpose of this study we decide to use the homozygous [5] because predicting gene expression level, by means the reads coverage of a gene model, require a low level of redundancy within the gene predictions.

The understanding of biological systems comprising large numbers of genes is a tough challenge but the tools available for transforming NGS data into knowledge and new hypotheses have improved over recent years. Traditional methods for gene expression analysis (northern blotting, qRT-PCR ...) miss important effects in biological processes, such as metabolic and signaling pathways and networks, because they required the pre-selection of single genes. The development of analog gene expression techniques such as microarrays, represented a critical breakthrough as the simultaneous measurement of the expression of many thousands of genes in a sample was finally possible. How-

[1] http://en.wikipedia.org/

ever, they have their limitations. For example, the data need to be normalized to remove spatial artefacts and systematic biases, appropriate statistical analysis must be used to reduce the number of false positives, furthermore, as the technique relies on hybridization, it brings a range of related potential problems. The information generated with hybridization arrays is also limited to the number of probes on the microarray slide and usually to genes with known sequence. Microarrays are also constrained in their ability to detect splice variants.[6]

The recent development of NGS and its use in transcriptomic analysis (RNA-Seq) now potentially enables the quantitative measurement of 'all' genes expressed in a sample [7] [8] [9]. Out of the currently dominating NGS technologies, Illumina and Solid platforms are better suited for RNA-Seq applications than Roche 454 Pyrosequencing. This is largely due to the much greater number of individual sequence reads and the resulting increased depth of coverage. A comparison of the Illumina sequencing platform with the Affymetrix microarray platform, showed that 81% of differentially expressed genes from arrays were detected with Illumina and more of these genes were true positives with the Illumina technology [10]. Additionally, comparison of relative RNA-Seq read densities to published qRT-PCR measurements for 787 genes in two reference RNA samples yielded a nearly linear relationship across five orders of magnitude, indicating that RNA-Seq read counts give accurate relative gene expression measurements across a very broad dynamic range [11].

3.3 Aim

Today, transcriptome analysis is performed most commonly using an NGS application called RNA-seq, in which some RNA pool-total RNA, messenger RNA or noncoding RNA, for instance is reversetranscribed into cDNA,

converted into a sequencing library, sequenced and analysed. Expression levels of specific genes, differential splicing, allele-specific expression of transcripts can be accurately determined by RNA-Seq experiments to address many biological-related issues. In this study, we sought to assess the contribution of the different analytical steps involved in the analysis of RNA-seq data generated with the Illumina platform, in order to perform a genomewide comparison between 10 different grapevine cultivars. This study also aims to give a survey of the RNA-Seq methodology, particularly focusing on the challenges that this application presents both from a biological and a bioinformatics point of view.

3.4 Material and Methods

3.4.1 Plant material

We selected 10 cultivars of *Vitis Vinifera* with different metabolic profiles:

- 7 varieties with black berry : *Pinot Noir, Teroldego, Alicante Bouchet, Sangiovese, Moscato Rosa, Lambrusco Salamino, Cabernet Franc.*
- 3 varieties with white berry : *Chardonnay, Inzolia e Poloskei Muskotaly (Kozma).*

Those cultivars were chosen to maximize as much as possible the genes expressed in grape berry tissue and the variability in term of flavonol profile as shown in Fig. 3.1 [1], in addition we have considered the commercial significance for wine makers in Italy and especially for the Trentino region.

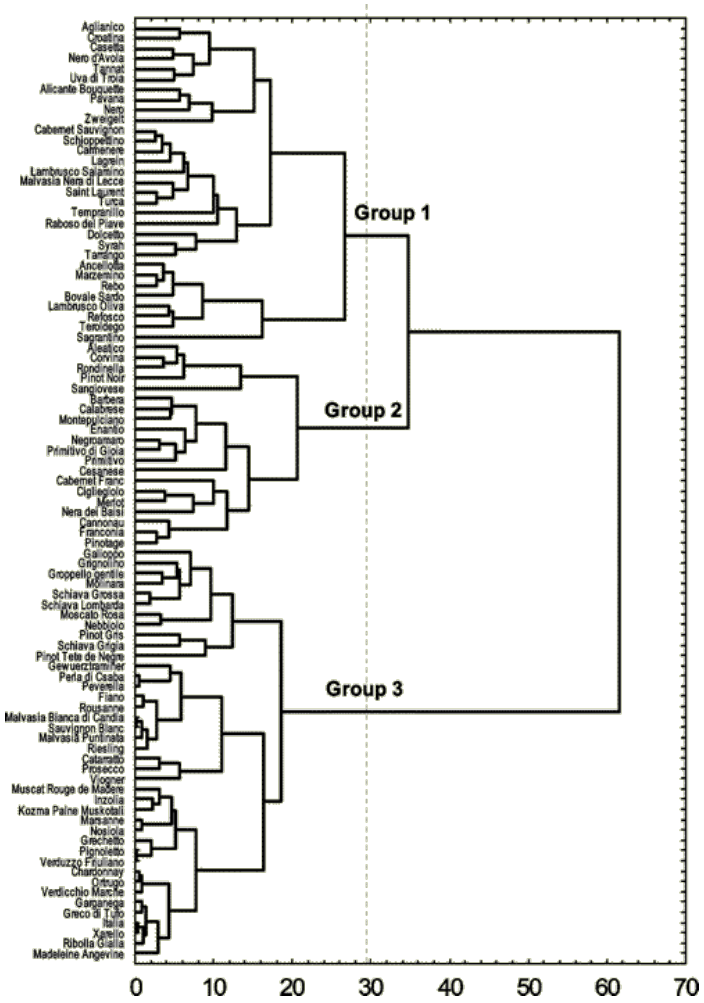


Figure 3.1: Flavonol variability at varietal level: Hierarchical tree plot showing the classification produced using cluster analysis of the flavonol profile (the percentage of each aglycon out of the total). The linkage distance is given on the X-axis. [1].

In the black varieties Pinot Noir has high level of free resveratrol and glycosilate on berry skin, this cultivar has also a shorter ripening time compared to the other studied varieties. Teroldego grapevine, peculiar of Trentino re-

gion, has an high content of anthocyanins either on berry skin and flesh. It differs from Pinot Noir for low level of resveratrols. Alicante Bouquette was selected because it peculiarly accumulates anthocyanins on berry flesh ('tintorea'). Moscato Rosa is highly aromatic. Sangiovese was selected because is one of the most cultivated grape of Italy. Lambrusco Salamino e Cabernet Franc were added during the second year of project since they have a metabolic features very different from the other cultivars. In the white varieties Chardonnay, Inzolia and Poloskei Muskotaly (Kozma) different aromatic profiles and different morphology of bunch (Fig. 3.2).

All the varieties in this study belong to the Mattivi's collection of 2006 [1]. They were of certain origin, checked, and named in agreement with existing literature and cultivated using a standardized system.

3.4.2 Sampling criteria

To facilitate the discrimination between differentially expressed genes in *Vitis Vinifera*, we generated 10 non normalized libraries. The total mRNA was extracted from a pool of berries for each cultivar under normal growth condition. According to the manufacturer's instructions, we have prepared 10 cDNA library with random primers using TruSeq Illumina Kit. Then we have obtained a global view of the grape berry transcriptome and gene expression, sequencing the resulting libraries using Illumina sequencing platform (85bp paired-end reads).

All varieties were sampled at technological maturity, defined as a content of soluble solids in the must between 17-18 °Bx [2].

[2]Degrees Brix (°Bx) is the sugar content of an aqueous solution. One degree Brix is 1 gram of sucrose in 100 grams of solution and represents the strength of the solution as percentage by weight (% w/w) (strictly speaking, by mass). If the solution contains dissolved solids other

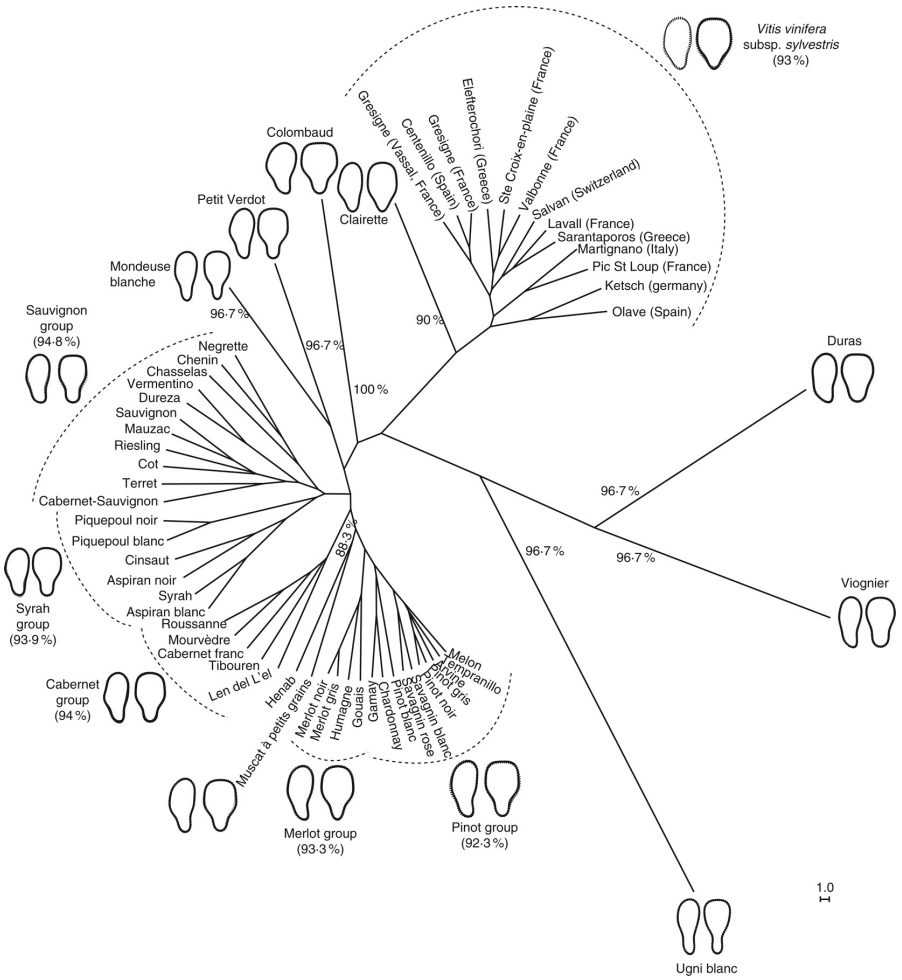


Figure 3.2: Shape variability at varietal level: UPGMA dendrogram based on the minimum Mahalanobis distance among wild grape populations and cultivars. Discriminant rate (%) and reconstructed seed outlines in dorsal and lateral view of morphoclades identified are also presented [3].

For each variety three independent samples were extracted for the RNA-

than pure sucrose, then the °Bx is only approximate the dissolved solid content. The °Bx is traditionally used in the wine, sugar, fruit juice, and honey industries.

seq analysis. To ensure a good representativeness of the sample, the bunches were taken from different plants of the same variety. Moreover the total RNA was extracted from a pool of berries.

The specific issues of mRNA extraction and library preparation was not taken into account for this PhD thesis.

3.4.3 Protocol consideration and workflow

Many issues should be considered when planning an RNA-seq experiment. No matter which method will be used or how many reads will be generated. Generally accepted experimental design principles must be used such as randomization samples and sufficient biological replication should be recommended. The aim of the study affects, the library type (normalised or not), how many reads are required, how long reads are required, what type of reads should be used (Fragment, Paired-end, Mate-pair), how samples should be prepared, what type of RNAs to be investigate, to preserve strand information or not. But most important is to have a good plan and a good question from the beginning.

In order to perform an RNA-seq experiment we should first of all choose the right sequencing platform, the NGS market is currently dominated by three different platforms: the FLX pyrosequencing system from 454 Roche, the Illumina Genome Analyser, and the SOLiD (Life Technologies). The second step is to use the cDNA library that is better for your purposes, the library preparation is a key step of RNA-seq, because it determines how closely the cDNA sequence data reflect the original RNA population. In the classic NGS protocols, adapters are ligated onto a double-stranded cDNA, but a substantial drawback of this approach, however, is the loss of information on transcriptional direction, because the adaptor is ligated to double-stranded cDNA. Library preparation and/or sequencing procedures can also introduce systematic

biases and artefacts such as over-amplification of GC-rich regions and generation of duplicate sequences [12].

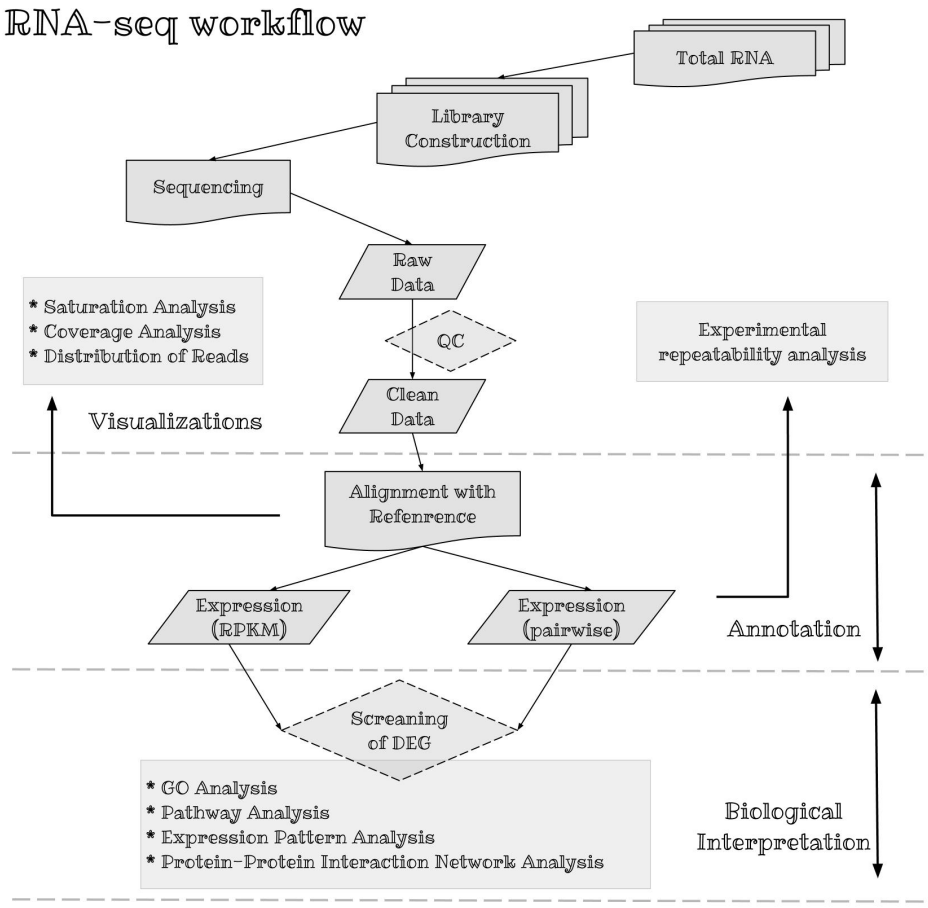


Figure 3.3: RNA-seq workflow

After image and signal processing, i.e. the process that convert a series of image from the sequencer in list of sequences stored in a text file, NGS data consist of a list of short sequences together with their base call qualities. These data are fundamentally different from microarray data. With hybridisation-based techniques, the scanner returns signal intensities for each probe on the

array. In the case of RNA-seq data, the number of reads mapping to any given region of the genome makes up the signal. Thus, RNA-seq data are countable and digital in nature.

The generation of reliable RNA-seq data therefore relies heavily on proper mapping of sequencing reads to corresponding reference genomes or on their efficient de novo assembly. Mapping NGS reads with high efficiency and reliability currently faces several challenges. First, the computing resources required to map huge numbers of small reads within a reasonable time can be limiting. The second challenge arises from the error rate of NGS data, meaning that non-perfect matches can be considered when mapping reads back to a genome.

Once the sequencing reads have been filtered and mapped (or assembled), it is possible to compute an expression score for every base in the genome and thus obtain transcriptome maps at the best possible resolution. The true resolution of this approach, however, depends on the amount of sequence coverage and therefore on the amount of sequences generated. Sequence coverage can be a limiting factor, especially when large genomes are analysed, due to costs and machine time required [13].

In this case we suppose to use this huge amount of data in order to explore transcriptome complexity of different grapevine cultivars not only in term of digital gene expression but also for developing, in the future, useful marker, such as SNP, for breading application. Moreover explore the complex relation between alternative splicing and gene regulation and how AS patterns were conserved between different cultivars as explained in the previous chapter. These kind of consideration and the available amount of money bound us to replace the biological replicates with statistical simulations in order to increase the number of varieties.

Some consideration and the project workflow are shown in the Fig. 3.3.

3.4.4 Sequencing and Pre-processing

We have obtained a global view of the grape berry transcriptome and gene expression, sequencing the cDNA libraries using the Illumina sequencing platform (85bp paired-end reads). A summary results with the available Illumina reads was shown in table 3.1

Table 3.1: A) Sequencing results for each different cultivars. The amount of pair reads generated by Illumina sequencing platform

Sample	# Raw Pairs
Alicante Bouquette	16148936
Cabernet Franc	19416930
Chardonnay	17816446
Inzolia	23344136
Kozma Palne Muskotali	21237030
Lambrusco Salamino	22357539
Moscato Rosa	24864531
Pinot Noir	22443561
Sangiovese	22181869
Teroldego	16583320
TOTAL	206394298

Quality control is also an important aspect of RNA-seq data analysis. For example, it is useful plot both the proportions of each nucleotide type, and the base quality score, for each sequence position.

Below we showed a series of characteristics that we had taken into account performing our quality checks to test the goodness of the individual sequencing: (i) first of all the total amount of sequence, i.e. the sample coverage; (ii) the average quality per position, basically we check the quality degradation of the sequences after every cycle i.e. base after base; (iii) % GC overall and per base; (iv) we also evaluated the nucleotide frequencies per base, in a random library we would expected that there would be little to no difference between the different bases of a sequence run, a bias consistent across all bases either indicates that the original library was sequence biased, or that there was a systematic problem during the sequencing of the library preparation; (v) the amount of total ambiguous nucleotide as overall statistic and per base frequency; (vi) sequence length distribution before and after the cleaning phase.

We have applied two stringent filters in order to remove reads with low base calling quality, a dynamic end trimming with 30 Phred as minimum quality level and a minimum read length of 50 bp.

Phred's base-specific quality scores examines the peaks around each base call to assign a quality score to each base call. In our illumina reads the quality scores range from 0 to 40, with higher values corresponding to higher quality. The quality scores are logarithmically linked to error probabilities, so quality score of 30 correspond to a probability of 0.1% that the base is called wrong, 20 to 1% and 10 to 10%.

An overview on the effect of the dynamic quality ends trimming is shown in figure 3.4. An extensive analysis of pre-precessing results is described in the Results section.

For these tasks we used manly python script developed during the first year

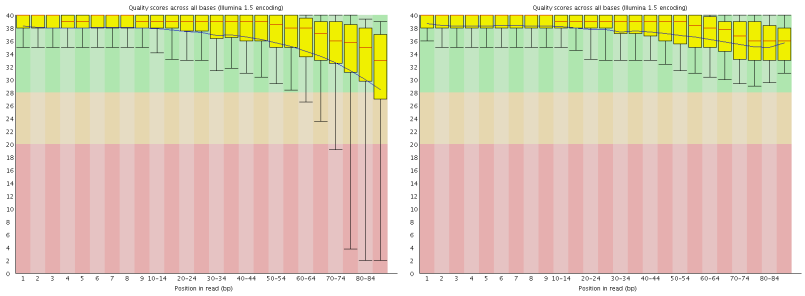


Figure 3.4: Ends Trimming: For each position a BoxWhisker type plot is drawn. The elements of the plot are as follows: (1) the central red line is the median value; (2) the yellow box represents the inter-quartile range (25-75%); (3) the upper and lower whiskers represent the 10% and 90% points; (4) the blue line represents the mean quality. The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read. Left: raw data; Right: clean reads

of PhD, in combination with some open source software such as FastX [3] and FastQC [4]. All the process is now more ore less completely automatic and the pipeline developed for these purposes can be adapted to many other RNA-seq purposes.

[3] FastX:http://hannonlab.cshl.edu/fastx_toolkit/index.html

[4] FastQC: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

3.4.5 Read alignment to the reference genome *Vitis Vinifera* cv. Pinot noir

Alignment of sequencing reads to a reference genome is a core step in the analysis workflows

In total, we generated more than 200 million paired-end reads (see Table 3.2a), on average 20 million for each cultivar. We have applied two stringent filters in order to remove reads with low base calling quality, a dynamic end trimming with 30 Phred as minimum quality level and a minimum read length of 50 bp. This filtering step produced as expected a strong reduction of the initial amount of reads (from 7% to 17%) as shown in table 3.2a but at the end of this process all the sequences that pass our filters could be safety mapped into the genome.

We used TopHat [14] to map the reads over the reference genome *Vitis Vinifera* cv Pinot Noir [5], and for the novel splice junction detection, using standard parameters except for the minimum intron lengths that was fixed at 25 nt. This cut-off is similar to other studies which sometime used smaller intron sizes (20 nt [15]; 1 nt [16]) but also bigger (60 nt [17]), moreover it has been estimated that an intron should be long approximately 30 bases to obtain a good intron removal [18]. The quality of the gene models is a mandatory requirement to obtain a good result on the AS detection. For the used version (12X) of the genome assembly, there are available two gene predictions, the first (12Xv0) is available from 2009 at the NCBI and also at Genoscope website [5], but there is also a second later version (12Xv1) that combines 12Xv0 with further predictions carried out at the CRIBI [6] in Padova, Italy. Our choice fell on the latter (12Xv1) for which recently a new functional annotation is

[5] http://www.cns.fr/vitis

[6] http://genomes.cribi.unipd.it/

Table 3.2: A) Tab.2.1a is copied here for convenience - Mapping and Sequencing results for each different cultivars. The amount of pair reads generated by Illumina sequencing platform, the read pairs that passed our filters and the amount of mapped reads.
B) Tab.2.1a is copied here for convenience - The frequencies of unique match, amount of reads with perfect match and 1 mismatch. All the percentage are referred to the total amount of alignments.

(a)

Sample	# Raw Pairs	# Cleaned (%Diff)		# Mapped (%Clean)	
Alicante Bouquette	16148936	14950506	-7,42%	9984939	66,79%
Cabernet Franc	19416930	17597497	-9,37%	7920727	45,01%
Chardonnay	17816446	15350304	-13,84%	6231975	40,60%
Inzolia	23344136	20008317	-14,29%	8324836	41,61%
Kozma Palne Muskotali	21237030	17546066	-17,38%	5131353	29,25%
Lambrusco Salamino	22357539	20214750	-9,58%	15541157	76,88%
Moscato Rosa	24864531	22585359	-9,17%	16159983	71,55%
Pinot Noir	22443561	20329915	-9,42%	16702563	82,16%
Sangiovese	22181869	20099091	-9,39%	16254295	80,87%
Teroldego	16583320	15014666	-9,46%	10950345	72,93%
TOTAL	206394298	183696471	-11,00%	113202173	61,62%

(b)

Sample	# Alignments	# Unique (%Align)		# Perfect (%Align)		# 1 Mismatch (%Align)	
Alicante Bouquette	18811078	18095946	96,20%	10983715	58,39%	4924774	26,18%
Cabernet Franc	14869982	14398933	96,83%	8854520	59,55%	3843744	25,85%
Chardonnay	11660580	11323991	97,11%	7767179	66,61%	2530131	21,70%
Inzolia	15518523	15114902	97,40%	10059622	64,82%	3618971	23,32%
Kozma Palne Muskotali	9613330	9305482	96,80%	6378477	66,35%	2124875	22,10%
Lambrusco Salamino	28961653	28176600	97,29%	19289975	66,61%	6322159	21,83%
Moscato Rosa	29998357	29140185	97,14%	20095659	66,99%	6698070	22,33%
Pinot Noir	31313762	30415625	97,13%	22510332	71,89%	5938681	18,97%
Sangiovese	30425400	29539182	97,09%	20567089	67,60%	6571565	21,60%
Teroldego	20284556	19772911	97,48%	14212485	70,07%	4088600	20,16%
TOTAL	211457221	205283757	97,08%	140719053	66,55%	46661570	22,07%

available [19]. In the current gene prediction each gene model is annotated with only one isoform without any knowledge about AS.

Various criteria were applied to evaluate alignments used for accurately discover novel splice junctions (SJs). As first step a number of maximum 8 mismatches were allowed, this value that permits us to cope with the uncertain genetic variability among grape cultivars and the reference genome. As second step only reads that mapped uniquely on the genome were retained and, third, only splitted reads with shortest side longer than 8 bp were kept . This filters were implemented in order to reduce the number of false positive.

TopHat [14] is a powerful freely available mapping program which can map reads allowing detection of novel SJs, TopHat make use of Bowtie [20] to perform the alignment, basically TopHat uses Bowtie as an alignment 'engine' and breaks up reads that Bowtie cannot align on its own into smaller pieces called segments. Often, these pieces, when processed independently, will align to the genome. When several of a read's segments align to the genome far apart from one another, TopHat infers that the read spans a splice junction and estimates where that junction's splice sites are [21].

We built the grape genome index converting each ambiguous bases to one of the corresponding A,C,G,T. We allowed TopHat to perform multiple alignments for each reads and a maximum of 2 mismatches for segments mapping (segment length 30 bp). All the alignments were than filtered with a maximum number of 8 mismatches and only 1 match in the genome (NH=1). We applied other two fail-safe filter, external mapped fragments must be longer than 7 bp.

3.4.6 Digital Gene expression

One of the problems that must be faced when dealing with analysis of short reads is that the quantification of expression depends on the length of the biological features under study (genes, transcripts or exons), as longer features will generate more reads than shorter ones [22].

Common normalization methods, including division by transcript length such as RPKM (Reads Per Kb of exon model per Million mapped reads) from Mortazavi et al. 2008 [23], mitigate but do not completely eliminate this bias [24].

Another drawback is the very nature of the sequencing technology, which is basically a sampling procedure from a population of transcripts, implying that differences in transcript relative distributions between samples will affect the assessment of differential expression. Furthermore, the ability to detect and quantify rare transcripts is obscured by the wide dynamic range of mapped reads and the concentration of a large portion of the sequencing output in a reduced number of highly expressed transcripts [25].

However, RNA-seq technology boasts a general high level of data reproducibility across lanes and flow-cells, which reduces the need of technical replication within these experiments [10].

An underlying factor that relates to several of the mentioned problems in RNA-seq analysis is the amount of reads generated in a given experiment. The more the target is sequenced, the more transcripts are identified and the higher the value of the expression level.

In this study we decided to use **NOISeq** [25] method in order to compute differential expression between two conditions given the expression level of

the considered features. The gene was used as the expression unit, The gene expression level is the number of reads or in the library mapping to a gene, i.e. the read counts. Let c^i_{gj} be the number of read counts for each gene i in the j-th sample (or replicate or lane) from the experimental condition or group $g(g = 1 \text{ or } 2)$, where j varies from 1 to the number of samples in group g.

Then, the library size or sequencing depth s_{gj} can be computed as the sum of counts c^i_{gj} over all the genes for the j-th replicate in experimental condition g.

In order to avoid library size bias, the NOISeq method corrects the counts by a factor closely related to the sequencing depth. The default option is taking the number of counts per million reads, so the corrected expression values would be:

$$x^i_{gj} = c^i_{gj} \times 10^6 / s_{gj}$$

Then we decide to use another implemented normalization technique, the Trimmed Mean of M values (TMM) from Robinson and Oshlack 2010 [26], instead of RPKM value for the further pairwise comparison, definitely more reliable for detecting differential expression between two genes. Regardless of the normalization procedure used, NOISeq permits applying a feature length correction which consists of dividing the expression level by a factor equal to any power of the feature length.

The differential expression statistics in NOISeq are the log-ratio (M) and the absolute value of difference (D). These statistics collect the information on fold-change and also the absolute pseudo-counts difference, thereby compensating the unstable behaviour of M at low expression values. They can be defined for a certain gene i as:

$$M^i = \log_2(\frac{x_1^i}{x_2^i})$$
and
$$D^i = |x_1^i - x_2^i|$$

Once M and D values have been obtained for each gene, a threshold for these values must be established in order to classify genes as differentially or non-differentially expressed. A gene is considered to be differentially expressed if the corresponding M and D values are very likely to be higher than noise values.

Let M^* and D^* be the random variables describing noise distribution. Let G^i be a random variable which takes the value 1 if gene i is differentially expressed between two experimental conditions and 0 when it is not. We are interested in determining the probability of G^i taking a value of 1. A gene i has been considered to be differentially expressed when the corresponding values for $|M|$ and $D(|m^i|$ and $|d^i|)$ are likely to be higher than in noise ($|M^*|$ and D^* values). Then, the probability of a gene being differentially expressed given the expression levels in both conditions can be written as follows:

$$P(G^i = 1|x_1^i, x_2^i) =$$
$$= P(G^i = 1|M^i = m^i, D^i = d^i) =$$
$$= P(|M^*| < |m^i|, D^* < d^i)$$

Thus, the probability of not being differentially expressed between the two conditions can be easily derived as:

$$P(G^i = 0|M^i = m^i, D^i = d^i) =$$

$$= 1 - P(|M^*| < |m^i|, D^* < d^i)$$

.

The odds:

$$\frac{P(G^i = 1|M^i = m^i, D^i = d^i)}{P(G^i = 0|M^i = m^i, D^i = d^i)}$$

may be used to decide whether a gene is differentially expressed between the two conditions or not. For instance, an odds value of 4:1 is equivalent to $P(G^i = 1|M^i = m^i, D^i = d^i) = 0.8$ and it means that the gene is 4 times more likely to be differentially expressed than non-differentially expressed. This is the probability threshold we used throughout the paper.

When there are no replicates for any of the experimental conditions, the algorithm can simulate them. The simulation relies on the assumption that read counts follow a multinomial distribution, where probabilities for each class (gene) in the multinomial distribution are the probability of a read to map to that gene. NOISeq [25] has been implemented in the statistical language R [7].

For each gene model in 12Xv1 annotation file, we counted the number of reads with an in house python script and we have calculated RPKM, FDR, the noiseq pairwise comparison and the hierarchical clustering with a mixture of in python and R script that I have developed for this PhD thesis by using on-line open source libraries.

[7] NOISeq is available at http://bioinfo.cipf.es/noiseq

3.5 Results

3.5.1 RNA-seq: form raw data to digital coverage

We obtain in average 20 million of paired-end reads for each cultivars with Illumina sequencing technology. We applied two stringent filters in order to remove reads with risky base calling error rate (dynamic end trimming to 30 Phred, minimum length of 50 bp) and the results as expected was a strong reduction of the amount of reads (from 7 to 17) as shown in table 3.2a and in the figure **??**. At the end all the sequences that pass our filters could be safety mapped into the genome.

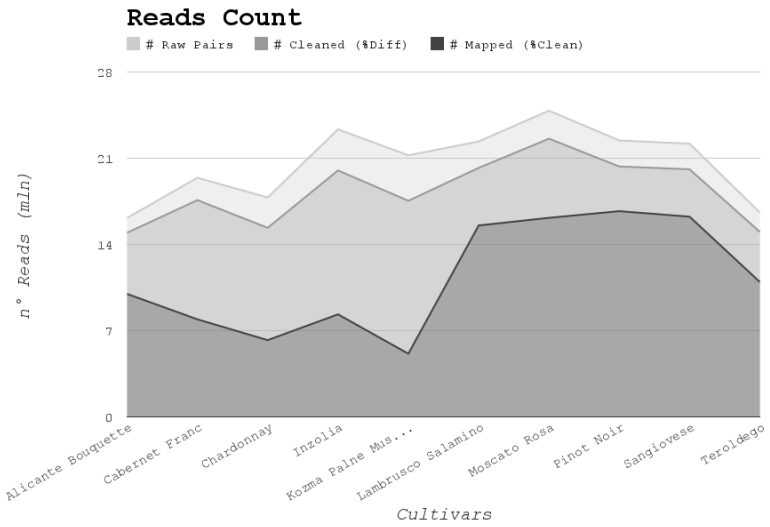


Figure 3.5: Reads Count: from raw data to mapping results. The upper line represents the total amount of reads generated with Illumina platform, the mean line represent the amount of reads for each cultivar after the pre-processing step. Finally the last darker line shows the amount of mapped reads.

We decided to map our reads with both the available genomes. The FEM-IASMA (Pinot Noir heterozygous) and Genoscope (Pinot Noir homozygous). Alignment results are shown in table 3.2b for the France genome. We obtained an high performances in term of unambiguous mapping, in both cases we were able to map uniquely more than 90% of the pair reads, with results a little bit better with the homozygous genome (more than 96%) as expected. Moreover about 60% of the alignments were obtained without mismatch.

Another important notice was about the very low performance in term of number of mapped reads in some of our cultivars. For example in *Paloskei Muskotaly*, *Cabernet Franc*, *Inzolia* and *Chardonnay* we was able to map only 30-40% of the available reads. These results are difficult to explain especially because the genetic variability of these cultivars nowadays is unknown. In term of mapping rate, some cultivars have shown a very low performance, especially for cultivar Kozma mapping rate was around 29%. We have investigated those results with an ab initio assembly of the entire sample using Trans-ABySS [27] (data not shown), the majority of reads clustered together inside 10 contigs all of them annotated as ribosomal RNA.

All the alignments were than filtered again. We have removed all the alignments with more than 8 mismatch, with an intron longer than 50kb and if it is spliced the shorter alignment portion must be longer than 7bp. Moreover we have taken into account only reads that map uniquely in the genome. The filters was set considering an high variability between our cultivars and the reference genome. We have removed more or less 4% of alignments from all the cultivars.

In conclusion, considering the similar results obtained mapping our reads over two different Pinot genomes, we decided to use only the homozygous version for the following analysis considering more reliable and less redundant its gene prediction and functional annotation.

3.5.2 Global Changes in Gene Expression

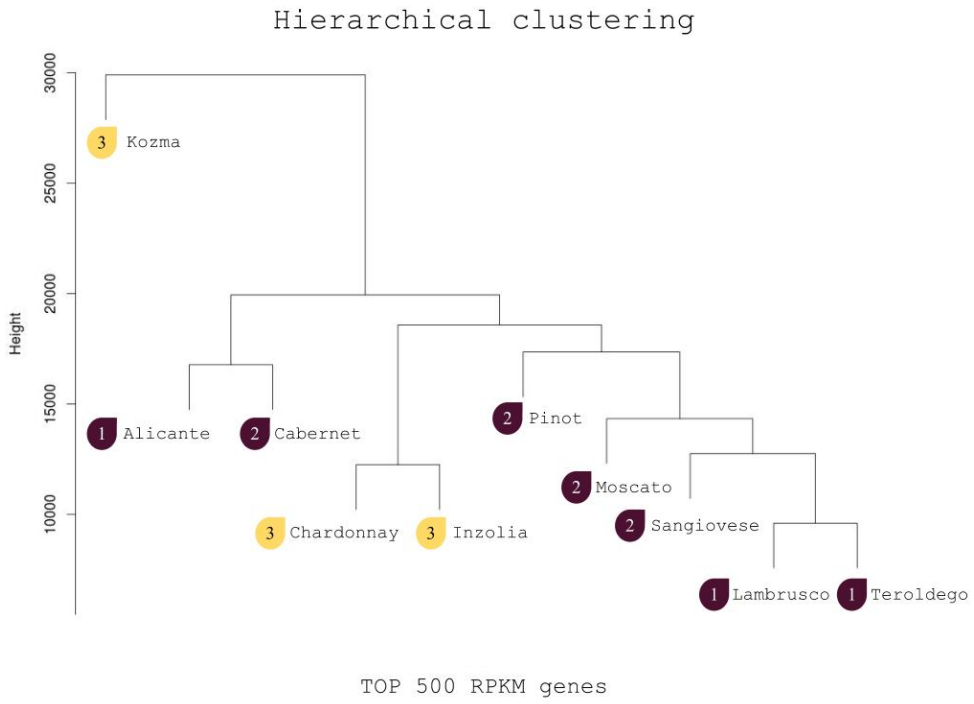


Figure 3.6: Hierarchical clustering: Using the higher 500 expressed genes we build this hierarchical clustering. In the figure on the Y-axis is represented the matrix distance, for each cultivar is also represented the berry color and the correspondente Mattivi's group, where group 1,2,3 represent three differet level of flavonol production.

The number of RNA-seq reads generated from a transcript is directly proportional to that transcript's relative abundance in the sample. However, because cDNA fragments are generally size-selected as part of library construction (to optimize output from the sequencer), longer transcripts produce more sequencing fragments than shorter transcripts. To calculate the correct expression level of each transcript, we must count the reads that map to each

transcript and then normalize this count by each transcript's length. Similarly, two sequencing runs of the same library may produce different volumes of sequencing reads. To compare the expression level of a transcript across runs, the counts must be normalized for the total yield of the machine. The commonly used fragments per kilobase of transcript per million mapped fragments (or FPKM , also known as RPKM) incorporates these two normalization steps to ensure that expression levels for different genes and transcripts can be compared across runs [28].

To characterize the differences of molecular response between the white and black cultivar, genes expression levels were calculated by RPKM (reads per kilobase per million reads) using the formula [10] RPKM = (109 ×C)/(N × L), where C is the number of reads that uniquely aligned to the gene, N is the total number of reads that uniquely aligned to all genes, L is the sum of the gene in base pairs. The RPKM method eliminates the influence of gene length and sequencing discrepancy in calculating gene expression, allowing direct comparison of gene expression between treatments.

RPKM were used to evaluate expressed value and quantify transcript levels. P-value and FDR (false discovery rate) were manipulated to determine differentially expressed genes [29]. In the present study, the top 500 differentially expressed genes were used as a data matrix in order to perform a hierarchical clustering for a comparative analysis with a similar clustering developed by Mattivi et al. with metabolomics data of flovonol production [1].

The dramatic expression profile suggested significant transcriptional complexities in *Vitis Vinifera* showing an high variability comparing the expression profile in different cultivar. This particular behaviour will be discussed in more detail in the further section, where we reported an extensive pairwise comparison with a more sensible methods.

RPKM and hierarchical clustering have been calculated by using a set of R script. The result is showed in Fig. 3.6, Kozma is the only one that cluster outside as a single group, ths is probably due to its parental origins not completely from Vinifera. The two other white berry cultivars are grouped together as well as class 3 and class 1 black berries. Class 1,2,3 in fig. 3.6 inside the berry picture, represent 3 classes of flavonol production inferred by Mattivi's study [1] and showed in Fig. 3.1. The class 1 as showed in figure 3.1 is composed only by white varieties, Kozma, Chardonnay and Inzolia, class 2 is composed by 4 black varieties, Pinot, Moscato, Sangiovese, Cabernet and finally the last group of cultivar Teroldego, Lambrusco and Alicante are part of the class 3.

Table 3.3: DGE pairwise matrix: 45 pairwise comparison between 10 cultivar have been done in order to detect differentially expressed gene between 2 cultivars with a probability higher than 90%. In this table each line represents a cultivar, and each column how many times a gene has been found differentially expressed within the 9 possible combination of pairwise. The last column is the amount of genes differentially expressed in at least 1 cultivar.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	DIFF. EXP.
Alicante Bouquette	6244	3828	3808	2918	1904	1155	690	377	244	112	15036
Cabernet Franc	6072	3450	2966	2734	2433	1607	1061	548	204	23	15026
Chardonnay	4368	2254	2170	2559	3464	2720	1871	904	317	127	16386
Inzolia	5827	3560	3695	3139	2313	1348	733	364	172	93	15417
Kozma Palne Muskotali	3506	1593	1546	1896	2773	3326	3312	2070	553	48	17117
Lambrusco Salamino	6285	2410	3558	3441	2873	1599	831	547	343	197	15799
Moscato Rosa	6190	1867	2703	3689	3471	2239	1077	529	295	163	16033
Pinot Noir	6496	2228	2933	3504	3064	1835	1005	536	363	202	15670
Sangiovese	6135	1957	2571	3403	3151	2305	1189	656	448	342	16022
Teroldego	7126	4394	4018	2666	1610	909	543	320	153	53	14666

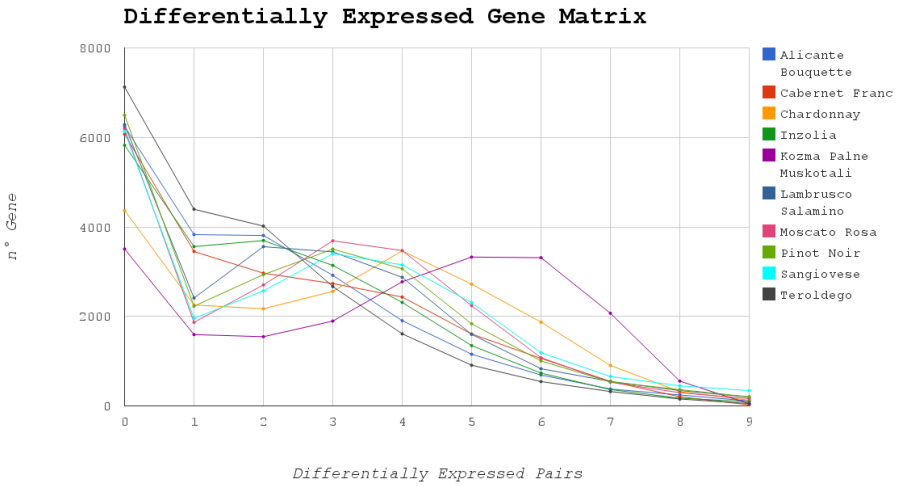


Figure 3.7: DGE pairwise matrix: 45 pariwise comparison between 10 cultivar have been done in order to detect differentially expressed gene between 2 cultivars with a probability higher than 90%. In Figure each line represents a cultivar, on Y-axis is showed the number of gene for each category on X-axis, from 0 to 9, that represent how many times a gene has been found differentially expressed within the 9 possible combination of pairwise

3.5.3 Pairwise comparison for digital gene expression

We compared the expression of all the genes corresponding to each cultivar in order to explore the role of different pathways. Once we had a sure set of alignments we were able to apply NOISeq to the reads count for each gene. Due to the fact that NOISeq allow only pairwise comparison we have done all the possible pairwise comparison, 45 different tables were obtained with the genes that are likely to be differentially expressed in the pair of cultivars with a probability higher than 90%.

Than we have count how many time a gene is differentially expressed referring only to the 9 possible pairs of each cultivars. The results was shown in table 3.3 and in figure 3.7. Just an example to explain better the meaning of the table, in the case of Pinot noir there were 202 gene that we have found differentially expressed in 9 pairwise comparison, i.e. against all the other cultivars. With this approach we have detected a list of candidate genes that probably are peculiar for each cultivar.

Each cultivar have been compared with the other 9 in a way that we obtain 9 list of gene differentially expressed with a probability higher than 90%. Considering 9 comparison all the cultivar showed more than 14000 gene differentially expressed in at least one cultivar, this high level of of differential expression is a good measure of the complexity that occur in the grape berries even in 9 cultivar of the same species.

In figure 3.8 are showed in the gray bar the number of expressed genes for each cultivar, Alicante Bouquette 21280, Cabernet Franc 21098, Chardonnay 20754, Inzolia 21244, Kozma Palne Muskotali 20623, Lambrusco Salamino 22084, Moscato Rosa 22223, Pinot Noir 22166, Sangiovese 22157, Teroldego 21792. Over 29971 total gene model our data showed a pervasive expression of the berries transcriptome and the differences between them are most probably due to a read coverage non uniform.

A very interesting discovery is the high level of consistency of gene regulation over the 9 possible comparison, in figure 3.8 the darker line represent the amount of gene differentially expressed in the same way, up or down regulated, within all the comparison where the gene is found differentially expressed. Basically all the gene showed the same direction of regulation in all the comparisons. Only few genes have been found up regulated in some cultivar and down regulated in some others.

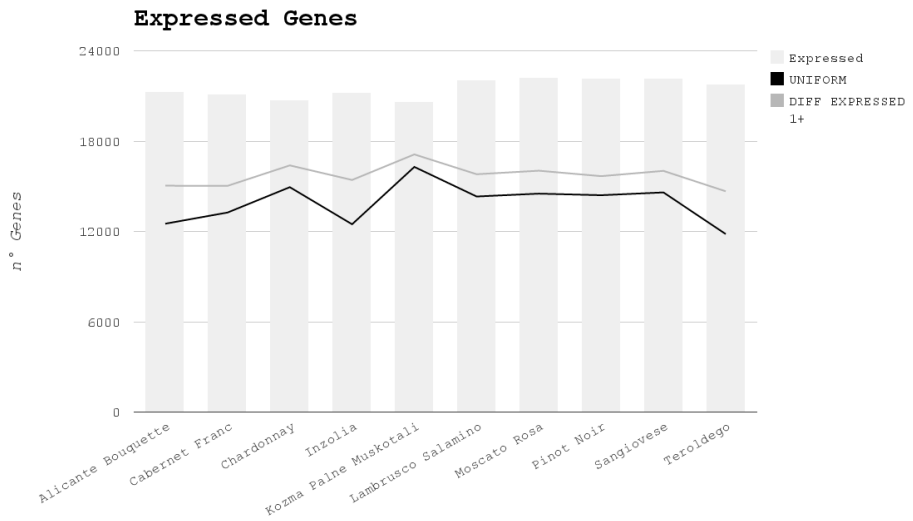


Figure 3.8: Expressed Genes: the gray bars in backgroud represent the number of expressed gene for each cultivar, i.e. with at least 1 reads. The gray line is the amount of gene differentially expressed at least in one pairwise comparison. The black line is the amount of gene differentially expressed in the same 'direction', up or down, within all the pairwise comparison where the gene is differentially expressed.

In figure 3.9 we have considered only genes completely uniform, i.e. regulated in the same 'direction' in all the comparison, surprising there is a dramatic difference in the number of genes up and down regulated between the cultivar with an high mapping rate an the cultivar with a low mapping rate. That probably suggest that there is an overestimation of up regulation in the bigger samples

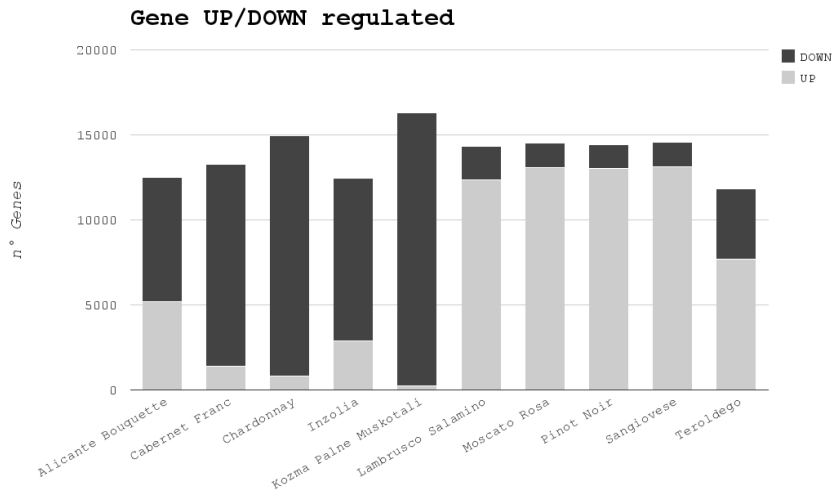


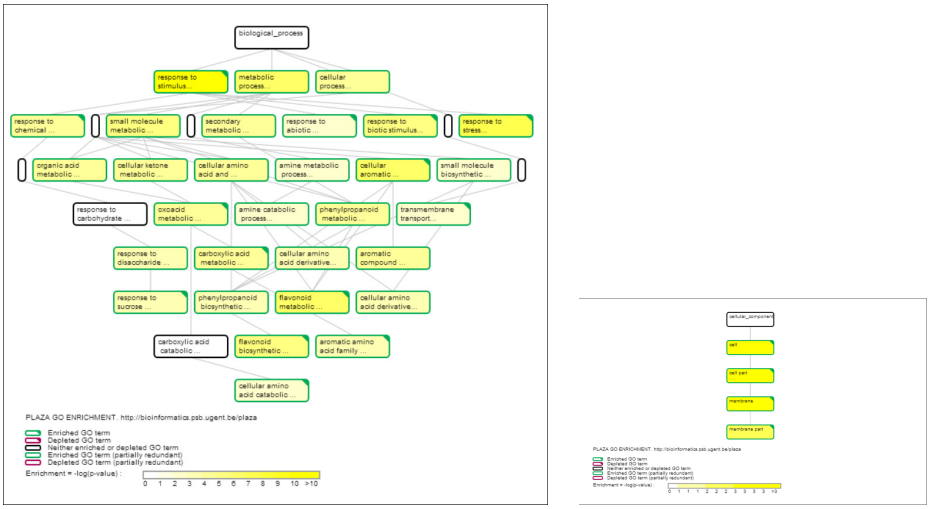
Figure 3.9: Genes Up or Down regulated: for each cultivar the gray part of the par is the number of gene tha we have found UP regulated in all the comparison where the gene have been found differentially

3.5.4 Functional Annotation and GO enrichment

To evaluate the potential functions of genes with significant transcriptional changes between our cultivars, all the gene differentially expressed in at least 8 other cultivars have been annotated using gene ontology terms. Gene Ontology (GO) categories were assigned to the significant genes based on the Argot2 [30] and Plaza [31] annotation. A GO enrichment analysis has been performed using Plaza platform of differential expression genes according to the cellular component, molecular function, and biological process.

As shown in Figures 3.10 3.11 3.12 3.13 3.14 3.16 3.15 3.19 3.17 3.18, significant GO biological process and the the correspondent enrichment P-vales terms were identified for those gene that are specific for the cultivars, i.e. that showed a differential expression in 8 or 9 possible pairwise comparison.

With regard to cellular component, the analysis revealed a high percentage of membrane activity spread over all the cultivars. For molecular function, differential expression genes specific of each cultivar showed only in Alicante the flavonoid biosynthesis, most probably those genes are the main actor of the Alicante's red pulp. In almost all the other cultivars are over represented GO related to stress or stimuli response. Within some of the black cultivars we have also identified peculiar activity of transport mechanisms and finally a common feature in 2 white cultivars is the Lipids biosynthesis.

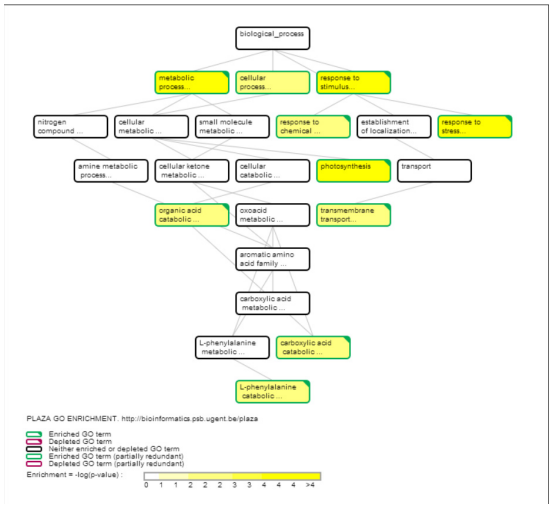


(a) Biological Process **(b)** Cellular Component

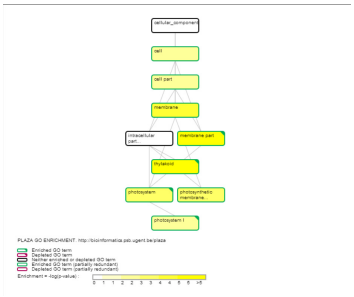
Figure 3.10: GO Enrichment: Alicante

3.6 Discussion and Conclusion

In this study we have found, comparing ten grapevine cultivars, an high level of differentially expressed genes. On average $\sim$ 15000 genes have been annotated at least in one comparative analysis as differentially expressed for each

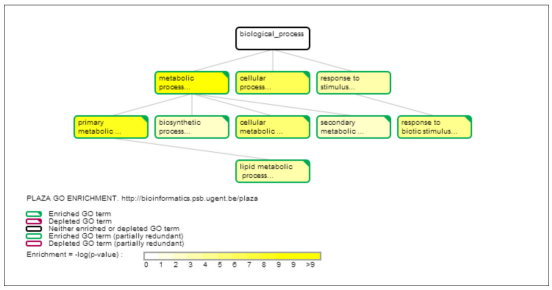


(a) Biological Process

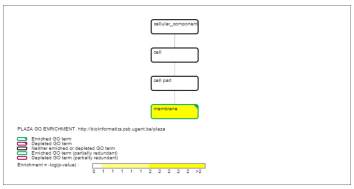


(b) Cellular Component

Figure 3.11: GO Enrichment: Cabernet



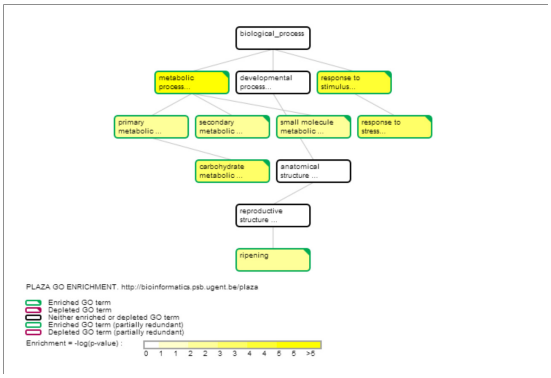
(a) Biological Process



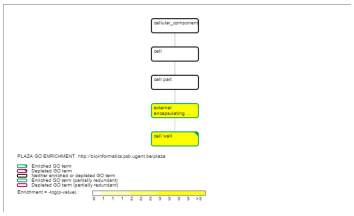
(b) Cellular Component

Figure 3.12: GO Enrichment: Chardonnay

cultivar. In conclusion the number of putative genes that probably play a role in determine the phenotype variability within our cultivars is about $\sim 70\%$ of the expressed genes ($\sim 50\%$ considering all the gene predictions). Our results confirm once again that RNA-seq is a suitable tool for deciphering the complexity of the gene expression mechanisms [32] and that for future analysis we should consider the transcriptome comparative from a network point of view

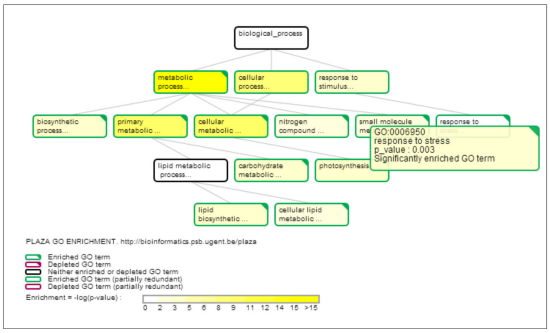


(a) Biological Process

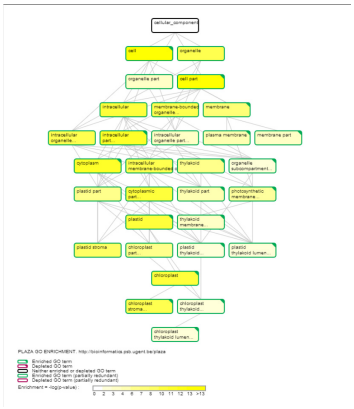


(b) Cellular Component

Figure 3.13: GO Enrichment: Inzolia



(a) Biological Process

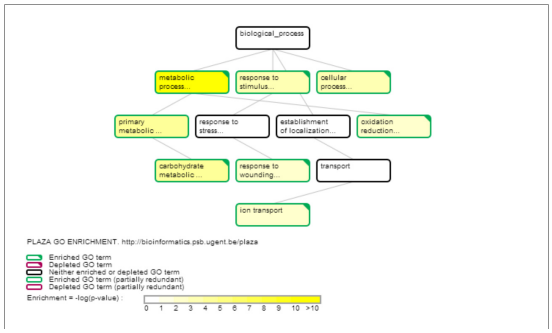


(b) Cellular Component

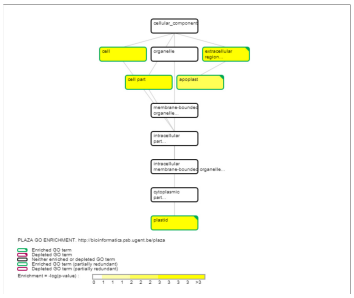
Figure 3.14: GO Enrichment: Kozma

much more than focusing on the effects of a single gene.

We have calculated a hierarchical clustering by using the top 500 expressed genes considering only their RPKM values. The resulting tree is surprisingly very close to the Mattivi's clusterization, even if the second have been made with metabolomic data. Kozma probably due to an interspecific crossing and

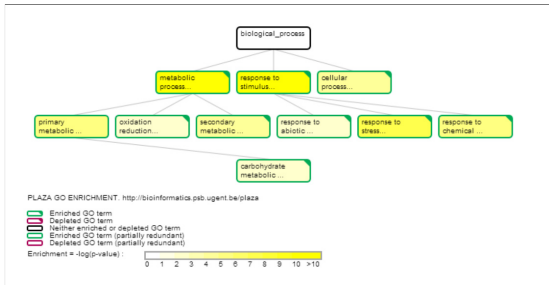


(a) Biological Process

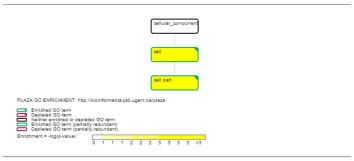


(b) Cellular Component

Figure 3.15: GO Enrichment: Moscato



(a) Biological Process

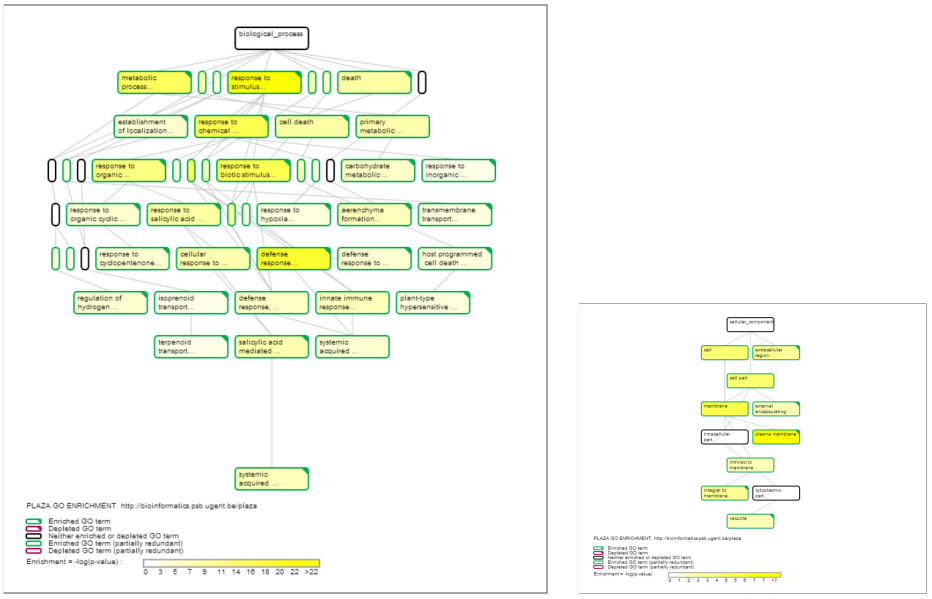


(b) Cellular Component

Figure 3.16: GO Enrichment: Lambrusco

the lack of coverage is very distance from all the others. Moreover white and black cultivars are grouped separately, showing a significant differentiation in term of gene expression between this two grape subcategories.

With the specific pairwise comparison that we have carried out in order to improve results specificity, a list of candidate genes for each cultivar are identified as ‘specific’ for the cultivar. 112 for Alicante, 23 for Cabernet, 127 for Chardonnay, 93 for Kozma, 197 for Lambrusco, 163 for Moscato, 202 for Pinot, 342 for Sangiovese and 53 for Teroldego are the gene we have annotated always as differentially expressed.

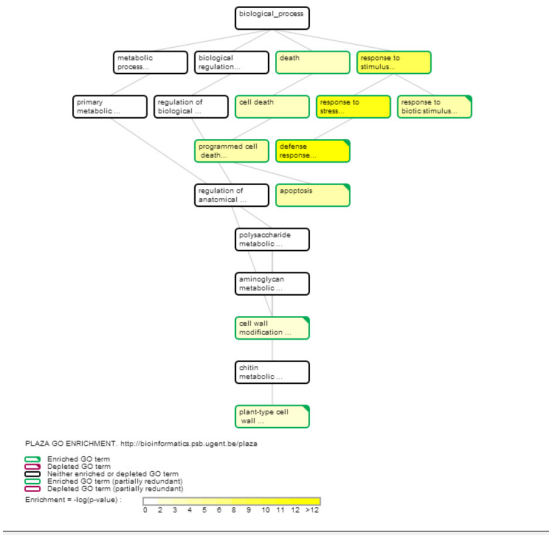


(a) Biological Process **(b)** Cellular Component

Figure 3.17: GO Enrichment: Sangiovese

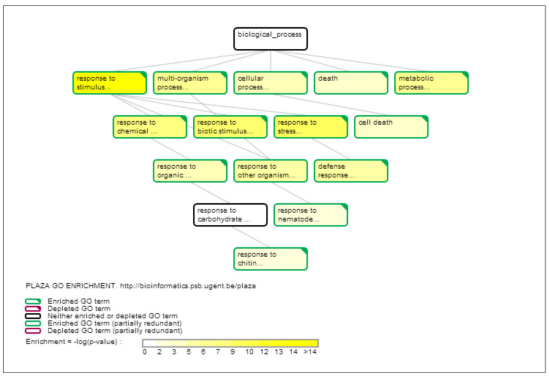
The functional annotation showed as main outcome that Alicante, famous cultivar for its red pulp due to a characteristic accumulation of anthocyanins, is the only one with a over-representation of GO term linked to the flavonol pathways.

Regarding the statistical methods for DGE analysis, we still see some gaps of knowledge [33]. Although the RPKM is a quick and easy system of comparison, it was already shown in the past that is also not very effective in the presence of a high variability of the coverage in the samples. For this reason we decided to rely on a second type of normalization, implemented in the R package Noiseq, accepting the limit of pairwise comparisons in favour of greater specificity in the results. The results, in general unsatisfactory, showed some peculiarities that may still due to a non-uniform coverage of our samples. For example, the low coverage cultivars have a prevalence of gene under

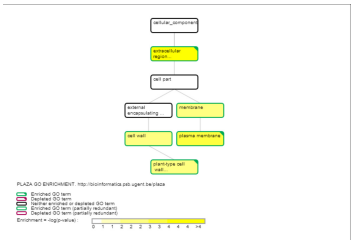


(a) Biological Process

Figure 3.18: GO Enrichment: Teroldego



(a) Biological Process



(b) Cellular Component

Figure 3.19: GO Enrichment: Pinot

expressed, the opposite behaviour for cultivars with high coverage.

Although a global analysis of gene expression based on a single replica-

tion does not allow a solid biological interpretation, this RNA-Seq analysis clearly provided a comprehensive view of the participation of several multigene families in berry development and ripening, identifying which members are expressed and characterizing their expression profiles in detail, showing which are likely to participate in the synthesis and accumulation of secondary metabolites in 10 different cultivars. In comparison to previous studies [34], the RNA-Seq method identified many additional transcripts, paving the way for a more accurate and more detailed description of the molecular processes involved in the development of grape berries and the basis of their organoleptic properties

3.7 Author Contribution

Being first author I played the lead role in designing and implementing the statistical analysis and comparisons. cDNA library preparation and sequencing was done by Elisa Asquini. I wrote the manuscript, though considerable contributions were made by dr. Alessandro Cestaro.

Prof. M.L. Racchi and dr. Alessandro Cestaro have supervised the project.

Bibliography

[1] F. Mattivi, R. Guzzon, U. Vrhovsek, M. Stefanini, and R. Velasco. Metabolite profiling of grape: flavonols and anthocyanins. *Journal of agricultural and food chemistry*, 54(20):7692–7702, 2006.

[2] S.K. Pond, S. Wadhawan, F. Chiaromonte, G. Ananda, W.Y. Chung, J. Taylor, A. Nekrutenko, et al. Windshield splatter analysis with the galaxy metagenomic pipeline. *Genome research*, 19(11):2144–2153, 2009.

[3] J.F. Terral, E. Tabard, L. Bouby, S. Ivorra, T. Pastor, I. Figueiral, S. Picq, J.B. Chevance, C. Jung, L. Fabre, et al. Evolution and history of grapevine (vitis vinifera) under domestication: new morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient european cultivars. *Annals of botany*, 105(3):443–455, 2010.

[4] R. Velasco, A. Zharkikh, M. Troggio, D.A. Cartwright, A. Cestaro, D. Pruss, M. Pindo, L.M. FitzGerald, S. Vezzulli, J. Reid, et al. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PloS one*, 2(12):e1326, 2007.

[5] O. Jaillon, J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467, 2007.

[6] N.C. Roy, E. Altermann, Z.A. Park, and W.C. McNabb. A comparison of analog and

next-generation transcriptomic tools for mammalian studies. *Briefings in functional genomics*, 10(3):135–150, 2011.

[7] Y. Zhang, J. Zhu, and H. Dai. Characterization of transcriptional differences between columnar and standard apple trees using rna-seq. *Plant Molecular Biology Reporter*, pages 1–9, 2012.

[8] G. Liu, W. Li, P. Zheng, T. Xu, L. Chen, D. Liu, S. Hussain, and Y. Teng. Transcriptomic analysis of'suli'pear (pyrus pyrifolia white pear group) buds during the dormancy by rna-seq. *BMC genomics*, 13(1):700, 2012.

[9] C. Feng, M. Chen, C. Xu, L. Bai, X. Yin, X. Li, A.C. Allan, I.B. Ferguson, and K. Chen. Transcriptomic analysis of chinese bayberry (myrica rubra) fruit development and ripening using rna-seq. *BMC genomics*, 13(1):19, 2012.

[10] J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, and Y. Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

[11] E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, and C.B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[12] M. Bainbridge, M. Wang, D. Burgess, C. Kovar, M. Rodesch, M. D'Ascenzo, J. Kitzman, Y.Q. Wu, I. Newsham, T. Richmond, et al. Whole exome capture in solution with 3 gbp of data. *Genome biology*, 11(6):R62, 2010.

[13] S. Marguerat and J. Bahler. Rna-seq: from technology to biology. *Cellular and Molecular Life Sciences*, 67(4):569–579, 2010.

[14] C. Trapnell, L. Pachter, and S.L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[15] S.A. Filichkin, H.D. Priest, S.A. Givan, R. Shen, D.W. Bryant, S.E. Fox, W.K. Wong, and T.C. Mockler. Genome-wide mapping of alternative splicing in arabidopsis thaliana. *Genome research*, 20(1):45–58, 2010.

[16] G. Zhang, G. Guo, X. Hu, Y. Zhang, Q. Li, R. Li, R. Zhuang, Z. Lu, Z. He, X. Fang, et al. Deep rna sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome research*, 20(5):646–654, 2010.

[17] Y. Marquez, J.W.S. Brown, C. Simpson, A. Barta, and M. Kalyna. Transcriptome survey reveals increased complexity of the alternative splicing landscape in arabidopsis. *Genome research*, 22(6):1184–1195, 2012.

[18] O. Jaillon, K. Bouhouche, J.F. Gout, J.M. Aury, B. Noel, B. Saudemont, M. Nowacki, V. Serrano, B.M. Porcel, B. Ségurens, et al. Translational control of intron splicing in eukaryotes. *Nature*, 451(7176):359–362, 2008.

[19] J. Grimplet, J. Van Hemert, P. Carbonell-Bejerano, J. Díaz-Riquelme, J. Dickerson, A. Fennell, M. Pezzotti, and J.M. Martínez-Zapater. Comparative analysis of grapevine whole-genome gene predictions, functional annotation, categorization and integration of the predicted gene sequences. *BMC Research Notes*, 5(1):213, 2012.

[20] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[21] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg, J.L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *nature protocols*, 7(3):562–578, 2012.

[22] A. Oshlack, M.J. Wakefield, et al. Transcript length bias in rna-seq data confounds systems biology. *Biol Direct*, 4(1):14, 2009.

[23] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[24] M. Young, M. Wakefield, G. Smyth, and A. Oshlack. Gene ontology analysis for rna-seq: accounting for selection bias. *Genome Biology*, 11(2):R14, 2010.

[25] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa. Differential expression in rna-seq: a matter of depth. *Genome research*, 21(12):2213–2223, 2011.

[26] M.D. Robinson, A. Oshlack, et al. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11(3):R25, 2010.

[27] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S.D. Jackman, K. Mungall, S. Lee, H.M. Okada, J.Q. Qian, et al. De novo assembly and analysis of rna-seq data. *Nature methods*, 7(11):909–912, 2010.

[28] Y. Deng, J. Yao, X. Wang, H. Guo, and D. Duan. Transcriptome sequencing and comparative analysis of saccharina japonica (laminariales, phaeophyceae) under blue light induction. *PloS one*, 7(6):e39704, 2012.

[29] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, and I. Golani. Controlling the false discovery rate in behavior genetics research. *Behavioural brain research*, 125(1):279–284, 2001.

[30] M. Falda, S. Toppo, A. Pescarolo, E. Lavezzo, B. Di Camillo, A. Facchinetti, E. Cilia, R. Velasco, and P. Fontana. Argot2: a large scale function prediction tool relying on semantic similarity of weighted gene ontology terms. *BMC bioinformatics*, 13(Suppl 4):S14, 2012.

[31] S. Proost, M. Van Bel, L. Sterck, K. Billiau, T. Van Parys, Y. Van de Peer, and K. Vandepoele. Plaza: a comparative genomics resource to study gene and genome evolution in plants. *The Plant Cell Online*, 21(12):3718–3731, 2009.

[32] V. Costa, C. Angelini, I. De Feis, and A. Ciccodicola. Uncovering the complexity of transcriptomes with rna-seq. *Journal of Biomedicine and Biotechnology*, 2010, 2010.

[33] P. Glaus, A. Honkela, and M. Rattray. Identifying differentially expressed transcripts from rna-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, 2012.

[34] N. Terrier, D. Glissant, J. Grimplet, F. Barrieu, P. Abbal, C. Couture, A. Ageorges, R. Atanassova, C. Leon, J.P. Renaudin, et al. Isogene specific oligo arrays reveal multifaceted changes in gene expression during grape berry (vitis vinifera l.) development. *Planta*, 222(5):832–847, 2005.

Difficult and impossible are cousins often mistaken
for one another, with very little in common.

Locke Lamora

Appendix A

findTools

We developed a dedicated software to carry out the alternative splicing detection, called findAS (available upon request), and findTools. findTools is a comprehensive python package that provides all the findAS libraries and other useful RNA-seq scripts.

A.1 findTools: Modules

findTools is also a python package with a set of libraries for Next Generation sequencing data manipulation, that we can also use for other scripts. Probably the most useful and well tested are:

- **libfindTools_GFF**, for gff and gtf iteration
- **libfindTools_FASTX**, for fasta and fastq iteration

An extensive Help should be provided within the package. A brief example usage:

GFF3 iteration

```
from findtools import libfindTools_GFF
for mRNA in libfindTools_GFF.gff_itr("file/path/"):
   print mRNA.start, mRNA.end, mRNA.exons #.CDSs
```

GFT3 iteration

```
from findtools import libfindTools_GFF
for mRNA in libfindTools_GFF.gtf_itr("file/path/"):
   print transcript.start, transcript.end,
   print mRNA.exons #.CDSs
```

FASTA iteration

```
from findtools import libfindTools_FASTX
for seq in libfindTools_FASTX.fasta_itr("path/to"):
   print seq.header, seq.sequence,
   print seq.sequence_mask, seq.lenght
```

FASTQ iteration and Sequence cleaning

```
from findtools import libfindTools_FASTX
for seq in libfindTools_FASTX.fastq_itr("path/to"):
   print seq.header, seq.sequence,
   print seq.quality, seq.offset

   # Quality Trim
   seq.qualityTrim(min,offset)
   print seq.sequence[seq.start, seq.end]
   '''clean seq, start end are the trimming
        position after qualityTrim'''

   # Adaptor Trim
   seq.leftTrimAdaptor(FastaRecord,
       ALLOWproportionOFmismatch)
   seq.rightTrimAdaptor(FastaRecord,
       ALLOWproportionOFmismatch)
   print seq.sequence[seq.start, seq.end]
   '''
   clean seq, start end are the trimming
        position after adaptorTrim'''
```

A.2 findTools: Scripts

findTools provide some script that we will find in the path after the installation (findXX -h for HELP!):

- **findAS**, splice site statistics and alternative inference
- **findDRAW**, display findAS results
- **countAS**, statistics over findAS results
- **findCLEAN**, Bam file filtering, such as number of hits or mismatch and so on and so forth
- **findCOUNT**, count the number of reads in a BAM file for each gene prediction in a GFF3, RKPM conversion is not mandatory.
- **findDEG**, management of NOISeq results.

A.2.1 Installation Guide

Manual:

- Download SomeWhere findTools_x.x.x.tar.gz
- Install

```
pip install findTools_x.x.x.tar.gz
```

This command install findtools module in the python path and also findXX tools in the /usr/local/bin/ folder. All the dependences will be automatically installed if necessary! You need Internet access! Probably Root privileges are required! Local Install:

```
pip install --user findTools_x.x.x.tar.gz
```

or

```
pip install --install-options= \
"--prefix=personal/path/" \
findTools_x.x.x.tar.gz
```

NOTE: If you have some error installing dependencies, try to manually download and install pysam 0.6 and psutil 0.4.1 with the follow command::

```
python setup.py install --prefix /home/user/.local/
```

If everything successfully, findTools script being installed in $HOME/.local/bin. Probably you have also to add $HOME/.local/bin to your $PATH and make findTools works!

Uninstall:

```
pip uninstall findtools
```

A.2.2 Requirements

With pip these packages will be automatically installed!

- $pysam >= 0.6$
- $psutil >= 0.4$

A.3 findAS

findAS is a tool for inspecting Splice Site and Intron Coverage with multiple source of information. We have to provide a gff file with the gene prediction that we want to explore, a sorted bam file with multiple read group (RG tag) and a text file with your RG. samtools must be available somewhere in the pc.

```
findAS [options] <RG> <gff3> <bamfile>
```

Mandatory:

- The <bamfile> MUST be sorted!
- Also a <bamfile>.bai file MUST be present in the <bamfile> location
- The <bamfile> MUST contain for each reads an RG tag (Read Group)This pipeline is written to use multiple RG. TAKE CARE!
- The <gff3> must contain CDS and UTR tag.
- The <RG> is a file with the READ GROUP that you want to use.

Example 1:

```
findAS -d -w workdir -j test001 rg.txt \
  chr1.gff3 alignment_sort.bam
```

Example 2:

```
nohup findAS -n -d -w workdir -j test001 \
 rg.txt chr1.gff3 alignment_sort.bam &
```

Note: Always use -n if you run this script with nohup! test001_start.cvs:

```
contig mRNA     pos   isNew EC ali   cab
chr1   VIT_020  4642  mRNA E  343535 8565
chr1   VIT_020  55516 mRNA A  456    9
chr1   VIT_020  55751 mRNA A  65477  6755
chr1   VIT_0201 923   mRNA A  788    755
chr1   VIT_0t04 26    mRNA A  8      8
```

Options:

```
--version             show program's version
                      number and exit
-h, --help            show this help message
                      and exit
-d, --debug           With this option findAS
                      switch the logging level
                      to DEBUGGING MODE
-n, --nohup           The progress bar have
                      some problem when you
                      redirect the stdout to
                      a file. Use -n to remove
                      all the progress bar from
                      stdout logging
-w WDIR, --outputDir=WDIR
                      Working Directory:
                      This directory MUST EXIST!
                      [Default: ./]
-j JOBNAME, --jobName=JOBNAME
                      findAS create a folder
                      inside the working
                      directory with this
                      jobName. Also each output
                      file will be tagged with
                      this jobName [Default: findAS]
-s SAMTOOLS, --samtools=SAMTOOLS
                      If you want to change
                      samtools location.
                      [Default: samtools]
```

A.4 drawAS

drawAS generate a navigable picture for a gene model that we have analysed with findAS. Using the same working directory and the same job name, drawAS is able to print the following picture A.1.

Usage:

```
drawAS -w workdir -j test001 -m GSVIVG01032864001 \
chr1.gff3
```

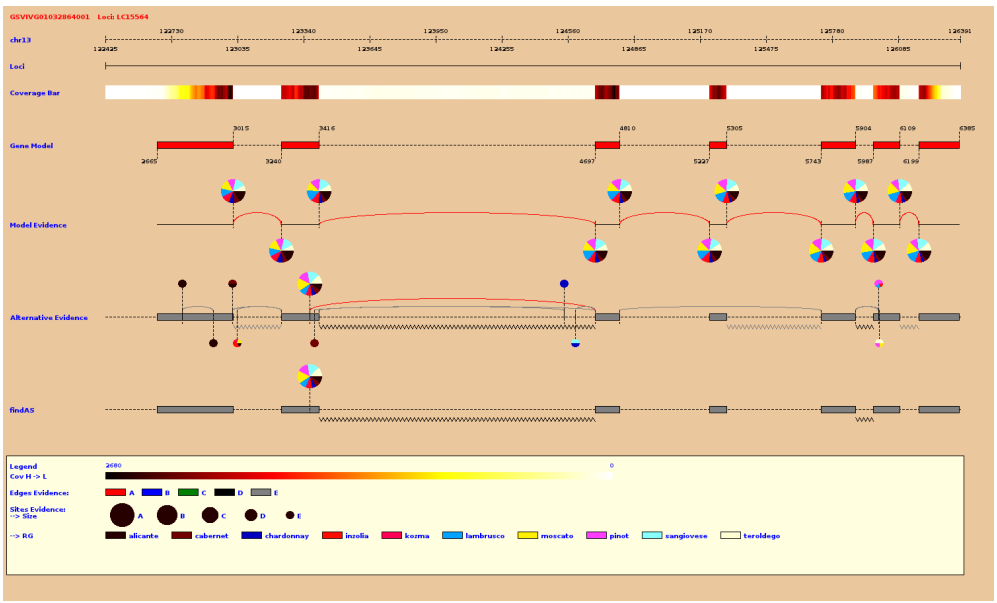


Figure A.1: drawAS: GSVIVG01032864001 alternative splicing predictions

In the first line is represented the genome coordinates, the second line shows the locus (or more then one if needed) coverage; the third line shows the

read coverage and then the gene model. In the last three lines are represented the splicing evidences in the gene model, the alternative splicing evidences and the AS prediction after findAS detection pipeline. A bigger pie correspond to a bigger coverage, the color of ES and IR correspond to the event coverage.

A.5 findCLEAN

findCLEAN is a tool for cleaning a BAM files. We have to provide only a sorted BAM file and set up the filters.

A.5.1 Filters

findCLEAN check all the reads (line by line). Available filters :

- **NH**, Number of hits
- **NM**, Number of Mismatch
- **EXONside**, Check the length of only external exons in spliced reads (–gapstep for set up the definition of exon and intron based on alignment gap length)
- **EXONall**, Check the length of all the exons in spliced reads (–gapstep for set up the definition of exon and intron based on alignment gap length) Len EXON $<= 10, > 5$ [Default None]
- **INTRON**, Check the length of all the introns in spliced reads (–gapstep for set up the definition of exon and intron based on alignment gap length)
- **OPT**, Filter for some other option that you must specify, see in samtools documentation for available options
- **PAIR**, Check if the reads is proper pair. A proper SAM flag must be specified in you BAM file

A.5.2 Results

- A cleaned BAM file, if you specify –sort and or –index you will obtain a sort BAM and its index (outuput file name = WDIR + JOBNAME + _clean.bam)
- Another BAM file with only trashed reads (trash file name = WDIR + JOBNAME + _trash.bam)

A.5.3 Basic Usage

Usage:

```
findCLEAN [options] <bamfile>
```

Mandatory:

- The <bamfile> MUST be sorted!
- Also a <bamfile>.bai file MUST be present in the <bamfile> location

NOTE: you have to set up the interval that you consider clean, findCLEAN remove all the rest

EXAMPLE 01 $--> $ discard all the reads with 3+ alignment:

```
findCLEAN --NH "<3" input_sort.bam
```

EXAMPLE 02 $--> $ discard all the reads with CC equal to chr16:

```
findCLEAN --OPT "CC:!=chr16" input_sort.bam
```

EXAMPLE 03 $--> $ discard all the reads with external exon shorter than 10 or longher than 70:

```
findCLEAN --EXONside ">10,<70" input_sort.bam
```

Filters Setup:

```
--NH=NH               Number of Hits
                      "<=10,>5"... [Default None]
--NM=NM               Number of Mismatch
                      "<=10,>5"... [Default None]
--EXONside=EXONSIDE
                      Len External EXON
                      "<=10,>5"... [Default None]
--EXONall=EXONALL     Len EXON "<=10,>5"
                      ... [Default None]
--INTRON=INTRON       Len INTRON "<=10,>5"
                      ... [Default None]
--OPT=OPT             MANUAL options
                      "XS:<=10,HI:>5"...
                      [Default None]
--PAIR                check if is proper pair...
                      [Default False]
--gapstep=GAPSTEP     min gap (insertion
                      or deletion) to call
                      exon [Default  20]
--sort                sort the clean file
                      [Default False]
--index               sort and index the
                      clean file [Default False]
```

Options:

```
--version             show program's version number
                      and exit
-h, --help            show this help message and
                      exit
-d, --debug           With this option findAS
                      switch the logging level to
                      DEBUGGING MODE
-n, --nohup           The progress bar have some
                      problem when you redirect
                      the stdout to a file. Use
                      -n to remove all the
                      progress bar from stdout
                      logging
-w WDIR, --outputDir=WDIR
                      Working Directory: This
                      directory MUST EXIST!
                      [Default: ./]
-j JOBNAME, --jobName=JOBNAME
                      findAS create a folder
                      inside the working directory
                      with this jobName. Also
                      each output file will be
                      tagged with this jobName
                      [Default: findAS]
-s SAMTOOLS, --samtools=SAMTOOLS
                      If you want to change
                      samtools location.
                      [Default: samtools]
```

Contents

List of Figures

List of Tables


variety
sequence
specii
order
NGS
number
alternative
le
also
one
identii
gene
biology
berry
sequencing
genes
alignment
comparison
among
multiple
Expression
Introduction
Figure
grapewine
coverage
called
base
tion
reads
function
results
Alternative
expressed
genome
Noir
data
Sangiovese
junctions
step
grape
library
splice
Chardonnay
complexity
two
line
acid
Kozma
expression
exon
several
evaluation
Material
showed
intron
Fig
example
alignments
GO
Vinifera
mRNA
RNA-seq
read
junction
RNAs
genetic
cultivar
di
form
events
shown
Vitis
found
Illumina
studies
must
length
Alicante
sites
grapevine
site
transcriptome
transcripts
many
new
Teroldego
regulation
development
using
total
i.e
functional
novel
evidence
Nature
within
cell
study
Protein
level
form
mapping
process
quality
amount
way
per
ed
used
SJs
generated
splicing
annotated
rst
DNA
differentially
complex
transcript
research
exons
analysis
available
different
Lambrusco
model
Gene
sequences
cDNA
types
cultivars
detection
long
transcription
Digital
ndTools
single
large
human
Pinot
possible
Cabernet
RNA
high
al
constitutive
ndAS
use
AER

If you put your mind to it,
you can accomplish anything.

Marty McFly