UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

Facoltà di Scienze MM.FF.NN.

Dipartimento di Biologia

Scuola di dottorato di ricerca in Biochimica e Biotecnologie

Indirizzo di Biotecnologie

CICLO XXII

Genomic analysis and identification of polymorphisms in grape by second generation sequencing

Direttore della Scuola : Ch.mo Prof. Giuseppe Zanotti

Supervisore : Ch.mo Prof. Giorgio Valle

Dottorando : Chiara Rigobello

A.A. 2009/2010

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.

Padova, 07/09/2010 Chiara Rigobello

A copy of the thesis will be available at http://paduaresearch.cab.unipd.it/

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.

Padova, 07/09/2010 Chiara Rigobello

Una copia della tesi sarà disponibile presso http://paduaresearch.cab.unipd.it/

Abstract

Recently, an extensive amount of genomic data has been collected for grapevine (*Vitis vinifera*), culminating with the complete sequencing of the genome (August 2007) of a highly homozygous lineage of Pinot Noir (PN40024). My group have focused its research on this cultivar of Pinot with the multiple goals of genome assembly, gene identification and annotation, transcriptome analysis and identification of polymorphisms. Genomic projects heavily depend on genome annotations and are limited by the current deficiencies in the published predictions of gene structure and function. For this reason an improved annotation will allow better data mining of the grape genome, and more correct planning and design of next experiments. Moreover in the genomics era, many of the experiments useful to confirm the identification of gene and their function can be achieved using high-throughput methods: for example, whole genome sequencing and massive parallel transcriptome analysis obtained by means of second generation sequencers (SOLiD, Applied Biosystem; Solexa, Illumina; 454, Roche). In addition, these methodologies are suitable for re-sequencing strategies in order to identify variations (polymorphisms) that could explain differences in phenotype. During my PhD, I was involved in the sequencing project of *Vitis vinifera* genome to gain a 2 X coverage of genome sequence via traditional Sanger method. In a second moment, with the introduction of a "new generation sequencer" (SOLiD$^{\mathrm{TM}}$, Applied Biosystem) in my lab, I was able to perform a new kind of DNA sequence analysis (through sequencing by ligation system) which produces a larger amount of data (Giga bases per run) in comparison with Sanger method. I have applied this new technology in testing the sequencing efficiency and in discovering polymorphisms in *Vitis vinifera* cultivars of Merlot and Prosecco. The sequencing of the homozygous lineage of Pinot noir has been achieved through a "whole-genome-shotgun" (WGS) approach. It implies the shearing of DNA in random fragments, the cloning in a vector, and at the end the sequencing of the cloned insert. Out of three tested amplification methods (PCR, Tepli 29 kit and Millipore mini- preparation), the miniprep was used for the majority of the template since it produces very reliable results in terms of reproducibility and template quality. The amplification process is a necessary step in Sanger sequencing to read the signal on automatic capillary electrophoresis machines. The data output ("reads") are given by electropherograms, 900-1000bp long, collected by a software. The amount of coverage for grape was 12 X, that means that the consortium produced 480 Mb (genome size) 12 times: 5.7 Gb. My lab contributed with a 2 folds genome coverage. A draft genome sequence has been obtain in August and led to the publication of a paper in Nature: "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla."

[Nature, 499, 463-468 (2007)]. The availability of the SOLiD technology in my lab have allowed a specific study on the discovery of polymorphisms through re-sequencing of Merlot and Prosecco cultivars. The sequencing of grape genome has been performed on a particular homozygous lineage in order to have a determined reference sequence. Pinot Noir in nature is highly polymorphic with two clearly distinguishable haplotypes revealing millions of SNPs. This represent a powerful resource for molecular breeding programs and QTL markers association studies. In fact, once the initial sequence for a particular genome is available, it is then possible to perform comparative sequencing or re-sequencing to identify polymorphisms, mutations, and structural variations between organisms. Whole genome re-sequencing requires a highly parallel system to provide the depth of coverage required for variant detection. Library preparation is also critical as the complexity and time involved are multiplied when analyzing multiple genomes.

The choice of these two particular cultivars resides in several aspects:

1. availability of source samples (supplied by prof. F. Lo Schiavo and prof C. Bonghi - University of Padua);
2. different growth conditions;
3. autochthonous cultivar origins (Merlot is cultivated in Monselice and comes from a French clone, while Prosecco is a real Veneto-grape);
4. sparse of genomic information on these two specific cultivars.

Assuming these information, mate pairs libraries were created for the two examined cultivars in order to possibly evaluate polymorphisms presence within the two cultivated varietas. These libraries were used in a standard sequencing run on SOLiD$^{\mathrm{TM}}$3. On the average, 7 Gb of sequences have been produced for Merlot and Prosecco and about 1.2 million SNPs and 2.2 million SNPs were identified respectively through bioinformatics analysis. These large amount of data produced will be analyzed to obtain further information. Variations in sequence will be tested via PCR of random sampled polymorphic sites to confirm bioinformatics suggestions. Moreover, analysis of specific gene sets will be useful in investigating differences within gene family or between families. All variations are going to be mapped in the *Vitis vinifera* GBrowse as SNPs Merlot and SNPs Prosecco entries. Each entry shows the modified base, the modified codon and the possibly modified amino acid. The last part of the research investigates structural variations (SVs). Preliminary results have been observed, indicating some interesting zones to be better understood. The limit of bioinformatics analyses is the "low" coverage obtained taking into account only the right (mates mapped with the right distance and orientation against the reference genome) positioned pairs of the mate-pairs library. The large amount of produced data offers the possibility to investigated several aspects of genes relationship and

regulatory mechanisms. In particular, a more accurate analysis of rearrangements in coding regions will be conducted to verify the nucleotide diversity and the mutation rate among cultivars.

Sommario

La quantità di dati genomici (ESTs, geni, proteine) disponibili per la vite (*Vitis vinifera*) è, ad oggi, molto ampia. Il risultato più importaante lo si è raggiunto nell'agosto del 2007 con il sequenziamento dell'intero genoma di una linea altamente omozigote, ed appositamente creata, di Pinot Noir (PN40024). Il mio gruppo ha incentrato la sua ricerca su questa cultivar di Pinot con l'intento di completare l' assemblaggio del genoma, di identificare i geni e annotarli (cioè di descriverne la composizione), di studiarne il trascrittoma e infine di identificare i polimorfismi. I progetti di genomica dipendono fortemente dall' annotazione e sono vincolati da eventuali carenze nelle predizioni sulla struttura e sulla funzione genica. Per questo motivo, un miglioramento nella fase di annotazione si riflette in una più precisa descrizione dei dati ottenuti dal sequenziamento del genoma e una conseguente pianificazione degli esperimenti più corretta. Inoltre, nell'era della genomica e grazie a dei metodi high-throughput, possono essere sviluppati in parallelo degli esperimenti di identificazione genica e/o tesi a descriverne la loro funzione con un output molto elevato: il sequenziamento di interi genomi o l'analisi del trascrittoma possono venir ottenuti grazie a singoli esperimenti con sequenziatori di seconda generazione (SOLiD (Applied Biosystems) Solexa (Illumina) e 454 (Roche)). Queste metodologie sono adatte per le strategie di ri-sequenziamento di interi genomi, con l'intento di identificare varianti (polimorfismi) genotipiche che potrebbero spiegare le differenze a livello del fenotipo. Durante il mio dottorato, sono stata inizialmente coinvolta nel progetto di sequenziamento del genoma di *Vitis vinifera* intrapreso da un Consorzio europeo (I.G.G.P.) con l'intento di sequenziare il genoma in modo che ogni base fosse rappresentata 12 volte (12 X coverage per base). Il mio gruppo ha partecipato al progetto di sequenziamento per una quota di 2 genomi equivalenti attraverso l'approccio Sanger. La disponibilità della sequenza genomica di vite potrebbe aiutare i ricercatori a comprendere meglio alcuni caratteri comuni ad altre piante da frutto. In particolare, considerando l'alto tasso di eterozigosità delle varie cultivar di *Vitis*, le differenze tra le varietà dovrebbero scaturire dalla valutazione dei polimorfismi condivisi e quelli specifici per la singola cultivar. In un secondo momento, con l'introduzione di un sequenziatore di nuova generazione (SOLiD$^{\text{TM}}$, Applied Biosystems che sfrutta il sistema di "sequenziamento per ligazione") nel mio laboratorio, ho avuto l'opportunità di applicare questa nuova tecnologia nell'identificazione di polimorfismi in due cultivar di *Vitis vinifera*: Merlot e Prosecco. L'obiettivo del progetto era quello di avere il maggior numero possibile di marcatori al fine di disegnare eventualmente una mappa genetica per la singola cultivar. E' ben noto che la disponibilità di marcatori genetici offre la possibilità di idagare i genotipi e valutare le differenze tra

le specie o le sottospecie. Le mappe genetiche consentono di facilitare le tecniche di allevamento delle piante (breeding) e la ricerca genomica, individuando gli alleli migliori associati a caratteri "positivi" o alleli che portano, ad esempio, alla suscettibilità rispetto ad alcuni patogeni o a determinate condizioni ambientali. Il sequenziamento della linea omozigote di Pinot nero è stato ottenuto attraverso un approccio "Whole genome shotgun" (WGS) che implica la frammentazione casuale del DNA, il clonaggio in un vettore, l'amplificazione e il successivo sequenziamento dell'inserto clonato. Dei tre metodi testati per l'amplificazione (PCR, Tepli 29 kit Millipore e mini-prep), la miniprep è stata scelta per amplificare la maggior parte dei templati. Questo perchè durante lo svolgimento di questa ricerca la tecnica della mini-prep ha prodotto dei risultati molto affidabili sia in termini di riproducibilità che di qualità dell'inserto. Il processo di amplificazione è un passo necessario per il sequenziamento Sanger. I dati (reads) sono prodotti sottoforma di elettroferogrammi, lunghi tra le 900 e le 1000 bp, che sono successivamente raccolti da un software. E' stato sequenziato un totale di 12 X coverage del genoma, che corrisponde a circa 5,7 Gb di sequenza.

Un primo consensus del genoma della vite corrispondente all' 8,4 X coverage è stato ottenuto nell'agosto del 2007 e un articolo è stato pubblicato su Nature: "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla." [Nature, 499, 463-468 (2007)]. La disponibilità della piattaforma SOLiD (Applied Biosystems) nel mio laboratorio, mi ha permesso di condurre un esperimento sull'identificazione dei polimorfismi attraverso il re-sequencing delle cultivar di Merlot e Prosecco. Il sequenziamento Sanger del Pinot nero è stato effettuato su un ceppo omozigote in modo da avere una precisa sequenza di riferimento priva ($< 3\%$) di siti in eterozigosi. Il Pinot nero in natura è altamente polimorfico, con due aplotipi ben distinguibili che rivelano milioni di SNP. Questo aspetto della vite rappresenta una potente risorsa per i programmi di miglioramento genetico e molecolare. Una volta che la sequenza di una particolare specie è disponibile, è possibile poi eseguire degli esperimenti di sequenziamento comparativo o ri-sequenziamento di altri genomi correlati per identificare polimorfismi, mutazioni e variazioni strutturali. Questo tipo di studi, però, necessita della disponibilità di una reference (un genoma a cui fare riferimento) ed un sistema ad alta processività che fornisca la copertura (numero di reads per base) necessaria per il rilevamento di una variante. Un altro punto critico del re-sequencing è la preparazione delle librerie di DNA che è molto complessa e impegna tanto tempo considerando l'analisi multipla dei genomi da confrontare. Per questi motivi l'uso dei sequenziatori di nuova generazione è innovativo: gli esperimenti di re-sequencing sono eseguiti in parallelo su diversi genomi con un notevole rispormio di tempo.

La scelta di queste due cultivar in particolare è dovuta a diversi aspetti:

1. la disponibilità di campioni (fornito dalla prof F. Lo Schiavo e dal prof

C. Bonghi - Università degli Studi di Padova);

2. le diverse condizioni di crescita;
3. l'origine autoctona delle cultivar (il Merlot proviene da una coltivazione in campo nei pressi di Monselice e deriva da un clone francese, mentre il Prosecco è un vero e proprio vitigno veneto);
4. l'esiguità di informazioni genomiche su queste due specifiche cultivar.

Considerando tutti questi aspetti, due librerie mate-pairs sono state create, una per ogni cultivar a cui è seguita una corsa di sequenziamento standard sulla piatttaforma SOLiD$^{\text{TM}}$ 3. Successivamente i dati prodotti sono stati analizzati per l'identificazione di eventuali polimorfismi. Sono state prodotte per il Merlot 8,4 Gb di sequenza genomica, mentre per il Prosecco 6,8 Gb. Grazie all'uso di un software specifico di allinemento di short reads, circa 1,2 milioni di SNP e 2,2 milioni di SNP sono stati identificati rispettivamente. Ulteriori studi sono necessari per approfondire questa prima analisi dei dati. Le varianti individuate saranno inoltre testate mediante una PCR di pool di SNP casuali per confermare le analisi bioinformatiche. L'analisi di specifici set di geni sarà utile per indagare le differenze all'interno di una famiglia genica o tra famiglie. Tutte le variazioni sono state mappate nel GBrowse della vite come SNP di Merlot e SNP di Prosecco. Ciascuna evidenza indica il cambiamento di base, il codone che nel caso viene modificato e l'amminoacido che eventualmente cambia. Durante questo studio ho cercato di identificare anche le variazioni strutturali (SVs). Sono stati ottenuti dei risultati preliminari che portano all'identificazione di alcune "aree" di particolare interesse, soprattutto per quel che riguarda le delezioni definite *large*. Il limite delle analisi bioinformatiche per il rilevamento delle differenze è spesso dovuto ad una bassa copertura del genoma. In questo caso, prendendo in considerazione solo le coppie corrette della libreria mate-pairs, cioè quelle coppie con corretto orientamento reciproco e che mappano ad una giusta distanza nel genoma di riferimento, si è ottenuta una buona copertura fisica (50 X per il Merlot e 141 X per il Prosecco) e una bassa copertura di sequenza (1,5 X Merlot e 3,5 X Prosecco). Quest'ultimo dato, in ogni caso, se preso in considerazione assieme al coverage fisico, fornisce alcune importanti indicazioni sui riarrangiamenti genomici. Si può quindi affermare che, la grande quantità di dati prodotti dai sequenziatori di nuova generazione offre la possibilità di studiare in parallelo diversi aspetti che riguardano le relazioni tra i geni e i meccanismi che regolano le loro funzioni. Il problema sorge nell'analisi ed interpretazione corretta dei dati stessi; infatti, una pianificazione della ricerca non corretta potrebbe portare ad un grosso spreco di risultati. Per quanto riguarda questo specifico studio, è neccessaria un'analisi più accurata dei riarrangiamenti nelle regioni codificanti per verificare la diversità nucleotidica e il tasso di mutazione tra le cultivar.

Contents

1 Introduction **3**

1.1 *Vitis vinifera* 4

1.2 *Vitis vinifera* genome 6

1.3 Pinot, Merlot and Prosecco: a brief description 8

1.3.1 Pinot noir 8

1.3.2 Merlot 9

1.3.3 Prosecco 10

1.4 Genetic variation in grapevine 12

1.5 Polymorphisms 12

1.6 SNPs 13

1.7 Structural variation 14

1.8 From the Sanger technology to next generation sequencing 17

1.8.1 The SOLiD$^{\text{TM}}$ system 20

1.9 Next generation sequences analysis 25

2 Aim of the research **27**

3 Methods **29**

3.1 Plant material 29

3.2 *Vitis vinifera* genome project 30

3.2.1 Library amplification 30

3.2.2 Sequencing reaction and run on ABI 3730*xl* 31

3.2.3 Assembly and annotation 31

3.3 Next generation sequencing of Merlot and Prosecco 32

3.3.1 Genomic DNA extraction 32

3.3.2 RNase treatment 32
3.3.3 SOLiD library: mate pair protocol 32
3.4 Computational analysis 33
3.4.1 Mate-pair alignment 33
3.4.2 SNPs and SVs analysis 35

4 Results and Discussion 39
4.1 Whole genome shotgun of Pinot noir 39
4.2 Sequencing by ligation and coverage 42
4.3 Organelles DNA contamination 44
4.4 The mate-pairs library performance 46
4.5 SNPs detection 48
4.6 Analysis of synonymous and non-synonymous SNPs 50
4.7 Mapping on the GBrowse 51
4.8 Identification of Structural Variants 52

5 Conclusions 55

A Plant culture 59
A.1 Merlot culture condition: spurred cordon 59
A.2 Prosecco culture condition 59
A.2.1 MS-$\frac{1}{2}$ medium preparation 59

B Abbreviation 61
B.1 **SI units** 61
B.2 **SI prefixes** 61
B.3 **Other abbreviations and terms** 61

Bibliography 70

Acknowledgements

I'd like to thank firstly Prof. Giorgio Valle for giving me the opportunity to do a PhD studentship.

A special thank to Dr. Alessandro Vezzi that during these three years supported me and gave me some practical helps. In addition, he has spent these last weeks in evaluating my thesis.

A particular thanks to Dr. Michela D'Angelo, Dr. Riccardo Schiavon and Dr. Stefano Campanaro for their ever accurate expertise.

I'd also like to thank my colleagues at CRIBI who have helped and encouraged me. I've surely found not only working people, but friends. Thanks to Andrea, Francesco, Elisa, Erika, Riccardo, Monica, Alessandra, Fabio, Elisa, Alessandro, Claudio, Nicola and Davide.

Thanks to my parents, who encouraged me during all my studies and to my parents in law, who helped me in many ways.

And last, but not least, I'd like to thank my husband, Daniele for his increased levels of patience. There are no words which can explain how I'm gratitude to him. Thanks, thanks, thanks! And finally, thank to my son, Alessandro, for offering me a free smile every time it was needed. He has the power to change my life.

CHAPTER 1

Introduction

The sequencing of grape (*Vitis vinifera* L.) genome [1] is one of the most important goals in plant genomics, not only from a biological-agricultural, but also from an economical point of view. Grape is, in fact, the first crop fruit to be sequenced and the most widely cultivated one in the world. Therefore, the knowledge of its genome could help researchers to better understand peculiar species-specific characters, as well as features that are common to fruit plants. To achieve this target, the IGGP (International Grape Genome Program) promotes collaborations among European groups; our group is part of a French-Italian consortium and it had the task of sequencing of 2 genome equivalents (on a total amount of 12 X coverage). In order to have a complete analysis of grapevine genomic sequences, the IGGP project also includes: markers discovery and mapping, BAC libraries construction, the build of physical maps, ESTs and transcriptional profiling, functional analysis, and bioinformatics. The recent availability of genomes of other important species (*Arabidopsis thaliana*, *Oryza sativa*, *Medicago truncatula*, *Lycopersicon esculentum* and *Populus trichocarpa*), in addition to the relatively small size (480 Mb) of the grape genome, will lead to the real prospect of making sense of these data in a reasonable amount of time.

My group has focused its research on the sequenced cultivar of Pinot Noir with the multiple goals of genome assembly, gene identification and annotation, transcriptome analysis and identification of polymorphisms. Genomic projects heavily depend on genome annotations and are limited by the current deficiencies in the published predictions of gene structure and function. For this reason an improved annotation will allow better data mining of the grape genome, and more correct planning and design of future experiments. Moreover in the genomics era, many of the experiments useful to confirm the identification of gene and their function (mutant phenotypes, examination of expression profiles, confirmation through biochemical assays) can be achieved using high-throughput methods: for example, whole genome sequencing and massive parallel transcriptome analysis obtained by means of second generation sequencers (SOLiD, Applied Biosystem; Solexa, Illumina; 454, Roche). In addition, these methodologies are suitable for re-sequencing strategies in order to identify variations (polymorphisms) that could explain differences in phenotype.

1.1 *Vitis vinifera*

Grape is the world's most economically important crop fruit, but it also has ancient historical connections with the development of human culture: the first appearance of *Vitis vinifera* (L.) has been dated between 60 to 70 million years ago [2].

Cultivated grapes were domesticated from the wild *V.vinifera* subsp. *sylvetris*, which have been found widely in the Northern Hemisphere. The wild grapevine is a heliophilous liana growing generally along river banks and in alluvial and colluvial deciduous and semi-deciduous forest. It was yielded by early farmers both for nutritional and therapeutic properties; the first written can be found in an ancient Sumerian text from the third millennium BCE ("*A carnelian tree was in fruit, hung with bunches of grapes, lovely to look on*" - Epic of Gilgamesh). There are several evidences of grape cultivation in all the major ancient cultures from Egyptian to Etruscan, Greeks and Roman (Figure 1.1).

Figure 1.1: *Winemaking in ancient Egypt.*

Vitis vinifera was firstly domesticated and then methods of harvesting and wine making were optimized to obtain a beverage which was considered divine, “a drink of the gods”: even Dionysus and Bacchus were dedicated to this beverage.

Between the fifth and tenth centuries, viticulture was sustained almost exclusively by religious orders in monasteries extending the grape growing and planting new vineyards. During the Middle Age and Renaissance, viticulture and its related economic activity knew different development, due to specific culture events. The expansion of Islam in European regions caused its decline, whereas population concentration in towns increased investments in wine production. In the Modern Age, traditional culturing techniques have been replaced by scientific methods based on microbiology, chemistry and ampelograpy due to social and economic changes. The main product, wine, drives the global market to the production of a selected number of these cultivars which are generally classified according to their final usage: wine grapes, table grapes and raisins.

The grapevine belongs to the botanical family *Vitaceae*, which consists of almost one thousand species, grouped into 17 genera and even used as ornamentals in gardens (*Parthenocissus quinquefolia* and *P. tricuspidata*). The genus *Vitis* consists of about 60 inter-fertile species and, among them, *V.vinifera* is currently present in two forms: *V.vinifera* subs. *sylvestris* (the wild one) and the cultivated *V.vinfera* subps. *vinifera* (or *sativa*). This separation is historically due to morphological differences [3] which have brought the mostly cultivation of the subspecies *sativa* with the resulting origin of thousands of different cultivars [4]. Belonging to the cultivated species of *Vitaceae*, *Muscadinia rotundifolia* has to be taken into account; it is one of the three species of the genus *Muscadinia* (Figure 1.2).

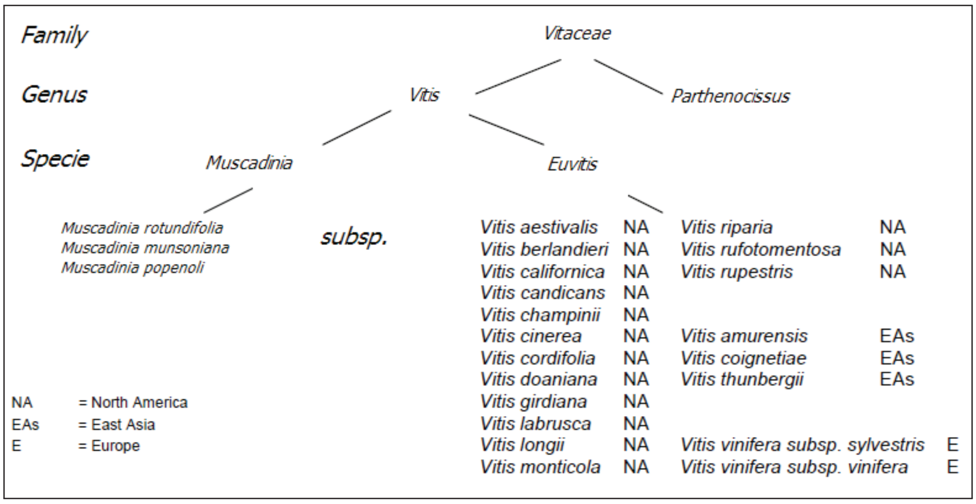


Figure 1.2: *Taxonomy of genus Vitis.*

In relation to the karyotype, the genus *Vitis* has 38 chromosomes that form 19 bivalents at meiosis (2n=2x=38), while the other *Vitaceae* genera have multiples of 10 chromosomes [5]. During the domestication, the biology of grape underwent numerous significant changes to ensure greater content for better fermentation, greater yield and more regular production. Without doubt, the changes in berry shape and the transformation from dioecious wild plants to hermaphrodite cultivated plants were crucial. The domestication process could have involved several independent events and a low number of sexual generations, including spontaneous cross hybridizations with wild populations [6] or it could be happened quickly through mutations, selection and subsequent propagation by vegetative multiplication. According with these characteristics, the grapevine genome is highly polymorphic.

Cytological observations of F1 hybrids between *Vitis vinifera* and *Muscadinia rotundifolia* (2n = 40), suggested an allopolyploid origin of the *Vitis* genome [7]. Jaillon et al. (2007) [1] proposed that the grapevine genome was closer to the common ancestor of dicotyledonous plants and their analysis suggested that all dicots arose from a hexaploid ancestor (three haploid genome equivalents). The lack of recent whole genome duplications in grapevine let to assign differences within each subfamily either to an ancestral polyploidization event predating the divergence of those three species or to later duplication events within each lineage.

The grape ancestor, more similar to the modern wild variety, diverges from current cultivars in numerous traits. The genetic relationship between the wild and the cultivated forms could enable the identification of regions of the genome that have undergone a strong selection during the domestication process, and thus identify genes controlling such traits. However, only one example of a direct relationship between a wild and cultivated individual has been published [8], suggesting the absence of gene flux between wild and cultivated compartments. Thanks to researchers, wild individuals have been identified among Europe, even if it's hard to check they have never undergone cultivation or they are hybrids between wild and cultivated forms. Therefore, more accurate information on the geographical and genetic origins of the genotypes is needed, together with the study of haplotypes.

1.2 *Vitis vinifera* genome

The sequencing of grape genome was performed on the quasi-homozygous genotype PN40024 (Pinot noir), created by INRA in Colmar [9] to obtain a 12 genome equivalents. A Whole Genome Shotgun (WGS) approach was employed with different plasmid libraries (size range 3-10 kb). A HindIII BAC library of 70,656 clones (80 kb) and a fosmid library (40 kb) were sequenced to obtain 3'ends used in the assembly to join contigs into scaffolds.

	Library type	**Insert size Reads** (millions of bp)	**Coverage**
Plasmid (high copy number)	3kb	3.04	4.6 x
Plasmid (low copy number)	10 kb	2.07	3.6 x
Fosmids	40 kb	0.03	0.04 x
BACs	100 kb	0.1	0.16 x
total		6.23	8.4 x

Table 1.1: *Sequencing overview.*

The assembly have produced a golden path of the genome of about 480 Mb in length, three times higher than Arabidopsis (125 Mb)[10], but relative small if compared to maize (2,8 Gb)[11]. The knowledge of genome sequence will lead to transfer molecular mechanisms regulating agronomic characters to grape and makes possible the characterization of germoplasm present in worldwide collections.

Once the reads were produced (August 2007), a *draft* genome of *Vitis vinifera* was created (assembly 8.4 X coverage: April 2008, (Table 1.1)). Scaffolds were positioned along the physical map of the chromosomes creating a "golden path" or consensus. The assembly is a crucial computational step, because many genomes contain large numbers of repeats present in different locations (about 40% of the grape genome is made of repetitive/transposable elements). According to the 8 X prediction [1], the 19 chromosomes contain about 30,000 protein-coding genes. The prediction on the 12 X assembly is still in progress and will be completed in a few months.

With the holding of genomic information, it is possible to ameliorate cultivation conditions to cope climate changes, the emergence of new diseases, environmental protection imperatives and consumer behavior. In addition, functional genomics approaches find homologous genes and reveal protein interactions, detecting alleles implicated in specific mechanisms. Information about genes and their relationship is necessary to understand biological qualities as aromatic compounds (synthesis of anthocyanins, flavonoids, polyphenols and other secondary metabolites), which have related rules in phenotypic appreciable traits and wine taste. On the other side, functional genomics is applied to investigate growth and maturation steps as berry quality, biology of reproduction, resistance to pathogens and growth conditions in relation to environment. In particular, the sequencing will make possible to obtain a great number of markers of the physiological state of the plant in the vineyard, in order to develop tools for a more precise viticulture which is more sustainable and of high quality. To date, new high-throughput sequencing technologies generate greater amounts of DNA sequence quicker and cheaper in comparison with standard Sanger sequencing. So, numerous studies are undergoing to better explain and understand genes and their function. In particular, targets for genome re-sequencing of grape might be: identification of nucleotide variation (SNPs) between the reference and other cultivars, profiling copy number variation (CNV), identification of unique sequences into each varietas and the generation of *de novo* assembly of analyzed cultivars. These data collection could be applied in phenotypic observation studies providing researchers the genetic basis of specific traits.

1.3 Pinot, Merlot and Prosecco: a brief description

1.3.1 Pinot noir

Pinot noir is a red wine grape variety (Figure 1.3). The name comes from the French words for "pine" and "black" referring to the clustered dark purple pine cone-shape. Pinot noir grapes are grown in the cooler regions around the world, but the Burgundy region of France is the most interested region. It produces some of the finest wines in the world, but is a difficult variety to cultivate. It is sensitive to light exposure; it has low yields; it relies on soil types and pruning techniques. The particularly thin skin makes it extremely susceptible to fungal diseases. In the broadest terms, the wine tends to be of light to medium body with an aroma reminiscent of black cherry, raspberry or currant. Pinot noir is an ancient variety that may be only one or two generations separated from wild vines and it has been proposed to be a cross between Pinot meunier and P.traminer [12]. Pinot meunier is a chimera, indeed; it has a mutation in epidermal cells that makes the plant a little smaller. On of the two layers of the epidermis is identical to Pinot noir, so Pinot meunier cannot be the parent of Pinot noir. During past years, other Pinot cultivars have been introduced in viticulture: Pinot gris, Pinot blanc, Pinot moure and Pinot teinturier. They all show a similar DNA profile to Pinot noir. The genome of Pinot noir seems to be particularly prone to mutation, maybe suggesting the presence of active transposable elements [13], which give rise to new clones. In Italy Pinot noir has traditionally been

Figure 1.3: *Pinot noir grape.*

cultivated in Alto Adige (since 1838) and Trentino regions.

1.3.2 Merlot

Merlot is a red wine grape (Figure 1.4). The origins of name Merlot reside in the Old French word for young blackbird, merlot, a diminutive of merle, the blackbird (*Turdus merula*). It is a progeny of Cabernet Franc and is a sibling of Cabernet Sauvignon. Merlot is cultivated in cooler regions. It is one of the most relevant *varietas* in the economic market (the third most cultivated vine in the world) and one of the primary grapes of Bordeaux wine. Merlot grapes have a thinner skin and tend to have higher sugar content. Water stress is important to the vine production since grape grows better in well drained soil.

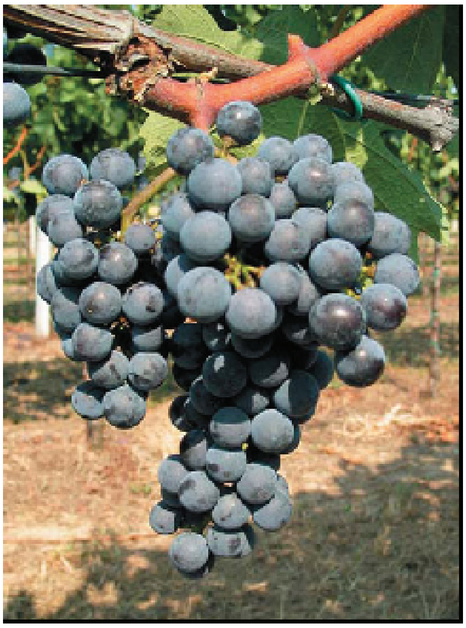


Figure 1.4: *Merlot grape.*

Further than France, it is also grown in the cooler portions of many regions in Italy. A large portion of Merlot is planted in the Friuli, whereas in Tuscany, it is often blended with Sangiovese to give a soft wine. The low acidity which characterizes Merlot wine is often used as a balance for the higher acidity in many Italian wines in Veneto, Alto Adige and Umbria [14]. The "Strada del Merlot" is a popular tourist route through Merlot wine countries along the Isonzo river. Italian Merlots are often characterized by their light bodies and herbal notes. Clonal selection in Italy produces 11 clones which differ in growth and yield depending on the final use of wine. In particular, the "Istituto Agrario di San Michele all'Adige" is analyzing numerous Merlot clones in order to test environmental aspects and genetic variability, allowing a better description of commercial clones.

1.3.3 Prosecco

Figure 1.5: *Prosecco grape.*

While the two above described cultivars comes from a French region, Prosecco is an autochthonous Italian wine (Figure 1.5). Its name is probably linked to Prosecco, a small village near Trieste and it is supposed to be similar to the grape Glera variety. The main area where Glera and Prosecco are produced is Veneto, traditionally near Conegliano and Valdobbiadene, the northern area of Treviso (Figure 1.6). Since the beginning of the XIX century, with the foundation of The School of Viticulture and Oenology in Conegliano, research into this vine variety has greatly increased and the Prosecco has spread throughout the area. The exact origins of this variety are still unknown, but some would have it that it was, in fact, already known as the 'Pucino' in the time of the Roman Empire.

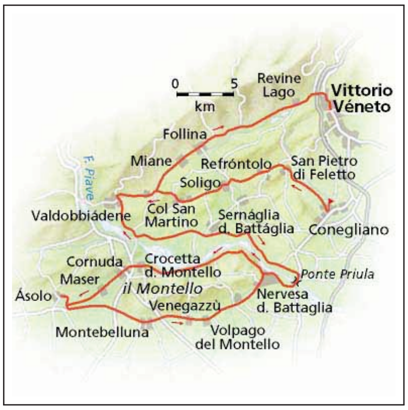


Figure 1.6: *Area of Prosecco, Veneto, Italy.*

Until the 1960s, its flavour was similar to other dry, sparkling wines [15], but modern production techniques, make it a high-quality cultivar producing one of the most popular wine in the world. Its specificity is protected as a DOCG within Italy, as Prosecco di Conegliano-Valdobbiadene, Prosecco di Conegliano and Prosecco di Valdobbiadene. For more than 200 years, Prosecco has been cultivated in this area. This region presents ideal climatic vineyard conditions: the Dolomites protect the vines from cold winds and the Adriatic Sea mitigates the climate for every season. The Prosecco is a vigorous and hardy vine, with nut-coloured shoots and quite large, loosely-packed winged clusters of beautiful golden yellow berries nestled amongst large bright green leaves.

Because of Prosecco is low in alcohol (11 to 12 percent by volume) in comparison with other sparkling wines, in Italy it is enjoyed for every occasion.

	Pinot noir	**Merlot**	**Prosecco**
Leaf	Average size tri-foiled	Average size tri or five-lobed	Average-medium size tri or five-lobed
Cluster	Small cylindrical and thick	Medium-size pyramidal and straggly	Medium-big size pyramidal and straggly
Grape	Small globular	Medium size, globular	Medium size, globular
Skin	pruinose, thick blu-noir coloured	Pruinose blu-noir coloured	Pruinose yellow-gold coloured
Environmental	Hilly freshly lands, sandy soil	Hilly freshly lands, moistly soil	Hilly lands, moistly soil
Pathogen agents	Rots Oidium, iron chlorosis	Rots Downy mildew, cochineal	Oidium Downy mildew, yellows
Wine	Soft red wines with plum flavors not usually high in alcohol. Vinification gives sparkling wines	Red wine, bitterish alcoholic, medium body with hints of berry plum, and currant	dry or off-dry sparkling wine with good acidity and a lightly creamy flavour

Table 1.2: *Relevant ampelograpyhic traits of Pinot noir, Merlot and Prosecco.*

1.4 Genetic variation in grapevine

Grape has a highly heterozygous genotype [16] and any progeny is a potential "new" seed, obtained through the combination of parental alleles giving rise to a phenotypic variation that segregate in descendants. Sexual reproduction, vegetative propagation and somatic mutations are the main processes that have permitted the development of cultivated grape and the new genotypes are obtained by sexual reproduction, either by crossing or self-fertilization.

The juvenile period (three-to-five years) of grapevine plants and the time necessary for evaluation of a trait important for wine production is a bond in selecting a particular phenotype (i.e. berry trait, aromatic compounds). Furthermore, many generations might be necessary to recover the desired traits. The clonal propagation could be a good method in maintaining genotype. However, the occurrence of somatic mutation in one cutting and not in another might eventually bring to plants of the same cultivar having a slightly different genotype ('clonal variation'). Moreover, these variants might exist in only one cell layer of the plant, resulting in genetic chimerism. For these reasons nowadays, cultivars are maintained by vegetative propagation.

Over the last 50 years, the cultivated grapevine has undergone a drastic reduction of diversity, due to the globalization of wine companies and markets, resulting in the emergence of the new familiar worldwide grown cultivars such as Chardonnay, Cabernet Sauvignon, Syrah (Shiraz) and Merlot, and the disappearance of old local cultivars. The sanitary selection of healthy disease-free clones has also induced a reduction in clonal diversity for these major cultivars. Thus, the diversity of existing grapes has been shaped by human history. Several thousand cultivars exist but most of these are largely confined to germoplasm collections [16].

1.5 Polymorphisms

Polymorphisms are different forms of a DNA sequence. These variations are a kind of genetic diversity within a population gene pool, which may or may not affect biological function. They can be used to locate (map) genes [17] and they can help the matching of two samples of DNA to determine if they come from the same source [18]. Their use is commonly applied to several areas, including agriculture. In particular, the identification of a polymorphism associated with a specific plant character represents a genetic marker. It can be used to unveil specific biological patterns and processes, by studying a group of related individuals so that differences in genotype can be determined. Markers identification is useful to manipulate and identify genes associated with advantageous agronomic and quality traits within breeding programs marker-assisted selection (MAS) and cultivars characterization [19], [20], [21]. Informative, abundant, high-throughput markers associated with genes are desirable both for breeding and genetic analysis.

Discovering individual desirable qualities is an important task to transfer information between individuals within the same specie (interbreeding) or among species and to improve plants yield. If enough polymorphisms are analyzed, it is possible to distinguish between individuals with a high degree of confidence, providing a DNA profiling.

These variations arise through mutation which may be due to a change from one nucleotide to another, an insertion or deletion (*indels*), or a rearrangement of nucleotides. Once formed, a polymorphism can be inherited like any other DNA sequence, allowing its inheritance to be tracked from parent to child. Although any polymorphism is a change in the DNA sequence, it can or cannot have an effect on phenotypic qualities depending on its mapping position. If it does not have any consequence on the organism is said to be selectively neutral. Usually, differences in gene portions are describes as different alleles within a population. However, polymorphisms are also found in the non-coding DNA and these regions tend to have more polymorphisms. This fact can be easily explained because of the importance in maintaining DNA sequences that encode for proteins: a variation in a coding region may have a deleterious effect on the individual that carries it.

Traditional methods relying on differential mobility in chromatography or electrophoresis have a high-throughput potential and can be applied to many individuals in the same population, but can identify only the presence of polymorphisms, not the type. On the other side methods based on hybridization on microarrays can discover the kind of variation between two genomes, but are very expensive and they identify less than 50% of polymorphic sites [22].

The development of tools for a more precise agriculture rely on the possibility to access and use the genetic variation present in germoplasm collection and in wild species. In particular, the economic relevance of grape drive to gain a more sustainable viticulture and of high quality. To do this, approaches of Next Generation Sequencing (NGS) are ameliorating to obtain a great number of markers [23].

1.6 SNPs

Single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) are the most abundant type of DNA sequence polymorphisms [24] They represent the finest resolution of DNA sequence and have a low mutation rate (the chance of a mutation occurring in an organism or gene in each generation). They can be used as genetic markers for many genetic applications such as cultivar identification, construction of genetic maps, assessment of genetic diversity, detection of genotype/phenotype associations, or marker-assisted breeding [25]. Furthermore, the development of high throughput genotyping methods makes single nucleotide polymorphisms (SNPs) highly attractive as genetic markers. The principal challenge in SNPs discovery remains the discrimination between true genetic polymorphisms and the often more abundant sequence errors.

SNPs have been found in many plant species through a systematic approach of well-studied model species (Arabidopsis, maize, barley) and in few woody species [24] or derived from large-scale re-sequencing projects thanks to next generation sequencing platforms. In grape, single nucleotide polymorphisms come from BAC and EST libraries and have been successfully employed in building genetic maps, their anchoring to physical map, analyzing genetic diversity and linkage disequilibrium (LD). Furthermore, the recent decoding of the grape genome sequence in the heterozygous Pinot Noir cultivar provided the grape research community with 1,700,000 SNPs from coding and non-coding regions [26].

Different strategies have been applied in grape for SNP detection and genotyping and the presence of many SNPs is a primary trade in the development of grape markers, but it is more challenging in taking advantages of short read sequences. Moreover, most of the SNP discovery and application have been limited to *V. vinifera*. The transferability of SNPs across *V. vinifera* has been restricted to a few cultivars [27] and to a few wild forms. Nevertheless, the knowledge on transferability is fundamental to allow the identification of useful alleles for diversity and association studies.

1.7 Structural variation

A structural variation (SV) is any polymorphism that changes the structure of the genome, including insertions, deletions and inversions: it often results in rearrangements, alterations or fusion in genes and in chromosomal aberrations. While the firsts are copy number variants (CNV), the last is count invariant. Most of these variations are due to genomic rearrangements and a few others contain novel sequences that are not present in the reference genome. Traditional approaches in SV discovering are:

- Whole genome comparative array genome hybridization (aCGH), which tests the relative frequencies of probe DNA segments between two genomes [28];

- SNP arrays using data from Hap Map projects: it measures the intensity of a probe at a known SNP locus, taking into account the allelic ratio at heterozygous sites [29];

- Paired-end mapping through Sanger sequencing which provides a better resolution than the former two [30].

These array-based methods are limited by the density of the probes in turn by the density of the array (for aCGH) or by the density of the known SNPs.

The progress of high throughput sequencing and array technologies has enabled the re-sequencing of entire genomes, especially in identifying and reconstructing the variants in an individual's genome compared to another one. The costs and sensitivities of these technologies differ considerably from each other, and even more technologies are expected to appear in the near future. On one side, Sanger sequencing of genomes leads to excellent and precise results, but at high costs. By contrast, short sequences generated by inexpensive new platforms can easily locate SNPs, but they could not be able to either unambiguously find out SVs in repetitive genomic regions or fully reconstruct many of the large SVs. In particular, the reconstruction of a large SV with paired-end reads needs the combination of data coming from Sanger sequencing or reads spanning a wide regions. Usually, sequencing-based methods use mate-pair or paired-end libraries to discover structural variation. In this approach, two "paired" sequences are generated at an approximately known distance in the genome. The sequenced reads are then mapped to a reference genome and the pairs mapping at a distance that is different from the expected length, can highlight structural variations. However, a single signature is insufficient to identify precisely a variation due to the noise in the signal. Sequencing errors and chimeric reads may result in wrong mapping, whereas chimeric clones will result in misleading information about the distance and orientation between two reads.

Discovering structural variation is a huge objective for heterozygosity, association studies, cancer genomics and molecular evolution. Characteristics of a good method in finding SV are specificity and sensitivity, the ability to discover accurately a breakpoint, setting the variant size and the change in copy count. Usually, to discover structural variants between species, or cultivars, an assembled genome (the reference) and another sequenced genome (the "donor") are both necessary. Reads obtained from the donor sequencing are mapped against the reference. The comparison relies on particular "signatures" (Figure 1.7) that are created by structural variation [31].

- Basic inversions, insertions and deletions: are the most common signatures; considering a deletion, the mapped distance is greater than the insert size, while if the event is an insertion, the distance is smaller. An inversion is supported by the inverted orientation of the reads on the reference genome.

- Linking: considers two separate regions of the reference genome that are adjacent in the donor (a deletion is a type of linking signature).

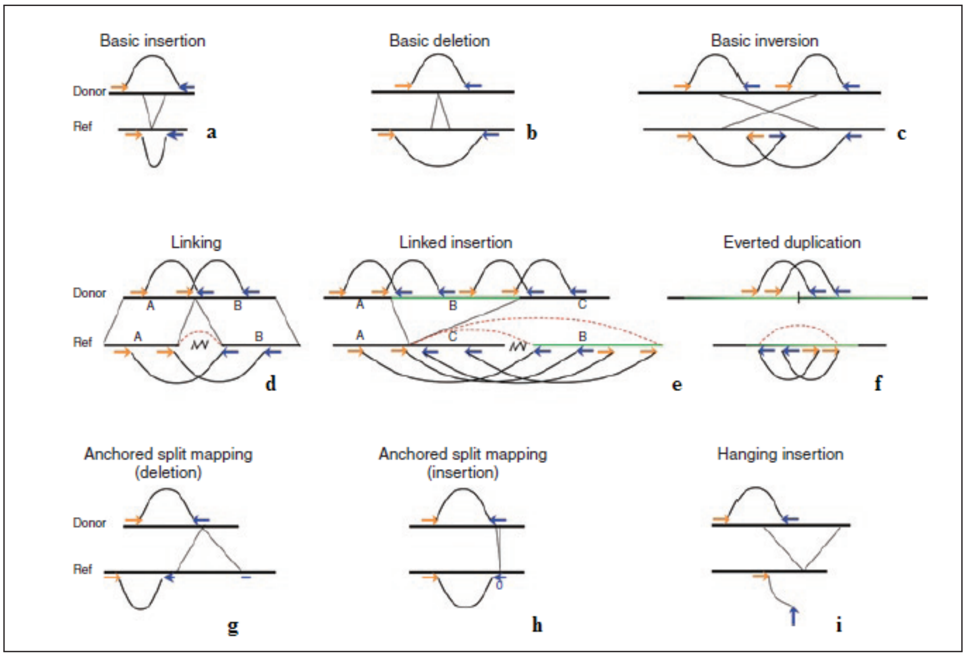


Figure 1.7: *Illustrations of PEM signatures. Basic signatures include (a) insertions and (b) deletions, where the mapped distance is different from the insert size, as well as (c) inversions, where the order of the two mates is preserved but one of them changes orientation (d). A linked insertion signature (e) is composed of two linking signatures and arises when the inserted sequence (green) is copied from another location in the genome. A tandem duplication (f) will create an everted duplication linking signature, with mates linking the end of the duplicated region to its beginning. In the anchored split mapping signature (g,h), one mate has a good mapping, whereas the other has a split mapping. For a deletion (g) the prefix and suffix surround the deletion, whereas for an insertion (h) the split read has the prefix and suffix mapped to adjacent locations, while a middle part does not map. When a novel genomic segment is inserted (i), a hanging insertion signature is created, in which only one of the mates has a good mapping.*

They are defining for distant parts of genome such as different chromosomes or genes.

- Breackpoint identification: uses of the reads mapping on the reference to discover the breakpoint. The read across this point leaves a prefix or suffix signature to different locations.
- Signatures based on depth of coverage: assumes that a uniform sequencing produces an equal amount of reads covering the whole genome, the number of reads mapping to a region should follow a Poisson distribution, proportional to the number of times the region appear in the donor. Thus, if a region has been deleted will have less reads mapping to it. This kind of signature is useful for large events and it's not able to identify small variants or localize breakpoints.

The first three kinds of signatures are called "*PEM signatures*" (paired end mapping) and are dependent on insert size that follows a distribution: depending on the tightness of this distribution is difficult to distinguish between a true paired end signature caused by a small *indel* from a mate pair with an insert size from the tail of the distribution. Whereas, the signature coming from the depth of coverage are affected by the technology which causes certain region of the genome under or over sampled.

To date, several algorithms have been developed and are available to identify genetic variations with short reads (table) and each one can be characterized in terms of two distinguishing factors: the signatures they detect and the way they cluster (Table 1.3) and each one can be characterized in terms of two distinguishing factors: the signatures they detect and the way they are grouped together (*clustering* [32] or *sliding windows* [33]).

1.8 From the Sanger technology to next generation sequencing

Determining the sequence of a DNA fragment has been possible through the development of chain-terminating inhibitor-based technology, introduced by Sanger [39] The determination of the DNA sequence has been a revolutionary event in biology: not only we can have information about nucleotide contents (GC percentage, AT- rich region, repetitive elements), but we can also reconstruct metabolic pathways or sub-cellular interactions through function analysis of groups of genes. Moreover, differences in genomes of individuals of the same species or related species could help researchers to better understand how phenotypic characters are related to genotypic variations.

The *whole genome shotgun* (WGS) approach is the most common strategy in genome sequencing [40] This methodology consists of a random fragmentation of genomic DNA (about 1000 - 2000 bp in length), an amplifica-

Name	Availability	Kind of signature
VarScan	Downlodable	-
PEMer	Downloadable	Basic deletion Basic insertion Basic inversion Linking Linked insertion
Variation Hunter	Downloadable	Basic deletion Basic insertion Basic inversion Linked insertion
MoDIL	Downloadable	Basic deletion Basic insertion
ABI tools	Downloadable	Basic deletion Basic insertion Basic inversion

Table 1.3: *Description of current methods detecting SV with NGS. VarScan [34]; PEMer [35]; VariationHunter [36]; MoDIL [37]; ABI tools [38].*

tion of fragments, by means of cloning DNA into bacterial vectors (*shotgun library*), a purification step and the sequencing of templates using vector-based primers (Figure 1.8) The data output (*reads*) are assembled by specialized analysis tools to obtain firstly *contigs* and, at the end, a possible *consensus* sequence. The amount of produced sequences determines the genome *coverage*, which depends on the length of the genome. It indicates the number of times each nucleotide has been sequenced, but it doesn't reflect the real coverage. In fact, whereas some regions could be covered by numerous fragments, others could be not represented. This is due to some problems involving the loss of fragments: toxic elements for plasmid during bacterial growth; the presence of long repeat or secondary structure within the plasmid sequence.

For about 30 years, the Sanger method has been the only one used, owing to obvious advantages in reducing handling of toxic chemicals and radioisotopes. Advances in this application lead to the production of ABI-3730XL instrument (Applied Biosystems), a platform which can sequence up to 96 fluorescently labelled samples in a single batch (run) and perform as many as 24 runs a day. The data output ("reads") are given by electropherograms, 900-1000 bp long, collected by software and in a single experiment thousands of bases are produced. This kind of DNA sequencer, allowing high-throughput analysis of samples, were utilized in many signif-

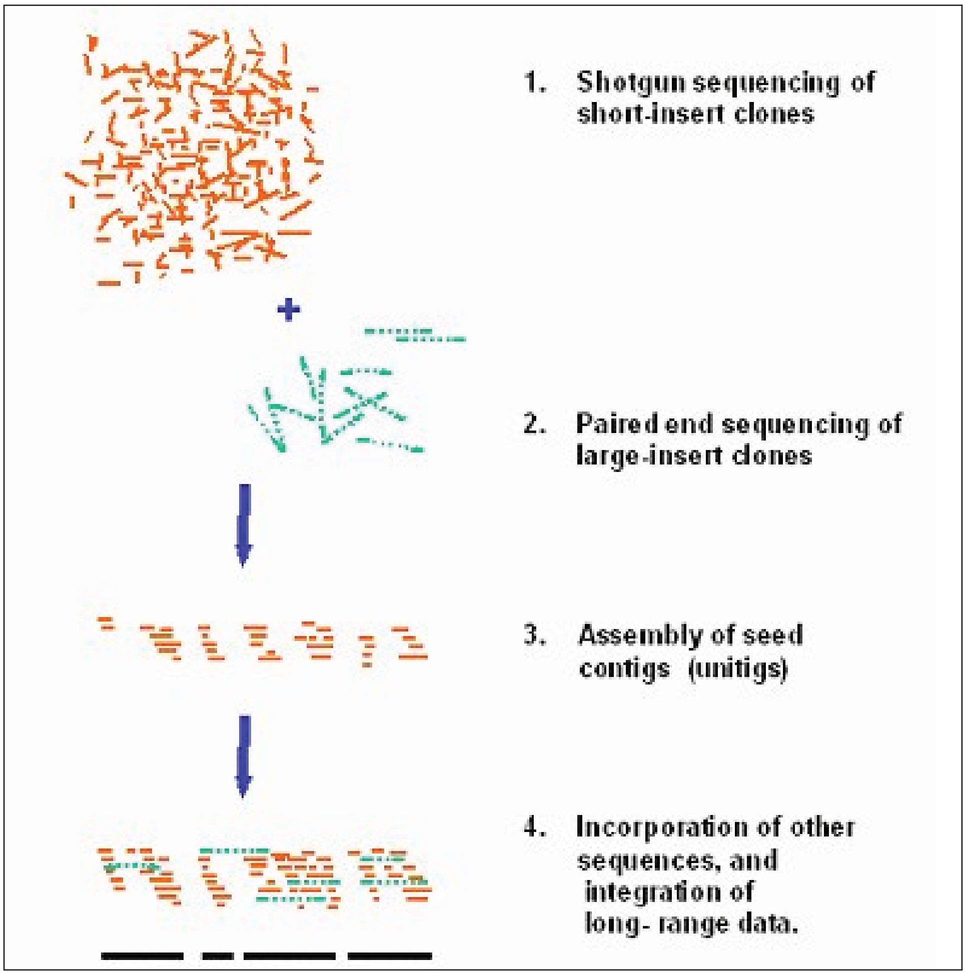

1. Shotgun sequencing of short-insert clones
2. Paired end sequencing of large-insert clones
3. Assembly of seed contigs (unitigs)
4. Incorporation of other sequences, and integration of long- range data.

Figure 1.8: *Schematic representation of Whole Genome Shotgun approach[41].*

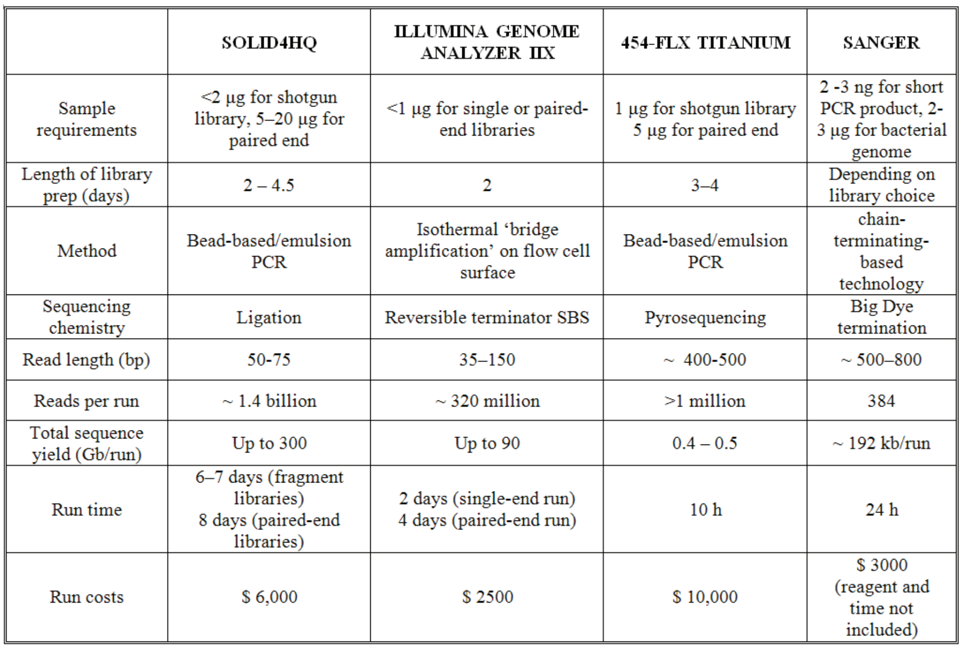
icant large-scale sequencing projects and it is considered the most accurate in terms of both read length and sequencing accuracy [42].

Starting from 2005, innovative methods to obtain DNA sequences were introduced; new sequencing strategies imply sequencing by pyrosequencing or sequencing by ligation. So far, different companies have built up new instruments allowing an automation of these procedures and the three main distributors are *Roche Applied Science, Mannheim, Germany* with the 454-FLX system [43], *Illumina Inc., San Diego, CA,USA* with the Solexa Genome Analyzer [44], and *Applied Biosystems, Foster City, CA, USA* with the SOLiD$^{\mathrm{TM}}$ [45].

These platforms are able to generate hundreds of thousands of sequencing reactions in parallel, allowing a great increase of throughput, permitting ultra-deep sequencing projects on large-size genome and are considerably less expensive than the Sanger method. All these next generation sequencing technologies are characterized by a substantial reduction in read-length ranging from 25 to 400 bases, however, this is an acceptable trade-off for many applications, particularly re-sequencing. With the availability of a reference genome, short reads result indeed very informative, considering that a read should be long enough and sufficiently accurate to align uniquely. Furthermore, there are several perspectives in uncovering nucleotide diversity at whole genome level in multiple lines.

Up to now, NGS gave good results in ChIP-sequencing to identify binding sites of DNA-associated proteins [46], RNA sequencing to profile transcriptomes [47], as well as whole plants genome sequencing [48].

Nowadays, the interest in single nucleotide polymorphism (SNP)-based association studies and structural variants is increasing [49]. On the other side, crucial tasks are the samples preparation complexity, the quality of single read and how to give sense to the large amount of data the sequencers produce. (Figure 1.9)



	SOLID4HQ	ILLUMINA GENOME ANALYZER IIX	454-FLX TITANIUM	SANGER
Sample requirements	<2 µg for shotgun library, 5–20 µg for paired end	<1 µg for single or paired-end libraries	1 µg for shotgun library 5 µg for paired end	2 -3 ng for short PCR product, 2-3 µg for bacterial genome
Length of library prep (days)	2 – 4.5	2	3–4	Depending on library choice
Method	Bead-based/emulsion PCR	Isothermal 'bridge amplification' on flow cell surface	Bead-based/emulsion PCR	chain-terminating-based technology
Sequencing chemistry	Ligation	Reversible terminator SBS	Pyrosequencing	Big Dye termination
Read length (bp)	50-75	35–150	~ 400-500	~ 500–800
Reads per run	~ 1.4 billion	~ 320 million	>1 million	384
Total sequence yield (Gb/run)	Up to 300	Up to 90	0.4 – 0.5	~ 192 kb/run
Run time	6–7 days (fragment libraries) 8 days (paired-end libraries)	2 days (single-end run) 4 days (paired-end run)	10 h	24 h
Run costs	$ 6,000	$ 2500	$ 10,000	$ 3000 (reagent and time not included)

Figure 1.9: *Differences in next-generation sequencing platforms. All information come from the respective company web sites.*

1.8.1 The SOLiD$^{\text{TM}}$ system

In the 2007, Applied Biosystems has produced its SOLiD$^{\text{TM}}$ sequencer, a highly accurate, massively parallel next-generation sequencing platform. The first application was of 20 - 35 bp short-reads, with read lengths increased to 50 bp in the SOLiD3 release. It is now upcoming the SOLiD4. The platform supports a wide range of applications for characterizing genomes and transcriptomes, including fragment and paired-end DNA and cDNA sequencing, expression level studies, methylation assays, small RNA sequencing, "barcoding" to permit subsample identification, and splice variant analyses. Furthermore, the flexibility of two independent flow cells allows carrying out different experiments in a single run.

The SOLiD relies on emulsion-PCR (em-PCR) to amplify fragmented DNA onto beads clonally. After em-PCR, amplified beads are recovered and the amplicon strands are modified at their 3' ends to allow covalent attachment to a glass slide [50]. The major steps of the sequencing process are:

Library Preparation: two different protocols describe how to make a fragment or a mate-pair library (Figure 1.10, [1]). Differences in the library reflect the kind of information produced. While a fragment library could be similar to a WGS approach, the mate-pair can produce two reads at an approximately known distance in the genome to generate reads from both sides of a segment of DNA (the insert). The genomic DNA is fragmented and size-selected inserts are circularized and linked by means of an internal adaptor. The circularized fragments are enzymatically modified and the adaptor with its flanking segments (the genomic mate-pairs) are purified. Two different universal adapters (P1/P2) are then ligated to the construct ends and used to clonally amplify the whole fragments. Finally, the mate pairs are generated by sequencing around the adaptor.

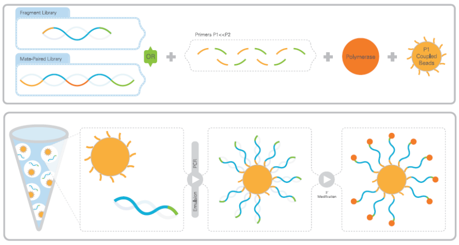


Figure 1.10: *Simplified version of emulsion PCR.*

Emulsion PCR/Bead Enrichment: the PCR is performed in an emulsion where microreactors contain template, PCR reaction components, primers and beads (Figure 1.11). After PCR, the templates are denatured and a bead enrichment step is carried out to separate beads with extended templates from undesired beads: beads with P1 and P2 adaptors linked to the DNA fragment are picked up, while other occurred events (P1-P1 or P2-P2 linking) are removed. The template on the selected beads undergoes a 3' modification to allow covalent attachment to the slide.

[1] `http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_081748.pdf`

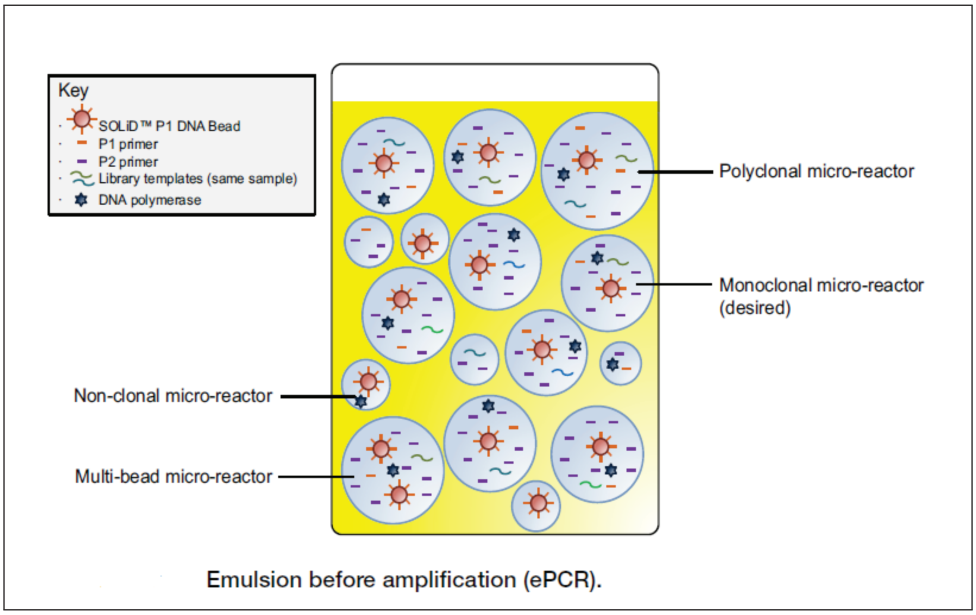


Figure 1.11: *Scheme of e-PCR. The PCR is perfomed in little drops of water, containig: SOLiD beads, template, DNA polymerase and a couple of primers.*

Bead Deposition: the beads are deposited onto a glass slide thanks to the 3' modification. Deposition chambers enable to segment a slide into one, four, or eight sections giving the opportunity to run several analyses in a single experiment. The SOLiD beads are considerably smaller ($1 \mu m$) than 454, i.e., allowing a higher density of beads to be collected into the same area. The current density upper limit is 700 million beads per sequencing run, even though a relevant portion of these beads are not analyzed because they have more than one template amplified onto them, giving a “mixed read”, or because they are on the glass borders.

Sequencing by Ligation: the platform uses sequencing-by-ligation, rather then pyrosequencing as 454-Roche. Primers hybridize to the P1 adaptor sequence on the beads (Figure 1.12) and 1024 random 8-mer probes are added (4 dyes, 4 dinucleotides, 256 probes per dye). These probes are labelled on the nucleotides at the first and second positions at the 3' end, using four fluorescent dyes and are complementary to the template strands. Once the oligo is linked to the sequencing primers, the slide is imaged. Then the probe is removed leaving only five bases associated to the sequencing primer, and a new random probe set is added. Multiple cycles of ligation, detection and cleavage are performed with the number of cycles determining the eventual read length. Following a series of ligation cycles, the extension product is removed and the template is reset with a primer complementary to the n-1 position for a second round of ligation cycles.

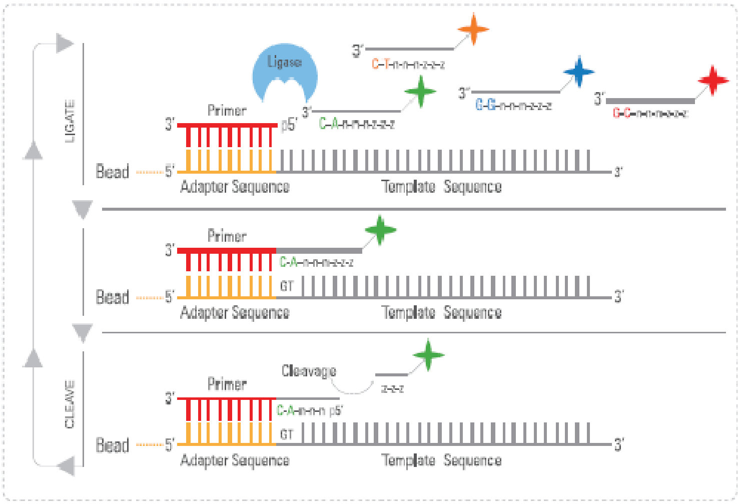
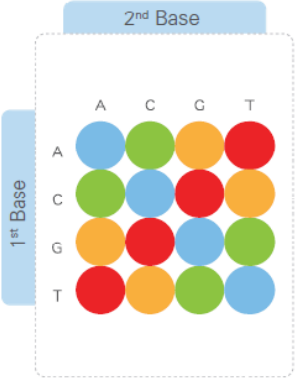


Figure 1.12: *Sequencing by ligation of 1024 random 8-mer labeled probes.*



	A	C	G	T
A	🔵	🟢	🟠	🔴
C	🟢	🔵	🔴	🟠
G	🟠	🔴	🔵	🟢
T	🔴	🟠	🟢	🔵

Figure 1.13: *SOLiD colors space table. Every di-nucleotide is represented by one color.*

Primer Reset: a total of five sequencing primers is used to obtain a series of colors for each bead (Figure 1.13).

Through the primer reset process, every base is virtually interrogated in two independent ligation reactions by two different primers. When a perfect match is formed between the template sequence and one of the probes, it is covalently ligated to the starting primer (preferentially with its 3'-end). The fluorescent dye that labels the 5'-end of the ligated probe is then excited and detected with a CCD camera. Once the imaging is completed, the last three nucleotides at the 5'-end are chemically removed and the released dyes are washed away. This cleavage step generates a new 5'-phosphorylated-end available for the next ligation cycle. Multiple cycles of ligation, detection and cleavage are performed, with the number of cycles determining the eventual read length. After all the cycles are completed, the newly synthesized DNA product is removed. In this way, the template is reset and becomes available for subsequent rounds of ligation cycles, which are primed from an oligonucleotide one position backward (n-1) of the previous starting nucleotide(n)(see Figure 1.12). This kind of interrogation is essential to the unmatched accuracy characterized by the SOLiD^{TM} System. Moreover, because only four colors are used and because each color represents four dinucleotides, it is not possible to deci-

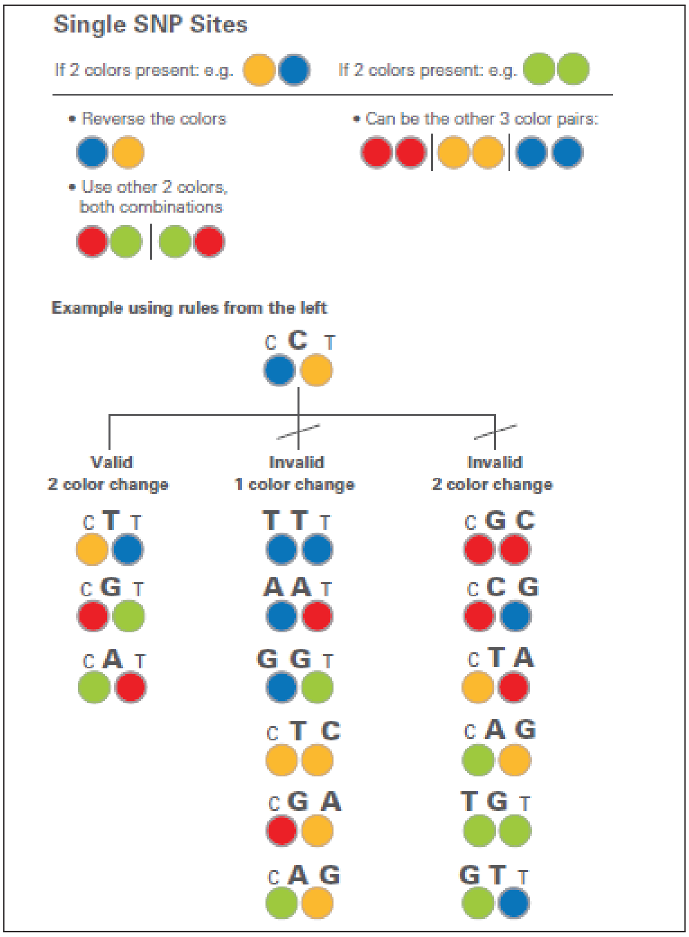
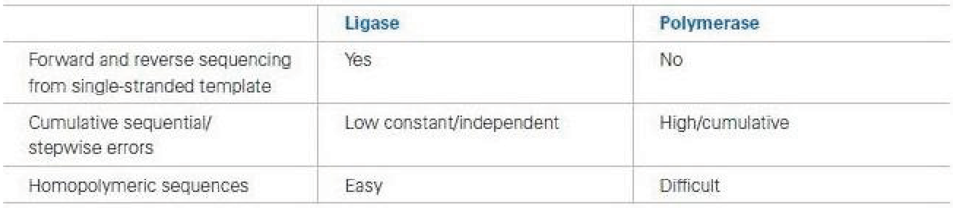


Figure 1.14: *SNPs vs Errors detection with SOLiD platform.*

pher the identity of the nucleotides without knowing the first base in the sequence. This is achieved by sequencing one base of the adapter. The conversion into nucleotide base-space is usually done after the sequence is aligned to the reference genome which is translated in the colorspace coding. This strategy provides higher sequencing accuracy and inherent error checking capability. The advantages of colorspace reside in the ligation-based chemistry which reduces errors rate in comparison with polymerase sequencing by synthesis (Table 1.15), thanks to the probe recognition of template instead of sequentially addition of dyes. Dual interrogation reduces sequencing errors, because a SNP, a true polymorphism will require a change in two adjacent position (see Figure 1.14). So that changes in a single position are considered as errors and can be removed by software data analysis. In addition, also surrounding bases of the two changed color undergo variation and may be filtered. For more complex variation (indels), analyzing tools need to be more accurate. All these advantages are well summarized at this link `http://www3.appliedbiosystems.com/cms/groups/mcb_` `marketing/documents/generaldocuments/cms_057511.pdf`. Moreover, a specific tool has been developed to transforms the reference to color spaces then maps the reads ([2]). Transforming colourspace to dnaspace may cause errors, because a single mismatch in colourspace affects all the following base calls. A mComparison of Sanger ismatch is defined as a colorspace difference between a mapped read and the reference.



	Ligase	Polymerase
Forward and reverse sequencing from single-stranded template	Yes	No
Cumulative sequential/ stepwise errors	Low constant/independent	High/cumulative
Homopolymeric sequences	Easy	Difficult

Figure 1.15: *Comparison of Sanger sequencing method and new generation sequencing by ligation.*

The SOLiD is currently capable of producing approximately 100 Gb of short-reads sequence data per run (25-50 bases) and so is more suited to re-sequencing than de novo assembly, although optimized protocols for long-insert read pairs up to 6 kb are available.

1.9 Next generation sequences analysis

Next-generation DNA sequencing platforms make available gigabases of data, however the size of the result presents computational, analytical and storage challenges. So, new bioinformatics approaches have been developed to align short sequences to a reference genome. Traditional alignment programs (i.e. BLAST or FASTA) are unable to perform the same job with millions of reads. To overcome this issue, software such as SOAP [51], ELAND, SHRiMP [52], ZOOM [53] have been implemented. In agreement to these algorithm, an extremely sensitive, efficient and fast algorithm for aligning millions of NGS reads allowing gaps and mismatches has been developed at the CRIBI laboratory and named PASS[3] [54]. PASS supports several data format (Solexa, SOLiD and 454 technology) carrying out fast gapped and un-gapped alignment onto a reference sequence and performing gap alignments more then 800 time faster than BLAST and several time faster than SOAP. PASS is able to align all NGS sequences in base-space and color-space and supplies modules for paired-end alignments, SNP and IN/DEL detection and spliced alignments. It is useful for single read mapping, paired-end re-sequencing, small RNA discovery and RNA-seq mapping. In particular, Pass executes paired-end alignments thanks to PASS_PAIR tool. It implies the recognition of adaptors and splitting reads in two sequences corresponding to the paired-end; the alignment of paired-end onto reference sequence; the check of the mutual paired-end distance and orientation. For the purposes of aligning couples of short sequences, the Pass_pair tool is suitable both for pair-end and mate-pair libraries. Furthermore, a specific option of

[2] `http://solidsoftwaretools.com/gf/project/mapreads/`

[3] `http://pass.cribi.unipd.it/cgi-bin/pass.pl`

the software (PASS_SNP tool) allows to minimize the effects of sequencing error to find true SNPs.

Up to this year, no software were available to investigate structural variations with SOLiD reads directly: a data elaboration was necessary to give right input files to different programs. On May 2010, a new application for investigating chromosomal rearrangements from paired-end and mate-pair sequencing data provided by the high-throughput sequencing technologies has been developed. SVDetect identifies structural variations applying both sliding-window and clustering strategies [55].

CHAPTER 2

Aim of the research

The agronomic relevance of grapevine and its main role in the economic world bring the knowledge of *Vitis vinifera* genome to be one of the most important goals to be achieved in plant genomics. The sequencing of a specific genome is the quickest method to obtain information from a particular DNA sample and associate the data to others from different experiments (arrays data, RFLPs, RAPDs, microsatellites). In addition, the sequencing strategy allows the identification of a great number of genetic makers related to particular phenotypic characters or physiological states of plants. The Marker Assisted Selection is used to choose a genetic determinant (or determinants) of a trait of interest (i.e. resistance to disease and abiotic stresses, increasing productivity and products quality) and the availability of new sequencing technologies is leading to an increase of produced markers within a single test. This advantage is suitable to improve vineyard yield, ameliorating berry qualities together and architecture of whole plants. Moreover, many grape varietas undergo natural mutations (base mutations, indels, transposable element insertions or excisions, and epigenetic modifications) that can lead to the appearance of desirable phenotypic variations.

My phD project can be divided into three main steps.

The first concerns the sequencing of the quasi homozygous genotype of Pinot noir (PN40024). My group is part of the IGGP Consortium which have promoted the investigation of *Vitis vinifera* genome in order to have a complete analysis of grapevine genomic sequence. I was involved in this project for the production of 2 genome equivalents of sequence reads. The total 12 X coverage of the genome obtained by the Consortium give rise to a huge amount of data which needs to be reordered. Numerous studies are undergoing to better explain and understand genes and their function, regulator elements and mutations (i.e. SNPs and SVs). Given the draft sequence of *Vitis vinifera* genome ([1]), it is possible to compare sequences obtained from other related species or subspecies. Moreover, the differences within the same subspecie lead to the possibility of characterizing a specific cultivars from another one.

The second part of this research takes advantage from the SOLiD (Applied Biosystem) Next Generation Sequencing platform which allows a massive parallel sequencing and is suitable for re-sequencing strategies in order to identify variations (polymorphisms) that could explain differences in phenotype. To perform the analysis of polymorphic sites within *Vitis* specie, I have chosen two cultivars of local interest (Merlot and Prosecco) and I have built a mate-pair library to perform a run on the SOLiD sequencer. The presence of a reference genome provide a backbone against which even short reads (25 nt) can be mapped uniquely.

The last part lies in the bioinformatics analysis of the produced data to unveil significant differences among the studied genomes.

Therefore, the aim of this research consists in evaluating these differences which are supposed to characterize a particular cultivar. The intent is to modelling the marker finding through next generation sequencers in order to obtain a useful and rapid method to discovery variations (markers) and have the possibility to associate a variant with a specific trait.

CHAPTER 3

Methods

3.1 Plant material

The three cultivars of *Vitis vinifera* employed during this experimental research are Pinot noir, Merlot and Prosecco (Table 3.1).

Cultivar name	Supplied as	Source of sample	Notes
Pinot noir PN40024	plasmid libraries	created by INRA in Colmar	-
Merlot clone	Leaves	open field	"spurred" cordon
Prosecco clone 10	Leaves	micropropagated	modified Murashinge-Skoog medium

Table 3.1: *DNA sample sources.*

Merlot leaves came from *Azienda vitinvinicola Borin Vini e Vigne* cultivated in Monticelli, Monselice, Padua and supplied by prof. Bonghi (University of Padua). Prosecco leaves came from C.R.A. (*Centro di Ricerca per la Viticoltura di Susegana*) and are supplied by prof. Lo Schiavo (University of Padua). Culture conditions are explained in Appendix A, page 59.

3.2 *Vitis vinifera* genome project

3.2.1 Library amplification

Material

TB medium:
Glycerol
Yeast extract
Bacto Triptone
K_2HPO_4 * $3H_2O$ 0,1 M
KH_2PO_4 1M
H_2O mQ
Ampicillin 50 mg/ ml

Resuspension solutions:
H_2O mQ Autoclaved
EDTA 1M pH 8
Tris HCl 1M pH 8
Glucose 1M
RNAse A 10 mg/ml

Lysis solutions:
H_2O mQ autoclaved
SDS 20%
Sodium hydroxide (NaOH) 10 M

Other reagents:
Isopropyl alcohol HPLC
Isopropyl alcohol standard
EtOH HPLC
Potassium Acetate (CH_3CO_2K) 3 M pH 5.5

384 well polypropylene plates not sterile (bacterial growth)
Millipore clearing plates (Montage Plasmid Miniprep Clearing Plates)

Instruments:
Beckman NX
Centrifuge Eppendorf 5810R
Heidolph Titramax
Micro Lab Star Hamilton

Method

The “Whole-Genome-Shotgun” (WGS) approach has been applied, to gain the 2 X coverage, obtaining sequences from about 1600 plasmid library plates (384 well) of Pinot noir (PN40024, INRA, Colmar, France).
As the project proceeded, three different protocols for DNA template preparation for the sequencing reaction have been used in our laboratory. By time: PCR, TempliPhi^{TM} HT DNA Amplification Kit (Amersham Biosciences) and Montage

Plasmid Miniprep Kit (Millipore). Since Miniprep produces very reliable results in terms of reproducibility and template quality, it has been applied to the (60%) of the plates. Template DNA was extracted from liquid bacteria culture using a procedure based upon alkaline lysis minipreps method adapted for high throughput processing in 384-well plate. Reagents were home-made and the dispensing operations have been accomplished using a robotic work-station Hamilton LabSTAR robot (Hamilton, Birmingham, UK).

3.2.2 Sequencing reaction and run on ABI 3730*xl*

Material

EtOH abs
Acetic Acid (NaAc) 3M pH 8.0
EtOH 70%
H_2O mQ autoclaved

Method

The sequencing reaction was performed with BigDye ®Terminator v3.1 Cycle Sequencing Kit chemistry on 384-Well GeneAmp^R PCR System 9700 (Applied Biosystem) according to manufacturing instructions. The DNA was subsequently purified through a EtOH-NaAc / EtOH 70% precipitation and re-suspended in 20 μl of sterile mQ H_2O.

The electrophoresis was performed on the DNA sequencer ABI3730*xl* (Applied Biosystem) according to manufacture directives. All the process had been developed thanks to an high-throughput system associated with the instruments which can sequence up to 96 fluorescently labelled samples in a single batch (run) and perform as many as 24 runs a day. The platform which had allowed the ht organization of the project was composed by Hamilton MicroLabSTAR robot (Hamilton, Birmingham, UK); Hamilton MicroLabSTAR Let (Hamilton, Birmingham, UK); Jouan GR 4 Auto centrifuge (Jouan Robotics, Saint-Herblain Cedex, France); Multimeck 96, Multimeck NX and Biomeck 2000 (Beckman Coulter, Brea, CA, USA); three ABI3730*xl* (Applied Biosystem, Carlsbad, CA, USA).

The amount of genome coverage obtained by the IGGP was 12 X, that means that the consortium produced 480 Mb (genome size) 12 times: ∼ 5.7 Gb of sequence. My lab has supplied two-fold genome equivalents: about 1 Gb.

3.2.3 Assembly and annotation

"The ensemble of the sequences obtained will be assembled using the ARACHNE assembler (Broad, Institute) and an automatic annotation will be performed."[1]

Gene prediction and genome annotation was executed by the bioinformatics group present in my lab.

[1] `http://www.cns.fr/spip/Vitis-vinifera-whole-genome.html` (Site du Genuscope)

3.3 Next generation sequencing of Merlot and Prosecco

Material

Chloroform ($CHCl_3$)
Isopropyl alcohol standard
EtOH abs TRIzol (Invitrogen)

3.3.1 Genomic DNA extraction

Genomic DNA was extracted from leaves of Merlot kept in freezer at -80 ℃ and fresh leaves of Prosecco. 3 g of plant tissue was ground in liquid nitrogen -80 ℃ inside a sterile mortar. Genomic DNA was isolated with Nucleon^{TM} PhytoPure^{TM} Genomic DNA Extraction Kit. This kit allows the nuclei enrichment during the extraction, avoiding the presence of organelles DNA. The main steps of the extraction protocol are:

1. Breaking of cell wall;
2. Cell lysis with potassium SDS;
3. Extraction with Nucleon^{TM}Phytopure resin and chloroform;
4. DNA precipitation;
5. DNA washing.

The DNA amount and quality were determined via spectrophotometer NanoDrop ND-1000 (NanoDrop Technologies) (2 μl) comparing the concentration values obtained at 230, 260 and 280 nm.

A measure was obtained also via fluorometer Qubit^{TM} Quantitation Platform (Invitrogen). The DNA concentration was expressed in μg of DNA for μl of H_2O.

3.3.2 RNase treatment

60 μg of DNA were treated with 10 U of RNase A (Sigma-Aldrich) placing the samples in a heat block at 37℃for 30 minutes. At the end of the RNase reaction, the samples were subjected to phenol-chloroform-isoamyl alcohol solution(25:24:1) and after that the DNA was precipitated as suggested by Sambrook et al. (1989) [56]. The pellet obtained were re-suspended in 50 μl of sterile mQ H_2O and the DNA amount was estimated as described previously.

3.3.3 SOLiD library: mate pair protocol

To set up the pair end library with an average insert size of 2000-3000 bp on the SOLiD^{TM} sequencer, 30 μg of DNA (Merlot and Prosecco) was sheared at speed SC9 for 20 cycles, using a HydroShear®Standard Shearing Assembly DNA Shearing Device (DIGILAB - Genomic Solutions). DNA was divided into three sub-samples of about 10 μg, performing a better shearing. The library was constructed from the eluted fraction of DNA using SOLiD^{TM} 2 × 25 bp Mate-Paired

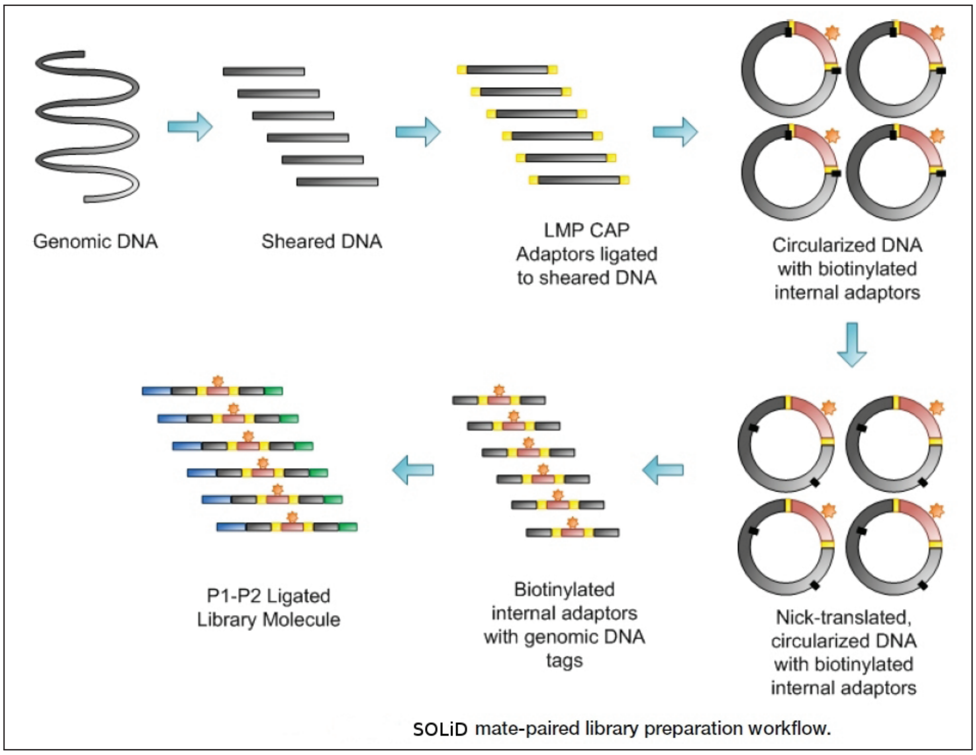


Figure 3.1: *Mate-pair workflow.*

Library Construction Kit, following their manufacture, with one minor modification: ethidium bromide was used for staining DNA in agarose instead of SYBR© Safe gel stain (Invitrogen). The library amplification was executed with 12 cycles for Merlot and 16 cycles for Prosecco samples (Figure 3.1).

On the Applied Biosystems SOLiD^{TM}3 a single standard run was performed, using a deposition chamber for Merlot library (1 well; ∼ 347 million beads deposited) and the other for Prosecco one (1 well; ∼ 274 million beads deposited (Figure 3.2)).

3.4 Computational analysis

3.4.1 Mate-pair alignment

The total amount of sequences produced by the SOLiD run were aligned against the reference genome (Pinot noir) with PASS through "Common Paired-End settings". **PASS** is based on the creation of a genome index, that is a structure containing the genome positions of all seed words (12 bases as default). After the genome index production, PASS tries to align each input read in three steps:

1. identification of the query seed words in the genome index;
2. check for the possibility to extend the alignment in the seed flanking regions;
3. refinement of the alignment with a modified Smith-Waterman algorithm.

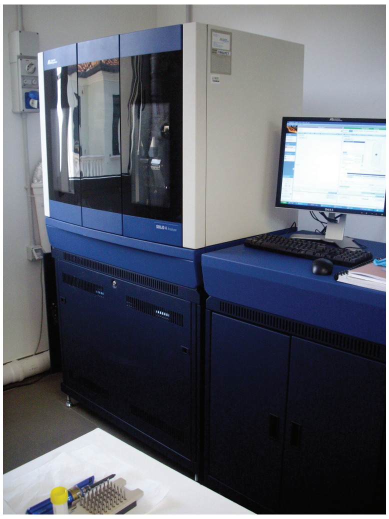


Figure 3.2: *SOLiD sequencing platform.*

In particular, the alignment extension uses a simple but effective approach that allows an immediate analysis of the flanking regions adjacent to seed words. It uses Pre-computed Score Table (PST) which analyzes all the possible short word alignments against each other. The length of these short words ranges from 6 to 8 bases, reflecting in different PSTs. Each alignment presents a score which is computed with Needleman and Wunsch algorithm [57], using different values for matches, mismatches and gaps. If these scores are higher than a pre-defined threshold, PASS performs an exact dynamic alignment of a narrow region around the match. In addition, it applies low-complexity regions filters.

A mate-pair represents the two extremities of the same insert, therefor they should be at a defined distance in the genome, depending on the insert length. For SOLiD technology, these two sequences are tagged R and F. With PASS_PAIR, the couples of R and F reads are aligned onto the reference genome taking into account the estimated insert length obtained during the library preparation.

In the output files, data are classified according these criteria:

- *where the reads align*: the pairs can be aligned in the same genomic location or can be aligned in two different positions;
- *how many times the reads align*: they can be aligned uniquely, (*unique*) with the best match score, or more that one time (*not_unique*);
- *distance value between coupled reads*: according to the distance between the reads, each couples can fall within a specific range explained below.

A brief summary about the number of couples that falls in each category is given in the output file. There are the estimated value for the library size (*L. Size*) and its standard deviation (*S. Error*). A graph shows the Gaussian curve, that represents the distribution of the mate-pair distances for the considered library. In the x axis there is the distance value, while in the y axis there is the obtained frequency for that distance (how many coupled reads have that distance).

To obtain a robust set of good quality reads to be used on PASS, bioinformatics group in my lab has implemented a series of preliminary filters. Short reads are trimmed off with 2 quality filters (*internal* and *at the border* of the read). The internal filter selects a region where the quality of a window (W) is always above an average quality (T). The 2nd filter is applied to the ends: it scans each end with a window (w) until it finds that all the bases are = to the threshold quality (t).If both filters are applied then the internal filter will be applied first, then the external filter will check and trim the resulting ends.
In a second moment the spectrum correction have been performed. The SOLiD^{TM} Accuracy Enhancer Tool (SAET) has been used to correct miscalls within reads prior to mapping or contig assembly. This tool acts in two steps:

- Spectrum Building: with all reads an ensemble of k-mers, is generated (*spectrum*). Each k-mer is obtained from the comparison of the reads that fall in the same position in the reference genome. Once the software identify the best alignment of these reads, it builds up a k-mer.
- Error Correction: each read in the input file which does not correspond to any k-mers is corrected to the most probable k-mer present in the spectrum.

Once the reads passed these two filters, they are aligned to the reference.

3.4.2 SNPs and SVs analysis

The identification of SNPs and indels was achieved through PASS_SNP, a specific tool of the PASS software which allows to resolve with the (%) of accuracy, the presence of a single nucleotide polymorphisms in the DNA. Suggested parameters for an optimal polymorphisms discovery were set. In a second moment, further filters were applied. The output file presents: a plot showing the number of sequences that confirm a SNPs; a plot represents the frequency of SNPs found related to read position and their relative quality. Generally, in terminal regions of short reads errors tend to increase. In this study, the two filters explained above, trimming the sequences, correct errors presence in the end of the reads.

The re-sequencing of a genome using mate - pair library can unveil some of the SVs that affects it. The reasons lie in the nature of the library: couple of reads align at a precise distance and orientation on the reference genome. Changes can underline a possible SV. In a second moment, the kind of variation can be determined analyzing how distance and orientation differ from the expected ones. The platform that finds out the SVs is a collection of C++ scripts organized by a supervisor perl script that manipulates input data and launches all the C++ scripts with the right options, in the right order. The results are written in four files, containing lists of different SVs ordered by chromosome and start position. Input files come from gff files created by PASS_PAIR:

- *UNIQUE_PAIR*: paired-ends with right orientation and distance;
- *UNIQUE_WRONG_D*: paired-ends with right orientation, but wrong distance (greater than expected one);
- *UNIQUE_SINGLE*: paired-ends for which only one read is aligned;
- *UNIQUE_WRONG_S*: paired-ends with right distance but wrong orientation.

As it can be understand from the suffix *unique*, all the paired-ends are sequences aligned with the best score in only one site of the reference genome.

Structural Variations Algorithm

Deletions and insertions (*in-dels*) in the donor genome are represented by a variation of the supposed distance in which the paired-ends should fall in. In particular, deletions show an higher value in length, while insertion a lower value than the expected one. The algorithm uses variations to find indels. In addition, it is necessary to take into account that mate-pairs map to the reference genome with a range of length distributions rather than a exact distance. To solve this issue, the program analyzes two distributions (expected and observed), to find variants. Indels can be found by comparison of expected and observed distributions: as a consequence of insertions, the distribution of lengths will move towards lower values, whereas deletions shift the curve towards higher values (see figure 3.3).

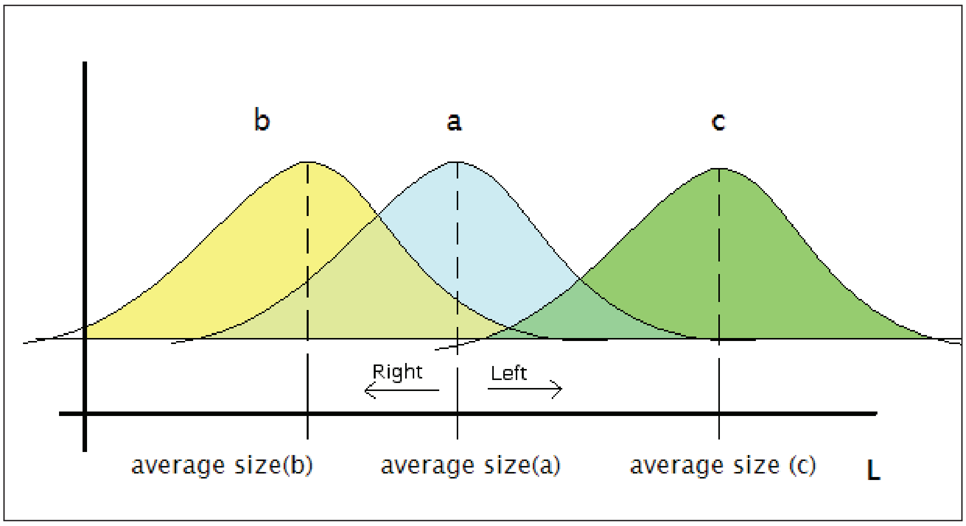


Figure 3.3: *Changes in library insert size distribution. a) Mean value of the "normal" distribution. b) Mean value of a distribution presenting insertions in the donor genome respect to the reference. c) Mean value of a distribution presenting deletions in the donor genome. The y axes represents the number of inserts.*

Regarding long structural variations discovery, the length of the SV results greater than the considered range of distribution (Average size +/- standard deviation) and therefore delineate a high stretched curve. Moreover, PASS_pair tool produces a specific file (Unique_pair_wrong_distance), which shows all the pairs that falls in a wrong range of distribution. Within this file, it is possible to identify large rearrangements that are discarded from the threshold established to build up the distribution curve.

Thanks to these evidences, each position in the reference genome is represented by a score which indicates the probability of assigning a variation at that position. In a second moment, a threshold is fixed, so that the location of the rearrangement can be extended in the adjacent bases with scores higher than this threshold. This allows the individuation of *zones* of variation. In this study the threshold is 5000 bp. There are three kinds of zone:

1. zones containing deletions

2. zones containing insertions
3. zones containing inversions

The localization of such areas does not imply the presence of a real SV. The three zones are potential structural variations, depending on the quality and quantity (coverage) of sequence obtained. Low standard deviation for distance distribution and an high coverage mean that nearly all the found zones are SVs, while an high number of false positives is a consequence of low quality reads or low coverage.

CHAPTER 4

Results and Discussion

4.1 Whole genome shotgun of Pinot noir

With the recent availability of genome sequences for many plants organisms, identification of sequences variation and understanding biological consequences has become a major aim of research. The I.G.G.P. consortium have promoted collaborations among different European groups in order to describe biological and genetic aspects of grape (*V. vinifera*). This plant is one of the major species for world agriculture and increasing in its quality production and adaptation to environment conditions is appreciable. The knowledge of the genomic sequence is bringing researchers new sources of information to develop genetic tools for its amelioration. In particular, it allows the description of specific alleles showing differences implicated in several mechanisms and it makes possible the discovery of genetic markers which can be used in a more precise cultivar characterization. In addition, the possible transfer of results from this organism to other plants fruit is desirable, using grape as model. The sequencing of the genome of *Vitis vinifera* (L.) has been performed on the quasi-homozygous genotype PN40024 created by INRA in Colmar [1]. The estimated size of the genome was 480 Mb [58] and the Whole Genome Shotgun approach was chosen with a twelve-fold average redundancy (coverage) of the consensus to be sure of obtaining a sufficient amount of data, describing each base. As a rule, with an average length of 700 high quality bases, both ends of 384-well plate are required every 100 kb of DNA to be sequenced.

The applied whole genome shotgun strategy implies the cloning of template into DNA vectors, the amplification of obtained fragments and the sequencing on the ABI3730xl platform. The sequencing takes place in a cycle reaction, where template denaturation, primer annealing and primer extension occur many times. The primer is complementary to sequences immediately flanking the region of interest. Each round of primer extension is terminated by the incorporation of fluorescently labeled dideoxynucleotides (ddNTPs). In the pool of end-labeled fragments, the label on the terminating ddNTP corresponds to the nucleotide identity. The total sequence is determined by high-resolution electrophoretic separation of the single-stranded in a capillary-based polymer gel. A laser excites the fluorescent labels at the end of the fragments and their detection produce a spectra which provides the Sanger trace. These traces are translated by a software into the DNA sequence. During the research, different protocols for DNA amplification have been tested, in order to have a good quality template for the sequencing reaction. The aim of the amplification was on one side to obtain long and high-quality reads, on the other side to have a method able to resolve homo/di-polymeric regions which cause gaps in the successive assembly step. Even if *Vitis* genome seems to have few low complexity regions as other plant genome (i.e. tomato), some problems in performing sequencing reaction arose due to the selected amplification methods. To amplify plasmid libraries supplied by INRA center,three different techniques were employed:

- **PCR (Polymerase Chain Reaction)**: the DNA template is obtained by performing a 384-well PCR using universal primers (M13 forward and reverse). Afterwards, an analysis on gel electrophoresis is required to identify positive PCRs which are chosen and pooled by a robotic work station (Microlab STAR, Hamilton Robotics). At the end, a purification step is necessary to perform the sequencing reaction. This method has several disadvantages:
  - it is time-consuming and requires laboratory personnel hands-on time for the selection of the right PCR products;
  - it produces sequences of low quality in presence of homo/di-polymers or GC-rich regions;
  - the low processivity of the Taq DNA polymerase through long stretch of TA (up to 100 bases) or regions making stable secondary structure causes the presence of gaps or low quality bases regions in the assembly;
  - even if the reads length is about 800 bp, the number of PCR failures is very high and it implies a mandatory pooling of the successfully amplified templates.
- **TempliPhi (Amersham Biosciences)**: this kit utilizes bacteriophage Phi 29 DNA polymerase enzyme and random hexamer primers to exponentially amplify DNA [59]. Phi 29 DNA polymerase has a proofreading activity with an error frequency of 1 X $10^{-6} - 10^{-7}$. This method has several advantageous features: it amplifies DNA directly from bacterial cultures, it avoids purification steps and the reaction can be performed on a heat block for 4 hours. However, low quality reads are still obtained for GC-rich regions and the average length is around 500 bp (Figure 4.1).
- **Plasmid minipreps**: the DNA is extracted from liquid bacteria culture, purified and concentrated using Montage Plasmid Miniprep Kit (Millipore) and the robotic work-station. Even if miniprep is a multi step procedure that requires overnight growth of bacterial replica plate and labor-intensive efforts, we have adopted this protocol to prepare DNA template for sequencing. In fact, it has significantly increased the length of sequencing reads (up to 800-900 bp) and base quality even within homo/di-polymer (up to 20 bases) and GC-rich regions (Figure 4.2).

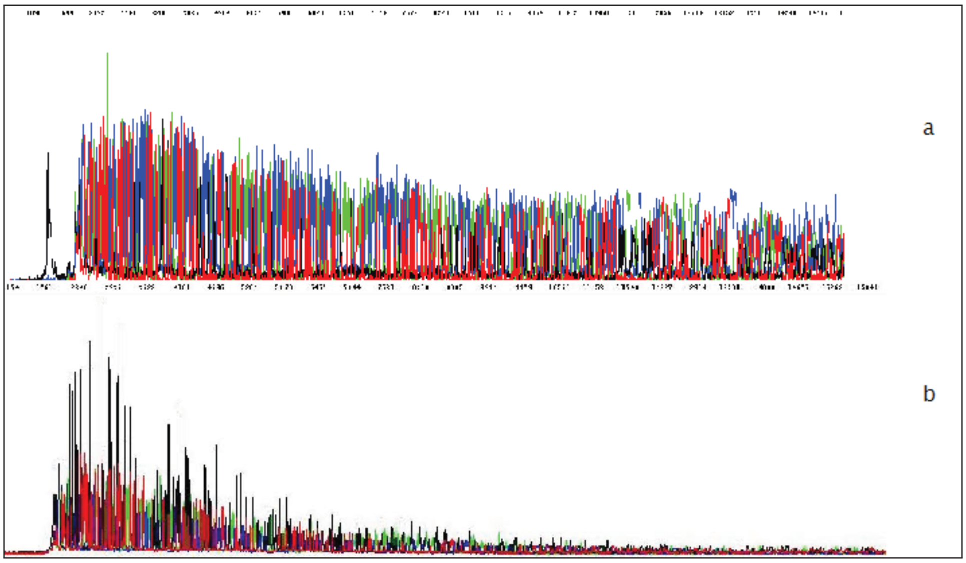


Figure 4.1: *Difference in sequencing performance of Templiphi amplification and Miniprep amplification. a) Raw graph of a standard sequencing reaction of Miniprep DNA template preparation (average length = 800 bp). b) Raw graph of a standard sequencing reaction of Templiphi DNA template preparation (average length = 500 bp).*

My lab gained the goal of the 2 genome equivalents (∼ 1 Gb) on July 2007. The first ensemble of the sequences has been assembled using the ARACHNE assembler [60] and an automatic annotation has been performed (see table 4.1). In parallel, a physical and genetic map comprising more than 1,500 markers have been constructed, permitting anchoring and orientation of the super-contigs on the 19 chromosomes. The estimated coverage obtained is 8.4 X and it is described in “The grapevine genome sequence suggest ancestral hexaploidization in major angiosperm phyla” Nature,2007.

	Number N50	**Length (kb)**	**Coverage**
Supercontigs	2,059	3,426	100.0%
Mapped ultracontigs	33	23,006	91.2%

Table 4.1: *8.4 X assembly summary `http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/`.*

To obtain the all amount of sequences (1 Gb) requested to end the project, more than one year was spent in preparing templates and performing sequencing reactions through Sanger method. This approach is undoubtedly accurate in determining the sequence of a template and allows a simplify assembly due to the length (about 800 bp) of the obtained fragments. Nevertheless, it is time consuming: the run performed on the SOLiD^{TM} machine gave me an output seven times higher within

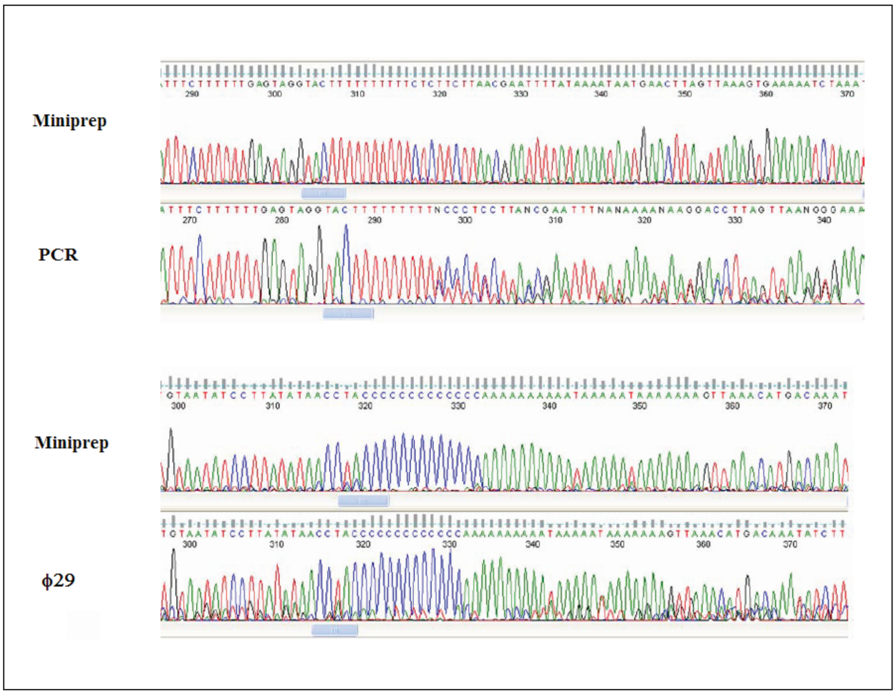


Figure 4.2: *Comparison of the same portion of reads obtained using different preparative.* ***A)*** *PCR vs miniprep: low quality base after the poly(T) for the PCR template.* ***B)*** *Phi29 vs miniprep: low quality base after the poly(C) for the Phi29 template (B).*

two weeks (considering library preparation and sequencing time).

4.2 Sequencing by ligation and coverage

The recent availability of alternative strategies for DNA sequencing including sequencing by ligation [45] and pyrosequencing [43] have dramatically change the approach to genomic, improving knowledge on the DNA sequence. The SOLiD platform (Applied Biosystems; Foster City, CA, USA) applied in this study, uses a sequencing by ligation method to sequence DNA fragments which, as a principle, should avoid sequencing errors through a double check on the single investigated base. The library preparation is accomplished by random fragmentation of DNA, followed by in vitro ligation of adaptor sequences. It is possible to have two alternative protocols generating random DNA fragments or mate-paired tags with controllable distance distributions. The amplification is performed with an emulsioned PCR, a PCR which works in micro-reactors, little drops of water containing all the components necessary for the amplification. The sequencing process consists of alternating cycles of enzyme-driven biochemistry and imaging-based data acquisition of the array at each cycle. Two mate-pairs libraries (25 + 25 nt) were synthesized from DNA of leaves tissues sampled from Merlot and Prosecco cultivars. A single standard sequencing by ligation run was performed using a whole slide of SOLiD^{TM} for Merlot and the other for Prosecco one. In total, 337.8 millions of reads for Merlot and

271.5 millions of reads for Prosecco were sequenced, corresponding respectively to 8.4 Gb and 6.8 Gb (see table 4.2). This variation of produced data between the two libraries, it is possibly due to contamination of unwanted DNA (organelles DNA) or to the quality of the sequences. The contamination of mitochondrial and plastidial DNA have been evaluated and results are shown in the next section.

Million of reads (25 nt)	**Merlot**	**Gb**	**Prosecco**	**Gb**
Tot output	337.8	8.4	271.5	6.8
First alignment	85.5	2.1	152.8	3.8
Final alignment	131.6	3.3	211.78	5.3
Not aligned	206.2	5.1	59.7	1.5

Table 4.2: *Overview of SOLiD output data and aligned reads.*

Tolerating up to five mismatches and no insertions or deletions (gap parameter = 0), I was able to align 85.5 and 152.8 million reads using PASS algorithm[54] to the reference genome of *Vitis vinifera* (Pinot noir reference assembly, based on 8.4 X WGS) ([1]). This alignment brought to the refusal of about 74.7% (Merlot) and 43.3% of the reads (Prosecco). To have a better assignment to the corresponding locations in the reference genome, the amount of produced reads was trimmed and corrected from the presence of sequencing errors (see Methods), obtaining a new set of reads. A further alignment was performed with these reads and the percentage of correctly positioned fragments was of about 39% (Merlot) and 78% (Prosecco) of sequenced reads. The improvement in mapping is shown in figure 4.3. In both alignments, pair reads and single reads (reads that have only one of the tags which map correctly to the genome) coming from the mate-pair library were aligned. A portion of the unmatched reads may arise from several parts of the genome that have been identified, but they are not correctly located yet. Furthermore, differences in genotypes may lead to the lack of a correct alignment to the reference.

Since the percentage of repeats and low complexity regions in grapevine is about 40% ([1]), I have considered only the 25 nt reads that contained a *unique* (best) match against the Pinot noir genome. Unique reads are those reads which are unambiguous located in the reference genome. As a consequence, 68.7 million reads were uniquely placed on the *Vitis vinifera* genome for Merlot and 111.2 millions reads for Prosecco cultivar. This *core data* of unique reads was used in the additional analysis for SNPs discovery. The average sequence coverage obtained with unique reads was 3.6 X for Merlot (figure 4.4 and table 4.3) and 5.8 X for Prosecco (figure 4.5 and table 4.4).

The obtained coverage is doubtless an underestimation of the SOLiD potentiality. Even if the throughput of the platform agrees with manufacture assumptions

[1] List of chromosomes deposited at the NCBI. NC_012025; NC_012024; NC_012023; NC_012022; NC_012021; NC_012020; NC_012019; NC_012018; NC_012017; NC_012016; NC_012015; NC_012014; NC_012013; NC_012012; NC_012011; NC_012010; NC_012009; NC_012008; NC_012007

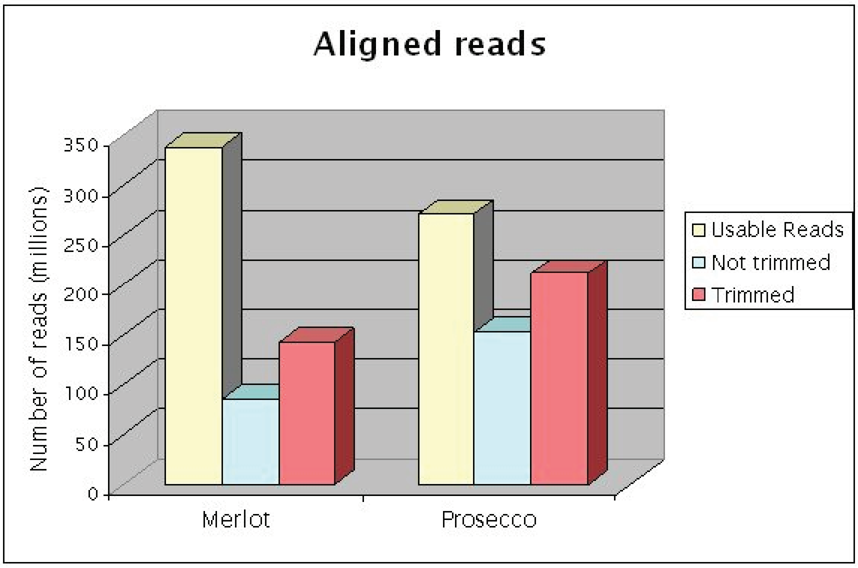


Figure 4.3: *Aligned reads. The plot shows the amount of produced - usable reads (yellow bar) and the increasing in mapping reads, using the above explained filters (green and red bars).*

	Merlot		
	Aligned reads (millions)	*Tot Mb*	*Coverage depth*
total	131.63	3,290	6.9 X
only unique	68.7	1,717	3.6 X

Table 4.3: *Through-put of Merlot alignment.*

(more than 6 Gb per run at that time), the aligned reads have produced a lower coverage than expected one. It has to be said that this run on Merlot and Prosecco is one of the firsts executed in my laboratory and it could be affected by some errors in producing libraries and/or depositing beads on the slide and/or by other general defects which can be overcome only acquiring experiences.

4.3 Organelles DNA contamination

It has been extensively demonstrated (dr. C. Ruberti personal communication) that the use of Nucleon^{TM} PhytoPure^{TM} Genomic DNA Extraction Kit on *Vitis vinifera* micropropagated samples avoids polysaccharides compounds and the contamination of organelles DNA (plastidial and mitochondrial genomic sequences) in DNA genomic extraction. This supposition is fundamental in short reads alignment against the nuclear reference genome of Pinot noir. The alignments of reads coming

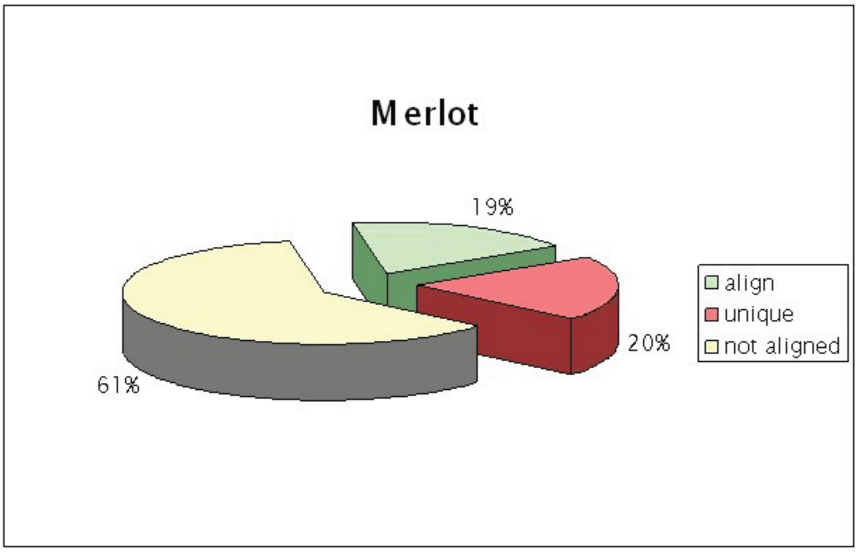


Figure 4.4: *Results of Merlot alignments: percentage of not aligned, aligned and uniquely aligned reads.*

	Prosecco		
	Aligned reads (millions)	*Tot Mb*	*Coverage depth*
total	211.9	5,295	11 X
only unique	111.2	2,780	5.8 X

Table 4.4: *Through-put of Prosecco alignment.*

from organellar DNA may affect the evaluation of discovered polymorphisms within nuclear sequences. Therefore, the mapping has been performed with PASS, using simultaneously as input file the *Vitis vinifera* genome sequences, the mitochondrial DNA (NC_012119; 773,279 bp) and the plastidial DNA sequences (NC_007957; 160,928 bp). Taking into account only uniquely mapped reads, about 2.3% of fragments have been positioned in mitochondrial DNA for Merlot sample and about 1% for Prosecco. Considering plastidial DNA, the percentage of contamination was respectively of about 4% and 1.3%. These values can be positively considered in the global analysis of polymorphisms discovery, since the presence of non-nuclear DNA is a small fraction of the produced reads. In any case, reads that had a best match with mitocondrion and plastids sequences were discarded for the further analyses. This analysis confirms the efficiency of the used kit in genomic DNA extractions avoiding the presence of exogenous DNA. In addition, the higher proportion of organelles DNA in Merlot than in Prosecco may explain, in part, the alignment values previously reported.

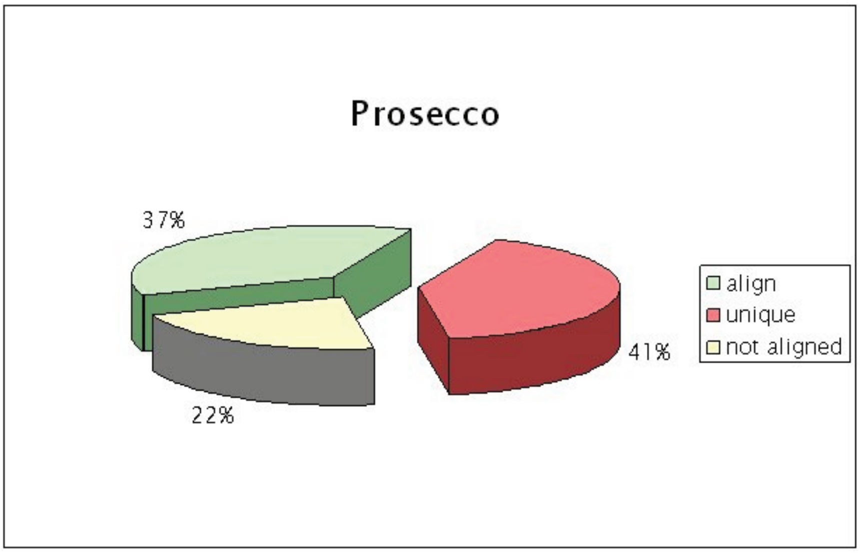


Figure 4.5: *Results of Prosecco alignments: percentage of not aligned, aligned and uniquely aligned reads.*

4.4 The mate-pairs library performance

Mate-pairs generated by sequencing were mapped against the reference genome using PASS_pair tool of PASS, which produced 14.3 million of Unique_pair reads of Merlot and 33.9 million of Unique_pair reads of Prosecco. This Unique_pair are reads which have both tags (F and R) mapping to the genome as uniquely placed pairs with the corrected insert size (see Methods). As previously reported [61], information gained from mate-paired libraries give a more comprehensive sampling of the genome with the uniquely placement of mate pairs than with the unique placement of each of the independent tags (i.e. fragment library). Taking into account only Unique_pair, PASS_pair calculates the length distribution of pair ends establishing the average size of the library. The reads distribution is plotted in figure(N) and the average value is around 1648 bp (± 601 bp) for Merlot and 1998 bp (± 611 bp) for Prosecco (Figure 4.6).

Considering that the supposed insert size was of 2000 - 3000 bp in length, the estimated value diverges from the expected due to two main aspects: the precision of the size selection step library preparation and the presence of unpaired reads (reads which have mapped only one tag). Reports of mate-pairs libraries produced in my lab on different species (human an tomato) delineate a similar divergent profile from the expected mean values for each library (dr. R. Schiavon personal communication). The size selection step on agarose gel during library preparation is possibly a critical point in insert size determination. This procedure can be ameliorated in order to obtain a narrow curve of distribution containing only desired fragments. This fact implies the inclusion in the library of undesired fragments which distend the distribution slope towards higher and lower length values.

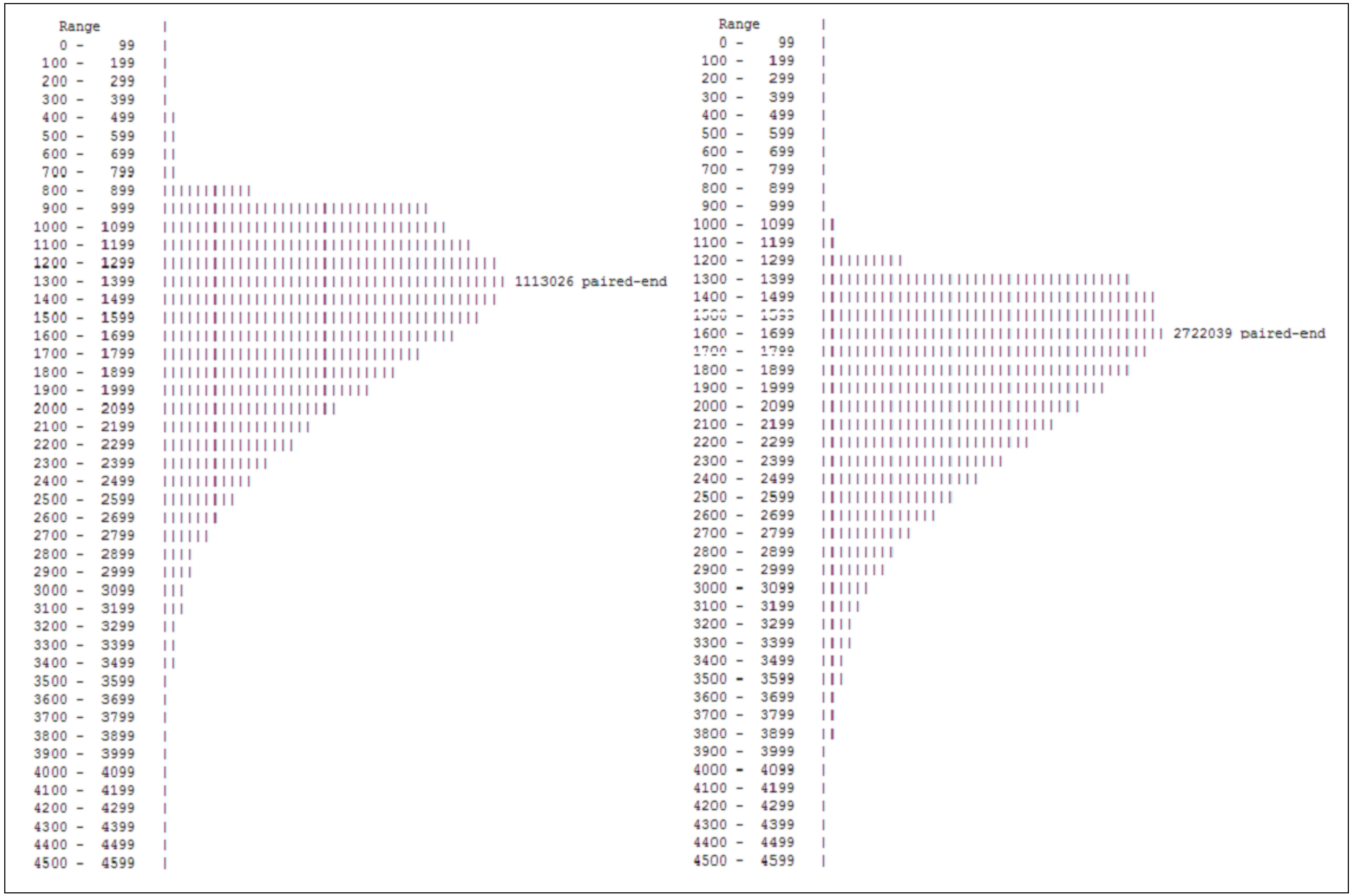


Figure 4.6: *Merlot insert library distribution. The mean value is of 1648 bp in length with a dev. st. of 601 bp (left). Prosecco insert library distribution. The mean value is of 1998 bp in length with a dev.st. of 611 bp (right).*

4.5 SNPs detection

Two different cultivars of *Vitis vinifera* were selected and sequenced using a mate-pair sequencing approach (Merlot and Prosecco), with the goal of producing significative SNPs maps which allow a possible characterization of the single cultivar. In fact, apart from the reliable estimation of relationships between varieties, it is also important to distinguish between genotypes. Usually, the majority of polymorphic information on the varietias is provided by RAPD and SSR markers discovery and investigations are made to determine the number of markers required for a reliable distinction. These traditional approaches are very accurate, but very expansive in term of money and time. With the introduction of new sequencing technologies (i.e. synthesis by ligation), a global view on the investigated genome is possible in a single experiment. In particular, Merlot and Prosecco has been chosen due to: their availability, their importance among the specie, their phenotypic characteristics, their significance in our own area (Veneto and North-East Italy) and because of the sparse of information on these cultivars.It has been previously demonstrated that mate-pairs sequencing strategy is very suitable to identify polymorphisms in a variety of species[62]. The PASS_snp tool of PASS have been used to identify polymorphisms along the entire genome of each cultivars, by aligning unique reads against chromosomes from the available assembly and comparing sequences with the reference. Due to low coverage for the single libraries (see above), I decided to use together, in SNPs mapping, reads coming from Merlot and Prosecco, increasing base coverage and assigning with more accuracy a polymorphism to a position. The figure(fig: 4.7)shows how the use of both libraries in the alignment, allows to confirm a larger number of SNPs that covered by a certain number of reads. To classify a difference in a sequence read as a true polymorphism, a minimum two reads aligning to the consensus must be present. Moreover, if there is the simultaneous presence of two alleles (indicating a possible heterozygosis), the ratio between the alleles must be less than 1:5. Even though these specifications turn down the sensitivity in detecting rare SNPs, the specificity of true SNPs detection increases, reducing the presence of false variants caused by alignment and sequencing errors. The presence of sequencing errors in the color-space SOLiD systems is evidenced by a *single* change in color-space sequence. In a second moment, the same procedure have been used for the each cultivar separately, in order to identify the polymorphisms belonging specifically to Merlot or to Prosecco.

Using these parameters, 1.2 million SNPs and 2.2 million SNPs were detected across the entire genome of Merlot and Prosecco. Among these SNPs, about 405,000 have been found to be in common between the two cultivars. The Venn diagram below shows the proportion of discovered SNPs (Figure 4.8).

Of these, 84,376 for Merlot and 131,745 for Prosecco are located in genes (UTRs, exons and introns); this values can be described as 2.5 SNPs per kb per Merlot and 4.5 SNPs per kb for Prosecco, confirming previous studies [26]. The proportion of transitions (59.5 and 60.6%) was greater than the proportion of transversions (40.5 and 39.4%) respectively for Merlot and Prosecco [5][24]. The nucleotide variation observed through the analysis of these sequences is summarized in (Table 4.5). The total amount of SNPs called by PASS can be divided into 4 categories according to the presence of one, two, three or four variant alleles (see table). The "one-called" allele indicates the presence of an homozygous variant, while the two calls indicate the heterozygous state. PASS detects the reference allele and the variant allele,

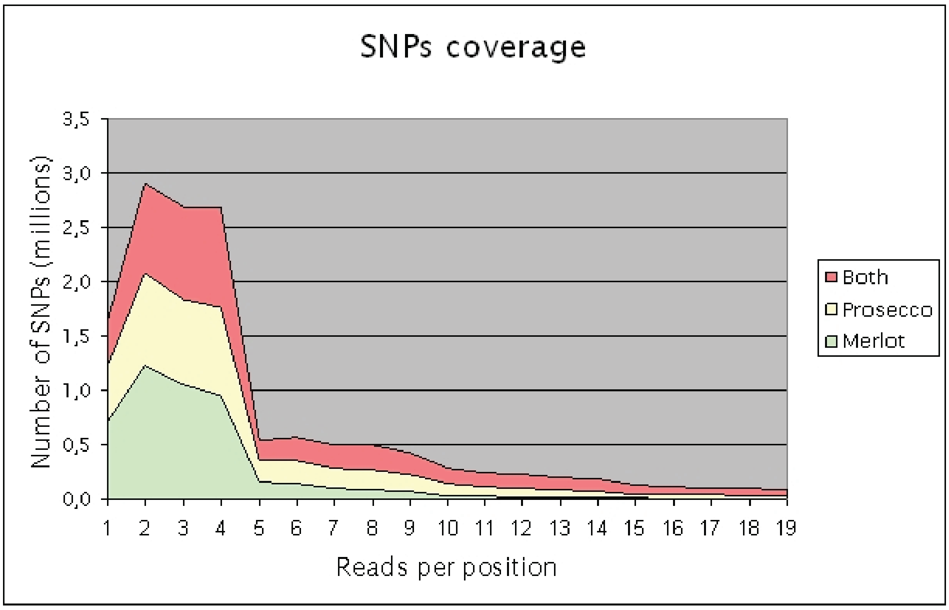


Figure 4.7: *Improvement in SNPs coverage using both libraries in mapping reads.*

calculating how much reads confirms the reference and how much the variant. A SNP is called as a homozygous variant when it is present as a "single call" which differs from the reference.

	Merlot	**Prosecco**
1 allele	666,643	934,510
2 alleles	535,140	1,264,133
3 alleles	22	53
4 alleles	13	40

Table 4.5: *Allele distribution across the two cultivars.*

SOLiD sequencing technology is very suitable for SNPs detection and the number of polymorphisms identified in a single run underlines its strength, but this kind of study, in an unknown genome doesn't address the understanding of real polymorphisms detection. As a result, I'm going to randomly choose 100 SNPs to be subsequently amplified by PCR from DNA extracted from the same leaves, using Sanger sequencing technology. A high coverage is required to sample two alleles rather than one and thus call a heterozygote rather than a homozygous locus. Reads that contain variant alleles show two colors in the color-space sequence produced by SOLiD, when a variant is present; therefore, these reads are allowed

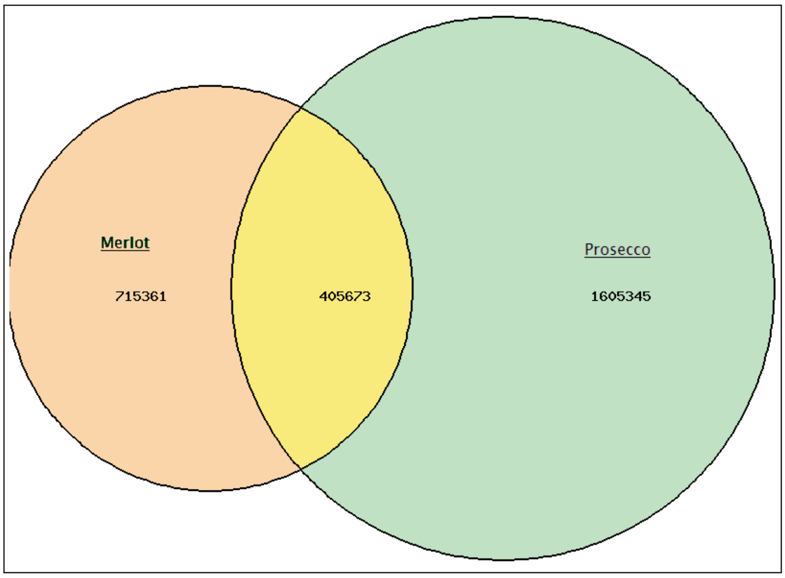


Figure 4.8: *Venn diagram of Merlot and Prosecco common SNPs. The proportion of polymorphisms only within Merlot is* 37.4%. *The proportion segregating only in Prosecco is* 83.8%, *while a* 21.2% *is in common between the two cultivars.*

fewer sequencing errors than reads with no variants. On the other side, errors lead to the presence of a single changed color, so that errors can be distinguish from SNPs. My goal is not only to detect SNPs (alternative alleles to the reference), but also to infer whether the sample is heterozygous or homozygous at a given position, the most challenging being the detection of a heterozygous state, given the sampling introduced by the shotgun process and the bias induced by mapping reads to a reference sequence. However, this task is facilitated by the error detection and correction scheme of the SOLiD sequencing chemistry, which reduces the average sequencing error rate to $< 0.1\%$. Merlot and Prosecco genomes have not been sequenced and a dbSNPs for Vitis vinifera is not yet existing, as a result, false-negatives cannot be evaluated as putative SNPs.

4.6 Analysis of synonymous and non-synonymous SNPs

Single nucleotide polymorphisms can occur in coding sequences of genes (exons), non coding regions of genes (introns) and in intergenic regions (between genes). Mutations that change amino acid sequence are called non-synonymous, those that do not, synonymous. Synonymous mutation are due to the degeneracy of the genetic code, which implies that a change in a single base will not necessarily change the amino acid codified by the triplet. The non-synonymous variations can be divided into missense variation (change an amino acid with a different kind of amino acid) or nonsense variation (change an amino acid with a stop codon). I determined whether SNPs introduce synonymous or non-synonymous mutations thanks to an ad hoc script (developed by dr. C. Forcato in my lab), which investigates the sequence reading frame, isolates codons containing SNPs, and compares the translated amino acids for each allele.

Unlike the expectation of finding higher values of synonymous rate within coding regions, the proportion of sense mutations is lower than non-synonymous one for both the cultivars. In Merlot cultivar there is an important occurrence of stop codons in the modified triplets (non-sense mutations), whereas this value is lower in the Prosecco. On the other side there is about half of SNPs that are sources of missense mutation 4.6.

%	**Merlot**	**Prosecco**
SENSE	41.6	43.1
NONSENSE	8.9	1.9
MISSENSE	49.5	55.0
transition	59.5	60.5
transversion	40.5	39.5

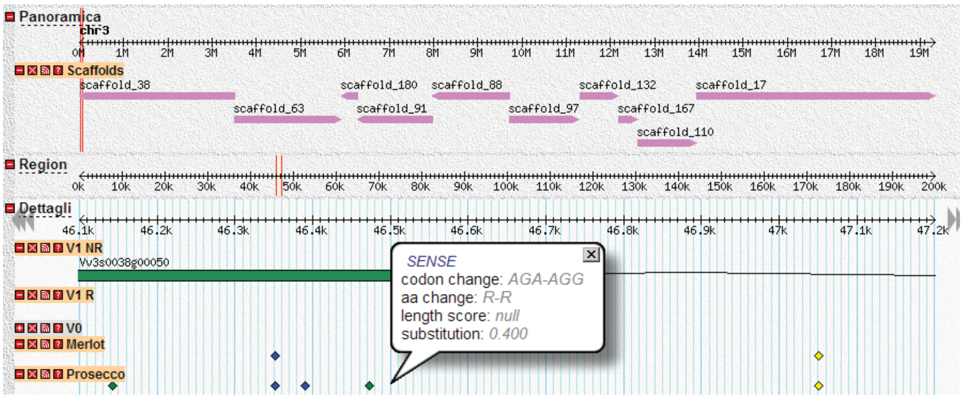
Table 4.6: *Summary overview of Merlot and Prosecco mutations.*

These findings of mutation rates are unusual, because changes in the codifying portion of the genome often imply dramatic consequences in gene working. Thus, it is necessary a deeper investigation on these mutations, taking into account the possibility of using such restricted filters in the SNPs selection. Another aspect is to analyze where this percentage is prevalent within the coding region, looking specifically for exons, introns and UTRs. A last choice may be due to the resources that heterozygosis constitute for plants. The need to respond to particular environmental changes may involve the use of different genomic states.

4.7 Mapping on the GBrowse

All the SNPs evidences produced were mapped on the *Vitis vinifera* GBrowse (`http://gbrowse.cribi.unipd.it/private/gbrowse/vitis_vinifera/`). The Gbrowse package allows to search for features on the genome assembly, zoom in and out, pan right or left, customize which features are displayed and their color and more. With this application, users can query for SNPs searching within a genomic location. The results can be restricted to confirmed SNPs, those with allele frequency data, or to those of a certain SNP function class, such as a coding or non-synonomous SNP.

All SNPs are shown as coloured points. Each point correspond to a specific kind of mutation (missense, sense and non-sense). Moreover, there is the triplet in the reference and the changed triplet in the observed cultivar. There is, at the end, the information on the amino acid. The picture below is a screenshot of the browser (4.9.


Panoramica
chr3
0M 1M 2M 3M 4M 5M 6M 7M 8M 9M 10M 11M 12M 13M 14M 15M 16M 17M 18M 19M
Scaffolds
scaffold_38
scaffold_63
scaffold_180
scaffold_91
scaffold_88
scaffold_97
scaffold_132
scaffold_167
scaffold_110
scaffold_17
Region
0k 10k 20k 30k 40k 50k 60k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 200k
Dettagli
46.1k 46.2k 46.3k 46.4k 46.5k 46.6k 46.7k 46.8k 46.9k 47k 47.1k 47.2k
V1 NR
Vv3s0038g00050
V1 R
V0
Merlot
Prosecco

SENSE
codon change: AGA-AGG
aa change: R-R
length score: null
substitution: 0.400

Figure 4.9: *Vitis vinifera genome browser. SNPs of Merlot and Prosecco are shown as coloured points on the lower part of the picture. A better explanation of the mutation can be reached going on the point.*

4.8 Identification of Structural Variants

Structural variants (SVs) consist of copy number variants (CNVs) and other rearrangements, as inversions, which do not involve a change in copy number. In plants, the majority of CNVs occurs as neutral polymorphisms, because of polyploidy (multiple sets of genes) brings an evolutionary advantage. Instead in animals variations in gene copy number often have negative consequences especially in some birth defects [63]. Array genomic hybridization is the traditional technique used to detect CNVs, but this technology does not detect SVs as inversions (variations which do not concern a variation in number). The NGS can identify structural variations, providing a better resolution than array genomic hybridization and allowing a easier genotype-phenotype correlation [64]. On the other side, short reads produced by massively parallel sequencing platforms limit their ability to map small indels (single base-pair level), because of the multi-occurrence of a short fragments in the genome, which does not allow to be mapped uniquely. Mapping a read to a unique location in the reference is necessary to recognize SVs and to count how many reads confirm a variation. The use of mate pairs permits the identification of genome rearrangements thanks to the simultaneously mapping of the two mates of the couple. Deletions map at a higher distance than expected length of the insert size, insertions map closer to each other, inversions have a wrong orientation from the original pairs. Unique mate-pairs obtained for Merlot and Prosecco have been used to map respectively 53,386 and 42,717 *indels*, that means about 1.1 indels per kb for Merlot and 0.9 for Prosecco. As expected, short indels are more likely to occur than long indels. Among these data in fact, the 18% of Merlot deletions and the 29 % of Prosecco are considered *large* deletions. I considered a deletion as *large*, if the variation falls beyond the established library standard deviation. Within this group of variations, I found five large deletions which fall in the same regions for the two considered cutivars. This event should be of a great interest and will be source of investigation for future data analyses. Regarding *insertions*, 6,100 variations have been found for Merlot and 7,773 for Prosecco. More than 50% of these SVs range between 1 to 300 bp in length. Confirming established library

insert length, no insertion larger than 2 kb were found in both cultivars.
The distribution of indels along chromosomes is shown in figure 4.10. Statistics on the mean and the median values of the ditributions indicates that the presence of indels is equally distributed in all chromosomes.

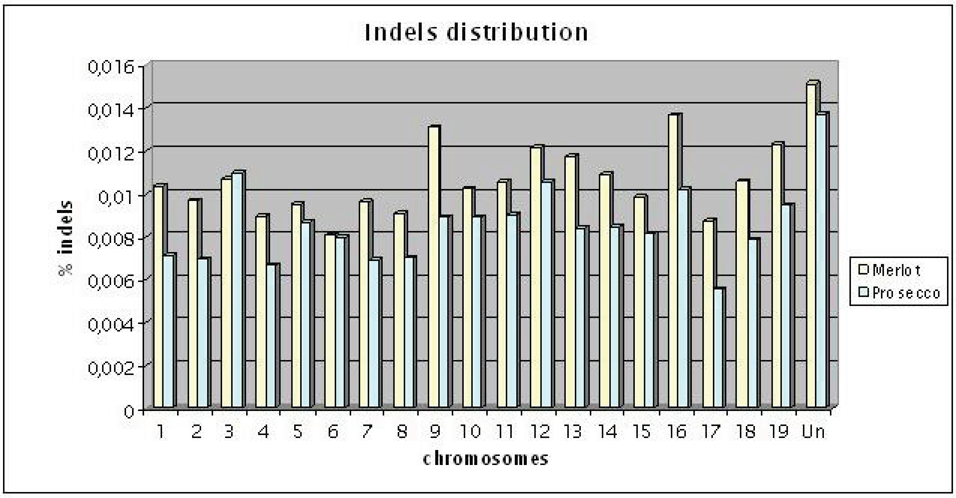


Figure 4.10: *Distribution of indels along chromosomes for Merlot and Prosecco. Merlot mean* = 0.011%*; Merlot median* = 0.010%*. Prosecco mean* = 0.009%*; Prosecco median* = 0.008% .

I successively analyzed all mate pairs with both ends mapped, but with sequences on opposite strands (Unique_pairs_wrong_d; where d stand for distance) to investigate *inversions*. SOLiD mate-pairs library creates two tags in which both sequence reads are normally on the same strand. If an inversions in the donor genome has been occurred, the tags will have a *wrong* orientation respect to the reference. PASS_pair looks for multiple mate pairs that show the same inverted orientation at the same coordinates in the reference genome. I observed 595 *inversions* for Merlot and 441 for Prosecco, with about 2 inversions per Mb for Merlot and 3 per Mb for Prosecco. The distribution of inversions along chromosomes is shown in figure 4.11 and statistics on the mean and the median values of the distributions indicates that the presence of inversions is equally distributed in all chromosomes.

Even if the physical coverage of the two libraries was thought to be higher than the observed, the main problem with this data is given by the low coverage of sequences.
Considering only reads that map uniquely to the reference, the physical coverage of Merlot library, calculate as:

$$Phy.cov. = \frac{(\overline{L}_{insert} + 2 * l_{read}) * N_p}{genome\ size}$$

where $\overline{L}_{insert}$ stands for average library length, l_{read} is the read length (25 nt) and N_p is the number of unique aligned mate-pairs.
is of 50.6 X, while the Prosecco phy.cov. is of 144.9 X.
Whereas, the coverage of sequences is calculate as

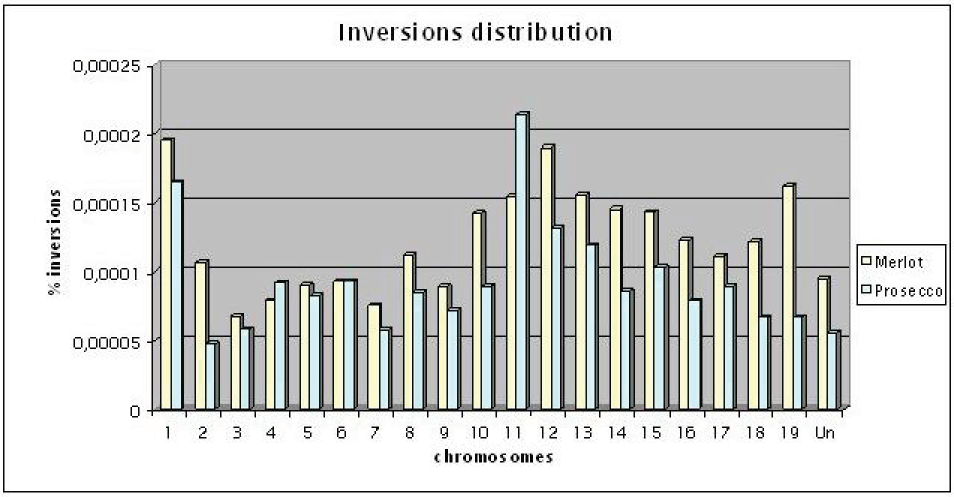


Figure 4.11: *Distribution of inversion along chromosomes for Merlot and Prosecco. Merlot mean* = 0.0001%*; Merlot median* = 0.0001%*. Prosecco mean* = $9.2 * 10-5\%$*; Prosecco median* = $8.3 * 10^{-5}\%$.

$$Seq.cov. = \frac{(l_{read} * 2 * N_p) + (l_{read} * N_s)}{genome\ size}$$

where N_p is the number of unique_single aligned
and is of 1.49 X equivalents for Merlot and 3.54 for Prosecco. The trimming of reads (see Methods) have produced reduced reads length, so that these coverage values are not precise, but give me the possibility to have an idea of the x-fold genome equivalents obtained through SOLiD sequencing. Physical coverage data will allows a further investigation of *large* indels and inversions, as already described. The finding of SVs in these two grapevine varietas, comparing to the Pinot noir, confirms the dynamic nature of the plant genomes, revealing that transposable element activity is an important source for genetic diversity.

CHAPTER 5

Conclusions

This work has pointed out that whole-genome re-sequencing with massively parallel platforms is very suitable to become the workhorse of genetic studies, because it allows a deep genome sequence coverage. These improvements in DNA sequencing technological innovations have increased the power of finding single variations (SNPs) and rearrangements (indels) along the entire genome sequence. However, little information exists on how to design these studies and there is the real risk to discard a lot of useful data.

The traditional Sanger approach [39] is doubtless very accurate in determining the DNA sequence, producing long fragments (800-1000 bp), nevertheless it is relatively expensive, time consuming and work intensive. In addition, the cloning of templates into vectors involves a series of problems which lead to the deprivation of some fragments in the library construction. As a matter of facts, by now it is unlikely to perform a *de novo* or re-sequencing strategy of multiple related genomes with this approach, because it requires a huge investment of large economical resources and the advances in new sequencing strategies. On the other side, it should be highlighted the importance of having a good reference genome sequence which is essential for comparative studies other sequences identifying mutations, polymorphisms and structural variations between organisms. In this study a high quality consensus sequence was obtained for grape with the Sanger method, ensuring a reference to be investigated.

In the last five years, several ultra high-throughput DNA sequencing technologies have transformed genomics research introducing new sequencing approaches, such as sequencing by ligation [44] and pyrosequencing [43]. Next generation sequencing (NGS) allows researchers to obtain in a single experiment complete genomes and, as a consequence, a large pool of possible nucleotide and structural variants. Whole genome sequencing studies, finding SNPs markers, lead to new sources of information which can be applied into breeding programmes and MAS (Marker Assisted Selection). The high heterozygosis of *V. vinifera* makes it very suitable for finding differences within cultivars.

To perform a comparative analysis on grape genomes, two cultivars have been selected for their specific characteristics: Merlot and Prosecco. The availability of the samples, the different growth conditions and the sparseness of genomic information on these varietas offered me the opportunity to investigate chromosomes for the discovery of SNPs and structural variations. The SOLiD (NGS) platform was applied in this study for the re-sequencing of *V. vinifera* genome. SOLiD^{TM} (Applied Biosystem, USA) uses a sequencing by ligation method and produces millions of DNA fragments in parallel (massive parallel sequencing). With this system problematic procedures, such as cloning, are eliminated. Sequence lengths generally range from 25-50 bp (short sequences) and this is sufficient for unique alignment to a reference genome. Because millions of fragments are sequenced in parallel, a fragment can be sequenced even if it exists in low abundance in the sample, increasing sequencing depth and enabling identification of single nucleotide polymorphisms (SNPs) with accuracy. The expected throughput of SOLiD platform used (version 2) was of about 6 Gb per run and this values have been confirmed by the sequencing of the two libraries. An average of seven times more data have been produced using the SOLiD system within fifteen days, in comparison with weeks of work using Sanger method. By now, these short SOLiD reads are not yet suitable for a *de novo* sequencing study, even if manufactures are working on this aspect to improve assembly efficiency with short reads. Even though, the availability of a reference genome makes SOLiD platform very appropriate for re-sequencing. The knowledge of *Vitis vinifera* genome brought this specie to be under further investigations in order to find difference among cultivars. A whole-genome sequencing with these technologies is appropriate to find SNPs and structural variations in a short time, producing an huge amount of data. In order to evaluate chromosomal rearrangements and differences among genomes, I created a mate-pairs library for each sample. The advantages of such a library reside in the possibility to look for variations in the tags mapping against the reference genome. These differences may correspond to structural variations. In addition, the reported results showed how nucleotide diversity can be sampled by high-throughput sequencing of different genotypes even if more accurate analyses need to be performed. This strategy is useful for detecting variants in a large number of genes that are in agreement with traditional sequencing projects. SNPs have been often seen as sources of deleterious mutations, but it is becoming clear that variations and rearrangements can have functional implications in gene integrity and function [32]. In this work, I have delineate a landscape of potentially variations by considering insertion and deletion events and inversion variants. This study provides guidance for future exploration of genetic variation in *Vitis vinifera* with ultra-high-throughput short-read sequencing technologies, such as SOLiD, and confirm that accuracy is an important factor that must be considered in determining the cost-effectiveness of the new sequencing methods in re-sequencing approaches. In fact, it is necessary to perform further developments, optimizing procedures as: the choice of the optimal starting DNA material, the size of genomic DNA fragments and the reduction of errors and biases.

The analysis of sequencing data requires the alignment to a reference sequence. The huge amount of information and shortness of the reads produced from NGS systems, makes bioinformatics particularly challenging. By now, several algorithms have been adapted or developed for short read alignment, including Newbler [43](Roche), SHRiMP [52] and PASS [54], but they often are platform dependent. In addition, variant discovery tools that use these alignment software are limited to a single platform. Very few tools are compatible with multiple data and aligner types and bioinformatics is still an important bottleneck, requiring a great amount of time. As a matter of fact, I spent the entire last year to collect, sort, filter and interpret data This work have pointed out the possibility of obtaining useful information from the SOLiD system. Even though these are preliminary results, the study have underlined the potential of NGS approach to investigate multiple genomes looking for similarity and/or appreciable differences to be moved between cultivars.

APPENDIX A

Plant culture

A.1 Merlot culture condition: spurred cordon

The vines have a short trunk, about 0.5m. A permanent branch, or '*cordon*', is trained along a wire on one side of the vine. The cordon, which is never pruned away, bears a number of spurs (how many often depends on appellation laws) which are subject to spur pruning (so called "spurred cordon"). The cordons may be one (unilateral cordon) or two (bilateral cordon) in number. The bilateral cordon is the most commonly encountered, but the unilateral method is becoming increasingly popular as a relatively easy method of vine training. A significant advantage of cordon training is its suitability to mechanical pruning, as the spurs are all at a very similar height along the cordon.

A.2 Prosecco culture condition

In vitro micropropagation uses the ability of plant to regenerate (*talea*). In a sterile chamber apical meristema of Prosecco with at least one leaf are cut and planted in MS$\frac{1}{2}$A.1 plates (100 ml). Plants raised in a growth chamber with a day/night period of 16/8 h, air temperature of 24 - 26 ℃. Ten days later roots are visible.

A.2.1 MS-$\frac{1}{2}$ medium preparation

The MS-$\frac{1}{2}$solution is prepared melting Murashinge-Skoog (MS)in distilled water (Duchefa - Micropoli)A.2. Afterwards saccarosio is added to reach the optimal (5.5) pH value. To obtain solid medium, 8g/l of plant agar are added. Finally, plates are autoclaved 20 minutes at 121 ℃ and 103.5kPa. These are the essential conditions to sterilize.

All information supplied by dr. C.Ruberti, lab Prof. Lo Schiavo, University of Padua, Italy.

Medium Name	**Medium source**	**Sucrose**	**pH**
MS 1/2	MS	3%	5.5

Table A.1: *MS modification.*

Macronutrients	**(mg/l)**
(CaCl2)	332.02
(KH2PO4)	170.00
(KNO3)	1,900
(MgSO4)	180.54
(NH4NO3)	1,650.00
((NH4)2PO4)	-
((NH4)2SO4)	-
Micronutrients	(mg/l)
(CoCl2*6H2O)	0.025
(CuSO4*5H2O)	0.025
(FeNaEDTA)	36.70
(H3BO3)	6.20
(KI)	0.83
(MnSO4*H2O)	16.90
(Na2MoO4*2H2O)	0.25
(ZnSO4*7H2O)	8.60
Comuni additivi organici	(mg/l)
biotin	-
Folic acid	-
Glicine	2.00
Myo-Inositol	100.00
Nicotinic Acid	0.50
Pyridoxine*HCl	0.50
Thiamine *HCl	0.10

Table A.2: *Murashinge-Skoog (MS).*

APPENDIX B

Abbreviation

B.1 SI units

g	g
L	litre
min	minute
mol	mole
s	second
℃	degrees Celsius

B.2 SI prefixes

n	nano- (10^{-9})
μ	micro- (10^{-6})
m	milli- (10^{-3})
k	kilo- (10^{3})

B.3 Other abbreviations and terms

A	adenine
aa	amino acid
abs	absolute
aCGH	Array comparative genomic hybridization
ATP	adenine triphosphate
BAC	Bacterial Artificial Chromosome
bp	base pair
C	cytosine
cDNA	complementary DNA

CNV	copy number variation
contig	set of overlapping DNA segments derived from a single genetic source
CHCl3	chloroform
C3H8O	Isopropyl alcohol
CH3CO2K	Potassium Acetate
CRIBI	Centro di Ricerca Interdipartimentale per le Biotecnologie Innovative
cultivar	cultivated variety
DEPC	diethylpyrocarbonate
DNA	deoxyribonucleic acid
DTT	dithiothreitol
donor	the genome investigated
EDTA	ethylene-diamine tetraacetic acid
EST	expressed sequence tag
EtBr	Ethidium bromide
EtOH	Ethyl Alcohol
G	guanine
Gb	Giga bases
h	hours
ht	high throughput
IGGP	International Grape Genome Program
indels	insertions - deletions
kb	kilo bases
KH2PO4	Potassium dihydrogen phosphate
L.	Linneus
Log	common logarithm
MAS	marker assisted selection
Mb	Mega bases
mRNA	messenger RNA
NaAc	Sodium acetate
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
PEM	Pair End Mapping
RNA	ribonucleic acid
RT	room temperature
SDS	sodium dodecyl-sulphate
SNP	single nucleotide polymorphism
SV	structural variation
T	thymine
UV	ultraviolet light
WGS	whole genome shotgun

Bibliography

[1] Olivier Jaillon, Jean-Marc Aury, Benjamin Noel, Alberto Policriti, Christian Clepet, Alberto Cassagrande, Nathalie Choisne, Sébastien Aubourg, Nicola Vitulo, Claire Jubin, Alessandro Vezzi, Fabrice Legeai, Philippe Hugueney, Corinne Dasilva, David Horner, Erica Mica, Delphine Jublot, Julie Poulain, Clémence Bruyère, Alain Billaut, Béatrice Ségurens, Michel Gouyvenoux, Edgardo Ugarte, Federica Cattorano, Véronique Anthouard, Virginie Vico, Christian Del Fabbro, Michaël Alaux, Gabriele Di Gaspero, Vincent Dumas, Nicoletta Felice, Sophie Paillard, Irena Juman, Marco Moroldo, Simone Scalabrin, Aurélie Canaguier, Isabelle Le Clainche, Giorgio Malacrida, Eléonore Durand, Graziano Pesole, Valérie Laucou, Philippe Chatelet, Didier Merdinoglu, Massimo Delledonne, Mario Pezzotti, Alain Lecharny, Claude Scarpelli, François Artiguenave, M. Enrico Pé, Giorgio Valle, Michele Morgante, Michel Caboche, Anne-Françoise Adam Blondon, Jean Weissenbach, Francis Quétier, and Patrick Wincker. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–7, 2007 Sep 27.

[2] Thierry Lacombe1 Patrice This1 and Mark R. Thomas. Historical origins and genetic diversity of wine grapes trends in genetics. Volume 22, Issue 9:511–519, September 2006.

[3] D. Zohary. *The Domestication of the Grapevine Vitis Vinifera L. in the Near East*, pages 23–30. P.E. Mc Govern et al., Editors. The origins and Ancient History of Wine, Gordon and Breach (1995), 1995.

[4] G Alleweldt and E Dettweiller. The genetic resources of vitis: World list of grapevine collections.

[5] Silvia Vezzulli, Diego Micheletti, Summaira Riaz, Massimo Pindo, Roberto Viola, Patrice This, M Andrew Walker, Michela Troggio, and Riccardo Velasco. A snp transferability survey within the genus vitis. *BMC Plant Biology*, 8(1):128, 2008.

[6] R Arroyo-Garcia, L Ruiz-Garcia, L Bolling, R Ocete, MA Lopez, C Arnold, A Ergul, G Soylemezoglu, HI Uzun, F Cabello, J Ibanez, MK Aradhya,

A Atanassov, I Atanassov, S Balint, JL Cenis, L Costantini, S Goris-Lavets, MS Grando, BY Klein, PE McGovern, D Merdinoglu, I Pejic, F Pelsy, N Primikirios, V Risovannaya, KA Roubelakis-Angelakis, H Snoussi, P Sotiri, S Tamhankar, P This, L Troshin, JM Malpica, F Lefort, and JM Martinez-Zapater. Multiple origins of cultivated grapevine (vitis vinifera l. ssp. sativa) based on chloroplast dna polymorphisms. *Mol Eco*, 15(12):3707–14, 2006.

[7] Williams LE Mullins MG, Bouquet A. Biology of grapevine. *Cambridge University Press*, 1992.

[8] MK Aradhya, GS Dangl, BH Prins, JM Boursiquot, MA Walker, CP Meredith, and CJ Simon. Genetic structure and differentiation in cultivated grape, vitis vinifera l. *Genet Res*, 81(3):179–92, 2003.

[9] A Bronner and J Oliveira. Creation and study of the pinot noir variety lineage. *Proceedings of the 5th Internnatioanl Symposium of Grape Breeding. St Martin/Pflaz, Germany*, pages 69–80, 1989.

[10] Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, December 2000.

[11] Patrick S. Schnable, Doreen Ware, Robert S. Fulton, Joshua C. Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A. Graves, Patrick Minx, Amy Denise Reily, Laura Courtney, Scott S. Kruchowski, Chad Tomlinson, Cindy Strong, Kim Delehaunty, Catrina Fronick, Bill Courtney, Susan M. Rock, Eddie Belter, Feiyu Du, Kyung Kim, Rachel M. Abbott, Marc Cotton, Andy Levy, Pamela Marchetto, Kerri Ochoa, Stephanie M. Jackson, Barbara Gillam, Weizu Chen, Le Yan, Jamey Higginbotham, Marco Cardenas, Jason Waligorski, Elizabeth Applebaum, Lindsey Phelps, Jason Falcone, Krishna Kanchi, Thynn Thane, Adam Scimone, Nay Thane, Jessica Henke, Tom Wang, Jessica Ruppert, Neha Shah, Kelsi Rotter, Jennifer Hodges, Elizabeth Ingenthron, Matt Cordes, Sara Kohlberg, Jennifer Sgro, Brandon Delgado, Kelly Mead, Asif Chinwalla, Shawn Leonard, Kevin Crouse, Kristi Collura, Dave Kudrna, Jennifer Currie, Ruifeng He, Angelina Angelova, Shanmugam Rajasekar, Teri Mueller, Rene Lomeli, Gabriel Scara, Ara Ko, Krista Delaney, Marina Wissotski, Georgina Lopez, David Campos, Michele Braidotti, Elizabeth Ashley, Wolfgang Golser, HyeRan Kim, Seunghee Lee, Jinke Lin, Zeljko Dujmic, Woojin Kim, Jayson Talag, Andrea Zuccolo, Chuanzhu Fan, Aswathy Sebastian, Melissa Kramer, Lori Spiegel, Lidia Nascimento, Theresa Zutavern, Beth Miller, Claude Ambroise, Stephanie Muller, Will Spooner, Apurva Narechania, Liya Ren, Sharon Wei, Sunita Kumari, Ben Faga, Michael J. Levy, Linda McMahan, Peter Van Buren, Matthew W. Vaughn, Kai Ying, Cheng-Ting Yeh, Scott J. Emrich, Yi Jia, Ananth Kalyanaraman, An-Ping Hsia, W. Brad Barbazuk, Regina S. Baucom, Thomas P. Brutnell, Nicholas C. Carpita, Cristian Chaparro, Jer-Ming Chia, Jean-Marc Deragon, James C. Estill, Yan Fu, Jeffrey A. Jeddeloh, Yujun Han, Hyeran Lee, Pinghua Li, Damon R. Lisch, Sanzhen Liu, Zhijie Liu, Dawn Holligan Nagel, Maureen C. McCann, Phillip SanMiguel, Alan M. Myers, Dan Nettleton, John Nguyen, Bryan W. Penning, Lalit Ponnala, Kevin L. Schneider, David C. Schwartz, Anupma Sharma, Carol Soderlund, Nathan M. Springer, Qi Sun, Hao Wang, Michael Waterman, Richard Westerman, Thomas K. Wolfgruber,

Lixing Yang, Yeisoo Yu, Lifang Zhang, Shiguo Zhou, Qihui Zhu, Jeffrey L. Bennetzen, R. Kelly Dawe, Jiming Jiang, Ning Jiang, Gernot G. Presting, Susan R. Wessler, Srinivas Aluru, Robert A. Martienssen, Sandra W. Clifton, W. Richard McCombie, Rod A. Wing, and Richard K. Wilson. The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, 326(5956):1112–1115, 2009.

[12] Ferdinand Regner, Alexandra Stadlbauer, Cornelia Eisenheld, and Herwlg Kaserer. Genetic Relationships Among Pinots and Related Cultivars. *Am. J. Enol. Vitic.*, 51(1):7–14, 2000.

[13] Jancis Robinson. *The Oxford Companion to Wine.* Oxford University Press, third edition edition, 2006.

[14] Oz Clarke. *Oz Clarke's Encyclopedia of Grapes.* Harcourt Books, 2001.

[15] Ron S ackson. *Wine Science.* Academic Press, second edition edition, 2000.

[16] P This, T Lacombe, and MR Thomas. Historical origins and genetic diversity of wine grapes. *Trends Genet*, 22(9):511–9, 2006.

[17] M Salmaso, G Malacarne, M Troggio, G Faes, M Stefanini, MS Grando, and R Velasco. A grapevine (vitis vinifera l.) genetic map integrating the position of 139 expressed genes. *Theor Appl Genet*, 116(8):1129–1143, 2008.

[18] Chen X Park WD Beachell HM Dilday RH Goto M McCouch SR. Olufowote JO, Xu Y. Comparative evaluation of within-cultivar variation of rice (oryza sativa l.) using microsatellite and rflp markers. *Genome*, 40:370–378, 1997.

[19] Delphine Van Inghelandt, Albrecht Melchinger, Claude Lebreton, and Benjamin Stich. Population structure and genetic diversity in a commercial maize breeding program assessed with ssr and snp markers. *TAG Theoretical and Applied Genetics*, 120:1289–1299, 2010. 10.1007/s00122-009-1256-2.

[20] R. Lande and R. Thompson. Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits. *Genetics*, 124(3):743–756, 1990.

[21] Vignani R Meredith CP. Bowers JE, Dangl GS. Isolation and characterization of new polymorphic simple sequence repeat loci in grape (vitis vinifera l.). *Genome*, 39:628–33, 1996.

[22] Andreas Dotsch, Claudia Pommerenke, Florian Bredenbruch, Robert Geffers, and Susanne Haussler. Evaluation of a microarray-hybridization based method applicable for discovery of single nucleotide polymorphisms (snps) in the pseudomonas aeruginosa genome. *BMC Genomics*, 10(1):29, 2009.

[23] C. Batley J. Edwards D. Imelfort, M. Duran. Discovering genetic polymorphisms in next-generation sequencing data.

[24] Diego Lijavetzky, Jose Cabezas, Ana Ibanez, Virginia Rodriguez, and Jose Martinez-Zapater. High throughput snp discovery and genotyping in grapevine (vitis vinifera l.) by combining a re-sequencing approach and snplex technology. *BMC Genomics*, 8(1):424, 2007.

[25] Sean Myles, Jer-Ming Chia, Bonnie Hurwitz, Charles Simon, Gan Yuan Zhong, Edward Buckler, and Doreen Ware. Rapid genomic characterization of the genus ¡italic¿vitis¡/italic¿. *PLoS ONE*, 5(1):e8219, 01 2010.

[26] Riccardo Velasco, Andrey Zharkikh, Michela Troggio, Dustin A. Cartwright, Alessandro Cestaro, Dmitry Pruss, Massimo Pindo, Lisa M. FitzGerald, Silvia Vezzulli, Julia Reid, Giulia Malacarne, Diana Iliev, Giuseppina Coppola, Bryan Wardell, Diego Micheletti, Teresita Macalma, Marco Facci, Jeff T. Mitchell, Michele Perazzolli, Glenn Eldredge, Pamela Gatto, Rozan Oyzerski, Marco Moretto, Natalia Gutin, Marco Stefanini, Yang Chen, Cinzia Segala, Christine Davenport, Lorenzo DemattÃ¨, Amy Mraz, Juri Battilana, Keith Stormo, Fabrizio Costa, Quanzhou Tao, Azeddine Si-Ammour, Tim Harkins, Angie Lackey, Clotilde Perbost, Bruce Taillon, Alessandra Stella, Victor Solovyev, Jeffrey A. Fawcett, Lieven Sterck, Klaas Vandepoele, Stella M. Grando, Stefano Toppo, Claudio Moser, Jerry Lanchbury, Robert Bogden, Mark Skolnick, Vittorio Sgaramella, Satish K. Bhatnagar, Paolo Fontana, Alexander Gutin, Yves Van de Peer, Francesco Salamini, and Roberto Viola. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE*, 2(12):e1326, 12 2007.

[27] S Vezzulli, M Troggio, G Coppola, A Jermakow, D Cartwright, M Stefanini, MS Grando, AF Adam-Blondon, MR Thomas, P This, and R Velasco. A functional integrated map for cultivated grapevine (vitis vinifera l.) from three pedigrees, based on 283 ssr and 501 snp markers. *Theoretical and Applied Genetics*, 117:499–511, 2008.

[28] Richard Sudar Damir Clark Steven Poole Ian Kowbel David Collins Colin Kuo Wen-Lin Chen Chira Zhai Ye Dairkee Shanaz H. Ljung Britt-marie Gray Joe W.Albertson Donna G. Pinkel, Daniel Segraves. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, 20:207–11, 10 1998.

[29] International HapMap Consortium. The international hapmap project. *Nature*, 437:1299–1320, 2005.

[30] Andrew J Bailey Jeffrey A Kaul Rajinder Morrison V Anne Pertz Lisa M Haugen Eric Hayden Hillary Albertson Donna Pinkel Daniel Olson Maynard V Eichler Evan E Tuzun, Eray Sharp. Fine-scale structural variation of the human genome. *Nat Genet*, 37:727 – 732, 05 2005.

[31] Monica Brudno Michael Medvedev, Paul Stanciu. Computational methods for discovering structural variation with next-generation sequencing. *Nat Meth*, 11 2009.

[32] Gregory M. Donahue William F. Hayden Hillary S. Sampas Nick Graves Tina Hansen Nancy Teague Brian Alkan Can Antonacci Francesca Haugen Eric Zerr Troy Yamada N. Alice Tsang Peter Newman Tera L. Tuzun Eray Cheng Ze Ebling Heather M. Tusneem Nadeem David Robert Gillett Will Phelps Karen A. Weaver Molly Saranga David Brand Adrianne Tao Wei Gustafson Erik McKernan Kevin Chen Lin Malig Maika Smith Joshua D. Korn Joshua M. McCarroll Steven A. Altshuler David A. Peiffer Daniel A. Dorschner Michael

Stamatoyannopoulos John Schwartz David Nickerson Deborah A. Mullikin James C. Wilson Richard K. Bruhn Laurakay Olson Maynard V. Kaul Rajinder Smith Douglas R. Eichler Evan E. Kidd, Jeffrey M. Cooper. Mapping and sequencing of structural variation from eight human genomes. *Nature*, pages 56–64, 05 2008.

[33] Derek Y. Chiang, Gad Getz, David B. Jaffe, Michael J. O'Kelly, Xiaojun Zhao, Scott L. Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S. Lander. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods*, 6(1):99–103, January 2009.

[34] Wylie T Larson DE McLellan MD Mardis ER Weinstock GM Wilson RK Koboldt DC, Chen K and Ding L. Varscan: variant detection in massively parallel sequencing of individual and pooled samples. bioinformatics (oxford, england). 25(17):2283–5, 2009.

[35] Mu XJ Carriero N Cayting P Zhang Z Snyder M Gerstein M Korbel J, Abyzov A. Pemer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*, 2009.

[36] Evan Eichler Fereydoun Hormozdiari, Can Alkan and S.Cenk Sahinalp. Combinatorial algorithms for structural variation detection in high throughput sequenced genomes. *Genome Research*, 19(7):1270–8, 07 2009.

[37] Can Alkan3 Seunghak Lee1, Fereydoun Hormozdiari2 and Michael Brudno. Modil: detecting small indels from clone-end sequencing with mixtures of distributions. *Nature Methods*, 6:473 – 474, 2009.

[38] Costa GL McLaughlin SF Fu Y Tsung EF Clouser CR Duncan C Ichikawa JK Lee CC Zhang Z Ranade SS Dimalanta ET Hyland FC Sokolsky TD Zhang L Sheridan A Fu H Hendrickson CL Li B Kotler L Stuart JR Malek JA Manning JM Antipova AA Perez DS Moore MP Hayashibara KC Lyons MR Beaudoin RE Coleman BE Laptewicz MW Sannicandro AE Rhodes MD Gottimukkala RK Yang S Bafna V Bashir A MacBride A Alkan C Kidd JM Eichler EE Reese MG De La Vega FM Blanchard AP McKernan KJ, Peckham HE. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. 19(9):1527–41, 09 2009.

[39] F. Sanger, S. Nicklen, and A. R. Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, December 1977.

[40] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science (New York, N.Y.)*, 269(5223):496–512, July 1995.

[41] Michele Morgante. High-throughput investigation of snps and genomic rearrangements through ngs data. September 2009.

[42] Olivier Harismendy, Pauline Ng, Robert Strausberg, Xiaoyun Wang, Timothy Stockwell, Karen Beeson, Nicholas Schork, Sarah Murray, Eric Topol, Samuel Levy, and Kelly Frazer. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome biology*, 10(3):R32+, 2009.

[43] Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, Michael S. Braverman, Yi-Ju J. Chen, Zhoutao Chen, Scott B. Dewell, Lei Du, Joseph M. Fierro, Xavier V. Gomes, Brian C. Godwin, Wen He, Scott Helgesen, Chun Heen H. Ho, Chun He H. Ho, Gerard P. Irzyk, Szilveszter C. Jando, Maria L. Alenquer, Thomas P. Jarvie, Kshama B. Jirage, Jong-Bum B. Kim, James R. Knight, Janna R. Lanza, John H. Leamon, Steven M. Lefkowitz, Ming Lei, Jing Li, Kenton L. Lohman, Hong Lu, Vinod B. Makhijani, Keith E. McDade, Michael P. McKenna, Eugene W. Myers, Elizabeth Nickerson, John R. Nobile, Ramona Plant, Bernard P. Puc, Michael T. Ronan, George T. Roth, Gary J. Sarkis, Jan Fredrik F. Simons, John W. Simpson, Maithreyan Srinivasan, Karrie R. Tartaro, Alexander Tomasz, Kari A. Vogt, Greg A. Volkmer, Shally H. Wang, Yong Wang, Michael P. Weiner, Pengguang Yu, Richard F. Begley, and Jonathan M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, September 2005.

[44] Kun Li Jin Billy Xie Bin Austin Derek Vassallo Sara L LeProust Emily M Peck Bill J Emig Christopher J Dahl Fredrik Gao Yuan Church George M Shendure Jay Porreca, Gregory J Zhang. Multiplex amplification of large sets of human exons. *Nat Meth*, 4, 11 2007.

[45] Jay Shendure, Gregory J. Porreca, Nikos B. Reppas, Xiaoxia Lin, John P. McCutcheon, Abraham M. Rosenbaum, Michael D. Wang, Kun Zhang, Robi D. Mitra, and George M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732, September 2005.

[46] Philippe Lefrancois, Ghia Euskirchen, Raymond Auerbach, Joel Rozowsky, Theodore Gibson, Christopher Yellman, Mark Gerstein, and Michael Snyder. Efficient yeast chip-seq using multiplex short-read dna sequencing. *BMC Genomics*, 10(1):37, 2009.

[47] P Kerr Wall, Jim Leebens-Mack, Andre Chanderbali, Abdelali Barakat, Erik Wolcott, Haiying Liang, Lena Landherr, Lynn Tomsho, Yi Hu, John Carlson, Hong Ma, Stephan Schuster, Douglas Soltis, Pamela Soltis, Naomi Altman, and Claude dePamphilis. Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, 10(1):347, 2009.

[48] Jeong-Hwan Mun, Soo-Jin Kwon, Tae-Jin Yang, Young-Joo Seol, Mina Jin, Jin-A Kim, Myung-Ho Lim, Jung Sun Kim, Seunghoon Baek, Beom-Soon Choi, Hee-Ju Yu, Dae-Soo Kim, Namshin Kim, Ki-Byung Lim, Soo-In Lee, Jang-Ho Hahn, Yong Pyo Lim, Ian Bancroft, and Beom-Seok Park. Genome-wide comparative analysis of the brassica rapa gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. *Genome Biology*, 10(10):R111, 2009.

[49] Antonio M. Ramos, Richard P. M. A. Crooijmans, Nabeel A. Affara, Andreia J. Amaral, Alan L. Archibald, Jonathan E. Beever, Christian Bendixen, Carol

Churcher, Richard Clark, Patrick Dehais, Mark S. Hansen, Jakob Hedegaard, Zhi-Liang Hu, Hindrik H. Kerstens, Andy S. Law, Hendrik-Jan Megens, Denis Milan, Danny J. Nonneman, Gary A. Rohrer, Max F. Rothschild, Tim P. L. Smith, Robert D. Schnabel, Curt P. Van Tassell, Jeremy F. Taylor, Ralph T. Wiedmann, Lawrence B. Schook, and Martien A. M. Groenen. Design of a high density snp genotyping assay in the pig using snps identified and characterized by next generation sequencing technology. *PLoS ONE*, 4(8):e6524, 08 2009.

[50] Devin Dressman, Hai Yan, Giovanni Traverso, Kenneth W. Kinzler, and Bert Vogelstein. Transforming single dna molecules into fluorescent magnetic particles for detection and enumeration of genetic variations.

[51] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

[52] Stephen M. Rumble, Phil Lacroute, Adrian V. Dalca, Marc Fiume, Arend Sidow, and Michael Brudno. Shrimp: Accurate mapping of short color-space reads. *PLoS Comput Biol*, 5(5):e1000386, 05 2009.

[53] Hao Lin, Zefeng Zhang, Michael Q. Zhang, Bin Ma, and Ming Li. ZOOM! Zillions of oligos mapped. *Bioinformatics*, 24(21):2431–2437, 2008.

[54] Davide Campagna, Alessandro Albiero, Alessandra Bilardi, Elisa Caniato, Claudio Forcato, Svetlin Manavski, Nicola Vitulo, and Giorgio Valle. PASS: a program to align short sequences. *Bioinformatics*, 25(7):967–968, 2009.

[55] Bruno Zeitouni, Valentina Boeva, Isabelle Janoueix-Lerosey, Sophie Loeillet, Patricia Legoix-né, Alain Nicolas, Olivier Delattre, and Emmanuel Barillot. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, 26(15):1895–1896, 2010.

[56] EF Sambrook, J Fritsch and T Maniatis. *Molecular cloning: a laboratory manual.* Cold spring Harbor Laboratory Press, second edition, 1989.

[57] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453, 1970.

[58] Guang-Ning Ye Norman F. Weeden Lodhi, Muhammad A. and Bruce I. Reisch. A simple and efficient method for dna extraction from grapevine cultivars, *Vitis* species and ampelopsis. *Plant Molecular Biology Reporter*, 12(1):6–13, 1994.

[59] Templiphi, phi29 dna polymerase based rolling circle amplification of templates for dna sequencing.

[60] Serafim Batzoglou, David B. Jaffe, Ken Stanley, Jonathan Butler, Sante Gnerre, Evan Mauceli, Bonnie Berger, Jill P. Mesirov, and Eric S. Lander. ARACHNE: A Whole-Genome Shotgun Assembler. *Genome Research*, 12(1):177–189, 2002.

[61] Kevin Judd McKernan, Heather E. Peckham, Gina L. Costa, Stephen F. McLaughlin, Yutao Fu, Eric F. Tsung, Christopher R. Clouser, Cisyla Duncan, Jeffrey K. Ichikawa, Clarence C. Lee, Zheng Zhang, Swati S. Ranade, Eileen T.

Dimalanta, Fiona C. Hyland, Tanya D. Sokolsky, Lei Zhang, Andrew Sheridan, Haoning Fu, Cynthia L. Hendrickson, Bin Li, Lev Kotler, Jeremy R. Stuart, Joel A. Malek, Jonathan M. Manning, Alena A. Antipova, Damon S. Perez, Michael P. Moore, Kathleen C. Hayashibara, Michael R. Lyons, Robert E. Beaudoin, Brittany E. Coleman, Michael W. Laptewicz, Adam E. Sannicandro, Michael D. Rhodes, Rajesh K. Gottimukkala, Shan Yang, Vineet Bafna, Ali Bashir, Andrew MacBride, Can Alkan, Jeffrey M. Kidd, Evan E. Eichler, Martin G. Reese, Francisco M. De La Vega, and Alan P. Blanchard. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, 19(9):1527–1541, 2009.

[62] Ada Ching, Katherine Caldwell, Mark Jung, Maurine Dolan, Oscar Smith, Scott Tingey, Michele Morgante, and Antoni Rafalski. Snp frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genetics*, 3(1):19, 2002.

[63] Bejjani B. A. Torchia B. Kirkpatrick S. Coppinger J. Shaffer, L. G. and B. C Ballif. The identification of microdeletion syndromes and other chromosome abnormalities: Cytogenetic methods of the past, new technologies for the future. *American Journal of Medical Genetics Part C*, 2007.

[64] Jerry Davison, Anand Tyagi, and Luca Comai. Large-scale polymorphism of heterochromatic repeats in the dna of arabidopsis thaliana. *BMC Plant Biology*, 7(1):44, 2007.