UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Scienze Biologiche e Molecolari

XXIV Ciclo

Bioinformatic tools for next generation genomics

Matteo Chiara mat. R08171

Tutor: Dr. David S. Horner

Anno Accademico 2010-2011

Index

Part I....1
Abstract....1
Premise....2
Introduction....3
From DNA sequencing to Next Generation Genomics....3
DNA and the code-book of life....3
The pioneering ages of DNA sequencing....5
Sequencing and the early genomics era....9
The early steps of bioinformatics....11
First generation HT DNA sequencing: Automation, Parallelization and Minaturization....12
Automated DNA sequencing and bioinformatics....14
The genomic era....16
Next Generation DNA sequencing technologies....25
Roche 454....26
Illumina Genome Analyzer....31
Applied Biosystems SOLiD....36
Next generation Costs....40
Third generation sequencing technologies....41
Next generation genomics and functional genomics....43
Genome assembly and characterization of genomic variants....46
Large-scale transcriptome analysis by RNA-Seq....49
ChiP-Seq....53
Small RNAs....56
Epigenomics studies....59
Part II....61
Aim of the project....61
Main Results....64
Deep sequencing and RNA editing of plant mitochondrial Transcripts....64
The mitochondrial genome of angiosperms: an overview....64
The mtDNA of Vitis vinifera....65
RNA editing in plants mitochodria....67
Studying the “editome....68
Assembly and annotation of the PN40024 mitochondrial genome....68
Reads mapping and comparison of sequencing technologies....71
Identification of edit sites....72
Identification of edit sites in Vitis vinifera....74
Development of a new software package for the determination of intra specific genomic diversity....77
Genomic variation in the human genome....77
Read-pair technologies....81
Read-depth methods. 82
Split-read approaches....82
Sequence assembly....83
NGS data and SVs detection: an overview....86

The 1000 Genomes project....87
Read pairs, small indels and 1000 GP....90
SVM2. Small indels and RP: a more comprehensive approach....94
Support Vector Machines....96
A time saving heuristic....97
Formulation of Genomic Windows....98
Features to describe read mapping patterns....99
Post-processing and Estimation of Event sizes and types....103
Training the SVM....103
Comparison with other tools....104
Accuracy of classification and genomic context of predictions....108
Assembly of fungal genomes and development of a custom scaffolder.....111
Genome assembly and scaffolding....111
WGS and its limitations....112
Next-generation assemblers....114
Next generation protocols for paired reads....117
Genome scaffolding and Next Generation scaffolders....120
De-novo assembly of two fungal genomes(Fusarium)....122
Sequencing and assembly of 2 closely related fusarium specimens....124
Fusarium sequencing data....126
Quality assessment trimming and assembly of PE data....127
Preliminary assemblies and comparison between different assemblers....132
PE data and difficulties in assembly 136
Explorative comparison of the assemblies....138
MP data and scaffolding....140
Development of a new scaffolder....143
Conclusion and final remarks....150
Appendix 1: Papers and manuscripts....174
Part III....215
Supplementary materials attached to the SVM^2 manuscript....216

Index of figures

Figure 1. The Sanger method....8
Figure 2: Automated DNA sequencing....14
Figure 3: Timeline of the human genome sequencing project....24
Figure 4 454 Pyrosequencing....30
Figure 5. Illumina Sequencing....35
Figure 6: Sequencing by ligation....39
Figure 7. Decrease in Sequencing cost 41
Figure 8: Identification of edit sites....74
Figure 9:Principal statistics of detected RNA editing sites in V. vinifera....76
Figure 10: Different approaches used for SV detection....85
Figure 11: 1000 Genomes Project....89
Figure 12: The Breakdancer approach....93
Figure 13: Expected patterns of reads mapping in the presence of different SV....95
Figure 14: Statistical test used by SVM2....101
Figure 15: Features used by SVM2....102
Figure 16:Sensitivity and specificity of SVM2....107

Figure 17: Overlap between the prediction by SVM2 BD and Pindel.....108
Figure 18: Empirical sensitivity in different genomic contexts.....110
Figure 19: Illumina PE and MP protocols.....120
Figure 20: Quality score plot.....130
Figure 21: Kmer graphs.....131
Figure 22: Positional quality plot after trimming.....132
Figure 23: Example of plots used to compare the assemblies.....140
Figure 24: Expected orientations of PE and MP.....142
Figure 25: Apparent conflict between RF and FR.....146
Figure 26:In house scaffolder.....147

Index of tables

Table 1: Overview of NGS sequencing technologies.....40
Table 2: List of NGS aligners.....46
Table 3: NGS sequence assemblers.....47
Table 4: Peak finding methods.....55
Table 5: Bioinformatic resources for small RNAs.....57
Table 6: PE and RP sequencing libraries.....126
Table 7: Different parameters used for the trimming.....134
Table 8: Performances attained by different assemblers.....135
Table 9: Results from the first assembly.....137
Table 10: Comparison between SSPACE and in house scaffolder.....148
Table 11: Analysis of P. syringae resequencing data.....149

Part I

Abstract

New sequencing strategies have redefined the concept of “high-throughput sequencing” [1] [2] [3] and many companies, researchers, and recent reviews use the term “Next-Generation Sequencing” (NGS) [4] instead of high-throughput sequencing. These advances have introduced a new era in genomics and bioinformatics [5] [6].

During my years as PhD student I have developed various software, algorithms and procedures for the analysis of Nest Generation sequencing data required for distinct biological research projects and collaborations in which our research group was involved. The tools and algorithms are thus presented in their appropriate biological contexts.

Initially I dedicated myself to the development of scripts and pipelines which were used to assemble and annotate the mitochondrial genome of the model plant *Vitis vinifera*. The sequence was subsequently used as a reference to study the RNA editing of mitochondrial transcripts, using data produced by the Illumina and SOLiD platforms.

I subsequently developed a new approach and a new software package for the detection of of relatively small indels between a donor and a reference genome, using NGS paired-end (PE) data and machine learning algorithms. I was able to show that, suitable Paired End data, contrary to previous assertions, can be used to detect, with high confidence, very small indels in low complexity genomic contexts.

Finally I participated in a project aimed at the reconstruction of the genomic sequences of 2 distinct strains of the biotechnologically

relevant fungus Fusarium. In this context I performed the sequence assembly to obtain the initial contigs and devised and implemented a new scaffolding algorithm which has proved to be particularly efficient.

Premise

During the last 60 years, numerous groundbreaking discoveries and advances have revolutionized biology and life science research. The progress made is so remarkable that nowadays words such as “DNA” or “genome”, which were originally restricted to the vocabulary of a small number of highly specialized scientists have become more and more commonplace and are often used by non specialists in everyday conversation.

Biology has experienced explosive growth in these last decades and is probably the fastest growing field of science. This is exemplified by the continuous development of new and specialized branches and technologies in the discipline. As a PhD student in the last 3 years I have tried to learn the fine art of bioinformatics, in order to provide a humble contribution in the field. A synthetic summary of my scientific activities, focused on development of bioinformatics tools for three particular applications of contemporary DNA sequencing technologies, is presented in this PhD thesis.

Introduction

From DNA sequencing to Next Generation Genomics

DNA and the code-book of life

It is universally acknowledged that the final proof in 1952 [7] that DNA constitutes the genetic hereditary material and the subsequent elucidation of its three-dimensional structure in 1953 [8] constitute a major milestone in the history of life sciences and probably the foundation of modern biochemistry and molecular biology. Since then many outstanding contributions have led to a significant increase in our ability to understand and catalog the complexity and diversity of living entities, with the final aim of “cracking” the so called code of life.

While some of the most striking discoveries of the last 100 years have been made in pursuit of this objective, even now, 60 years after the resolution of the structure of DNA, we are not close to the ultimate goal, and life and its code are proving themselves to be much more complex than previously imagined.

It is evident to the contemporary cell biologist how the information to build, maintain, differentiate and replicate cells and organisms is encoded at different levels and that the “secret of life” itself resides in a series of processes and responses to stimulation which take place in an ensemble of complex and partially ordered systems residing within a cell.

In other words the flexibility and capability to adapt to different environments and situations that characterize living systems, seems to

be hard coded in their genetic material and cellular machinery, making cells - the basic structural unit of life - so astonishingly resilient and yet so fascinating and complex to study.

Apart from these philosophical considerations, it should be noted that modern molecular genetics has solid foundations, and that more than 60 years of investigation has provided an insightful and sound basic knowledge of the cellular machinery.

As mentioned previously, the foundation of molecular biology is that DNA constitutes the genetic hereditary material, or to say it in a more colorful language: that the book of life is written in an alphabet of 4 letters. Many considerations could be suggested on how each page of such book should be read or interpreted and which rules and general conclusions could be drawn from it, but it is self evident that the first requirement in order to investigate the “code of life” is to have access to it, which in biochemical terms means being able to read a DNA sequence.

DNA sequencing is surely one of the techniques that has contributed most to the acceleration of biological research and discovery in recent decades.

Knowledge of DNA sequences has become indispensable for basic biological research and in numerous applied fields, such as medicine, biotechnology, forensic biology and agricultural sciences.

The pioneering ages of DNA sequencing

The first sequences of DNA were obtained in the early 1970s, however due to the complex and laborious nature of the methodologies in use, their practical impact was relatively low and the accessibility to such technologies rather limited. The landmark year in DNA sequencing research is without any reasonable doubt 1977 when two research groups, lead by Franck Sanger at the university of Cambridge and Walter Gilbert and Allan Maxam at Harvard, enlightened the world of biochemistry by demonstrating the application of new, fast and efficient DNA sequencing techniques [9] [10].

The methods originally proposed by Sanger and Maxam-Gilbert both generate a nested set of single stranded radioactively labeled DNA fragments, which can be separated according to their size by an electrophoresis procedure on a high-resolution polyacrylamide gel. The sequence is then inferred (read) from the gel.

The main difference between these methods lies in the procedure used to generate the DNA fragments: the Sanger method is also known as “enzymatic” method as it takes advantage of an understanding of the physiological chemistry of cellular DNA synthesis and uses enzymes along with “engineered” reagents, while Maxam Gilbert method, using a series of chemical reactions, became known as the chemical method.

In the chemical method a radioactively labeled DNA strand is subjected to hazardous chemical reagents that randomly cleave DNA at one or two specific nucleotides (A ,A+G, C+T, T) in each of four reactions. For example, purines are depurinated using formic acid, the guanines (and to some extent adenines) are methylated by dimethyl sulfate (DMS) and pyrimidines are methylated using hydrazine. The addition of salt (sodium

chloride) to the hydrazyne reaction inhibits the methylation of C.
The modified DNAs are then cleaved by hot piperidine at the position of the modified base. The concentration of the modifying chemicals is controlled in order to induce on average one modification per DNA molecule. Finally the fragments in the four reaction are electrophoresed side by side in a denaturing polyacrylamide gel for size separation and the sequence read from bottom (5') to top (3') of the gel.
This method rapidly became popular after its publication, due to the fact that unlike the initial formulation of the Sanger method it didn't require any cloning step, and purified DNA could be used directly. However with the improvement of the chain terminator method (see below), Maxam-Gilbert sequencing fell out of favor due to its technical complexity prohibiting its use in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties in the scale up. Nowadays this technique has a less practical application than the Sanger approach which remains widely used. However we can imagine that in some limited cases the chemical method could still be advantageous, for example in the determination of the sequence of DNA stretches, which due to a particular sequence or secondary structure, cannot be sequenced with ease by the Sanger method.

The Sanger method utilizes a DNA polymerase alongside dideoxy nucleoside triphosphate (ddNTPs) chain terminators to synthesize a complementary copy of a single-stranded DNA template. The key principle and main advantage of the enzymatic method with respect to the Maxam-Gilbert technology resides in the use of the ddNTPs, a modified form of the standard deoxy nucleosides, which, if incorporated

into a nascent DNA chain, prohibit its elongation due to their chemical properties. The use of ddNTPs as chain terminators make the procedure developed by Sanger and co-workers, more efficient and safer than chemical-sequencing, as fewer chemicals and lower amounts of radioactivity are needed.

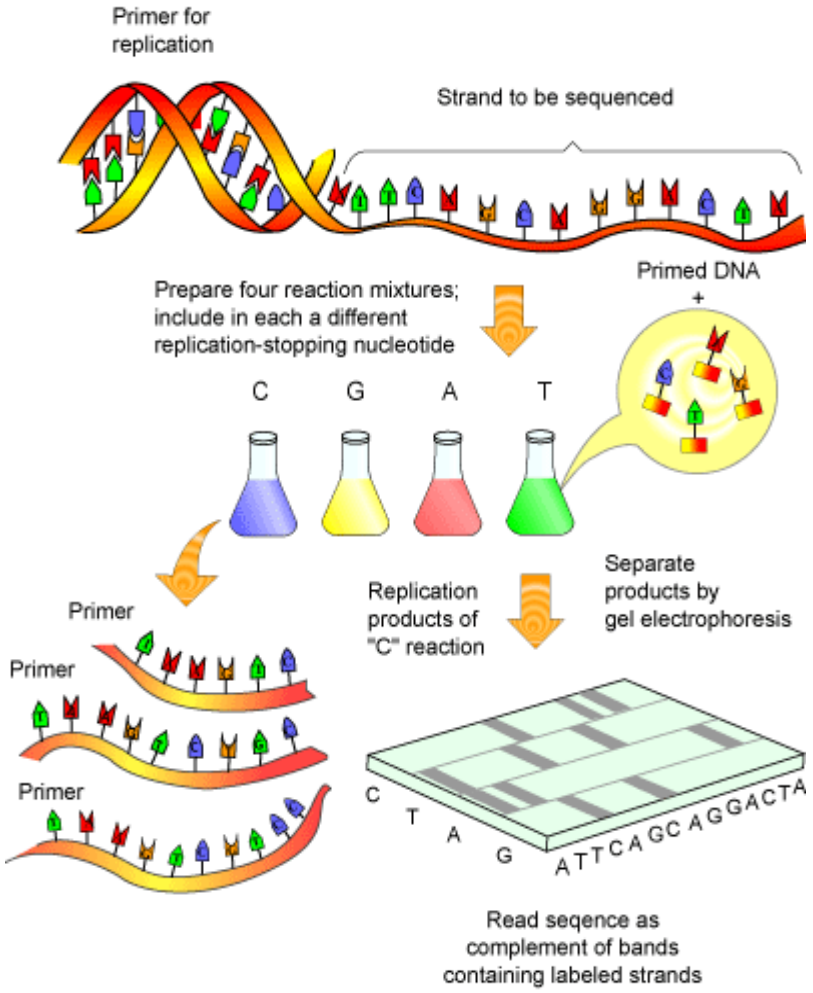
In its classical formulation the chain termination method requires a single stranded DNA template, a DNA primer, a DNA polymerase, and the four molecular species of the “normal” deoxy nucleosides as well ATP and the radioactively labeled chain terminating ddNTPs. Similarly to the Maxam-Gilbert 4 distinct biochemical reactions are prepared, each one containing a distinct chain terminator as long as all the necessary reagents and enzymes for the chain elongation.

In the reading phase the newly synthesized and labeled DNA fragments are heat denatured and separated by size with a resolution of just one nucleotide by gel electrophoresis on a denaturing polyacrylamide gel, with each one of the four reactions run in an individual lane.

Chain-termination methods have greatly simplified DNA sequencing. For example, chain-termination-based kits are commercially available that contain the reagents needed for sequencing, pre-aliquoted and ready to use. The main limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence.

Technical variations of chain-termination sequencing underly the development of several modern high throughput DNA sequencing techniques. Among these the “dye-sequencing” technology, which uses fluorescently and not radioactively labeled ddNTPs (or primers in the early formulation), is probably the most prominent [11].

Figure 1. The Sanger method



In the classical formulation of the Sanger sequencing method 4 distinct sequencing reactions are used. Each reaction contains all the necessary sequencing enzymatic apparatus and chemicals. In each reaction mixture a different replication stopping nucleotide is used in order to produce truncated DNA fragments of different length. The truncated fragments are separated by gel electrophoresis and the sequence is read directly from the gel

Sequencing and the early genomics era

Concurrently with the development and refinement of their groundbreaking sequencing technology in 1977 Sanger and co-workers also determined the first complete sequence of a genome, by applying their technology to the sequencing of the bacteriophage ϕX174 [12].

The complete sequence of ϕX was a revelation because, to the surprise of many, it turned out to be extremely interesting. Unlike the amino acid sequences of proteins, the DNA sequence of the ϕX genome could be interpreted to tell a fascinating story based upon interpretation of the sequence in terms of the genetic code. Analysis of mutations in genes identified by traditional phage genetics, combined with amino acid sequence information for protein components of the ϕX virion, allowed phage genes to be located on the DNA sequence. For the first time translation of a DNA sequence in all possible reading frames identified long open reading frames that could be assigned to genes identified by traditional genetic methods. And, most surprising, it was clear that significant portions of the genome were translated in more than one reading frame to produce two different protein products. These pairs of 'overlapping genes' had not been detected by recombination mapping of the ϕX genome but their existence was indisputable when the sequence was analyzed in light of genetics and protein sequence information.

The sequence of the simian virus SV40 followed quickly in 1978 [13]. Sequencing was rapidly completed after publication of the Maxam–Gilbert method. With the introduction of the gel-based sequencing methods, the rate of DNA sequencing accelerated. Progress in the methodology was incremental and was driven by the selection of sequencing targets of increasing complexity. The Sanger group

determined the sequence of the 16.5 kb human mitochondrial genome [14] the 48.5 kb complete phage lambda genome. Following Sanger's retirement his protégé Bart Barrell led sequencing of the 172 kb Epstein–Barr virus [15] and then the 237 kb human cytomegalovirus genome [16]. During this period the useful read length of dideoxy sequencing increased from about 100 up to about 400. Sequencing capacity was also improved by the adoption of gels with narrower lanes. Owing to these technical improvements in the middle 80s a single person could run 8 gels in a day and obtain some 30 Kb of primary sequence data, but this could be hardly done more than twice a week.

The early steps of bioinformatics

Beginning with the genomic sequence of the ϕX phage, the management and analysis of sequencing data became a major problem. The original ϕX sequencing data were stored in the notebooks of nine different workers each concerned with particular portions of the molecule. Michael Smith, on sabbatical in the Sanger group, had a brother-in-law named Duncan McCallum who was a business computer programmer in Cambridge. He wrote the first programs to help with the compilation and analysis of DNA sequence data (in COBOL). The manually deduced sequences, translated by each researcher in paper form, were entered in blocks of 60 on punched cards. The programs then

- compiled and numbered the complete sequence,
- allowed the editing of a previously compiled sequence,

- searched the sequence for specific short sequences or families of sequences, for example restriction sites and
- translated the sequence in all reading frames.

Though invaluable, the programs did not produce output suitable for publication, so the original figure displaying the ϕX sequence with its genes and their translation products annotated was hand typed by Peggy Dowding. Roger Staden, who had helped with computer analysis of the original ϕX sequence, wrote the first suite of interactive bioinformatics programs 'designed specifically for use by people with little or no computer experience'. These programs developed into the Staden Package, still in use today [17] [18].

Subsequently, with the proliferation of DNA sequence data, came the need for a DNA sequence database. Margaret Dayhoff was an early pioneer in this area. She had previously established a protein sequence database and published the first collection of nucleotide sequence information in 1981. Shortly thereafter, in 1982 GenBank was created by the NIH to provide a 'timely, centralized, accessible repository for genetic sequences' [19] Concomitant with continuous improvements in the yield and speed of sequencing technologies, these biological databases started to grew in size and number, to the point that comparing, retrieving and aligning the sequences soon became a rate limiting step. Again computer scientists assisted with the development of rapid search programs like FASTA and BLAST [20] [21].

By the early 1990s, the world wide web and affordable personal computers were becoming available and all the foundations for the "genomic revolution" were in place.

First generation HT DNA sequencing: Automation, Parallelization and Miniaturization

While high-throughput DNA sequencing is nothing more than an ensemble of highly efficient and automated DNA sequencing procedures, the precise quantification of what exactly is considered to be ”high” throughput or “high” efficiency is constantly changing with the progress of sequencing technologies (indeed what was considered astonishing less than a decade ago is now routine). As in other human activities the major factors contributing to the ability to increase the yield and speed are improvements in engineering and automation.

The idea of automating the sequencing process through the use of dedicated machines, traces back to the early 80s. Initially these machines did not automate much of the process apart from the reading and the base calling. Gels were still prepared and loaded manually.

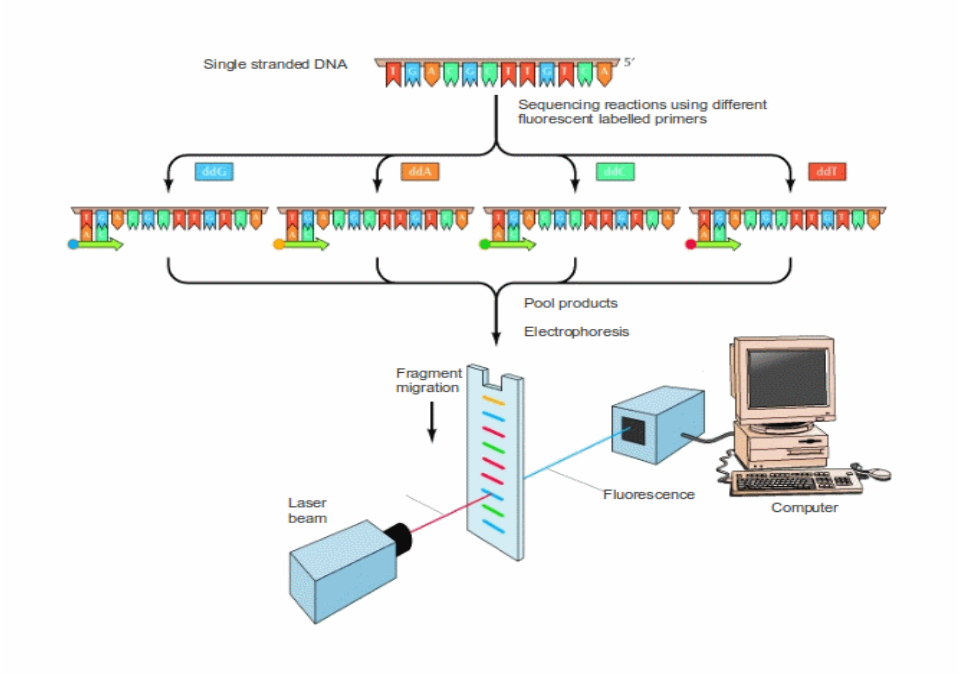
The first report of automation of DNA sequencing dates to 1986. In the laboratory of Leroy Hood at Caltech, in collaboration with Applied Biosystems (ABI), a completely automated sequencing process was set up, in order to demonstrate for the first time, how it was possible to record sequencing data directly to a computer without auto-radiography of the sequencing gel [22].

One prominent step towards the development of automatic sequencing machines was the development of the dye-sequencing technology. Dye sequencing is a substantial improvement on the original Sanger method, whose main innovation consists of the use chain terminators ddNTPs fluorescently labeled by the means of fluorescent dyes, emitting at different wavelengths (colors) [11].

The sequencing thus takes place in a single reaction where the fluorescent ddNTPs are used along with a modified (engineered) DNA-polymerase. Fluorescent DNA fragments of different lengths and “color” are generated and then separated by size in a reaction monitored by a charge-coupled-device (CCD) fluorescence detector. The final output consists of a time “trace” data or chromatogram of the fluorescent peaks that pass the CCD point. The order of the individual bases is determined by a careful examination of the fluorescence peaks, which is typically recorded into a computer.

Different protocols and variants for dye sequencing have been developed, which adopt a slightly modified principles (fluorescent primers), and take advantage of technical advances in the fields of biochemistry and molecular biology (continuous improvements in speed and efficiency in the techniques used to separate fragments).

Figure 2: Automated DNA sequencing



The first automated DNA sequencers took advantage of the development of the dye sequencing technology. The sequencing reaction generates DNA fragments of different lengths emitting light at different colors. The fragments are separated by electrophoresis in a reaction monitored by a charge-coupled-device (CCD) fluorescence detector. The time trace of the fluorescence peaks (chromotagram) is then recorded by a computer. Finally base calling is performed either by hand or by the means of computer programs from the chromatogram.

Automated DNA Sequencing and Bioinformatics

The development of bioinformatics is inevitably linked to the advent of the “sequencing era” in biology, we have already discussed how the first bioinformatics applications were developed to store maintain and analyze biological data. However it is with the development of high capacity/efficiency sequencing that the application of information theory

principles and procedures to biological data moved from desirable and useful to practical and almost necessary.

With the refinement and development of more and more sophisticated sequencing machines it quickly became evident that sequence data could be produced at a significantly higher rate than it could be processed thereby creating a bottleneck. The first DNA sequencing machines did not automate much of the sequencing process apart form the collection of data: gel preparation, reaction set up and sequence reading were still performed by hand. Sequence reading, or base calling, in particular turned out to be the most limiting step, especially in the presence of ambiguous or inconsistent trace data. The obvious solution to the problem was the development of automated base calling procedures. The first computer programs capable of transforming a fluorescence “trace” into a DNA sequence were developed within the ABI facilities in the late 80s [23], but the performance and accuracy of these early algorithms were rather unsatisfactory and their utility rather limited. Indeed it was common opinion at that time that completely automated sequencing could hardly compete with dedicated (graduate student) base callers. It was only in 1998, with the development and assessment of the Phred [24] base caller program (conceived in the early 90s by Phil Green, and developed by Brent Erwing and co-workers) that automated base calling became indisputably the de-facto standard for any sequencing project. Phred proved how automated base calling could be not only as accurate as its manual counterpart but also more consistent given the adoption of a quality scoring system [24][25][26].

A quality score is a numerical value which reflects the probability of a

base being wrong, and therefore its quality. The score is calculated through the means of complex equations from the position, height, width and intensities of the peaks on the chromatogram. Phred is acknowledged to be more accurate than other base callers (40-50% less errors) and can confidently, rapidly and consistently evaluate its calls (based on the relative strength of different base signals at a given position in a sequence and also incorporating empirically derived observations about sequence contexts that can promote erroneous calls), something that is not attainable by manual base calling. Following its great success Phred (or Phed-like) quality scores are now routinely used as standard measure to characterize the quality of DNA sequences, and to compare the efficacy of different sequencing methods.

Even if Phred adopts a complex mathematical formalism for calculating the error probability for each base, the quality scores themselves are calculated by the means of a simple mathematical formula:

$$Qs = -10\log\left(P_{error}\right)$$

Therefore a quality score of 40 corresponds by example to a wrong call every 10,000 bases , a score of 30 to 1 every 1,000 and so on.

The genomic era

Until 1995, the only completely sequenced DNA molecules were viral and organellar genomes. That year Craig Venter's group at TIGR, and their collaborators, reported the complete genome sequences of two bacterial species, *Haemophilus influenzae* [27] and *Mycoplasma*

genitalium [28]. The *H. influenzae* sequence gave the first glimpse of the complete instruction set for a living organism. The *M. genitalium* sequence provided an approximation of the minimal set of genes required for cellular life.

The methods used to obtain these sequences were as important for subsequent progress as the biological insights they provided. Sequencing of *H. influenzae* introduced the whole genome shotgun (WGS) method for sequencing cellular genomes. In this method, genomic DNA is fragmented randomly and cloned to produce a genomic DNA library in that can be propagated in *E. coli*. Clones are sequenced at random and the results are assembled to produce the complete genome sequence by a computer program that compares all of the sequence reads and aligns matching sequences. Sanger and colleagues used this general strategy to sequence the lambda phage genome (48.5 kb), published in 1982. However, no larger genome was shotgun sequenced until *H. influenzae* (1.83 Mb). Venter and colleagues introduced critical improvements that made it feasible, for the first time, to shotgun sequence complete cellular genomes. Among the others the most relevant was the adoption of the 'paired ends' strategy [29].

The sequencing procedure used in the H. influenzae project used melted double-stranded DNA as template. With double-stranded templates it was convenient to sequence each clone from both ends. Because the randomly sheared DNA was carefully size selected before cloning, the distance between the reads from the ends of each clone could be estimated with a certain degree of confidence. The assembly

program used this information to construct ‘scaffolds’ from the blocks of completely overlapped sequence (‘contigs’). When two contigs contained sequences from opposite ends of a single clone, then the two contigs could be linked, although a ‘sequence gap’ was said to exist between them. Sequence gaps remaining at the end of the shotgun phase of sequencing could be closed by sequencing from a specific primer for a site internal to a clone bridging the gap. Gaps between scaffolds are ‘physical gaps’ that contain sequences, which do not occur within any of the sequenced clones. Other measures, such as PCR between the ends of scaffolds using a genomic DNA template, were used to close physical gaps.

Another critical factor in the application of shotgun sequencing to cellular genomes was the TIGR assembler [30] and the rapid increase in amounts of RAM available in computers. Previous assembly programs were not designed to handle thousands of sequence reads involved in even the smallest cellular genome projects.

Once these initial sequences were reported the floodgates were open and a steady stream of completed genome sequences has been appearing ever since. Here it is possible only to touch on a few of the most significant landmarks.

Eventual sequencing of the human genome had become an imaginable goal at the outset of the sequencing era 30 years ago. Formal discussions of the idea began in 1985 when Robert Sinsheimer organized a meeting on human genome sequencing at the University of California, Santa Cruz [31]. That same year Charles DeLisi and David A. Smith commissioned the first Santa Fe conference, funded by the

DOE, to study the feasibility of a Human Genome Initiative. Discussions continued and in 1988 reports recommending a concerted genome research program were issued by committees of the congressional Office of Technology Assessment and the National Research Council. In 1990 the DOE and NIH presented a joint 5-year US Human Genome Project plan to Congress. It was estimated that the project would take 15 years and cost ~3 billion US$.

The US Human Genome Project established goals of mapping, and in some cases sequencing, several model organisms as well as humans. These included *E. coli*, yeast (*S. cerevisiae*), the worm (*C. elegans*), drosophila (*D. melanogaster*) and mouse (laboratory strains of *Mus musculus domesticus*). Several of the projects mentioned above received funding from the Human Genome Project. The publicly funded effort became an international collaboration between a number of sequencing centers in the United States, Europe and Japan. Each center focused sequencing efforts on particular regions of the genome, necessitating detailed mapping as a first step. Indeed the first strategy adopted in the sequencing of the human genome and in satellite projects was based on the generation of a physical map in order to drive the assembly. A physical map of a genome is a map determining where a given DNA marker is physically located on the DNA of a chromosome. In the context of the human genome the so called BAC by BAC mapping approach was initially adopted to perform the assembly. In this approach a BAC library of the human genome (more than 20000 BAC clones) is constructed. Then sequence tagged sites (STS), that is short sequence unique in the genome under study are determined by sequencing portions of the BAC clones. Once such unique sequences

are determined, a southern blot hybridization is used to map physically the STS on a chromosome. BACs are then fingerprinted by the use of restriction enzymes: the peculiar cutting pattern of the BACS are characterize by gel electrophoresis and each BAC is associated to its presumable cognates showing a similar pattern. Finally this clusters of fingerprinted BAC, anchored to the genome by the means of STSs are assembled by the use of a computer program. As in contrast to the WGS method described above this procedure is more laborious however it was common opinion in the scientific community that given the size and complexity of the human genome the adoption of the WGS protocol would have lead to poor results.

Because of its position as the pre-eminent model organism of molecular biology, sequencing of the genome of *E. Coli* (4.6 Mb) had been proposed by Blattner as early as 1983. Sequencing was started with manual methods and finished in 1997 with automated sequencers. Early sequences covering ~1.9 Mb, were deposited starting in 1992, and were obtained from an overlapping set of cosmid clones. The final ~2.5 Mb was obtained by shotgun sequencing of ~250 Kb I-Sce I fragments [32]. This *E. coli* genome sequence, along with several other strains sequenced subsequently has yielded a wealth of information about bacterial evolution and pathogenicity [33].

Meanwhile, another model for large-scale genome sequencing projects had emerged: the international consortium. The first genome sequence to be completed by this approach was the yeast *S. cerevisiae*(12.0 Mb),

in late 1996 [34]. This was the also the first eukaryotic organism to be sequenced. The project involved about 600 scientists in Europe, North America and Japan. The participants included both academic laboratories and large sequencing centers .

The first animal genome sequenced was that of 'the worm' *C. elegans* (97 Mb), in 1998 [35]. The authorship of this work was simply 'The *C. elegans* Sequencing Consortium', which was a collaboration between the Washington University Genome Sequencing Center in the United States and the Sanger Centre in UK.

In 1996, ABI introduced the first commercial DNA sequencer that used capillary electrophoresis rather than a slab gel (the ABI Prism 310), and in 1998 the ABI Prism 3700 with 96 capillaries was announced. For the first time DNA sequencing was truly automated. The considerable labor of pouring slab gels was replaced with automated reloading of the capillaries with polymer matrix. Samples for electrophoresis were automatically loaded from 96-well plates rather than manually loaded as the previous generation of sequencers had been. in May 1998 Celera Genomics was found by Craig Venter and the Applera Corporation (the parent company of ABI) to exploit these new machines by applying Venter's methods for WGS sequencing to the human genome, in direct competition with the publicly funded Human Genome Project.

Celera chose the *D. melanogaster* genome to test the applicability of the WGS approach to a complex eukaryotic genome [36]. This involved a scientific collaboration between the scientists at Celera and those of the

Berkeley and European *Drosophila* Genome Projects. These projects finished 29 Mb of the 120 Mb of euchromatic portion of the genome. (About one-third of the 180 Mb *Drosophila* genome is centromeric heterochromatin). Using the WGS approach, data was collected over a 4-month period that provided more than 12× coverage of the euchromatic portion of the genome. The results validated the data produced by the ABI 3700s, the applicability of the WGS approach to eukaryotic genomes, and the assembly methods developed at Celera [37]. This was a nearly ideal test case because the WGS data could be analyzed separately and then portions of it could be compared with finished sequence already produced by the *Drosophila* Genome Projects. At the same time the sequence information provided a valuable resource for *Drosophila* genetics. More than 40 scientists at an 'Annotation Jamboree' made an initial annotation of the sequence. These scientists, mainly drawn from the *Drosophila* research community, met at Celera for a 2-week period to identify genes, predict functions, and begin a global synthesis of the genome sequence information.

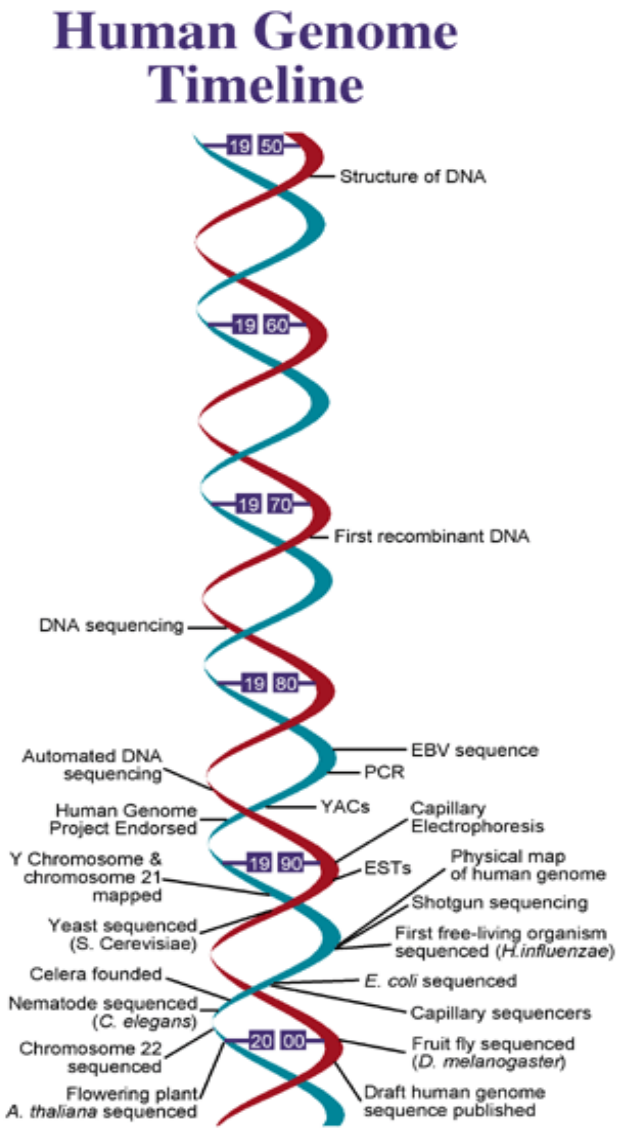
In 1998 the public Human genome sequencing project, now in a race with Celera, also adopted the new ABI Prism 3700 capillary sequencers. In 1999 the Human Genome Project celebrated passing the billion base-pair mark, and the first complete sequence of a human chromosome was reported (chromosome 22 [38]).

Meanwhile at Celera, human genome sequencing was underway using the WGS strategy. Human genome sequencing began in September 1999 and continued until June 2000, when data collection was completed and an initial assembly was achieved. The Celera data

provided approximately 5-fold coverage of the genome. An additional 3-fold coverage of unordered and un-oriented BAC sequences from the public effort was included in the assembly. The power of the WGS strategy was amply demonstrated. This first available rough draft assembly of the genome was completed by the Genome Bioinformatics Group at the University of California, Santa Cruz, primarily led by then graduate student Jim Kent.

On 25 June 2000 at the White House, President Clinton with British Prime Minister Tony Blair publicly announced draft versions of the human genome sequence from both the publicly funded project and from Celera. In February 2001 the Celera [39] and the public [40] draft human genome sequences were published the same week in Science and Nature. The race was officially a tie, but it was clear to all that the entry of Celera had speeded the process by several years. Both projects ended up needing the other to make the progress that was made. The Celera assembly benefited from data produced in the public project and the public project quickly adopted some of Celera's methods, in particular the paired-end strategy. Celera's basic methods have now been adopted by all publicly funded genome projects.
Ongoing sequencing led to the announcement of the essentially complete genome in April 2003, 2 years earlier than planned

Figure 3: Timeline of the human genome sequencing project



The most relevant events and achievements which lead to the sequencing of of the human genome are shown in chronological order

Next Generation DNA sequencing technologies

Using current Sanger sequencing technology, it is technically possible for up to 384 [41] sequences of between 600 and 1,000 nt in length to be sequenced in parallel [42] . However, these 384-capillary systems are rare. The more standard 96-capillary instruments yield a maximum of approximately 6 Mb of DNA sequence per day, with costs for consumables amounting to about $500 per 1 Mb.

Over the last decade, alternative sequencing strategies have become available which force us to completely redefine “high-throughput sequencing” [1] [2] [3]. These technologies outperform the older Sanger-sequencing technologies by a factor of 100–1,000 in daily throughput, and at the same time reduce the cost of sequencing one million nucleotides (1 Mb) to less than 1/40th of that associated with Sanger sequencing. To reflect these huge changes, many companies, researchers, and recent reviews use the term “Next-Generation Sequencing” (NGS) [4] instead of high-throughput sequencing, yet this term itself may soon be outdated considering the speed of ongoing developments.

The development of these new massively parallel sequencing technologies has sprung from recent advances in the field of nanotechnology, from the availability of optical instruments capable of reliably detecting and differentiating millions of sources of light or fluorescence on the surface of a small glass slide and from the ingenious application of classic molecular biology principles to the sequencing problem.

Currently available next-generation sequencers rely on a variety of

different chemistries to generate data and produce reads of differing lengths, but all are massively parallel in nature and present new challenges in terms of bioinformatics support required to maximize their experimental potential. Independently from the underlying sequencing chemistry for the evaluation of the quality of the data each method adopts Phred like quality scores.

Three distinct NGS platforms have already attained wide diffusion and availability:

- Roche/454 GS FLX Titanium sequencer
- Illumina Genome Analyzer II/IIx
- Applied Biosystems SOLiD.

Roche 454

The 454 sequencing platform was the first of the new high-throughput sequencing platforms on the market (released in October 2005). It is based on the pyrosequencing approach developed by Pål Nyrén and Mostafa Ronaghi at the Royal Institute of Technology, Stockholm in 1996 [43]. In contrast to the Sanger technology, pyrosequencing is based on iteratively complementing single strands and simultaneously reading out the signal emitted from the nucleotide being incorporated (also called sequencing by synthesis, sequencing during extension).
Electrophoresis is therefore no longer required to generate an ordered read out of the nucleotides, as the sequence data is now compiled simultaneously with the extension phase.

In the pyrosequencing process, one nucleotide at a time is washed over several copies of the sequence to be determined, allowing the polymerase to incorporate the nucleotide if it is complementary to the template strand. The incorporation stops if the longest possible stretch of complementary nucleotides has been synthesized by the polymerase. In the process of incorporation, one pyrophosphate per nucleotide is released and converted to ATP by an ATP sulfurylase. The ATP drives the light reaction of luciferase present in the reaction site and the emitted light signal is measured, allowing estimation of the number of consecutive identical bases present in the template.

In 2005, pyrosequencing technology was parallelized on a picotiter plate by 454 Life Sciences (later bought by Roche Diagnostics) to allow high-throughput sequencing [2]. The sequencing plate has about two million wells – each of them able to accommodate exactly one 28-µm diameter bead covered with single-stranded copies of the sequence to be determined. The beads are incubated with a polymerase and single-strand binding proteins and, together with smaller beads carrying the ATP sulfurylases and luciferases, gravitationally deposited in the wells. Free nucleotides are then washed over the flow cell and the light emitted during the incorporation is captured for all wells in parallel using a high-resolution charge-coupled device (CCD) camera, exploiting the light-transporting features of the plate used.

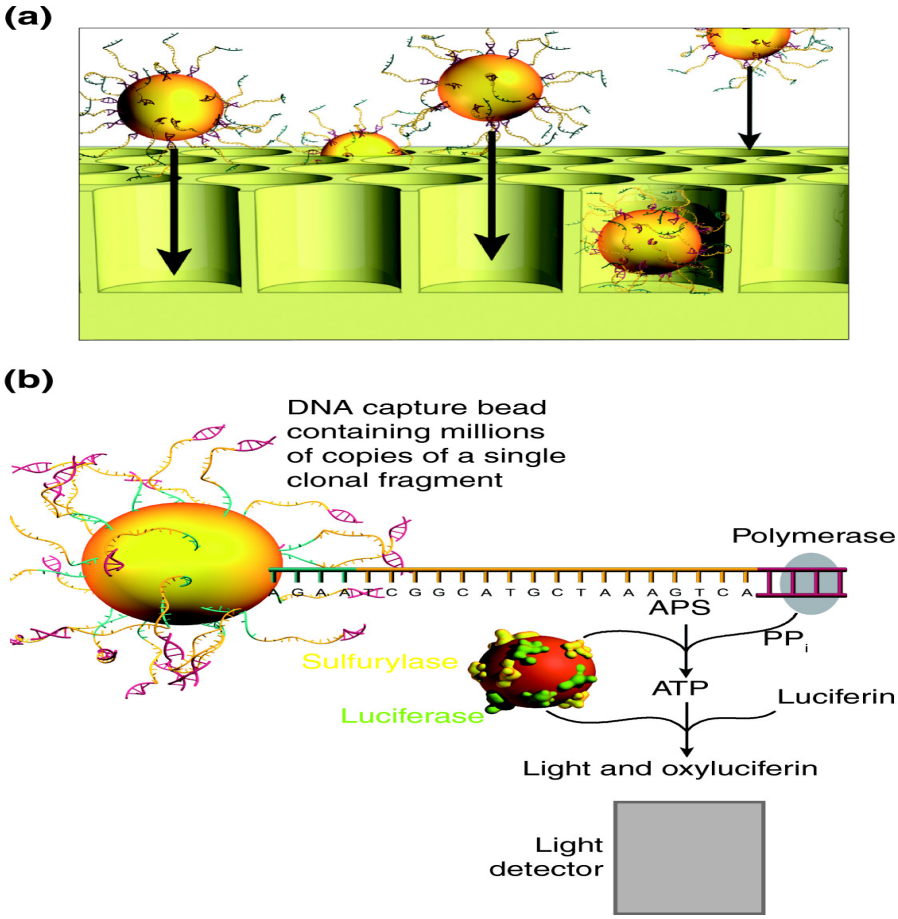
One of the main prerequisites for applying this array-based pyrosequencing approach is covering individual beads with multiple copies of the same molecule. This is done by first creating sequencing libraries in which every individual molecule gets two different adapter

sequences, one at the 5′ end and one at the 3′ end of the molecule. In the case of the 454/Roche sequencing library preparation [2], this is done by sequential ligation of two pre-synthesized oligos. One of the adapters added is complementary to oligonucleotides on the sequencing beads and thus allows molecules to be bound to the beads by hybridization. Low molecule-to-bead ratios and amplification from the hybridized double-stranded sequence on the beads (kept separate using emulsion PCR) makes it possible to grow beads with thousands of copies of a single starting molecule. Using the second adapter, beads covered with molecules can be separated from empty beads (using special capture beads with oligonucleotides complementary to the second adapter) and are then used in the sequencing reaction as described above. The average substitution (excluding insertion/deletion, InDel) error rate is in the range of 10^{-3}–10^{-4} [2] [44] which is higher than the rates observed for Sanger sequencing, but is the lowest average substitution error rate of the new sequencing technologies discussed here. In bead preparation (i.e., emulsion PCR) a fraction of the beads end up carrying copies of multiple different sequences. These “mixed beads” will participate in a high number of incorporations per flow cycle, resulting in sequencing reads that do not reflect real molecules. Most of these reads are automatically filtered during the software post-processing of the data. The filtering of mixed beads may, however, cause a depletion of real sequences with a high fraction of incorporations per flow cycle. A large fraction of the errors observed for this instrument are small indels, mostly arising from inaccurate calling of homopolymer length, and single base-pair deletions or insertions caused by signal-to-noise thresholding issues [44]. Most of these

problems can be resolved by higher coverage. For long (>10 nt) homopolymers, however, there is often a consistent length miscall that is not resolvable by coverage [44] [45][46].

Strong light signals in one well of the picotiter plate may also result in insertions in sequences in neighboring wells. If the neighboring well is empty, this can generate so-called ghost wells, i.e., wells for which a signal is recorded even though they contain no sequence template; hence, the intensities measured are completely caused by bleed-over signal from the neighboring wells. Computational post-processing can often correct these artifacts [47]. As with Sanger sequencing, the error rate increases with the position in the sequence. In the case of 454 sequencing, this is caused by a reduction in enzyme efficiency or loss of enzymes (resulting in a reduction of the signal intensities), some molecules no longer being elongated and by an increasing phasing effect. Phasing is observed when a population of DNA molecules amplified from the same starting molecule (ensemble) is sequenced, and describes the process whereby not all molecules in the ensemble are extended in every cycle. This causes the molecules in the ensemble to lose synchrony/phase, and results in an echo of the preceding cycles to be added to the signal as noise.

Figure 4 454 Pyrosequencing



In the pyrosequencing process, one nucleotide at a time is washed over several copies of the sequence to be determined, allowing the polymerase to incorporate the nucleotide if it is complementary to the template strand. The incorporation stops if the longest possible stretch of complementary nucleotides has been synthesized by the polymerase. In the process of incorporation, one pyrophosphate per nucleotide is released and converted to ATP by an ATP sulfurylase. The ATP drives the light reaction of luciferase present in the reaction site and the emitted light signal is measured, allowing estimation of the number of consecutive identical bases present in the template

Illumina Genome Analyzer

The reversible terminator technology used by the Illumina Genome Analyzer (GA) employs a sequencing-by-synthesis concept that is similar to that used in Sanger sequencing, i.e. the incorporation reaction is stopped after each base, the label of the base incorporated is read out with fluorescent dyes, and the sequencing reaction is then continued with the incorporation of the next base [1] [48].

Like 454/Roche, the Illumina sequencing protocol requires that the sequences to be determined are converted into a special sequencing library, which allows them to be amplified and immobilized for sequencing [47]. For this purpose two different adapters are added to the 5′ and 3′ ends of all molecules using ligation of so-called forked adapters. The library is then amplified using longer primer sequences, which extend and further diversify the adapters to create the final sequence needed in subsequent steps.

This double-stranded library is melted using sodium hydroxide to obtain single-stranded DNAs, which are then pumped at a very low concentration through the channels of a flow cell. This flow cell has on its surface two populations of immobilized oligonucleotides complementary to the two different single-stranded adapter ends of the sequencing library. These oligonucleotides hybridize to the single-stranded library molecules. By reverse strand synthesis starting from the hybridized (double-stranded) part, the new strand being created is covalently bound to the flow cell [1] [49].

If this new strand bends over and attaches to another oligonucleotide complementary to the second adapter sequence on the free end of the

strand, it can be used to synthesize a second covalently bound reverse strand. This process of bending and reverse strand synthesis, called bridge amplification, is repeated several times and creates clusters of several thousand copies of the original sequence in very close proximity to each other on the flow cell.

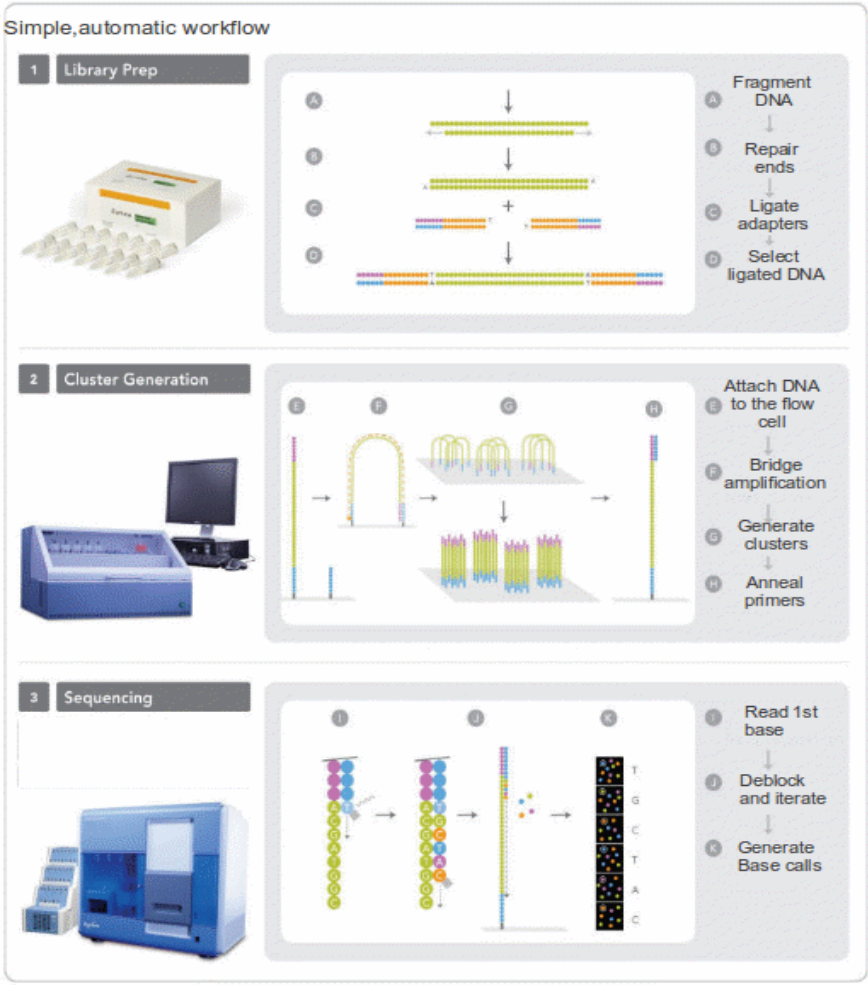
These randomly distributed clusters contain molecules that represent the forward as well as reverse strands of the original sequences. Before determining the sequence, one of the strands has to be removed to prevent it from hindering the extension reaction sterically or by complementary base pairing. Strands are selectively cleaved at base modifications of oligonucleotides on the flow cell. Following strand removal, each cluster on the flow cell consists of single stranded, identically oriented copies of the same sequence; which can be sequenced by hybridizing the sequencing primer onto the adapter sequences and starting the reversible terminator chemistry.

“Solexa sequencing”, as it was introduced in early 2007, initially allowed for the simultaneous sequencing of several million very short sequences (at most 26 nt) in a single experiment. In recent years there have been several technical, chemical, and software updates. The current instrument, known as the Illumina Genome Analyzer II, has increased flow cell cluster densities (around 120 million clusters per lane with 2 flowcells, each made up of 8 lanes run simultaneously), a wider range of the flow cell is imaged, and sequence reads of up to 150 nt can be generated. A technical update also enabled the sequencing of the reverse strand of each molecule. This is achieved by chemical melting and washing away the synthesized sequence, repeating a few bridge amplification cycles for reverse strand synthesis, and then selectively

removing the starting strand (again using base modifications of the flow cell oligonucleotide populations), before annealing another sequencing primer for the second read. Using this “paired-end sequencing” approach, approximately twice the amount of data can be generated. The Illumina library and flow cell preparation includes several in vitro amplification steps, which cause a high background error rate and contribute to the average error rate of about 10^{-2}–10^{-3} [50][51]. Further, the flow cell preparation creates a fraction of ordinary-looking clusters that are initiated from more than one individual sequence. These results in mixed signals and mostly low quality sequences for these clusters. Similar to the 454 ghost wells, the Illumina image analysis may identify chemistry crystals, dust, and lint particles as clusters and call sequences from these. In such cases the resulting sequences typically appear to be of low sequence complexity. As is the case for the other platforms, the error rate increases with increasing position in the determined sequence. This is mainly due to phasing, which increases the background noise as sequencing progresses. While the ensemble sequencing process for pyrosequencing creates uni-directional phasing, reversible terminator sequencing creates bi-directional phasing [50] [52] as some incorporated nucleotides may also fail to be correctly terminated – allowing the extension of the sequence by another nucleotide in the same cycle. With increasing cycle numbers, the intensities extracted from the clusters also decline [50] [53] [52]. This is due to fewer molecules participating in the extension reaction as a result of non-reversible termination, or due to dimming effects of the sequencing fluorophores. In early versions of the chemistry, one of the fluorophores could become stuck to the clusters creating another source

of increased background noise [50]. The simultaneous identification of four different nucleotides is also an issue. The GA uses four fluorescent dyes to distinguish the four nucleotides A, C, G, and T. Of these, two pairs (A/C and G/T) excited using the same laser, are similar in their emission spectra and show only limited separation using optical filters. Therefore, the highest substitution errors observed are between A/C and G/T [50] [51].

Figure 5. Illumina Sequencing



Purified DNA is sonicated and fragmented. Different adapters are added to the 5′ and 3′ ends of all molecules using ligation . The library is then amplified using longer primer sequences, by the means of “Bridge amplification” which creates clusters of several thousand copies of the original sequence in very close proximity to each other on the flow cell. These randomly distributed clusters contain molecules that represent the forward as well as reverse strands of the original sequences. Before determining the sequence, one of the strands has to be removed to prevent it from hindering the extension reaction sterically or by complementary base pairing. Strands are selectively cleaved at base modifications of oligonucleotides on the flow cell. Following strand removal, each cluster on the flow cell consists of single stranded, identically oriented copies of the same sequence; which can be sequenced by hybridizing the sequencing primer onto the adapter sequences and starting the reversible terminator chemistry.

Applied Biosystems SOLiD

The prototype of what was further developed and later sold by Life Technologies/Applied Biosystems (ABI) as the SOLiD sequencing platform, was developed by Harvard Medical School and the Howard Hughes Medical Institute and published in 2005 [3]. Distinct from its competitors the SOLiD technology doesn't rely on a sequencing by synthesis approach, their technology is based on a sequence by ligation strategy.

The principle behind sequencing-by-ligation is very different from the approaches discussed thus far. The sequence extension reaction is not carried out by polymerases but rather by ligases. In the sequencing-by-ligation process, a sequencing primer is hybridized to single-stranded copies of the library molecules to be sequenced. A mixture of 8-mer probes carrying four distinct fluorescent labels compete for ligation to the sequencing primer. The fluorophore encoding, which is based on the two 3′-most nucleotides of the probe, is read. Three bases including the dye are cleaved from the 5′ end of the probe, leaving a free 5′ phosphate on the extended (by five nucleotides) primer, which is then available for further ligation. After multiple ligations (typically up to 10 cycles), the synthesized strands are melted and the ligation product is washed away before a new sequencing primer (shifted by one nucleotide) is annealed. Starting from the new sequencing primer the ligation reaction is repeated. The same process is followed for three other primers, facilitating the read out of the dinucleotide encoding for each start position in the sequence. Using specific fluorescent label encoding, the dye read outs (i.e. colors) can be converted to a sequence

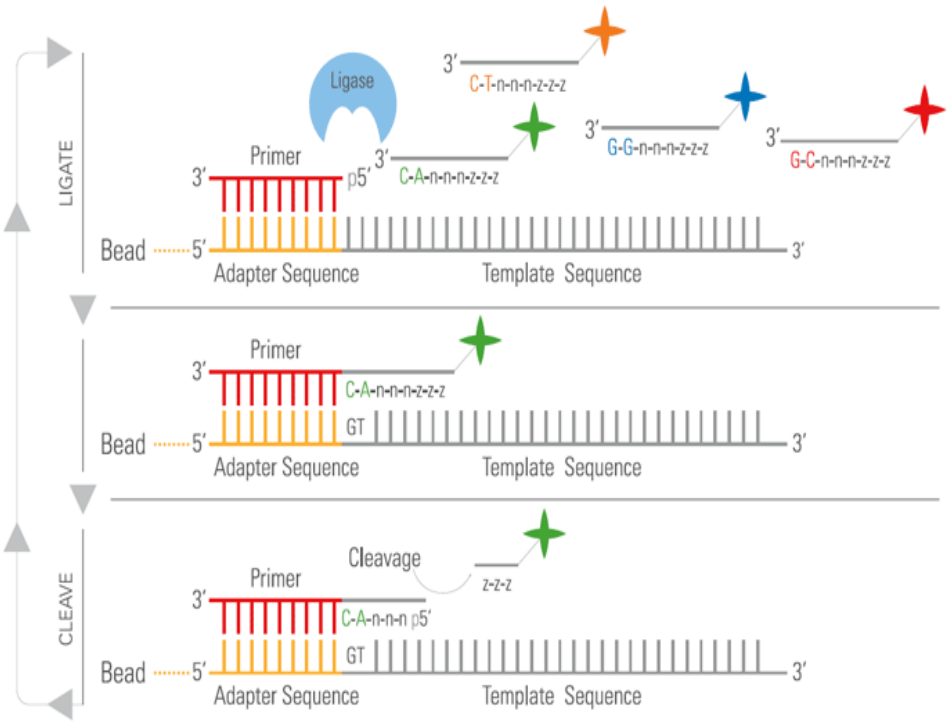
[54]. This conversion from color space to sequence requires a known first base, which is the last base of the used library adapter sequence. Given a reference sequence, this encoding system allows detection of machine errors and the application of an error correction to reduce the average error rate. In the absence of a reference sequence, however, color conversion fails with an error in the dye read out and causes the sequence downstream of the error to be incorrect. Despite this limitation, the fact that each base effectively read twice allows efficient error detection and the theoretical accuracy of sequence reading (99.47%) is higher than that attained by the Illumina technology.

For parallelization, the sequencing process uses beads covered with multiple copies of the sequence to be determined. These beads are created in a similar fashion to that described earlier for the 454/Roche platform. In contrast to the 454/Roche technology, the SOLiD system does not use a picotiter plate for fixation of the beads in the sequencing process; instead the 3′ ends of the sequences on the beads are modified in a way that allows them to be covalently bound to a glass slide. As for the Illumina GA system, this creates a random dispersion of the beads in the sequencing chamber and allows for higher loading densities. However, random dispersion complicates the identification of bead positions from images, and results in the possibility that chemical crystals, dust, and lint particles can be misidentified as clusters. Further, dispersal of the beads results in a wide range of inter-bead distances, which then have different susceptibility to be influenced by signals from neighboring beads.

Types and causes of sequence errors are diverse: first, the in vitro

amplification steps cause a higher background error rate. Secondly, beads carrying a mixture of sequences and beads in close proximity to one another create false reads and low quality bases. Further, signal decline, a small regular phasing effect, and incomplete dye removal result in increasing error as the ligation cycles progress [55]. Phasing, as described earlier, is a minor issue on this platform as sequences not extended in the last cycle are non-reversibly terminated using phosphatases. Since hybridization is a stochastic process, this causes a considerable reduction in the number of molecules participating in subsequent ligation reactions, and therefore substantial signal decline. On the other hand, given the efficiency of phosphatases the remaining phasing effect can be considered very low. However, incomplete cleavage of the dyes may allow cleavage in the next ligation reaction, which then allows for the extension in the next but one cycle. This causes a different phasing effect and additional noise from the previous cycle's dyes in the dye identification process.

Figure 6: Sequencing by ligation



In the sequencing-by-ligation process, a sequencing primer is hybridized to single-stranded copies of the library molecules to be sequenced. A mixture of 8-mer probes carrying four distinct fluorescent labels compete for ligation to the sequencing primer. The fluorophore encoding, which is based on the two 3′-most nucleotides of the probe, is read. Three bases including the dye are cleaved from the 5′ end of the probe, leaving a free 5′ phosphate on the extended (by five nucleotides) primer, which is then available for further ligation. After multiple ligations (typically up to 10 cycles), the synthesized strands are melted and the ligation product is washed away before a new sequencing primer (shifted by one nucleotide) is annealed. Starting from the new sequencing primer the ligation reaction is repeated. The same process is followed for three other primers, facilitating the read out of the dinucleotide encoding for each start position in the sequence. Using specific fluorescent label encoding, the dye read outs (i.e. colors) can be converted to a sequence

Next generation sequencing costs

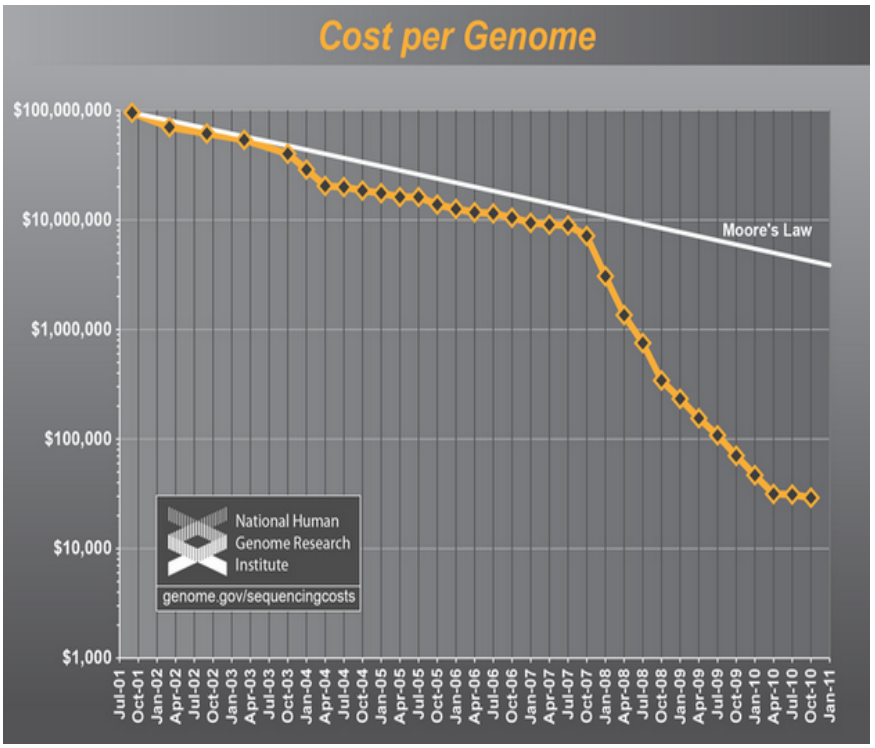
In an effort to illustrate the true cost of complete genome sequencing, the National Human Genome Research Institute (NHGRI) has compiled data from their sequencing centers to appropriately estimate the overall costs of sequencing a human genome [56]. Their calculations take into account labor, three-year amortization of sequencing instruments, data processing, and sample preparation. Figure 7 illustrates the cost associated with sequencing a human-sized haploid genome (3,000 Mb) over time since the initial draft of the human genome was published in 2001. The dramatic drop in cost seen in 2008 is the result of transitioning from first-generation Sanger sequencing to second-generation platforms installed in sequencing centers (i.e., 454, Illumina, and SOLiD). The second-generation technologies yield lower contiguous read lengths and require greater genome coverage for assembly; however, their high throughput reduces consumable costs and the number of sequencing runs. Detailed statistics upon the read-length and sequences capabilities of each method are found in Table 1

Table 1: Overview of NGS sequencing technologies

	Read-length	Sequence per day	Cost per base
454	400-500	1 Gb	18$/Mb
Illumina	100-150	6.5 Gb	0.4 $/Mb
SOLiD	50-80	5 Gb	0.5 $/Mb

The raw amount of data, length of the reads and sequencing cost for each NGS technology are displayed in the table

Figure 7. Decrease in Sequencing cost



The National Human Genome Research Institute, has compiled extensive data on the costs of sequencing DNA over the past decade and used that information to create two truly jaw-dropping graphs. NHGRI's research shows that not only are sequencing costs plummeting, they are outstripping the exponential curves of Moore's Law. By a large margin.For The Costs per Genome, NHGRI considered a 3000 Mb genome (i.e. humans) with appropriate levels of redundancy necessary to assemble the long strain in its entirety.

Third generation sequencing technologies

Although PCR amplification has revolutionized DNA analysis, in some instances it may introduce base sequence errors or favor the amplification of certain sequences over others, thus changing the relative frequency and abundance of various DNA fragments that

existed before amplification. To overcome this, the ultimate miniaturization into the nanoscale and the minimal use of biochemicals, would be achievable if the sequence could be determined directly from a single DNA molecule, without the need for PCR amplification and its potential for distortion of abundance levels. This sequencing from a single DNA molecule is now known as the “third generation of HT-NGS technology”. The concept of sequencing-by-synthesis without a prior amplification step, i.e., single-molecule sequencing is currently pursued by a number of companies. A comprehensive summary of this subject is beyond the scope of this thesis, however a detailed review on the forthcoming third generation sequencing technologies was recently presented [57].

Despite differences in the fine details and technicalities of proposed methodologies, these forthcoming sequencing technologies all promise the same advances/advantages with respect to their PCR based counterparts: a reduced error rate due to the overcoming of issues related to the biases introduced by PCR amplification and dephasing, and longer reads - thanks to the capacity to exploit more fully the high catalytic rates and high processivity of DNA polymerases. Noticeably the developers of these so called third generation sequencing technologies maintain that in the near future they will be able to both greatly increase the throughput and decrease sequencing time and costs. Implying that we may be on the verge of a new sequencing revolution.

Next generation genomics and functional genomics

With the reduction of sequencing costs by orders of magnitude and the opportunity to generate “epic” quantities of data in a matter of days, the advent of NGS sequencing techniques has introduced a new methodological era in contemporary genomics research [5] In addition to the conventional objectives of genome resequencing/SNP discovery, the characteristics of these technologies permit them to be efficiently applied to a number of other applications. For example, NGS of cDNA can be used to provide a comprehensive snapshot of the transcriptome [58], facilitating gene annotation and identification of splicing variants [59]. These novel technologies have also been extensively applied to the characterization of small RNA populations [60], the identification of microRNA targets in plants [61] and more recently in animals [62], the characterization of genomic regions bound by transcription factors (Tfs) [63] and other DNA binding proteins, the identification of genome methylation patterns [64], the characterization of RNA editing patterns [65] and metagenomics projects [66].

At the core of this methodological revolution NGS technologies have gradually but constantly replaced high throughput DNA hybridization arrays, with considerable gains in terms of time and money. In this context, even small research groups can now conduct genome-wide scale analyses at affordable costs.

The drastic change in the nature of the data respect to the “old” sequencing technologies, however posed new and intriguing challenges to bioinformaticians [6]. Indeed, the scale of these new datasets precludes their analysis/interpretation by any means other than

algorithms implemented in computer programs.

As a first consideration, the reduced length of individual sequence reads reads (initially only 25-35 bp for SOLiD and Illumina), limited the usability of NGS data for large de-novo genome sequecing projects, and required the development of completely new algorithms and software even for resequencing applications (the search for Single Nucleotide Polymorphisms and structural variants between closely related genomes. To a lesser extent this assertion is valid even 4 years later, notwithstanding the remarkable and rapid progress in the refinement of the sequencing chemistries and optimization of the sequencing media and the resulting improvements in read lengths (For detailed explanations see the section on genome assembly). Indeed the availability of a reference genome from the same or a closely related species is highly desirable in the context of NGS based experiments (apart from De novo genome and transcriptome assembly – see below), especially in the case of higher eukaryotes with large and complex genomes.

It follows that the first and arguably most crucial step of most NGS analysis pipelines is to map reads to sequences of origin. Mapping of reads is a distinctive manifestation of perhaps the oldest bioinformatics problem, sequence alignment. However, classical methods such as pure Smith–Waterman dynamic programming, or indexing of longer k-mers in the template sequence (BLAT [67]) , or combinations of the two (e.g. BLAST [21]) are not well suited to the alignment of very large numbers of short sequences to a reference sequence. To avoid the need for expensive dedicated hardware, the overall goal of short read

mapping is to obtain satisfactory results as efficiently (in terms of time and memory requirements) as possible. As a result, many methods are based on the similar principles and algorithms, but differ in the 'programming tricks' or ad hoc heuristics used to increase speed at the price of minimal loss of accuracy [68]. Research in this field is booming and new, or modified mapping tools currently appear on an almost weekly basis [69].

The principle of creating an index of the positions of all distinct k-mers in either the sequence reads or the genome sequence underlies most short read mapping tools. The most fundamental differences between available mapping algorithms are, arguably, whether the genome or the sequence reads are indexed, and the indexing method applied. Additionally, different methods may or may not allow the presence of indels in alignments, the reporting of only unique best matches or of all matches within a defined maximum Hamming—or edit—distance. As mentioned previously, various heuristics have also been introduced to accelerate searches, for example 'quality scores' indicating the confidence of base calls can be used to limit the search space. Thus, mismatches can be confined only to those tag nucleotides that are deemed to be 'less reliable', or reads containing low-quality base calls can simply be excluded. Alternatively, since less reliable base calls are often located near the end of reads, one could require exact matching for the beginning of the reads and allow for mismatches in the rest. When entire tags do not generate a satisfactory mapping, the last bases (more likely to include sequence errors) can be trimmed away and the matching can be repeated for the shorter reads. As they are crucial in many if not all NGS based experiments, recent years have witnessed

huge efforts in the production of aligners, and yet the same basic strategies are implemented a plethora of different software is now available. A (non comprehensive) list is found in Table 2.

Table 2: list of NGS aligners

Algorithm	**algorithm**	**indels**	**Author**
Galign	Hash	+	Shaham S. [70]
GSNAP	Hash	+	Wu T. and Nacu S [71].
RazerS	Hash	+	Weese D. [72]
RMaP	Hash	+	Smith A. et al. [73]
rNA	Hash	+	Policriti A. et al. [74]
soap2	Hash	+	Li R. et al. [75]
SHRiMP2	Hash	+	Matei D. et al. [76]
GnuMap	Hash	+	Clement N. [77]
Novoalign	Hash	+	Krawitz P. et [78]
AGILE	Hash	+	Misra S. et al. [79]
BWA-SW	Suffix Tree	+	Li H. and Durbin R [80]
BOWTIE	Suffix Tree	-	Langmead B. et al.[81]
CLC-BIO	Unknown	+	commercial
SSAHA2	Hash	+	Ning Z et al. [82]
BWA	Suffix Tree	+	Li H. and Durbin R [83].
Eland	Hash	-	commercial

A list of publicly available NGS sequence assemblers is reported. Under the column algorithm the algorithmic structure used to index the genome/sequencing reads is reported. The column “indels” indicates if alignment gaps are allowed by the software

Genome assembly and characterization of genomic variants

As with aligners, progress in the development of de novo assemblers has been rapid and is ongoing [84]. At least most of the available de

novo short read assemblers utilize a common underlying principle, being based on a modified version of De Bruijn Graphs [85]. De Bruijn Graphs are directed graphs representing overlaps between sequences of symbols [86] . Given a set of reads, NGS assemblers build a de Bruijn Graph by dividing all the reads in all possible k-mers, associating k-mers to nodes, and then connecting nodes. This construction has the double advantage that no overlap has to be computed and the amount of memory needed is proportional to the number of distinct k-mers and not to the number of distinct reads. An overview of available NGS sequence assemblers is shown in Table 3.

Table 3: NGS sequence assemblers

Name	Algorithm	Author
Abyss	DBG	Simpson, J. et al. [87]
SOAPdenovo	DBG	Ruiqiang Li, et al. [88]
SUTTa	DBG	Narzisi G., et al. [89]
SGA	OLC	Simpson. J.T. et al. [90]
PE-assembler	OLC	Pramila, N.A. et al. [91]
Euler	DBG	Pevzner, P. et al [92]
Euler-SR	DBG	Chaisson, MJ. et al. [93]
SSAKE	Prefix-tree	Warren, R. et al. [94]
QSRA	Prefix-tree	Douglas W. et al. [95]
Velvet	DBG	Zerbino, D. et al. [96]
Allpaths	DBG	Butler, J. et al. [97]
CLC	DBG	commercial

A list of publicly available NGS De novo sequence assemblers and the main computational algorithm adopted by each to calculate the overlaps between the reads is reported. DBG= De Brujin Graph, OLC= overlap/layout/consensus, Prefix tree=prefix tree

Although NGS technologies can produce impressive amounts of data, and despite the development of efficient and highly reliable protocols for the generation of paired end and “mate-pair” data, de novo assembly of higher eukaryotes genome from NGS data alone still proves to be problematic, highly demanding (in terms of computational resources) and costly. The main problem remains the limited length of the reads resulting in low “resolution” in the presence of high copy number repeats, transposable elements, long low complexity regions and high heterozygosity rates [98].

Despite the inherent limitations this technology has proven extremely successful for the de novo assembly of relatively small genomes (size in order of tens of megabases) and is routinely applied in the sequencing of lower eukariotes, fungi in particular, and microbes[99].

The application to larger genomes such as human or higher plants has proven feasible but less effective [100] [101] [102]. As with the “old” Sanger data the usage of mated reads is crucial to attain good assemblies, and therefore dedicated tools have been developed which can improve the results achieved by the assemblers by “scaffolding” the original contigs [103] [104] [105] and see section on genome assembly and scaffolding in the results.

Apart from complete genome assemblies from scratch, NGS data are routinely used for the characterization of intra specific genomic diversity [106]. SNPs and small indels can be determined by an accurate parsing of the output of the aligners. Whilst more complex or longer variants can be accommodated by analyzing local statistical properties of the read mapping patterns (read-depth and insert size statistics –again see the decicated section in results-). More complex and longer variants

can also be determined by partial or complete de novo assembly of the reads and subsequent alignment [107].

HT-NGS platforms have also found application in high throughput mutation detection and carrier screening using a method called functional genomic fingerprinting (FGF). The method implies a selective enrichment of functional genomic regions (the exome, promoterome, or exon splice enhancers) approach to address the discovery of causal mutations for disease and drug response [108]. The target enrichment approach, based on microarray or bead technologies has also allowed the parallel, large-scale analysis of complete genomic regions for multiple genes of a disease pathway, and for multiple samples simultaneously, thus providing an efficient tool for comprehensive diagnostic screening of mutations [109].

Large-scale transcriptome analysis by RNA-Seq

Traditionally, microarray expression profiling has been recognized as the premier tool for correlating gene activity and phenotype and allowing rapid discovery of gene pathways involved in biological processes and pathological states. Despite these successes, the capacity of array-based technologies to put recently found transcriptional complexity into a biological context has been constrained by the limitations in array technology. These include:

- the requirement of extensive prior knowledge of the transcriptome for successful chip design,
- the requirement for suitable sequence content in target

sequences to ensure clear hybridization results

- the challenge that homologous target sequences cross-hybridize to give non-specific signals
- the difficulty in identifying changes in exon usage as seen with alternative splicing and
- the limits of sensitivity for rarely expressed transcripts.

With the advent of massively parallel next-generation sequencing, it is now possible to assay transcription at a level not previously practicable. For example, RNA quantification based on RNA-Seq is thought to have a greater dynamic range compared to array-based approaches because read counts do not suffer from the same saturation and sensitivity limitations as array fluorescence signals [110] [111]. Furthermore, compared to array-based approaches, RNA-Seq has the advantage that novel mRNAs, alternative start and polyadenylation sites and splicing events such as exon skipping, alternative 5′ and 3′ splice sites and novel exon usage can be detected [112] [113] [114].

In RNA-Seq experiments, transcript expression levels are typically inferred by the number of tags that describe a certain transcript sequence. How to accurately quantify transcript expression levels from RNA-Seq data is an active area of research and different tools for RNA quantification are currently being developed [115]. To describe the average transcript activity within a sample, RPKM, or the number of mapped reads per kilobase of exon per million mapped reads, has become a common approach [116]. More recently, modified RPKM values or read counts that take the mappability of different transcript regions into account have been used [117]. This approach removes biases in RNA quantification that are purely a result of the different level

of ‘uniqueness’ of different transcripts.

Several factors have been shown to affect quantification, such as the depth of sequencing and the length of the transcript to be quantified [116]. The depth of sequencing will determine the ability to detect and quantify rare transcripts, while the length of the transcript affects the probability with which tags are detected and hence the statistical power to detect differential expression [118]. Specifically, longer transcripts are over-represented among differentially expressed transcripts compared to shorter transcripts [118]. Given that some gene classes tend to be composed of genes of longer length, this transcript length bias may affect downstream interpretation of the types of pathways disregulated in comparisons of different experimental treatments. Development of more-sophisticated statistical analysis approaches for RNA-Seq data will thus be of paramount importance.

These challenges left aside, RNA expression quantification from RNA-Seq is typically performed in the following analysis steps. Initially, RNA-Seq reads are mapped to the genome as well as to a library of exon junctions. Novel exons may be identified by the presence of a cluster of tags that map outside known exons. Subsequently a profitable measure of trancripts level such as the RPKM is calculated. Finally relevant statistical procedures are applied in order to detect alternative expressed transcripts [119].

Given our incomplete knowledge of eukaryotic transcriptome diversity, the library of exon junctions against which sequence reads are mapped in a first instance commonly includes all theoretically possible exon combinations for a given gene locus, so that novel combinations of known exons can be identified in the sample. Nevertheless, mapping

will still fail if the read spans a junction that is not represented in the junction library or one that involves one or two novel exons. To overcome these challenges, the software QPALMA [120] was developed for de novo junction mapping. This software uses information from known splice sites, including intron length models, to train a support vector machine that can then be applied to identify novel junctions in a test sample. However, the software requires the availability of a set of known exon junctions for training and has long computational run times. More recently, an ab initio method for the detection of splice sites was developed, TopHat [121], that does not rely on a training set of known splice sites and which has favorable computational run times.

An alternative approach to transcriptome discovery is based on the de novo assembly of transcriptomes and would appear particularly useful for the identification of exon skipping, intron retention and novel, alternative splicing events. Recent advances in this area, analogous to those seen in de novo genome assembly, use de Bruijn graphs and overlapping k-mers to assemble short reads into contigs. However, both sequencing errors, and the widespread occurrence of alternative splicing in higher eukaryotes, significantly complicate the analysis and also result in very long computational analysis times even when paired end RNAseq data are employed [122].

ChiP-Seq

Given that a regulatory sequence is accessible (not in heterochromatin), the time, amount, and duration of transcription of a gene is under the control of various specific (TF) and general transcription factors (GTF) that might bind to the regulatory sequence. GTFs are usually cofactors of the RNA polymerase complexes, while specific TFs represent 'classical' transcription factors such as NFkB, SP1 or AP1. Currently more than 760 specific TFs are known for the human genome alone (MatBase 8.2, Genomatix Software, Munich). The interaction of these TFs with their respective TFBSs in regulatory regions determines the major part of direct transcription control as they form the activator complexes on promoters and enhancers that subsequently attract the pol II complex, which in turn initiates transcription.

Chromatin Immunoprecipitation (CHiP) [123] refers to the isolation of genomic fragments bound to proteins through the use of crosslinking agents and specific antibodies to identify genomic regions bound to histones or specifically by DNA binding proteins such as TFs. This technology is rapidly becoming the method of choice for the large-scale identification of TF–DNA interactions, or, more broadly, of the characterization of chromatin packaging—how genomic DNA is packaged into histones and in correspondence with which histone modifications. Chip-Seq implies the characterization of isolated DNA by NGS approaches (as opposed to the search for specific sequences by PCR, or the identification of isolated DNA through microarray-based approaches). Genomic fragments may be subjected to single or paired end sequencing strategies and reads are mapped to the genome to

identify enriched regions—in principle those that contain functional binding sites for the factors of interest.

Once reads have been mapped to the reference sequence, it is necessary to determine which regions are flanked by a sufficient number of reads to discriminate them from 'background' noise due to sequence errors, contamination of isolated protein–DNA complexes, non-specific protein binding and other stochastic factors.

One way to filter out noise is to use a negative control to generate a pattern of noise to be compared to the read map generated from the real data (either using an antibody which does not recognize any TF, or by using a cell type that does not express the factor of interest). It is clear that genomic regions enriched only in the positive experiment should be those of interest.

In the absence of control experiments, background read levels must be estimated using stochastic methods. If we assume that in a completely random experiment each genomic region has the same probability of being extracted and sequenced, given the overall number of tags, and given the size of the genome, then the probability of finding one tag mapping in a given position is given by t/g. The same idea can be applied by dividing the genome into separate regions (for example, the chromosomes or chromosome arms), since for experimental reasons different regions can have different propensities to produce reads. Thus, global or region-specific 'local' matching probability can be calculated, and the expected number of tags falling into any genomic region of defined size can be estimated for example using Poisson or negative binomial distributions. Finally, the significance of tag

enrichment is computed, by using sliding windows across the whole genome. If a 'control' experiment is available, the number of tags it produced from a given region can serve directly as 'background' model. Several 'peak-finding' methods have been published [115], an overview is reported in Table 4.

Table4: Peak finding methods

Name	Peak criteria	FDR	Author
CisGenome v1.1	1:Number of reads in window 2: Number of ChIP reads minus control reads in window	1: Negative binomial 2: conditional binomial	Ji H et al [124]
FindPeaks	Height threshold	Monte Carlo simulation	Fejes et al [125]
MACS	Local region Poisson P value	control/ChIP	Feng J et [126]
PeakSeq	Local region binomial P value	1: Poisson background 2: binomial for sample plus control	Rozowsky et al [127]
QUEST	Height threshold, background ratio	control/ChIP as a function of profile threshold	Jiao et al [128]
SICER	P value from random background Enrichment relative to control	From Poisson P values	Garmire et al [129]

List of Peak finding software for Chip-Seq analysis. The statistical criteria used to call the peaks as well as those used to estimate false discovery ratios are reported.

The general concordance of conclusions drawn from ChIP-Seq and ChIP on chip approaches has been shown to be extremely high [130].An analogous approach (RNA Immunoprecipitation Sequencing or RIP-

Seq) [131] has been used to tentatively identify sites of binding of proteins involved in mediating mRNA stability and splicing.

Small RNAs

Recent years have seen number of important discoveries relating to the regulated expression of small (typically 18–25 base) RNAs in eukaryotic cells and their important roles, principally as regulators of stability or availability for translation of mRNAs, with which they can interact by means of base complementarity e.g. [132] but also as guides for genome methylation [133] and potentially in other processes. Deep sequencing of small RNAs has become the method of choice for small RNA discovery and expression analysis [134]. Unlike oligonucleotide array studies, deep sequencing requires no a-priori knowledge of the nature of small RNAs, is less subject to the lack of specificity of short probes sometimes associated with oligonucleotide arrays and expression levels can be followed over a wider range with deep sequencing [135]. Indeed, even the shortest sequencing reads will yield the complete sequence of a 'small RNA', making these molecules ideal targets for characterization by NGS technologies.

Many classes of small RNAs exist as families present as multiple highly conserved copies within a single genome and often conserved between related organisms. Clustering of observed sequences and comparison with databases of annotated small RNAs (e.g. miRBase [136]) allows the identification of members of conserved families and provides

indications as to their relative expression levels. Analysis of the size distribution of reads can also prove informative as to the nature of small RNAs present. For example, microRNAs tend to be ~21 bases in length as are the transactivating small RNAs tasi-RNAs) of plants, other siRNAs in plants typically being 24 bases in length while piRNAs of animals tend to be between 25 and 33 bases in length.

Several specific bioinformatics tools have been developed to identify members of different classes of small RNAs from deep sequencing data, an overview is reported in Table 5. While expression profiling can be carried out using principles similar to those used in typical RNA-seq experiments.

Table 5: Bioinformatic resources for small RNAs

Name	Function	Author
deepBase	Capture, storage and retrieval of largescale genomic data	Barret et al [137]
miRBase	Search for miRNA Database Analyze genomic coordinates and context Mine relationships between miRNAs	Griffiths-Jones et al. [136]
mir2Disease	Collection of microRNA-disease relationship information	Jiang et al., 2009 [138]
MicroRazerS	Small RNA reads alignment	Emde et al. [139]
MirDeep	Detect known and novel miRNAs	Friedlander et al [140]
MirTrap	Detect novel miRNAs	Hendrix et al [141]
mirTools	Small RNA read alignment Comparative analysis of two or more miRNA expression data Classification and annotation of known miRNAs Detect novel miRNAs	Zhu et al [142]

miRanalyzer	Small RNA read alignment Detect known and unknown miRNAs Stand-alone Detect undetected mature-star miRNAs	Hackenberg et al [143]
UEA sRNA toolkit	Predict miRNAs and their targets Compare expression levels in sRNA loci	Moxon et al [144]
miRNAkey	Small RNA read alignment Comparative analysis of miRNA expression data	Ronen et al [145]
miRExpress	miRNA expression profiling	Wang et al [146]

A list of algorithms commonly used in bioinformatics pipelines for the analysis of short RNA NGS data is displayed. For each program a synthetic description of the functionality is also reported

Recently, several innovative, second generation sequencing based, approaches to the identification of mRNAs targeted by miRNAs have been proposed [147] and [61] independently developed similar methods to isolate and sequence the 5' ends of mRNA degradation products in plants. Addo-Quaye subsequently proposed a bioinformatics strategy to reconcile over-represented degradation products to predicted miRNA target sites, complementing the experimental approach [148] Alternatively, a manifestation of the RIP-Seq methodology can be used to identify smallRNAs and their mRNA targets that are incorporated in the RNA Induced Silencing Complex (RISC) that mediates RNA silencing [62].

Epigenomics studies

Epigenetics refers to the mechanisms that regulate the cell type or tissue specific transcription or gene expression levels without altering the DNA sequences, through biochemical modifications such as the addition of a methyl group to cytosines, and post-translational modifications of histone proteins. These epigenetic mechanisms play a critical role in the normal stages of cellular developmental and processes such as embryogenesis, cell differentiation (cell lineage specification), inactivation of the X chromosome and genomic imprinting through modulation of transcriptional regulation in a tissue specific manner. Abnormalities in these epigenetic mechanisms have been linked to a wide range of diseases. The importance of exploring the epigenetics of human complex diseases and traits is now being increasingly recognised. [149] [150] [151] .

One of the most popular methods of characterizing the methylation state of genomic DNA has been the targeted sequencing of particular genomic regions after treatment of isolated DNA with bisulfite which converts unmethylated cytosines to uracil, but does not modify 5′ methylated cytosines [152]. More recently, and analogously to the situation with ChiP experiments, specifically designed microarrays have allowed the identification of methylated and non-methylated regions though hybridization with bisulfite treated genomic DNA. The development of NGS technologies has provided an alternative approach whereby bisulfite treated DNA is directly sequenced and mapping of reads to the genomic sequence allows identification of methylated sites and quantification of the frequency with which such sites are methylated

DNA [153]. While genome-wise studies of histone modifications can be carried out in a relatively straightforward manner, by the means of dedicated Chip-Seq experiments [154].

Part II

Aim of the project

During my years as PhD student I spent most of my time developing software, algorithms and procedures for the analysis of sequencing data, with a particular focus on data produced by Next-Generation Sequencing technologies. The development of each of the software pipelines and tools that have constituted the major part of my doctoral studies was prompted by the needs of distinct biological research projects and collaborations in which our research group was involved. The tools and algorithms are thus presented in their appropriate biological contexts.

Initially I dedicated myself to the development of scripts and pipelines which were used to assemble and annotate the mitochondrial genome of the model plant *Vitis vinifera*. The sequence was subsequently used as a reference to study the RNA editing of mitochondrial transcripts, with data produced using the new Illumina and SOLiD RNA-seq protocols and published in the international peer reviewed journal Nucleic Acids Research.

The primary aim of this study was to identify and characterize the editing profile of Vitis mitochondrial transcripts, to study tissue specific patterns of editing, and to compare the pros and cons of the competing sequencing technologies. Within this project I also developed ancillary scripts for the visualization and comparison of data.

I subsequently developed a new approach and a new software package for the detection of structural variants between a donor and a reference genome, using NGS paired-end (PE) data. This work aimed to demonstrate an enhanced approach, based on machine learning algorithms, for the detection of relatively small indels. In particular, I was able to show that, suitable Paired End data, contrary to previous assertions, can be used to detect, with high confidence, very small indels in low complexity genomic contexts. An associated manuscript is currently under review at Nucleic Acids Research.

Given that associated manuscripts have been published (or are under review for publication), rather synthetic summaries of the biological issues and bioinformatics approaches employed are presented and the manuscripts themselves are provided in the appendices of this thesis

Finally I have also participated in a project which is aimed at the reconstruction of the genomic sequences of 2 distinct strains of the biotechnologically relevant fungus Fusarium. In this context I performed the sequence assembly to obtain the initial contigs and devised and implemented a new scaffolding algorithm which has proved to be particularly efficient. A manuscript associated with this software is currently in preparation.

Main Results

Deep sequencing and RNA editing of plant mitochondrial Transcripts

The mitochondrial genome of angiosperms: an overview

For historical reasons, the angiosperm mitochondrial genome is usually described as a single circular DNA molecule that houses a complete set of genes, called the 'master chromosome' [155].

Angiosperm mitochondrial genomes show much greater variation in size than their animal counterparts and have been described as varying from 90 to 800 kb in size [155]. The gene content is also variable between species, but the most striking feature is the fluidity of intergenic regions, where species-specific sequences predominate [156].

Master circular chromosomes of angiosperm mtDNA have been generated by restriction mapping followed by shotgun or mapped-cosmid sequencing, although some groups reported difficulty in generating circular configurations [157] [158]. The multipartite structure of plant mitochondrial genomes generated by recombination within master circles was used to explain the heterogeneity that is observed when plant mitochondrial (mt) DNA is examined by electron microscopy and gel electrophoresis [159]. However, the apparent morphology of carefully isolated mtDNA was not found to be consistent with the circular model [160]. The observed molecules appeared to be linear and circular

and of various sizes (including structures that exceed estimated genome sizes). In addition, Y-, H-, and theta-shaped branched forms were seen, which presumably represent recombination intermediates. Therefore, the entity of an angiosperm mitochondrial genome is likely to be a mixture of various DNA molecules[161].

It has been suggested that plant mtDNA replicates via a recombination-dependent mechanism [160][161]. It should be remarked that the concept of the master chromosome remains because the question as to how multipartite and branched DNA molecules are transmitted properly to the next generation is unsolved.

The number of mitochondrial genes in angiosperms is 50–60 (not considering copy number). The differential number of genes is due to the differential gene content for the subunits of Complex II, and especially, ribosomal proteins and tRNAs. When the content of ribosomal protein genes is compared among angiosperms, one can realize how often genes have been lost from the mitochondrial genome during angiosperm evolution. Most of the genes that are lost from the mitochondrion appear to have been transferred to the nuclear genome, but this is not always the case [156].

Some of the mitochondrial genes in angiosperms are interrupted by introns. In each of the sequenced genomes, the total number of the introns is 20–24, constituting 4–13% of the genome. All the introns in the sequenced mitochondrial genomes are classified as group II type; however, a horizontally transferred group I intron has also been documented [162].

Outside of genes, which themselves are highly conserved, and after

accounting for chloroplast, nuclear and plasmid DNA insertions, the majority of the DNA in the sequenced plant mitochondrial genomes is of unrecognizable origin. Considering the compact and conserved nature of animal mitochondrial genomes, it was truly surprising to find that, for the first sequenced angiosperm mitochondrial genomes, over half of each genome showed no obvious homology to any sequences in the public databases [157] [163] [164][165].

The mtDNA of *Vitis vinifera*

Recently two distinct and complementary sequencing project, aiming to produce the complete sequence of the economically/agriculturally relevant model angiosperm *Vitis vinifera* have been undertaken. An italian project, lead by scientists from the IASMA center [166], and a French-Italian consortium [167] have successfully elucidated the genomic sequence of two different cultivars (ENTAV115 and PN40024) from the Pinot noir strain. The two cultivars are closely related yet still agronomically different, and thus have been chosen as to maximize the insights from comparative genomics analysis.

Within the IASMA project the proposed sequence for the master circle of the *Vitis vinifera* mtDNA , was been produced as a part of the whole genomic sequencing (WGS) from shotgun sequencing libraries with average insert sizes of 2, 3, 6, 10, and 12 kb [168].

The proposed sequence is more than 770 kB long and at the time of publication represented the longest angiosperm mtDNA ever reported. Intergenic spacers constitute the largest part (90.21%, 697591 bp.) of

this molecule (where promiscuous DNA is considered as part of the spacer sequences). The protein-coding sequences comprise only 4.98% of the molecule length (38,529 bp). RNA genes constitute 0.91% of the mtDNA of Vitis and introns 3.89% (30,100 bp). Gene content in the mitochondrial genome is similar to that of other published angiosperm mtDNAs.

The large size of the genome is due to the expansion of the spacer regions. These regions contain 1,338 repeated sequences ranging in size from 30 to 651 bases (reaching 52,861 bp in total, which corresponds to 6.84% of the genome length), of which 645 are direct repeats (25,325 bp in total length, 3.28% of the genome length). Most of the genome sequence has no similarity to the sequences of other mitochondrial genomes of angiosperms.

The mtDNA of grape also contains 30 fragments of chloroplast-like DNA ranging in size from 62 to 9,106 nt. The total extent of chloroplast DNA sequences present in the mtDNA of *V. vinifera* is 68,237 bp, corresponding to 8.8% of the whole mitochondrial genome length and to 42.4% of the grape chloroplast genome. This is the largest proportion of chloroplast-like DNA sequences observed in a plant mitochondrial genome, both in absolute and relative terms. Most of the insertions are unique to the grape mtDNA, as evident from the observation that only 9 out of 30 chloroplast-like insertions have full-length homologs in the mtDNA of other plant species.

RNA editing in plants mitochodria

RNA editing is a widespread post-transcriptional molecular phenomenon that can increase proteomic diversity by modifying the sequence of completely or partially non-functional primary transcripts, through a variety of mechanistically and evolutionarily unrelated pathways [169].

'Substitution' editing by simple base modification is the most frequent type of editing and is seen both in plant organelles and in the nucleus of higher eukaryotes as well as in sequences of viral origin. In land plant organelles, RNA editing consists almost exclusively of C-to-U substitutions (rarely reverse U-to-C conversions) mostly at first or second positions of codons—typically leading to conservative amino-acid changes and increasing similarity to non-plant homologs [170].

Some plant organellar RNA editing events create translation initiation or termination codons while several known editing events in tRNA or introns improve the stability of functionally relevant secondary structure motifs [169].

Moreover, the alteration of the RNA editing pattern in plant mitochondria can lead to male sterility, also known as the CMS phenotype [171].

Classically, RNA editing events were identified experimentally by comparing cloned cDNA sequences with their corresponding genomic templates [172]. This procedure allows the study of a relatively small number of sequences and does not take into account potential cloning artifacts.

Studying the “editome” of a model angiosperm

As part of the French Italian consortium for the sequencing of the PN40024 Pinot noir cultivar, our laboratory had access to a quantity of NGS RNAseq data which had been produced for transcript annotation and transcriptome characterization. These reads were obtained from the polyA+ fraction of the cellular RNA and should in theory include only low levels of organellar transcripts (for which the degree of polyadenylation is considered to be limited). However, it is expected that a degree of contamination from organellar transcripts will be present. In the context of a fruitful and longstanding collaboration with the bioinformatics group from the university of Bari, we decided to take advantage of these data and Sanger reads generated during the sequencing of the nuclear genome to reconstruct the PN40024 mitochondrial genome and test whether NGS RNA-seq data could be used for the determination of the RNA editing profile of the mitochondrial genome of this model plant.

Assembly and annotation of the PN40024 mitochondrial genome

Although the reference mitochondrial genome from a closely related cultivar was already available [168], and plant mitochondrial coding regions tend to show extremely high level of conservation; as RNA editing sites are usually identified by direct comparison of transcribed sequences with their related templates, we wished to compare transcriptome reads to genomic templates derived from the same

cultivar. Accordingly I devised ad hoc strategies and custom scripts to retrieve an adequate number of Sanger sequencing reads of presumable mitochondrial origin (mt-like) from the PN40024 genome sequencing project trace archive.
To automate the similarity searchers against the trace archive database I made use of the blast url api, a dedicated web based programming interface which provides access to the whole ensemble of databases and functionalities enclosed within the NCBI facilities. While for the sequence and quality trace retrieval, I used the 'query_tracedb' script provided by NCBI trace archive [173].

Initially I used overlapping windows of 10 Kb from the ENTAV115 mitochondrial genome and automated blast based similarity searches to retrieve a large number of mt-like sequencing trace. I retained only traces showing at least 95% identity to the ENTAV115 mitochondrial genome and with no better mapping solution on the nuclear or plastidic genome of the PN40024 genotype.
After a preliminary assembly of these data, using the software PCAP [174] I obtained 22 non-overlapping contigs. Subsequently in an attempt to close the gaps between the contigs and establish their order and orientation, I applied a similar strategy but using the ends of the contigs generated as blast queries and adopting and iterative "search and assemble" approach.
In brief, in each similarity search we tried to identify mt-like reads spanning the contig ends, and then ran a completely new assembly from scratch including these new data along with the previously identified mt-like contigs. At this stage to maximize the information retrieved by each

search phase, for any putative mitochondrial read spanning the end of a contig I also retrieved the corresponding mate-pair sequence.
When contigs failed to be extended at both ends I removed them from the set of the queries for the forthcoming search phase.
This strategy enabled me to merge the most of my initial contigs, and to obtain the read set used to construct what is considered to be the final assembly of the PN40024 mtDNA.

To obtain the final assembly of the PN40024 mitochondrial genome I used 16789 putative mitochondrial sequences of which 13682 were identified as mate pairs. The average read length was 785 bases, implying a hypothetical redundancy of greater than 20 fold (if the mt genome of pn40024 is of the same size as that of ENTAV115).
The final assembly of the PN mtDNA consisted of 4 contigs of 339 264, 132 252, 202 123 and 76 068 nt and covers 96.37% of the ENTAV sequence with which it showed 99.92% identity.

I also performed annotation of the PN40024 mt genome using similarity with the ENTAV115 mt genome. Similarity searches using the ENTAV115 annotation allowed the identification of all of the genes of mitochondrial origin proposed by Goremykin et al [168]. In addition, support for mitochondrial origin of each coding gene was confirmed by comparing grape ORFs to genomic and unedited mitochondrial genes downloaded from the specialized REDIdb database [175].

Reads mapping and comparison of sequencing technologies

In total, 205 435 765 short reads were obtained by sequencing cDNA obtained from four tissue samples with the Illumina technology: leaf (11 lanes), root (9 lanes), callus (9 lanes), stem (14 lanes) (Sequencing performed by Illumina inc), while 139 467 080 short reads from leaf and 188 742 647 short reads from root were produced by the SOLiD RNA seq technology (sequencing performed by the group of Prof. Giorgio Valle at the University of Padova, Italy). The read lengths ranged from 33 to 35 bp for the Illumina reads, while all SOLiD short reads were 35-nt long.

Short tags, pooled from all tissues, were mapped to the assembled *V. vinifera* mitochondrial genome using version 0.5 of the PASS [176] software with a seed length of 12, a minimum identity of 90% and a minimum alignment length per read of 30 nt. Similar to a BLAST [21] approach, PASS seed sequences (called long word anchors) are extended on the flanking regions using DNA words of predefined length (typically 6 or 7 bases) for which the alignment scores are pre-computed according to Needlelman–Wunch. Significant matches are then refined to improve the global alignment quality.

The procedure recovered 939,554 unique Solexa/Illumina alignments and 5,207,827 unique SOLiD alignments. The different fraction of uniquely aligned reads (0.45 and 1.59% for Solexa/Illumina and SOLiD, respectively) also reflect quite different coverage patterns, which were much more biased for SOLiD. Despite the much higher overall fold coverage of SOLiD (158×) than SOLEXA (35×) both platforms provided a similar percentage of covered nucleotides in the coding regions, 96.9

and 96.6%, respectively. Furthermore, 16 out of 37 annotated mitochondrial coding genes were fully covered by Solexa/Illumina reads while only 11 were fully supported by SOLiD data. Looking at reads distribution along the reference sequence, we also noted local maxima in SOLiD reads in which several mitochondrial regions appeared deeply covered.

While the patterns of coverage seem to indicate a notable bias in the per-site distribution of the coverage depth across coding genes for the SOLiD data, a moderate, but highly significant ($r = 0.25$, $P < 0.0001$) correlation was observed between per base coverage by SOLiD and Illumina sequencing for individual positions in the coding sequences of the 37 genes of mitochondrial ancestry—possibly due to a known dependence of recovery of fragmented cDNA (by gel elution) on GC content. However, distinct coverage patterns by these different sequencing strategies contribute to a substantially higher coverage when both technologies were combined—complete coverage of 25 genes out of the 37 and an overall coverage of 98.3% of all coding nucleotides.

Identification of edit sites

Solexa/Illumina and SOLiD mapping results in GFF format were used to identify C-to-U changes due to RNA editing in the grape mitochondrial genome of the cultivar PN40024 by means of ad hoc custom scripts.

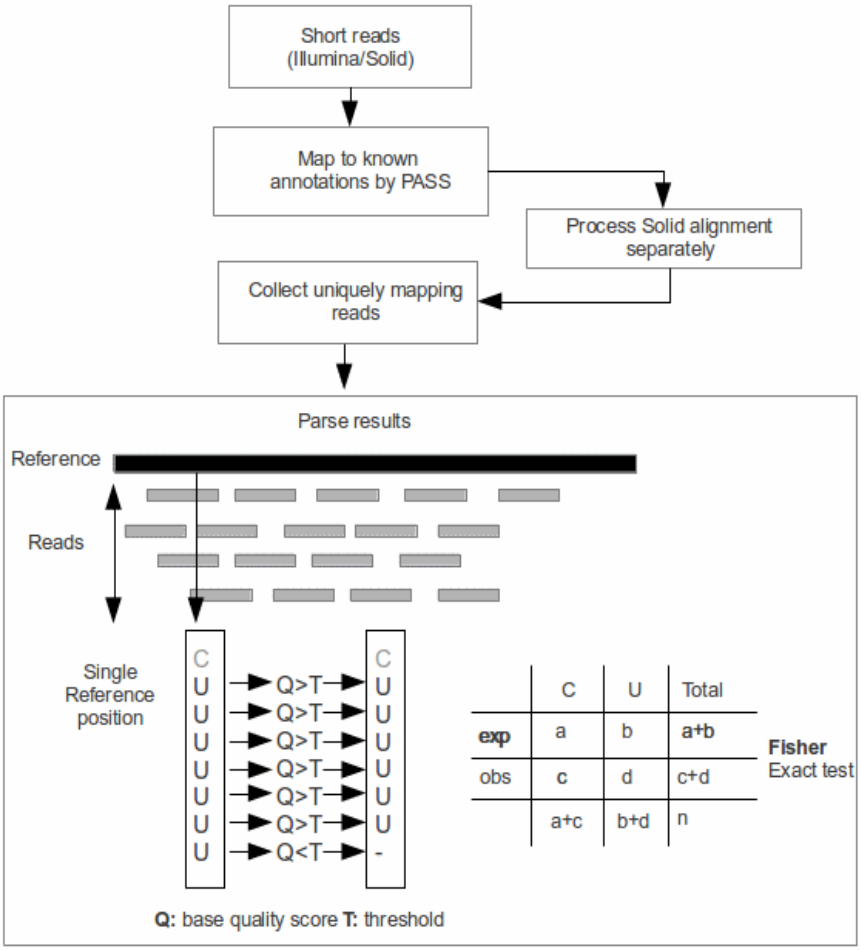
The main script, in particular, used as input a GFF file, the reference sequence of the grape mitochondrial genome in FASTA format and a

textual file containing protein-coding annotations. It collects all uniquely mapping reads (with at most two mismatches and no indels) falling in annotated genes and for each reference position calls the corresponding read nucleotide if the associated quality score is above the fixed threshold of 15. Finally, for each reference position, the script calculates the frequency of the modified nucleotide (if any) over the total recorded signal (sum of modified and not modified nucleotides).

RNA editing sites due to C-to-U changes were detected separately for each platform and tissue. Rates of sequencing errors were estimated for each sample as the total frequency of non-C↔U substitutions. Among the potential editing sites, corresponding to sites where a genomic C was aligned to one or more U from RNA-Seq data, statistically significant editing sites were determined by applying the Fisher's exact test to compare the observed and expected C and U occurrences in the aligned reads. A confidence level of 0.05 (also with FDR or Bonferroni correction) was used as cut-off.

A putative editing site is classified as 'conserved' if one or more homologous sites in other plants are experimentally known to be edited or if a fully conserved U is observed in all homologous sites, according to the data collected in the REDIdb database [175]. RNA editing sites in non-coding grapevine genes and group II introns were detected according to the same computational strategy. Statistically significant edited sites have been classified fully or partially edited depending on if the observed fraction of RNA-Seq aligned U was above or below 90%.

Figure 8: Identification of edit sites



Graphical overview of the computational methodology used to detect RNA editing sites by short sequencing reads of next generation platforms.

Identification of edit sites in *Vitis vinifera*

In total 401 significantly supported editing sites were identified in grapevine mitochondrial coding regions with a 5% confidence level in

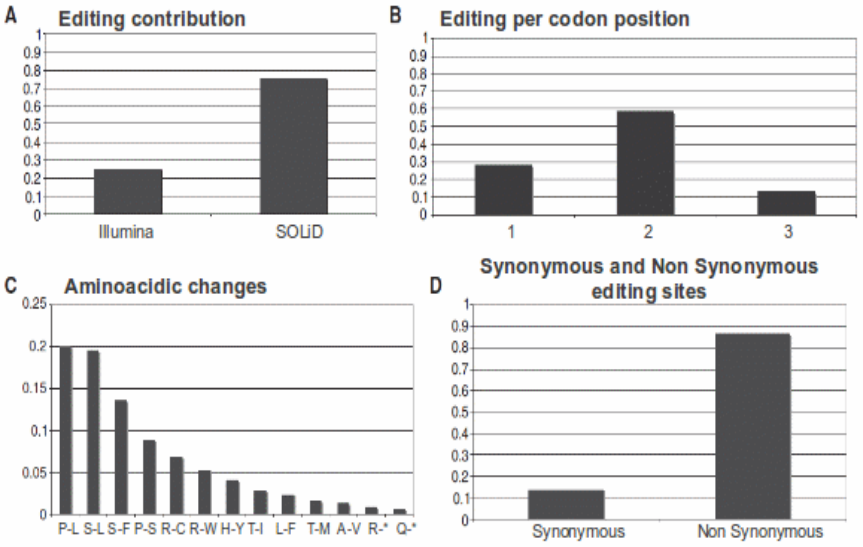
the Fisher's exact test. To evaluate the effectiveness of the statistical assessment we determined the percentage of conserved edited sites of putative editing sites. Interestingly, >90% of significantly detected edited sites were conserved, supporting the reliability of the statistical test. Indeed, only a slight increase was observed with more stringent cut-offs (5% confidence level with FDR or Bonferroni correction). It should be noted that a remarkable level of conservation was also observed for putative editing sites filtered out by the statistical test. It is highly likely that the read coverage at these positions is not deep enough to provide statistical support. Including all 314 additional putative edited sites with conserved homologous counterparts in other plants, more than 700 sites may be edited in the grapevine mitochondrion.

87% of the 401 editing modifications occurred at the first and second positions of codons, almost invariably resulting in replacement of the encoded amino acid. Indeed, only 1 out of 114 events affecting the first codon position resulted in synonymous changes. All non-synonymous editing conversions could modify the biochemical nature of the affected proteins. As observed in mitochondria of *A. thaliana* [131], the most frequent amino acid changes induced by RNA editing in grapevine were P-to-L (20.0%), S-to-L (19.4%) and S-to-F (13.5%) increasing the proportion of hydrophobic amino acids and suggesting a real functional role for RNA editing through protein modifications in predominantly membrane-localized proteins. Additionally, S-to-L or S-to-F substitutions potentially increase the hydrophobicity of interface residues while P-to-L conversions occurring in secondary structures can contribute to protein functionality by avoiding defects in 3D structures.

Besides the non-random distribution of editing with respect to codon

positions, a preference of RNA editing towards specific codons was observed. In particular, the three most frequently edited codons were UCA, CCA and UCC, accounting for 32.7% of all edited codons. The only C-containing codons never affected by editing were GGC, AGC and UGC in which editing could only lead to synonymous substitutions.

Figure 9:Principal statistics of detected RNA editing sites in *V. vinifera*



(A) The contribution of each sequencing platform to editing detection; (B) distribution of C-to-U editing conversions across codon positions; (C) distribution of amino acids changes induced by detected RNA editing; (D) frequencies of synonymous and non-synonymous editing changes.

Twenty four percent of the 401 C-to-U conversions were classified as fully edited sites while 76% were considered partially edited sites supporting the hypothesis that partial RNA editing is common in higher plant mitochondria. A proportion of partial editing might be due to transcripts where editing was not yet complete, while other partial

events might derive from tissue-specific edits derived from mixed tissue samples.

Tissue specificity accounts for a fraction of the observed partial RNA editing. Tissue specific editing might be required to modulate protein functionality in response to cell-type specific requirements. The high depth of coverage afforded by the SOLiD data resulted in the recovery of the majority of the significantly tissue specific edits by this technology. In summary, using the information from both sequencing technologies we discovered that 71% of all tissue-specific C-to-U changes occurred in leaf, whereas only a small fraction (0.4%) occurred in stem. Tissue specific editing events occurring in root and callus, instead, constituted 21 and 7.6%, of the total, respectively.

Detailed methods and the full results of this project are presented in the published paper “Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing” (Picardi et al. Nucleic Acids Res. 2010 Aug;38(14):4755-67) which is also provided in appendix 1 of this thesis.

Development of a new software package for the determination of intra specific genomic diversity

Genomic variation in the human genome

The spectrum of intra-specific genetic variation ranges from the single base pair substitutions (SNPs) to large chromosomal events (deletions, inversions, recombination events etc), but it has become apparent that human genomes, for example) differ more as a consequence of structural variation (small indels in particular) than of single-base-pair differences [177] [178] [179]. [180]. Structural variation was originally defined as insertions, deletions and inversions greater than 1 kb in size [181]. With the sequencing of human genomes now becoming routine, the operational spectrum of structural variants (SVs) and copy number variants (CNVs) has widened to include much smaller events [182]. The challenge now is to discover the full extent of structural variation and to be able to genotype it routinely in order to understand its effects on human disease, complex traits and evolution.

The discovery and genotyping of structural variation has been central to understanding the molecular basis of some of the most severe diseases [183] [184] [185]. Systematic and comprehensive assessment of structural variation has been problematic owing to the complexity and multifaceted features of SVs. Ideally, SV discovery and genotyping requires accurate prediction of three features: copy, content and structure. In practice, this goal has remained elusive because SVs tend to reside within repetitive DNA, which makes their characterization more difficult. SVs vary widely in size and there are numerous classes of structural variation: deletions, translocations, inversions, mobile

elements, tandem duplications and novel insertions.

Before the advent of NGS sequencing DNA hydridization arrays were the most common approach used in SV discovery and genotyping. These are represented primarily by Comparative Genomic Hybridization arrays (CGH arrays) and SNP microarrays [186] [187]. Both hybridization-based technologies infer copy number gains or losses compared to a reference sample or population, but differ in the details and application of the molecular assays.

Array CGH platforms are based on the principle of comparative hybridization of two labelled samples (test and reference) to a set of hybridization targets (typically long oligonucleotides). The signal ratio is then used as a proxy for copy number.

Currently, Roche NimbleGen and Agilent Technologies are the major suppliers of whole-genome array CGH platforms. One key advantage of array CGH platforms is the availability of custom, high-probe-density arrays from both major manufacturers. This has led to their widespread adoption in clinical diagnostics, essentially replacing karyotype analysis as the primary means of detecting copy-number alterations [188].

SNP microarray platforms are also based on hybridization, with a few key differences from CGH technologies. First, hybridization is performed on a single sample per microarray, and log-transformed ratios are generated by clustering the intensities measured at each probe across many samples[189][190]. Second, SNP platforms take advantage of probe designs that are specific to single-nucleotide differences between DNA sequences, either by single-base-extension

methods (Illumina) or differential hybridization (Affymetrix) [189] [190]. The key advantage of the SNP arrays is the use SNP allele-specific probes to increase the sensitivity to allele specific CNV and enabling an higher resolution. The key disadvantage is that as the data collection is performed on separate slides and almost identical probes are used the signal to noise ratio is tendentially higher.

The major reason behind the success of SV detections microarrays lies in the fact that they are are economical and practical. Determining the pathogenic significance of any particular event in a rare-variant disease model requires screening of thousands of affected individuals and controls. Given the low cost of array CGH and SNP platforms and the large collection of public SNP data available from genome-wide association studies, microarray data provide an opportunity to assay the CNV landscape of large data sets [191].

Conversely microarrays are limited to detecting copy-number differences of sequences present in the reference assembly used to design the probes, provide no information on the location of duplicated copies and are generally unable to resolve breakpoints at the single-base-pair level [192]. Perhaps the most important limitation is the use of hybridization-based assays in repeat-rich and duplicated regions. Array CGH and SNP platforms assume each location to be diploid in the reference genome, which is not valid in duplicated sequence. The signal for a 5 to 4 copy ratio, or other complex patterns, will not fit the expected results for a diploid reference sequence and may drop below the assay's sensitivity to discriminate signals [193].

The advent of next-generation sequencing (NGS) technologies have

enabled applying sequence-based approaches for mapping SVs at a fine scale. However, NGS approaches present substantial computational and bioinformatics challenges.

Most of the current algorithms for SV discovery are modelled on computational methods that were first developed to analyse capillary sequencing reads and fully sequenced large-insert clones [194].

There are four general types of strategy, all of which focus on mapping sequence reads to the reference genome and subsequently identifying discordant signatures or patterns that are diagnostic of different classes of SV [195] [196].

Read-pair technologies

Read-pair methods assess the span and orientation of paired-end reads and cluster 'discordant' pairs in which the mapping span and/or orientation of the read pairs are inconsistent with the reference genome. Most classes of variation can, in principle, be detected. Read pairs that map too far apart define deletions, those found too close together are indicative of insertions, and orientation inconsistencies can delineate inversions and a specific class of tandem duplications [178] [179] [197][198]. Read pairs in which only one end clusters and the others do not map to the reference have been used to flag variant sequences not included in the reference genome (novel insertions). The read-pair method is the most widely applied approach and was first demonstrated using BAC end sequences generated from the breast cancer cell line MCF-7 [199]. It was subsequently applied to germline genetic variation

using a fosmid end sequence library. Later, it was applied to next-generation, paired-end data generated by the 454 FLX platform [197]. There are now many computational tools based on a read-pair approach, including PEMer [200], VariationHunter [201], BreakDancer [202], MoDIL [203], and Corona [197].

Read-depth methods

Read-depth approaches assume a random (typically Poisson or modified Poisson) distribution in mapping depth and investigate the divergence from this distribution to discover duplications and deletions in the sequenced sample [204]. The basic idea is that duplicated regions will show significantly higher read depth and deletions will show reduced read depth when compared to diploid regions. Read-depth approaches using NGS data were first applied to define rearrangements in cancer [183] [205] and segmental duplication [206] and absolute copy-number maps in human genomes [207]. Methods that attempt to discover smaller deletions and duplications at better breakpoint resolution include the event-wise-testing (EWT) [208] algorithm and CNVnator [209]

Split-read approaches

Split-read methods are capable of detecting deletions and small insertions down to single-base-pair resolution and were first applied to longer Sanger sequencing reads [210]. The aim is to define the

breakpoint of a structural variant on the basis of a 'split' sequence-read signature (that is, the alignment to the genome is broken; a continuous stretch of gaps in the read indicates a deletion or in the reference indicates an insertion). Extensions of this approach may also detect mobile-element insertions (MEIs) if the reads are sufficiently long to span the mobile element (for example, >400 bp for Alu elements) to characterize the full sequence content [196]. Alternatively, if the read length is shorter but the MEI breakpoint is in a unique sequence, a split-read approach can be used to anchor the insertion. Application of this method to NGS data sets is currently limited owing to the difficulty in aligning shorter reads; however, the Pindel [211] algorithm uses paired-end reads to reduce the search space for potential split reads, thus reducing the computational overhead of the local gapped alignment of short sequences to the reference genome.

Sequence assembly

In theory, all forms of structural variation could be accurately typed for copy, content and structure if the underlying sequence reads were long and accurate enough to allow de novo assembly. In practice, sequence-assembly approaches are still in their infancy and typically use a combination of de novo and local-assembly algorithms to generate sequence contigs that are then compared to a reference genome. Local sequence assembly of fosmid clones with discordant read pairs has been used to systematically discover structural variation in 17 human genomes [178] [192] [198]. Approaches that involve library construction,

clone array and end sequencing are too laborious and prohibitively expensive to be widely adopted. Ideally, complete genome sequencing followed by de novo assembly and comparison to a high-quality reference could identify thousands of structural variants. For example, a genome assembly from capillary sequence reads from a human individual has been used to characterize 12,178 structural variants [212] [213] [214]. Well-known de novo assembly algorithms for next-generation whole-genome shotgun AbySS [87], SOAPdenovo [88], Velvet [96] and ALLPATHS-LG [97] (or see Table 3).

Figure 10: Different approaches used for SV detection

Read-pair methods analyse the mapping information of paired-end reads and their discordance from the expected span size and mapped strand properties. Sensitivity, specificity and breakpoint accuracy are dependent on the read length, insert size and physical coverage. Breakpoints are indicated by red arrows. Read-depth analysis examines the increase and decrease in sequence coverage to detect duplications and deletions, respectively, and predict absolute copy numbers of genomic intervals. Split-read algorithms are capable of detecting exact breakpoints of all variant classes by analysing the sequence alignment of the reads and the reference genome; however, they usually require longer reads than the other methods and have less power in repeat- and duplication-rich loci. Assembly algorithms have the most power to detect SVs of all classes at the breakpoint resolution, but assembling short sequences and inserts often result in contig/scaffold fragmentation in regions with high repeat and duplication content.

NGS data and SVs detection: an overview

None of the four main approaches to discovering structural variation using sequence data is comprehensive. When many algorithms and experimental methods are applied to the same DNA samples, a significant fraction of the validated variants remains unique to a particular approach. Each method has different strengths and weaknesses in detection, depending on the variant type or the properties of the underlying sequence at the SV locus. Although read depth is the only sequencing-based method to accurately predict absolute copy numbers, the breakpoint resolution is often poor [206] [207]. Read-pair approaches are powerful, but resolving ambiguous mapping assignments in repetitive regions is challenging and accurate prediction of SV breakpoints depends on very tight fragment size distributions, which can make library construction difficult and costly [195]. Similarly, split-read algorithms can be devised to detect a wide range of SV classes with exact breakpoint resolution; however, split read is currently reliable only in the unique regions of the genome . Sequence assembly promises to be the most versatile method by facilitating pair-wise genome comparisons; however, it has been shown to be heavily biased against repeats and duplications owing to assembly collapse over such regions [215] [216]. Its application to SV detection is not routine and will require substantial development.

Perhaps the greatest problem in using NGS to discover structural variation is the nature of the data. Sequence reads generated by the NGS platforms are considerably shorter than those produced by the capillary-based methods. Owing to the complex nature of human

genomes (for example, widespread common repeats and segmental duplications), there is considerable read-mapping ambiguity. Longer reads and inserts are needed to ameliorate this bias by increasing the specificity in read mapping. It is estimated, however, that >1.5% of the human genome cannot be covered uniquely even with read lengths of 1 kb [217]. Another concern is sequence coverage, defined as the average number of times each base pair in the genome is represented in an aligned read. Sequence coverage is an important factor in achieving high sensitivity and specificity in SV detection. Some projects may opt to sequence samples at low coverage for cost efficiency (for example, the 1000 Genomes Project uses two- to sixfold coverage); however, this reduces the power to discover structural variation.

The 1000 Genomes project

The 1000 Genomes Project (1000 GP) is the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation [182]. The goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in the populations studied. The 1000GP recently generated 4.1 terabases of raw sequence in two pilot projects targeting whole human genomes [182]. Sequence data produced by the 1000 GP provide an unprecedented opportunity to generate a comprehensive SV map. These studies comprise a population-scale project, termed ‘low-coverage project’, in which 179 unrelated individuals were sequenced with an average coverage of 3.6×, including 59 Yoruba individuals from

Nigeria (YRI), 60 individuals of European ancestry from Utah (CEU), 30 of Han ancestry from Beijing (CHB), and 30 of Japanese ancestry from Tokyo (JPT; the latter two were jointly analysed as JPT+CHB). In addition, a high-coverage project, termed the 'trio project', has been performed, with individuals of a CEU and a YRI parent-offspring trio sequenced to 42× coverage on average [182].

In the effort to generate a comprehensive catalogue of human diversity from this incredible amount of data, researchers from the 1000 GP have applied an highly selected ensemble of dedicated bioinformatic tools, based on different yet complementary approaches, for the detection of SVs from NGS resequencing data.

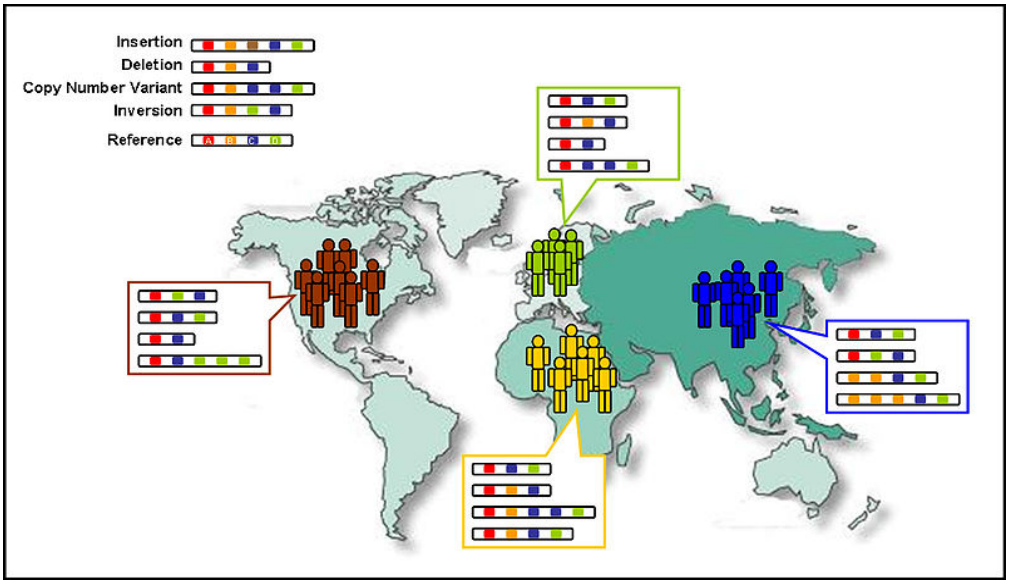
Apart from producing an outstanding catalogue of more than 1.32 millions distinct structural variations, the project succeeded in performing and extensive comparison of different approaches and tools therein used.

Comprehensive validation analysis based on customized hybridization arrays, and to a lesser extent on PCR were used to asses the validation rates of the individual methods. In this context methods designed explicitly to deal with pooled samples and combining the read-depth and paired read approaches like DinDel [218] and genome-Strip [219] achieved the better performances. This is unsurprising, considering that the major part of the data are obtained from low coverage sequencing of pooled individual. Furthermore it has to be remarked how, the validation strategy adopted, which is mainly based on custom validation arrays, is likely to have a major impact on the estimation of the validation rates and especially considering the inherent and systematic limitations of such approaches. This notwithstanding the

1000GP claims that the approach therein adopted enabled the consortium to characterize 95% of all the human SV, located in the genomic regions accessible to the NGS sequencing technologies.

Interestingly a recent survey based on careful and thorough analyses of Sanger resequencing reads, came to question this last proposition, showing how by the means of an “old-school” bioinformatics pipeline, it was possible to recover more than 2 millions indels from 98 millions of resequencing Sanger reads [179]. Intriguingly these SVs showed little coincidence (less than 25%) with the 1000 GP catalogue, demonstrating how far away we are from the complete characterization of genomic human diversity. In this light, the improvement of individual predictors of SVs from NGS data is of course desirable.

Figure 11: 1000 Genomes Project



In the 1000 GP 179 individuals were sequenced at low (avg 3.6X) coverage, including 59 Yoruba individuals from Nigeria (YRI), 60 individuals of European ancestry from Utah (CEU), 30 of Hanjuang ancestry from Beijing (CHB), and 30 of Japanese ancestry from Tokyo (JPT). In addition, a high-coverage project, termed the ‘trio project’, has been performed, with individuals of a CEU and a YRI parent-offspring trio sequenced to 42× coverage

Read pairs, small indels and 1000 GP

The read pair (RP) approach takes advantage of information such the expected span and orientation of mated resequencing reads to infer the presence of structural variants with respect to a reference genome.
This approach was originally designed for the detection of large SV (what exactly can be considered large depends on the initial insert size of the DNA library) using Sanger mated data. However, the principle recently been reimplemented in a plethora a programs for the analysis of NGS RP data (see above). Obviously the performance attainable by this type of approach is strongly influenced by quality and in particular the tightness of insert size distribution of the resequencing library.
This issue is particularly relevant with NGS data. Indeed as the size of NGS library inserts are typically short, and accurate gel separation of such short molecules is difficult, insert size distributions display great variability.
As this extreme variability hinders the application of simple and standard statistical tests, the sensitivity of NGS RP based approaches is rather limited in the detection of short indels. Indeed it is no coincidence, that with a few natable exceptions, all the RP NGS tools developed to date are intended to be used for the detection of long SV only.
The classic approach adopted is simple and can be effective: RPs are mapped on the reference genome and an arbitrarily cut-off value (typically the mean of the insert size +/- 2 units of standard deviation) is used to discriminate between RPs mapping at expected or aberrant distances. Finally, locations of indels are inferred from genomic clusters of aberrantly mapping RPs.

The first program specifically developed for detecting short indels from NGS data was MoDiL [203]. MoDil redefined the usual protocol for the analyses of RP data by introducing the concept of local insert-size distributions. MoDil does not use a cut-off value is to identify “anomalous pairs”, rather, “local” distributions of insert -size are calculated from the ensemble of RP reads mapping to a given locus (defined by a sliding genomic window). Under the assumption that the insert-size distributions are Gaussian, MoDil then compares each local distribution to the global distribution of insert-size (calculated from all the RPs) using the a Z-test. When significant differences are found, indels are called accordingly. Apart from identifying indels, MoDil implements a procedure for discriminating between homozygous and heterozygous events. To this end, it adopts an expectation maximization algorithm (maximum likelihood) and a log-likelihood statistical test. MoDil calculates the likelihood of the data underlying a local insert size distribution to be derived from a single (homozygous) or double (heterozygous) Gaussian distribution, and then compares these probabilities with a log-likelihood test. If the two distribution model is significantly more likely the indel is classified as heterozygous.
MoDil uses sliding windows of length equal to an insert-size and overlapping by 20 bp. Local insert size distributions are calculated and analysed within each window - taking into account the insert length of all the RPs mapping within each window.

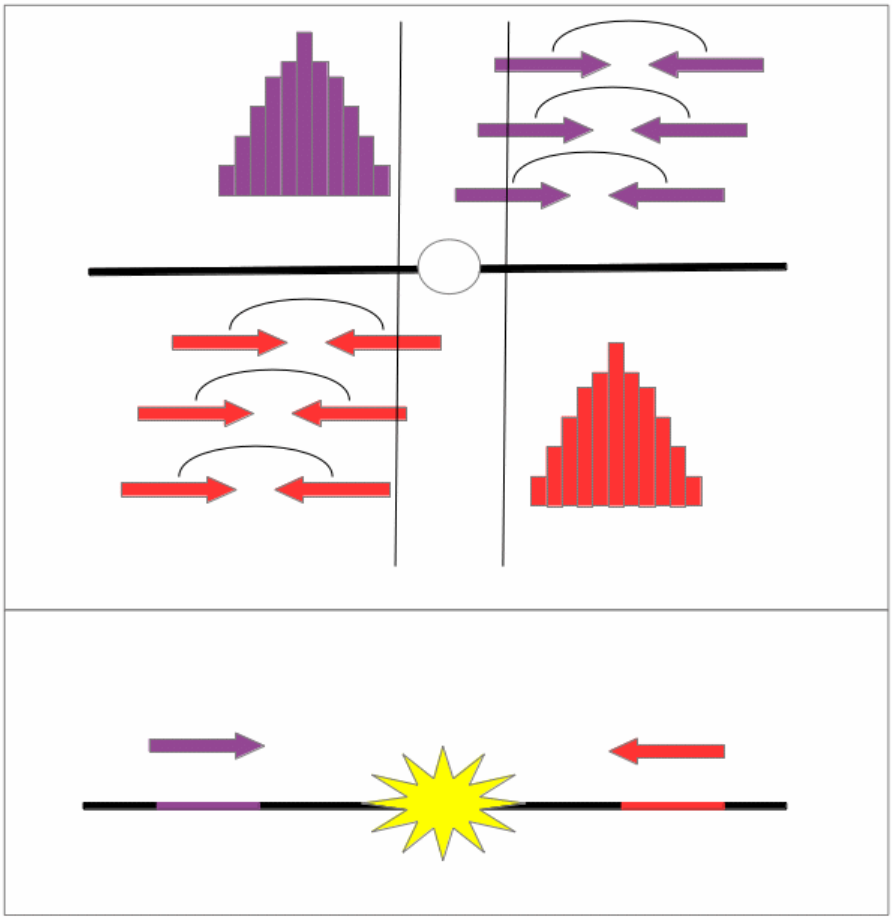
Instead of assuming any particular shape of the insert-size distribution, BreakDancer [202] uses a more appropriate, non parametric, Kolmogorov Smirnov (KS) test. More importantly it introduced the

concept of RP directionality. Indeed for a given genomic position (or genomic window) 2 separate local distributions are calculated, one for RPs pointing downstream and the other for RPs pointing upstream of the position in question. To asses the presence of indels, both of the local distributions are tested again the global one. As a result of this procedure each assayed window is marked with an orientation (upstream or downstream) and a "classification" (aberrant or normal). Indels are called when aberrant windows indicating the same type of event (insertion or deletion) are situated at an amenable distance from each other (defined by the presumed size by the indel) and pointing towards the other (as exemplified in Figure 12). While this approach enables the detection of some small indels, BreakDancer, like other methods also implements a classical "cut-off" based statistical approach. Breakdancer was demonstrated to be much more effective than MoDiL in the detection of small SVs [202], however the application of the approach is limited by the fact that the threshold for the KS test must be set empirically for each different RP library, in order to achieve results consistent with those reported in the original paper [202].
Furthermore even though BreakDancer attains greater sensitivity than MoDil, its overall performance in the detection of indels, and in particular those shorter than 20-30 bp, were not good enough to justify its usage for this specific task in large scale SV-detection projects [182].

Considering the limited applicability of both MoDil and BreakDancer, neither of this programs were to used to detect small indels in the 1000GP. Indeed, the fact that RP based approaches have not been used for the detection of small SV in the most important project for the

characterization of human genomic diversity reflects the widespread idea that – given the aforementioned variability in insert size distribution - they are suitable only for the detection of larger variants.

Figure 12: The Breakdancer approach



For any given genomic windows (white circle) BreakDancer discriminates between the MP mapping with forward (purple) and reverse (red) orientation. Purple reads are used to calculate the insert size downstream from the circle. While Red reads are related to the insert-size distribution downstream. Each distribution is tested against the global insert size independently. When 2 genomic windows supporting the presence of an indel are found with proper orientation and distance the appropriate indel is called

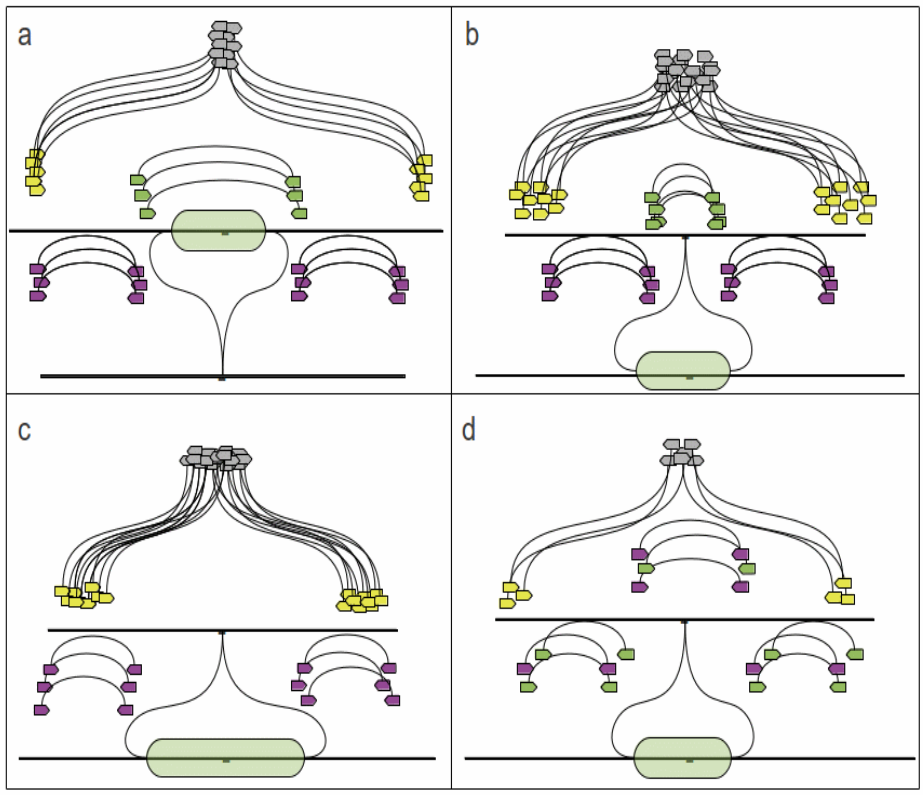
SVM^2. Small indels and RP: a more comprehensive approach

Breakdancer, MoDil and other programs for the detection of SV from RP data, use measures of insert size perturbation as their only source of information. However, it is apparent (see Figure 13) that different types of genomic rearrangements, even those involving only a few base pairs of DNA, are expected to generate complex and particular signatures of patterns of mapping of RPs and not only perturbations in the insert size.

The presence of Broken Pairs (BPs), reads for which only one of the 2 mates could be mapped on the reference genome, is possibly the most prominent of these “additional” signatures. BPs arise from the generation of new genomic junctions (breakpoints) as a consequence of genomic rearrangements and from the insertion of novel sequences in the “donor” genome (or from sequencing errors). Indeed any read covering a new junction in the “donor”, by definition will not map in the expected position on the reference genome - contrary to its mate, which is still expected to map. Importantly the “surviving” mate is not expected to map anywhere on the reference but more or less exactly at a distance of 1 insert-size from the breakpoint itself. Therefore identifiable enrichment in BP reads should be a good indication of the presence and position of a rearrangement - especially as different types of structural variants are likely to be associated with different patterns of BP reads. For example, in the presence of a deletion in the donor we expect a narrow peak of BP reads, as only reads mapping on the rearrangement junction will fail to map, while in the case of an insertion in the donor genome, this peak will extend the length of the insertion towards the rearrangement junction (as shown in Figure 13).

Furthermore when a “long” deletion occurs in the donor, a gap in coverage of proportional length is expected in the reference.

Figure 13: Expected patterns of reads mapping



Schematic representation of the expected pattern of mapping of reads on a reference genomic sequence in the case of a deletion (a) an insertion shorter than the insert-size (b) an insertion longer than the insert-size (c) and in the presence of a particularly variable region (d). Each SV event (green bubble) generates broken-pairs (yellow arrows) in a specific fashion: in the case of a deletion (a) we expect a sharp peak, while for short insertion (b) we expected a broader one and eventually whence the insertion becomes too long, all we can see is a peak of broken pairs as broad as the insert-size. Furthermore, by looking at their orientation, we can distinguish between RP (purple) mapping upstream or downstream respect to an hypothetical breakpoint. Finally (d) illustrates that there can be some misleading signals in the case of particularly variable and localized regions, which can also lead both to the generation of peaks of broken pairs and to subtle shifts in apparent insert size distributions (although without the directional specificity observed for indels).

While BPs are routinely used in the detection of novel sequence insertions [182], they are not incorporated in existing RP-based tools. I hypothesized that their integration (along with other information regarding mapping patterns), could assist such instruments in the identification of short indels.

The main project in my PhD studies was the development of a new and more comprehensive approach for the detection of small indels from RP data. The aim was to demonstrate that, if carefully designed, RP methods can be useful even in the detection of very short indels.

In this context I developed the SVM^2 software package. SVM^2 (Structural Variation Mapping using Support Vector Machines), is a novel SV finder based on an highly efficient supervised learning approach: Support Vector Machines. The core idea behind SVM^2 is that avoiding the use of stringent statistical cutoff values by employing a series of *ad-hoc* descriptors of read mapping patterns and supervised learning it might be possible improve sensitivity of SV detection without loss of specificity. As the full details about the implementation and functionality of the software are reported in the attached manuscript, a rather simple overview of the principle and statistics adopted will be presented in the main body of this thesis.

Support Vector Machines

The core of the SVM^2 program is a multi-class SVM classifier, which can be trained to learn from known examples, how to recognize different classes of SVs. In its current implementation the software is trained to discern between normal positions and 4 different classes of events:

small insertions, long insertions (longer than the insert size), deletions, and variable regions. The necessity of discriminating between long and short insertions, arises because, as shown in Figure 13, donor genome insertions that are longer than the library insert size must be identified from only the BP information - it is evident that in such cases, no RPs spanning the novel inserted sequence are expected to map to the reference genome. Variable regions are used to discern real indels from SNP rich regions.

Support Vector Machines are an ensemble of statistics/computational techniques that have been widely employed in biological classification problems. SVM uses a series of training data points, each known to belong to one of two (or more) classes of origin and described by a number of quantitative features, and, having transformed them into a higher dimensionality than allowed by the number of associated features and through the use of a kernel function, identifies the hyperplane that maximizes their separation by class in a multidimensional space. Once the optimal discriminating function has been established, it is used to classify unknown instances. Several software libraries implementing SVM are freely available and the method can be adapted to function in multiple category classification problems.

A time saving heuristic

In any genome resequencing project, the vast majority of the 2 genomes under study are expected to be identical, or to contain only a few SNPs. Even if SVM^2 adopts an efficient and quick implementation of the SVM algorithms, it is evident that applying the program to each position on a big genome, such as the human, would result in very long

execution time. To avoid useless calculations and attain reasonable execution times (1 day for a human genome at 40X coverage), SVM^2 coarse filters to identify, a-priori, regions that are potentially anomalous with respect to the expected pattern of mapping. Potentially anomalous position are identified as those showing an increase of BP reads above background (BP to RP ratio in the highest 5% of the genome) or those displaying an obvious perturbation in insert size (more than 1.5 s.d). The full analysis pathway is only invoked for such positions.

Formulation of Genomic Windows

The underlying rationale of my approach is use a series of sliding windows, “centered” upon each position along the genomic sequence and to calculate statistics describing distributions of RP mapping distances, Broken Pairs and overall read coverage around the position, with the expectation that these dynamics should change as the position considered approaches an SV event. The different features used to describe these patterns are measured in different portions of the window, according to the expectations previously described.

The windows used to assay the read mapping pattern are of variable length because any anomalous position identified by the coarse filters represents an hypothesis of a structural variant. SVM^2 tries to identify a site in the reference genome that is beyond the presumed SV event and corresponds to the expected position of mapping of the partners of reads mapping to original anomalous position. To this extent longer windows are used while assaying potential deletions, as it is expected that

in this case the mates will fall more distant apart on the reference genome (see the SVM^2 manuscript in Appendix 1).

Features to describe read mapping patterns

Perhaps the most crucial task is to formulate an informative set of features for discriminating between instances of the different classes of event under study. To discriminate between "normal" sites and each of the 4 different types of SV described above, SVM^2 utilizes 74 distinct features which are designed to measure: the presence/absence and position of BPs, insert size perturbations (according to different statistical measures) and the presence, length and number of resequencing coverage gaps.

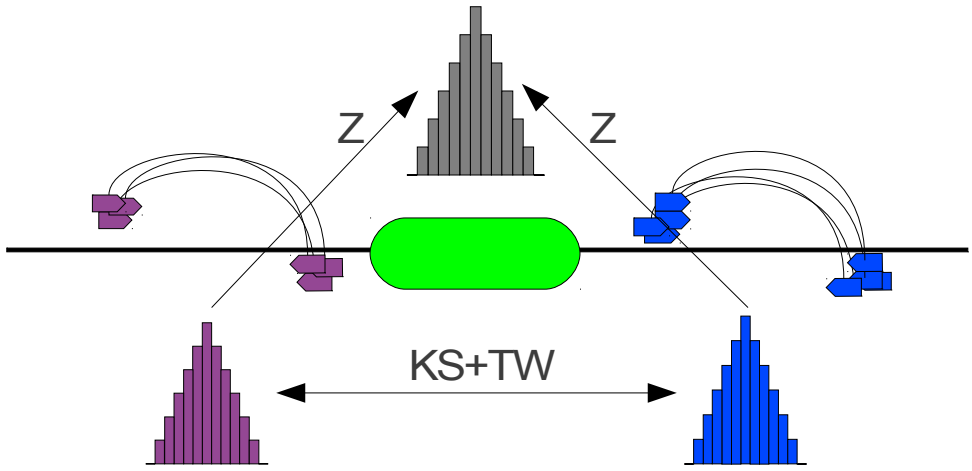
To measure insert size perturbations, as shown in Figure 14, 3 different statistical tests are applied at different levels of significance for all sites in a window spanning from the "starting position" to, effectively, a position one insert size downstream of the alerting position (all operations are performed in a strand-specific manner). To take full advantage of the "directionality of the data" SVM2, like BreakDancer, calculates the upstream and downstream distributions of insert size for any genomic position and compares these to the global insert size distribution, under the expectation that that the distribution of insert sizes of reads pointing towards an SV event will be different form the global one, but that those on the opposite strand (pointing away from the event) will resemble the global distribution. However, unlike BreakDancer, the 2 distributions are also compared to each other by the

means of a Student T Welch and a Kolmogorov Smirnov test (it is expected that these will differ in the vicinity of SV). The adoption of these multiple and independent statistical tests allows the identification of “random” local insert size perturbations and those caused by systematic biases but not SV.

For any genomic window, SVM2 performs the aforementioned tests for each position For each of 2 separate frequency histograms (one for insertions and another for deletions), four bins corresponding to different Pvalue ranges are stored. As the Z-test is applied twice (see above) the net result is that 32 features derive from these calculations. Each bin from each histogram is then used as a feature for the SVM.

To account for the presence and number of eventual gaps in coverage SVM^2 uses 2 additional features to record the length of the longest coverage gap and the number of gaps in coverage encountered in the same window that is used for the insert size statistics.

Figure 14: Statistical test used by SVM^2

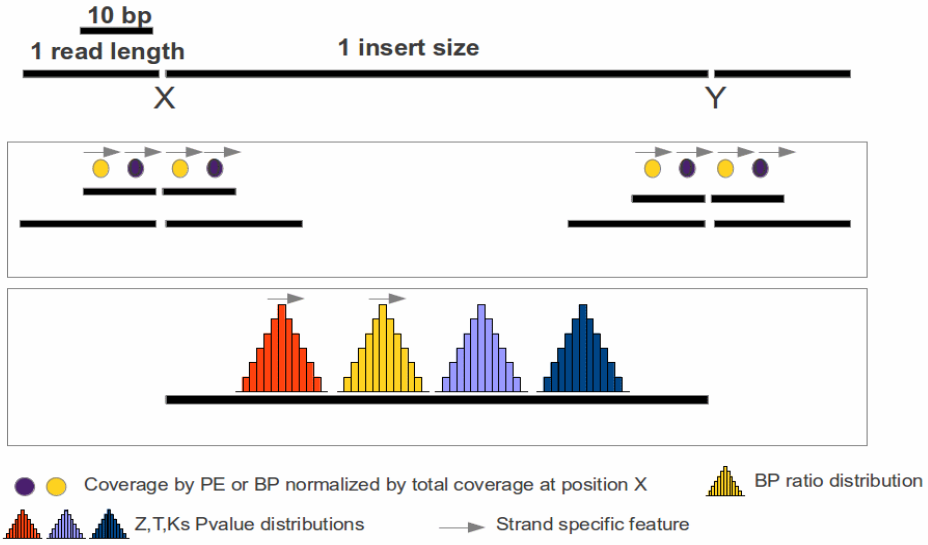


To take full advantage from the directionality of paired reads SVM^2 uses multiple test to compare insert size distributions. The upstream distribution (in purple) ad the downstream distribution (in blue) are compared to the global insert size distribution (in grey) by the means of a Z test. While to compare the 2 local distributions the one to the other a Kolgorov Smirnov (KS) along as a Student T Welch test are used.

Most SVs are very short [220] (too short to actually perturb the insert size distribution in a highly significant way) and the objective of the current method is to increase sensitivity by augmenting insert size data with BP-derived signal. To assay if the mapping pattern of such reads is amenable with the presence of an indel, 36 different features are used. To verify the presence of BP reads in the vicinity of the starting position 4 arbitrarily chosen windows are used. Such windows are used to measure for each strand the average number of RP and BP. In particular 2 windows of 10 bp in length as well as 2 windows of the same length as the resequencing reads, positioned immediately upstream and downstream of the starting position are used to calculate

the average number of BP and RP downstream and upstream from the invoking positions. Thus resulting in 16 features. Four windows, similar to those based on the starting position, but defined by the position 1 insert size downstream, are also used to measure the mapping pattern of RP and BPs - resulting in 16 features. Finally the ratio between BP and RP is assayed along the whole window used for the coverage in insert size statistics for each strand (direction) and two histograms with 4 bins are used to record the number of visited positions showing a BP to RP rate particular intervals. Again each of the bins is used as a feature in the SVM (8 features total). A graphical overview of the features used by SVM2 is depicted in Figure 15.

Figure 15: Features used by SVM[2]



Representation of the localization and strandness (arrow) of the features used by SVM[2]. X is the position invoking the SVM, while Y is the genomic position at which mates of X are expected to be found PE= paired end, BP=broken pairs Z= Z test, T=T-Welch test KS= Kolmogoroff Smirnov test. Features with an arrow on top are calculated on both strand.

Post-processing and Estimation of Event sizes and types.

Once the features have been calculated, the SVM at the core of SVM2 program is used to classify the invoking position. When the all the genome has been scanned post processing procedure is applied to clusters contiguous genomic positions belonging to the same class of events. Finally when clusters of the same type are found within an amenable distance and with proper orientation an SV of that type is called. For a detailed description of the post-processing steps, the reader is referred to the attached manuscript (appendix 1, SVM^2 manuscripts, methods).

The size of the event is estimated as the difference between the mean mapping distance of reads spanning the predicted event and the global mean mapping distance and a test of heterozygosity, similar to the one used by MoDil is also implemented.

Training the SVM

To train SVM^2 to discriminate between different classes of SVs, regions presumably conserved between the donor and reference genomes are identified after the mapping of resequencing reads on the reference genomes as windows of 10 Kb or longer where the overall BP rate is under 10% (for each position), no gaps in coverage are found, and the coverage depth is more than a quarter and less than 4 times the expected depth. SVs of know size and type are inserted into these conserved sequence contexts *in silico* to simulate the effect of the different types of

SVs. Sequence reads are then mapped back to these modified regions and a series of features describing read mapping patterns are calculated and used to train the SVM.

Comparison with other tools

To compare the performance of SVM^2 to other tools using real RP resequencing data we have taken advantage of publicly available RP resequencing data from an anonymous human donor generated with the Illumina technology [38]. The peculiarity of this dataset is that a large and consistent set of SV was previously detected and validated using low coverage (0.3X) longer insert (Sanger + 40Kb fosmids) from the same individual [178] thus it has been widely used as a benchmark to compare different SV detection tools [160] [161]. Indeed, the Kidd et al. data [178] was recently subjected to a second analysis [179] and here we consider the union of both sets of predictions as a validated indel set (265264 events).

We compared the performance of our tool with that of BreakDancer [158] that, in previous studies of the same dataset, exhibited the highest sensitivity and specificity among RP-based tools in detecting relatively small indels (indicatively greater than 10bp) and PinDel, a popular split mapping approach [170].

The sensitivity (the proportion of indels in the validation set that was recovered by each method, as a function of the validated size of the indel) of each method is shown in Fig17 (and supplementary Table 2). Under this criterion, SVM^2 outperforms BreakDancer in all size categories, overall recalling 4.5 times as many events. As expected, the

split mapping method (PinDel) is more sensitive in the detection of very small indels (up to 5bp) although SVM^2 recalls a larger proportion of events over this threshold.

The number of predictions and apparent specificity by predicted event size (proportion of predicted indels of coinciding with any indel in the validation set as a function of the predicted size of the indel) for each method is shown in Figure 16 (and supplementary Table 3). It should be noted that the genome coverage of the Kidd et al. data, 0.3x, represents the maximum theoretical specificity in this benchmark. All of the evaluated methods demonstrate similar overall performance. PinDel in particular shows a marginally better specificity with respect to the smallest events (<10bp) while the size/specificity profile of SVM^2 and BreakDancer are relatively uniform at around 26-27% “validation” for each size bin. Both SVM^2 and BreakDancer suffer an apparent loss in specificity with regard to predicted events greater than 30bp or more. This last observation is likely a stochastic effect due to the fact that larger rearrangements constitute a very small minority of SVs. To partially ameliorate the low genome coverage of the validation set, we compared predictions to all events in dbsnp130 which contains more than 4.2 million known rearrangements derived mostly from Sanger sequencing data. 81.5%, 80.6% and 80.4% of the predictions made by BreakDancer, PinDel and SVM^2 respectively correspond to known human SV events. The specificity by size profile strongly resembles that observed with the Kidd SVs (Supplementary Figure 3 and Supplementary Table 4). Cross referencing the predictions from the various methods with the collection of human genomic SVs provided by the 1000 genomes project, derived from NGS data(1.32 million events)

showed that 61% of BreakDancer predictions, 69% of SVM^2 predictions and 80.7% of PinDel predictions were coincident with events present in that database. 54% of the Kidd/Sanger based validation set events were present in the 1000 genomes database (Supplementary Figure 3 and Supplementary Table 4).

The Venn diagram in Figure 17 shows the overlap of validated calls made by SVM^2, BreakDancer and PinDel. The union of all methods identified 108158 of the 265264 events recovered from the Sanger data (41%). 24842 (23%) are found by PinDel and SVM^2, 9122 (8.5%) are identified by BreakDancer and SVM^2. Only 1730 (1.5%) are found by BreakDancer only while 49972 (46%) are unique to PinDel and 20974 (19%) are unique to SVM^2. 87% of validated BreakDancer predictions are also made by SVM^2. Taken together, these observations confirm that the incorporation of additional mapping information in SVM^2 allows a great increase in sensitivity over methods that use only mapping distance information. Furthermore, it is evident that a notable proportion of events are recovered by SVM^2 but not other methods. When compared to the sensitivity profile by event size (Figure 16) it is evident that SVM^2 identifies a significant number of small events not detected by PinDel.

Figure 16: Sensitivity and specificity of SVM^2

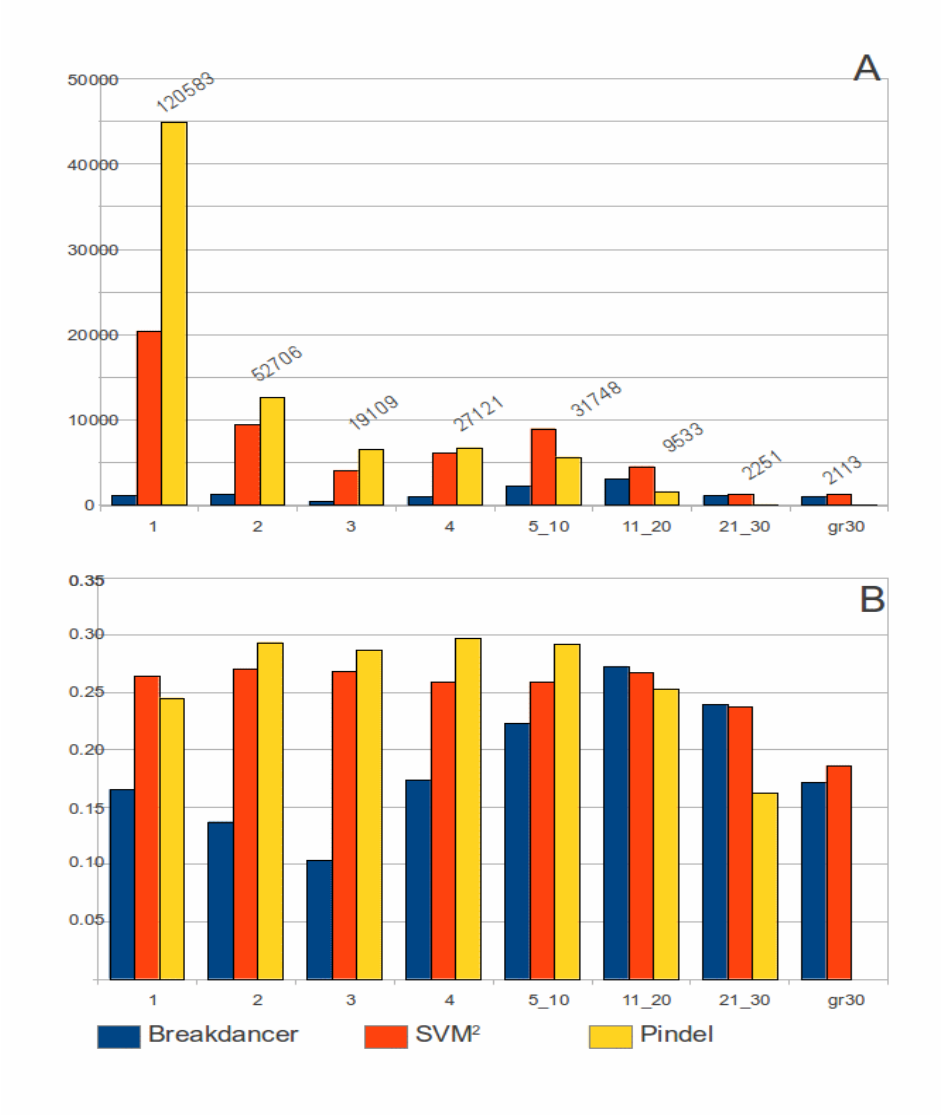
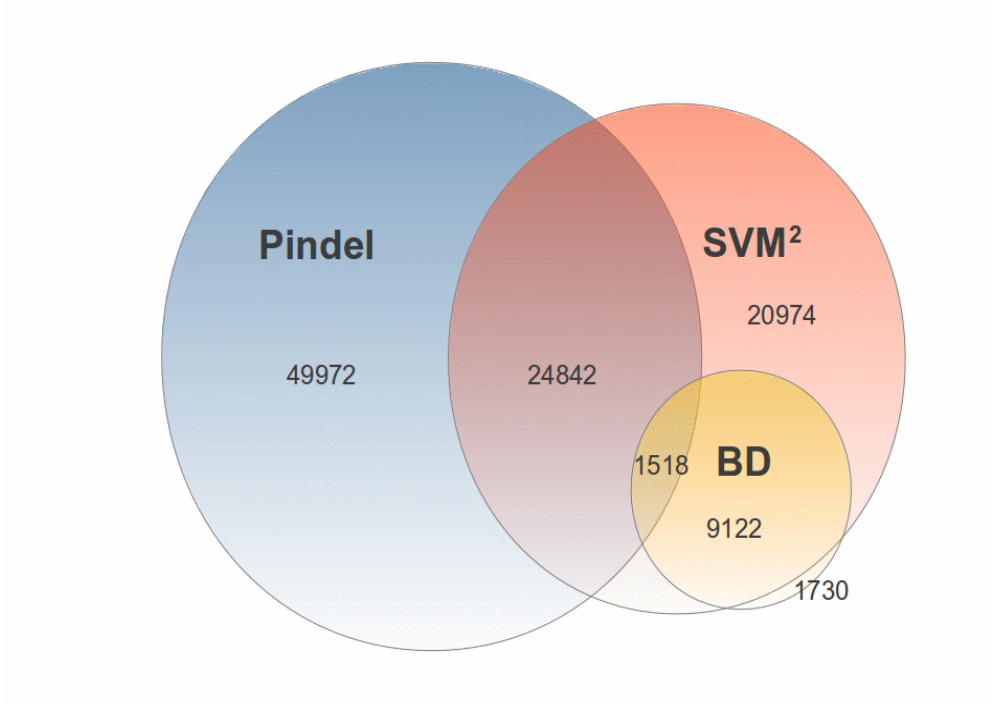


Fig 16A Number of indels from the Kidd dataset (binned by size of event in bp) recalled by each method.
Fig 16B Proportion of predicted indels (binned by predicted sizes) that are validated by an indel in the kidd et al (0.3X theroetical coverage).
Size bins: size≤1,size≤2,size≤3,size≤4,5≤size≤10,10<size≤20,20<size≤30,size>30

Figure 17: Overlap between the prediction by SVM^2 BD and Pindel



Venn diagram showing intersection between validated (by kidd) predictions by each method

Accuracy of classification and genomic context of predictions

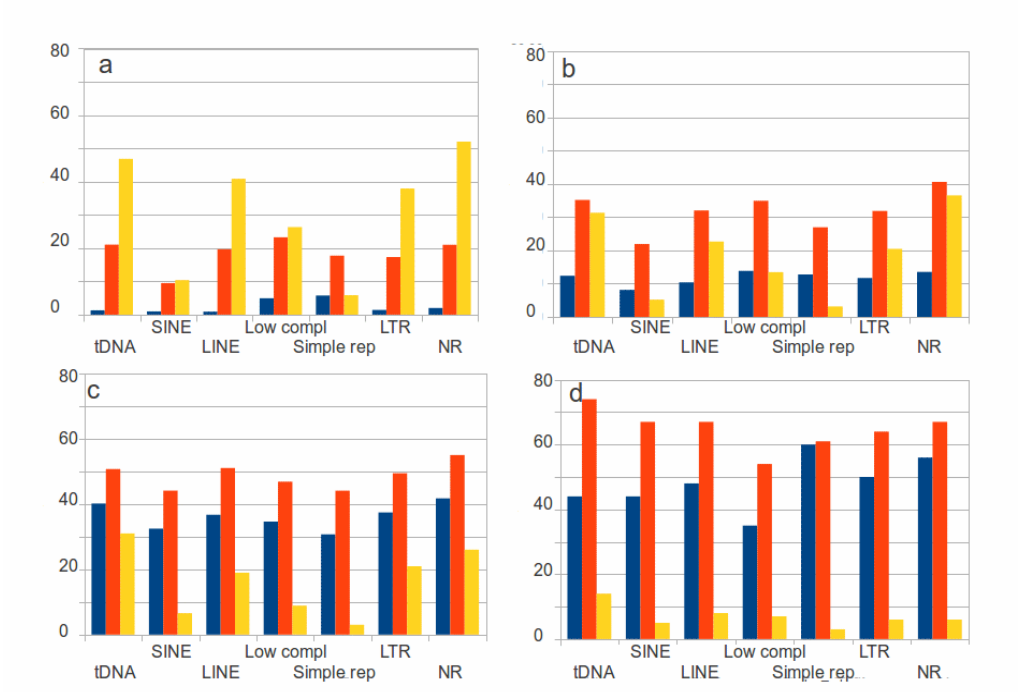
Next, we asked whether, for a series of size range bins the sensitivity by genomic context showed obvious differences between methods. Figure 19 confirms that for the smallest events (≤5bp), PinDel outperformed the other methods in most genomic contexts. However, the sensitivity of SVM^2 in SINEs and low complexity regions was comparable to that of PinDel, while in simple repeats SVM^2 outperformed PinDel). As expected – given the small number of predictions by BreakDancer in this

size range, the sensitivity was low. For events of between 6 and 10bp in size, SVM^2 was the most sensitive method dramatically outperforming BreakDancer in all genomic contexts. PinDel was almost as sensitive as SVM^2 in DNA transposons and non-repetitive DNA. As event size increases, PinDel shows decreasing sensitivity particularly in low complexity regions and simple repeats (an inevitable property of split mapping methods). Even for larger (>20bp) events, which BreakDancer was designed to detect, SVM^2 is more sensitive in all genomic contexts. It is notable that, overall, SVM^2 and BreakDancer seem to show much less dependence on genomic context than PinDel.

We were intrigued by the difference of apparent specificities between methods previously observed when using the 1000 genomes SV catalog (but not when using dbSNP or the Kidd et al. data) as a validation set and by the relatively large proportion of the small (<10bp) events found by SVM^2 but not PinDel that fall in low complexity and simple repeat regions (10037/19274, 52%). We reasoned that these observations might be linked by the fact that the 1000 genomes catalog used split mapping to identify small events, and showed that a notable proportion (>97%) of the part of the genome deemed “inaccessible” by their low coverage data, fell in regions annotated as “high copy repeats or segmental duplications”. Accordingly, we investigated the genomic distribution of predictions validated by Sanger sequencing but not by the 1000 genomes catalog by event size and method. We observed that a relatively small proportion of the small events (<10bp) validated by the Kidd et al. data and predicted by PinDel but not supported by the 1000 genomes dataset, fall in low complexity regions and simple repeats (1483/16081, 9.25%), while the equivalent numbers for SVM^2 were

(5991/18450, 32%), suggesting that SVM², or similar methods, might effectively complement existing tools and pipelines in the detection of very short SVs, particularly in repetitive and low sequence complexity areas of the genome.

Figure 18: Empirical sensitivity in different genomic contexts



Numbers of events in the Kidd dataset, in different genomic contexts (tDNA=DNA transposon, LTR = long terminal repeats, NR= non repetitive), recalled at different size ranges ((a) size≤5, (b) 5<size≤10, (c) 10<size≤20, (d) size>20) by different methods.

Assembly of fungal genomes and development of a custom scaffolder program

Genome assembly and scaffolding

The emergence of next-generation sequencing platforms led to a resurgence of interest in the development of whole-genome shotgun assembly algorithms and software. DNA sequencing data from the Roche 454, Illumina/Solexa, and ABI SOLiD platforms typically present shorter read lengths, higher coverage, and different error profiles compared with Sanger sequencing data.

An assembly is a hierarchical data structure that maps the sequence data to a putative reconstruction of the target genome. It groups reads into contigs and contigs into scaffolds. Contigs provide a multiple sequence alignment of reads plus the consensus sequence. The scaffolds, sometimes called supercontigs or metacontigs, define the contig order and orientation and the sizes of the gaps between contigs. Scaffold topology may be a simple path or a network. Most assemblers output, in addition, a set of unassembled or partially assembled reads. The most widely accepted data file format for an assembly is FASTA, wherein contig consensus sequence can be represented by strings of the characters A, C, G, T, plus possibly other characters with special meaning. Dashes, for instance, can represent extra bases omitted from the consensus but present in a minority of the underlying reads. Scaffold consensus sequence may have N's in the gaps between

contigs. The number of consecutive N's may indicate the gap length estimate based on spanning paired ends.

Assemblies are measured by the size and accuracy of their contigs and scaffolds. Assembly size is usually given by statistics including maximum length, average length, combined total length, and N50. The contig N50 is the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the assembly. The N50 statistics for different assemblies are not comparable unless each is calculated using the same combined length value. Assembly accuracy is difficult to measure. Some inherent measure of accuracy is provided by the degrees of mate-constraint satisfaction and violation. While a high N50 is of course indicative of an assembly which contains many large contigs/scaffolds, it is not necessarily and indication of the correctness of an assembly. Meaningful evaluation of correctness relies on alignment to reference sequences, but often de novo assembly (rather than building an assembly using a reference sequence) is necessary because a suitable reference is not available.

WGS and its limitations

From the assembly point of view, all the sequencing technologies developed to date, from "old" Sanger to NGS, suffer from the same inevitable limitation: read-lengths are much shorter than the smallest genome. To overcome this limitation the WGS (whole-genome-shotgun) approach over-samples the target genome with short reads from

random positions. The reconstruction of the target sequence is then carried out by a dedicated assembler program. However, even the more sophisticated assembler program suffer from inherent limitations, which in the end are ascribable to the limited lengths of the reads and the properties of the genome under reconstruction.

Genomic regions that share perfect repeats can be indistinguishable, especially if the repeats are longer than the reads. For repeats that are inexact, high-stringency alignment can separate the repeat copies. Careful repeat separation involves correlating reads by patterns in the different base calls they may have [221].

Repeat separation is assisted by high coverage but confounded by high sequencing error. For repeats whose fidelity exceeds that of the reads, repeat resolution depends on “spanners,” that is, single reads that span a repeat instance with sufficient unique sequence on either side of the repeat [221].

Repeat resolution is made more difficult by sequencing errors. Software must tolerate imperfect sequence alignments to avoid missing true joins. However, error tolerance leads to false positive joins. This is a problem especially with reads from inexact (polymorphic) repeats.

WGS assembly is confounded by non-uniform coverage of the target. Coverage variation is introduced by chance, by variation in cellular copy number between source DNA molecules, and by inherent bias of amplification and sequencing technologies. Very low coverage induces gaps in assemblies. Coverage variability invalidates coverage-based statistical tests, and undermines coverage-based diagnostics designed to detect over-collapsed (or over expanded) repeats [221].

WGS assembly is also made more difficult by the computational

complexity of processing large volumes of data. For efficiency, all assembly software relies to some extent on the concept of a K-mer [98]. This is a sequence of K bases. In most implementations, only consecutive bases are used. Intuitively, reads with high sequence similarity must share K-mers in their overlapping regions, and shared K-mers are generally easier to find than overlaps. Fast detection of shared K-mer content greatly reduces the computational cost of assembly, especially compared to all-against-all pairwise sequence alignment [98]. A tradeoff of K-mer based algorithms is lower sensitivity, thus missing some true overlaps. The a potential overlap spans sharing K-mers is really a true overlap depends on the value of K, the length of the overlap, and the rate of error in the reads [85]. An appropriate value of K should be large enough that most false overlaps don't share K-mers by chance, and small enough that most true overlaps do share K-mers. The choice of k-mer length should be robust to variation in read coverage and accuracy and can vary according to read length, error rate and the nature of the genome under assembly [85].

Next-generation assemblers

The most commonly used and profitable approach for de-novo genome assembly of short reads relies on K-mer graphs [85]. The K-mer graph does not require all-against-all overlap discovery, it does not (necessarily) store individual reads or their overlaps, and it compresses redundant sequence. Conversely, the K-mer graph does contain actual sequence and the graph can exhaust available memory on large

genomes. Distributed memory approaches can ameliorate this constraint. The K-mer graph approach dates to an algorithm for Sanger read assembly [92]. The approach is commonly called a de Bruijn graph (DBG) approach or an Eulerian approach [92]. In this kind of graph, each node represents a k-mer and is connected to nodes that represent other k-mers that are shifted by one base. Given perfect data – error-free K-mers providing full coverage and spanning every repeat – the K-mer graph would be a de Bruijn graph and it would contain an Eulerian path, that is, a path that traverses each edge exactly once. The path would be trivial to find making the assembly problem trivial by extension. Of course, K-mer graphs built from real sequencing data are more complicated.

Thus, if the data is ideal, assembly is a by-product of the graph construction. The graph construction phase proceeds quickly using a constant-time hash table lookup for the existence of each K-mer in the data. Although the hash table consumes extra memory, the K-mer graph itself stores each possible K-mer at most once, no matter how many times the K-mer occurs in the reads. In terms of computer memory, the graph is smaller than the input reads, given that some reads share K-mers.

Three factors complicate the application of K-mer graphs to DNA sequence assembly.

DNA is double stranded. The forward sequence of any given read may overlap the forward or reverse complement sequence of other reads. One K-mer graph implementation contains nodes and edges for both strands, taking care to avoid output of the entire assembly twice [92]. Another implementation stores forward and reverse sequence together

as cognate half-nodes with the constraint that paths enter and exit the same half [96]. Yet another implementation represents alternate strands in a single node with two sides, constraining paths to enter and exit opposite sides [87].

Real genomes present complex repeat structures including tandem repeats, inverted repeats, imperfect repeats, and repeats inserted within repeats. Repeats longer than K lead to tangled, complex K-mer graphs that complicate the assembly problem. Perfect repeats of length K or greater collapse inside the graph, leaving a local graph structure that resembles a rope with frayed ends; paths converge for the length of the repeat and then they diverge. Successful assembly requires separation of the converged path, which represents a collapsed repeat. If the repeat is perfect, the graph, by definition, contains insufficient information to disambiguate the repeat. Assemblers typically consult the reads (and possibly the paired end partners of reads falling within the repeat), to attempt to resolve these regions.

A palindrome is a DNA sequence that is its own reverse complement. Palindromes induce paths that fold back on themselves. At least one assembler avoids these elegantly; Velvet [96] requires K, the length of a K-mer, to be odd. An odd-size K-mer cannot match its reverse complement.

Real data includes sequencing errors: DBG assemblers use several techniques to reduce sensitivity to this problem. First, they pre-process the reads to remove error. Second, they weight the graph edges by the number of reads that support them, and then remove the poorly supported paths. Third, they convert paths to sequences and use sequence alignment algorithms to collapse nearly identical paths. Many

of these techniques derive from the Eulerian family of assemblers.

Next generation protocols for paired reads

The development of dedicated protocols for the generation of “paired reads” from NGS technologies, represents an immediate and widely recognized solution to the critical limitations of these new methods. The advantage of the “pairing” technology is its ability to uncover linkages between the two ends of DNA fragments. Using this unique feature, unconventional fusion transcripts [223], or genome structural variations [196], can be unraveled by paired reads analysis. Additionally the distances between the two ends of size templates may be used to relate discrete contigs in assembling genomes [103] [104] [105], as long as the “insert size” of the sequencing library is well characterized. Furthermore, genomic regions containing repeats can be oriented and positioned by their connectivity to sequence specific regions offered by paired sequences.

The underlying principle used in paired sequencing is an old yet simple one and this technology has been used profitably since its first theorization in the early 80s [224]. In brief the idea is to sequence both ends from a DNA/RNA molecule of known length.

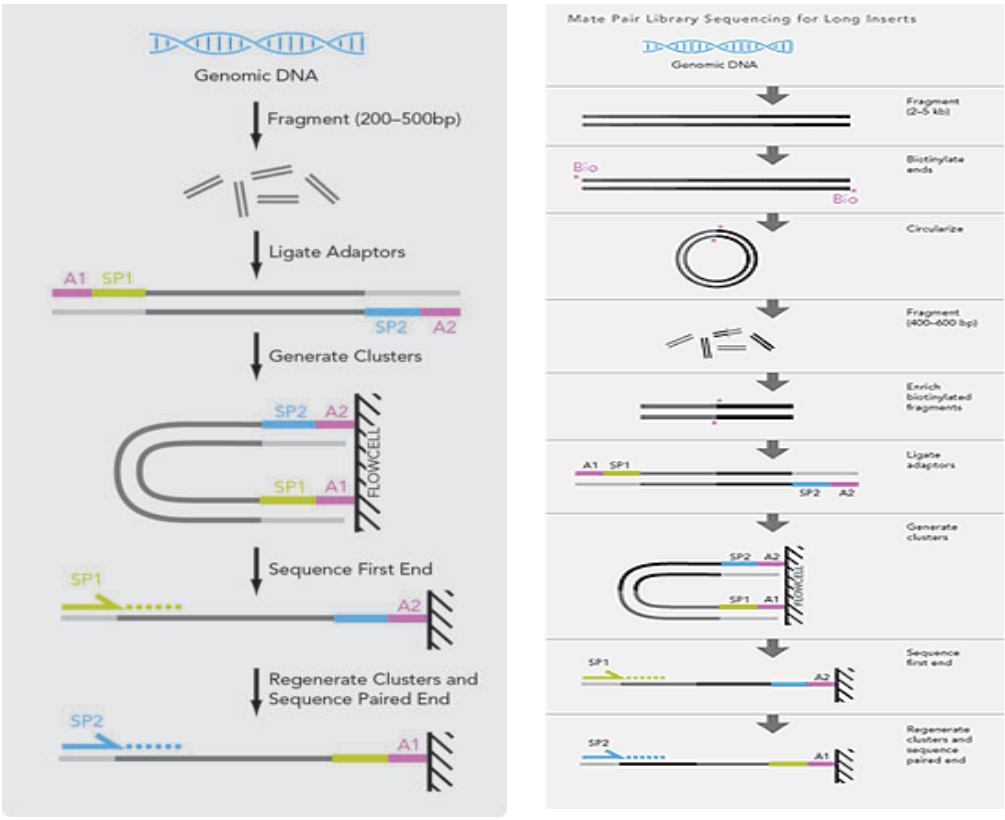
Every NGS technology now on the market devised its own protocols for the generation of paired reads. In this summary I will focus only on the protocols used by the Illumina technology (as they are the most commonly used, and produce similar data to those employed by other platforms).

Illumina uses two distinct protocols for the generation of read pairs. Usually the term “paired end” (PE) is used to refer to paired reads generated from short-insert libraries (200-600 bp), while data generated from libraries with a larger insert size (2 to 5 Kb) are referred to as “mate pairs” (MP). The two protocols differs in the library preparation, but are both based on the same technical device: the so called Illumina “Paired end module” (PEM). According to the manufacturer description, ”The Paired-End Module is a fluidistics station that attaches to the Genome Analyzer”. This device directs the regeneration and amplification operations to prepare the templates for a second round of sequencing. First, the newly sequenced strands are stripped off and the complementary strands are bridge amplified to form clusters. Once the original templates are cleaved and removed, the reverse strands undergo sequencing-by-synthesis. Then the second round of sequencing occurs at the opposite end of the templates.

The Illumina protocol for the generation of paired end (PE) data is relatively straight-forward, with just some slight modifications respect to the single end protocol. After a size selection step, typically performed by the means of a 2D gel electrophoresis, the data are fed to the machine and paired sequences are obtained by the use of the PEM.
The protocol for the generation of longer MP reads is more laborious.
After the size selection, 2-5 Kb fragments are end-repaired with biotin labeled dNTPs. The DNA fragments are then circularized, and non-circularized DNA is removed by digestion. Circular DNA is fragmented and fragments biotin labels (corresponding to the ends of the original DNA ligated together) are affinity purified. Purified fragments are end-

repaired and ligated to Illumina Paired-End sequencing adapters. The final prepared libraries consist of short fragments made up of two DNA segments that were originally separated by several kilobases. These libraries are finally sequenced utilizing the “standard” paired end Illumina procedure. As the affinity purification of biotinilated DNA is not completely efficient, it expected that contaminat PE reads will natively be present in any Illumina MP library, in a proportion that according to the manufacturer varies between 10 and 15%. Resulting reads from the final fragment ends are inverted with respect to their original genomic orientation.

Figure 19: Illumina PE and MP protocols



Differently from the PE protocol the MP protocols uses DNA circularization to create physical fragment of DNA end which where originally distant by 3 to 5 Kb. The sequencing of the mates is then performed using the standard Illumina protocol and the Illumina PE module.

Genome scaffolding and Next Generation scaffolders

To take full advantage of “paired” sequencing data, usually in any de-novo genome assembly project, the information relative to the expected distance and orientation of the mates is used to resolve the orientation and order of the contigs, in a process known as scaffolding.

The inputs to the scaffolding step are the contigs produced by an

assembler, the mappings of the paired reads to the contigs and the insert sizes of the PE/MP libraries. The objective is to find a linear and optimal ordering of the contigs based on “linking” evidence supplied by the paired reads. In practice, the linear ordering of all the contigs is not achieved, because the data may be incomplete and the organism may have several chromosomes. Instead each contig is assigned to a scaffold and given an orientation and position within the scaffold. Kececioglu and Myers [225] have shown that even determining the orientation of the contigs is NP-hard. Therefore, all practical methods to solve the scaffolding problem use heuristics and achieve only an approximate solution.

Many assemblers like Velvet [96], Allpaths [97] and SOAPdenovo [88] contain a scaffolding module. Some stand-alone scaffolders have also been developed. Bambus [226] is designed for Sanger data, and SOPRA [103], SSPACE [104] and MIP [105] are developed for second-generation sequencing data. Bambus and SSPACE are based on a greedy method, whereas SOPRA and MIP relies on statistical optimization and partitioning the scaffolding problem.

The greedy method adopted by SSPACE and Bambus adopts simple user supplied cut-off values for defining the minimum number of mates, the maximum distance between them and how the resolve conflicts (contigs with ambiguous links), necessary to link two contigs. While MIP and SOPRA construct a complex scaffolding graph, considering all the possible pairing and orientation and solve the scaffolding problem by finding the optimal cut-off values (maximum allowed distance and

number of supporting links) in the light of the graph itself. As the exploration of the graph constructed in this way may be computationally hard, suitable heuristics are used. In particular the graph is not explored at glance, but a finite number of partitions are calculated and solved independently.

De-novo assembly of two fungal genomes(Fusarium)

Fusarium is a large and widely distributed genus of filamentous fungi. Most species are harmless saprobes, and are relatively abundant members of the soil microbial community. Some species produce mycotoxins in cereal crops that can affect human and animal health if they enter the food chain. The main toxins produced by these Fusarium species are fumonisins [227] and trichothecenes [228]. Thus, the genus Fusarium collectively represents the most important group of fungal plant pathogens, causing various diseases on nearly every economically important plant species. The health hazard posed to humans and livestock by the plethora of Fusarium mycotoxins is of equal importance [229] [230]. Besides their economic importance, species of Fusarium also serve as key model organisms for biological and evolutionary research [231].

Among the Fusarium species, *F. oxysporum* is a ubiquitous soil inhabitant and one of the most important plant pathogenic species in the Fusarium genus [232]. Although they are predominantly harmless as soil saprophytes, many subspecies are found within the F. oxysporum complex cause disease in only a narrow range of plant species. Host

adaptation and specificity within subspecies have been studied extensively, but the evolutionary origin of the host specificity genes is unknown [233]. Comparison of the genomes of related species such as *Fusarium graminearum*, *Fusarium verticillioides*, *F. solani* and *F. oxysporum f.sp. lycopersici* showed the presence of a core set of chromosomes with a high level of synteny. Additionally, four lineage-specific (LS) chromosomes in *F. oxysporum* are rich in transposons and contain genes encoding proteins involved in signal transduction, and effector proteins involved in pathogenicity and virulence [231]. *Fusarium oxysporum f.sp. lycopersici and F. solani* each have LS chromosomes that are distinct with regard to repetitive sequences and genes involved in pathogenicity, indicating that LS chromosomes may have a distinct evolutionary origin compared with the core chromosomes. Interestingly, among the LS chromosomes, the 2-Mb chromosome 14 of *F*. oxysporum f.sp. lycopersici is enriched in genes encoding secreted effectors such as SIX1, SIX2, SIX3, SIX5, SIX6 and SIX7, of which some have proven to be virulence factors . This suggests that chromosome 14 might carry the main determinants for adaptation of *F. oxysporum* towards tomato [231]. Chromosome 14 and another smaller strain-specific chromosomes can undergo transfer between pathogenic and nonpathogenic strains during co-cultivation, resulting in new pathogenic lineages. LS regions are highly enriched in transposable elements as they contain >74% of the identifiable transposable elements present in the genome, including 95% of all DNA transposons. Only 20% of the predicted genes in the LS regions could be functionally classified on the basis of homology to known proteins. In addition to effector genes, these regions are enriched for a variety of cell

wall-degrading enzymes, genes for lipid metabolism, transcription factors and proteins involved in signal transduction, but are deficient in genes for housekeeping functions. Codon usage and codon adaptation index analysis indicated that the LS-encoding genes exhibit distinct codon usage and have a higher G+C content compared with the conserved genes on core chromosomes, supporting distinct and recent evolutionary origins [231].

It is also hypothesized that horizontal transfer of chromosome 14 from *F. oxysporum f.sp. lycopersici* to nonpathogenic *F. oxysporum* strains confers pathogenicity of those strains towards tomato and dedicated experiments also demonstrated that simple co-cultivation of genetically distinct strains can easily generate new pathogenic genotypes. Such events might have also occurred in nature in the past. This finding may also explain the rapid emergence of new pathogenic lineages in distinct non-pathogenic genetic backgrounds [231].

Sequencing and assembly of 2 closely related fusarium specimens

In a project aimed to gain a deeper insight on the evolutionary dynamics within the Fusarium genus, the laboratory where I worked during my PhD studies has been involved in a collaboration with prof. Chris Toomajian from Kansas state University. The objective of this project was to sequence, assemble and eventually annotate 2 closely related specimens of fusarium: the rice pathogen *Fusarium fujikuroi* and a putative fujikuroi-proliferatum hybrid, with the final goal of characterizeing their diversity through comparative genomics studies.

This work also served as a pilot study for assessing the pros and cons of the approach therein adopted, and evaluate the feasibility of a large scale fungal sequencing project in the context of this collaboration.
In this work we initially used Illumina PE reads (insert-size library of 400 bp) to assess if it was possible to attain a good assembly based on this type of data alone. Subsequently as the assembly based on PE alone was not completely satisfactorily, we decided to integrate MP Illumina reads (insert-size 3 Kb) in an attempt to enhance the scaffolding and improve the overall N50.
My role in this collaboration was to perform quality assessment and assembly of the data, while the group of prof. Toomajian had the responsibility of isolating the DNA samples and performing the sequencing and annotation of the genomes.
The project is still ongoing and is now in the annotation phase as we have recently finished the assembly and scaffolding. However preliminary results are encouraging as we achieved good assemblies, and comparative genomics studies are ongoing which are aimed to develop and asses consistent methods for the clarification and assembly of hybrid fungal isolates.

Importantly, during the course of this work, some limitations in available scaffolding tools became clear, and I dedicated some considerable energy to the development of alternative algorithms which show considerable promise.

Fusarium sequencing data

DNA samples were been collected from clonal cultures of the fungal

isolates in the laboratory of prof Toomajian. For each isolate a PE library (size 400 bp) as long as a MP library (size 3 Kb) have been prepared and sequenced within the sequencing facilities of the University of Missoury using the Illumina GA II. Paired end (2*100bp) reads were generated. Table 7 contains the full details about the sequencing libraries. From now on we will refer to the PE library as l1 (fujikuroi), l2 (hybrid) and to the MP libraries as l3 (fujikuroi) and l4 (hybrid).

The expected size of the 2 genomes is about 42 MB, each library constituting almost a 300X theoretical coverage of the genome (120X after the trimming).

Table 7: PE and RP sequencing libraries

Library	**N° of mates**	**Insert-size (mean and sd)**	**Theoretical coverage**	**Mates removed by trimming**	**Theoretical coverage trimming**
F1 PE (l1)	52460000	389 (33)	231X	26030000	128X
F1 MP (l3)	62123000	3.18 kb (389bp)	275X	21354000	146X
F2 PE (l2)	55010000	395(41)bp	244X	26754000	135X
F2 MP (l4)	61543000	3.21 kb (401 bp)	271	20321000	144X

For each Fusarium sequencing library the original number of reads along with the number of reads removed by the quality filters (and the corresponding theoretical coverage of the genome before and after the trimming) are reported.

Quality assessment, trimming and assembly of PE data

In initial efforts to assemble the genomic sequence of the 2 Fusarium specimens, only of PE data were available. Therefore in this first part I will describe only the analyses and assembly of this type of data.

Quality assessment and accurate trimming of the data is a crucial step in any genomic sequencing project. The use of high quality data reduces the risk of mis-assemblies and data of good quality can be used with high confidence for the disambiguation of imperfect repeats.
A typical measure used to asses the reliability of a sequence is the quality score. Quality values or quality scores state the uncertainty of the data, or the likelihood that a base call is incorrect. For example, the Phred algorithm assigns a quality value for each base in a Sanger read in which larger numbers designate smaller error probabilities. A Q20 value, for example, corresponds to a 1 in 100 error probability, and a Q30 value to a 1 in 1,000 error rate. In our case considering the extreme theoretical coverage provided by each library (more than 200X) we decided to apply very strict criteria for the quality trimming:

- As first criterion I decided to discard from each read in the library the final bases (a number of bases to be established as described below).
- To remove low quality sequence contexts I decided to remove from the dataset the part of each sequence following two or more bases with quality lesser or equal a user specified cutoff, or five or more bases with a quality lesser or equal to a second user specified cutoff.

- To account for the overall quality of the sequence, I iteratively calculated an empirical error rate as the sum of the theoretical error rate for each base and decided to truncate any sequence whence the empirical rate surpassed a specified level.
- Finally, I decided to discard every sequence which consequent to the trimming was shorter than 40 bp or whose median quality score was lower than a specified cutoff. In the presence of pairs where just one of the mates passed these quality filters, I retained the “good” mates only and used them in the assembly as single end reads.

While the aforementioned procedure is conceptually simple, the volumes of sequence data that need to be subjected to trimming are large. According I developed an efficient program written in ANSI C++ to allow rapid processing of raw sequence data under different parameter combinations.

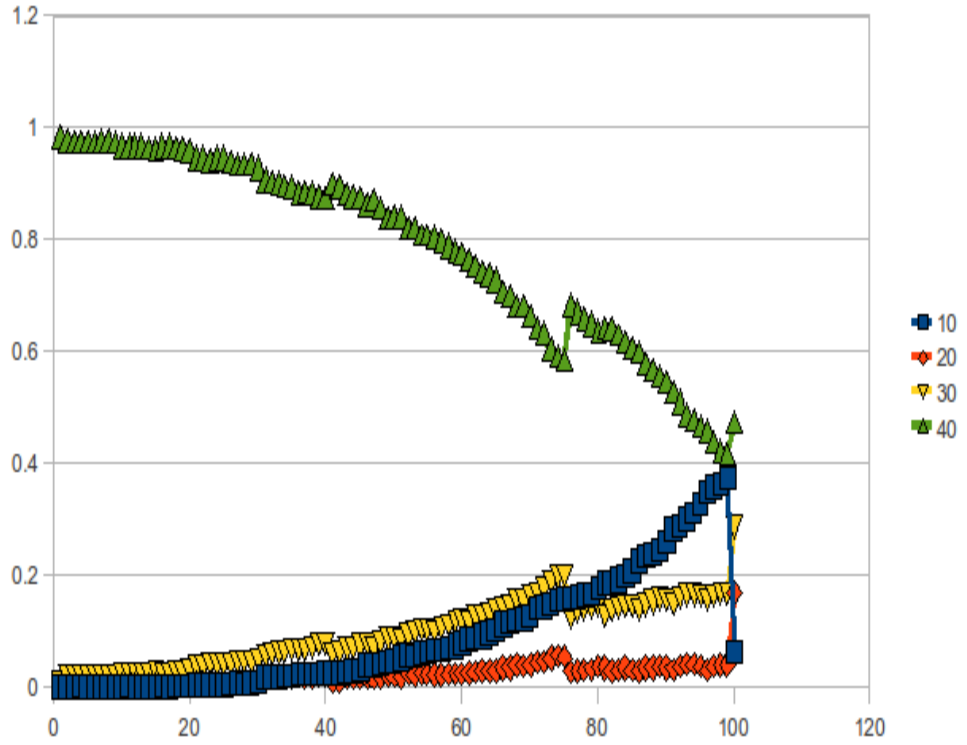
To estimate optimal parameter combinations for final formulation of the trimming criteria I used an empirical evaluation procedure. I applied the same criteria as above but using different combinations of the cut off values, in order to produce slightly different trimmed data sets and measured the effectiveness of each combination of parameters using simple but powerful statistical measures.

To understand the overall quality of the sequencing libraries and the effects of the trimming, I produced empirical position specific quality scores plots. To assess the effects of the trimming on the data structures used by the assemblers (i.e kmer graphs) I evaluated the kmer histograms.

Position specific quality score plots are a simple and informative way for showing the sequencing quality pattern and gaining a better understanding on the overall quality of the data. It is universally acknowledged that independently from chemistries, any sequencing method available to date suffers from an inevitable and progressive loss in quality of the data with the increase of the sequencing cycles. Quality score plots allow the estimation and visualization of the severity of this phenomenon.

A sample graph of this simple statistic is depicted in Figure 20, which shows the position specific quality score distribution of one of the libraries before the trimming (the equivalent picture for the other library is identical and enclosed in the supplementary materials). It is apparent from the graph that quality scores suffer from a inevitable decay along the sequence. Notably, downstream of position 70 we can observe a distinct increase in the proportion of very low quality bases (Quality scores below 10 ie 1 error every 10 calls) is observed (corresponding to a marked decrease in the proportion of high quality bases).

Figure 20: Quality score plot



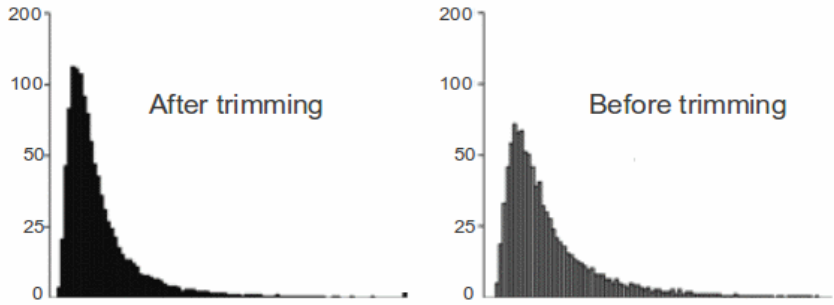
Example of quality score plot. Each line represents the average quality score (Y axis) per position (X axis) before the trimming (Qs<10,10<Qs<20, 20<Qs<30, 30<Qs<40). A constant decrease in the highest quality score bin is observed. While the last 30 position show a constant increase in the frequency of bases with the lowest quality.

The kmer coverage is an important parameter for evaluating the quality of the data in a WGS project. As the WGS strategy is supposed to sample the sequences randomly, in the absence of sequencing errors or biases, each kmer, apart from those falling in highly repetitive regions, is expected to be represented (covered) more or less uniformly in the sequencing library. Kmer coverage histograms are a convenient and simple way for generating a snapshot of quality of the data. Indeed in the absence of sequencing biases the perfect scenario would be to have a sharp and spiked kmer coverage distribution. While in the presence of

systematic biases in the sampling of the genome or high error rates, a broader and eventually multimodal histogram is expected.

In our trimming optimisation pipeline, we evaluated the kmer graphs before and after the trimming procedures for each trim set (with different trimming parameter combinations). An example of such a graphs is presented in Figure 21. From the picture it is evident how the Kmer coverage distribution is benefiting from the removal of low quality data as we observe a sharpest and more consistent distribution.

Figure 21: Kmer graphs

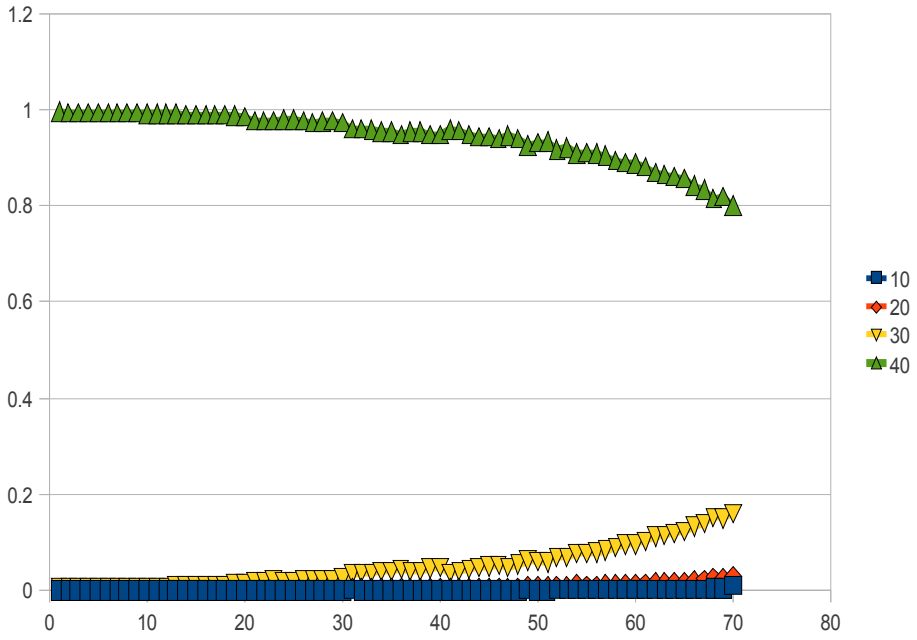


The observed frequency of the distinct kmers from library l1 is displayed before and after the trimming. It is possible to observe how the trimmed distribution is more tight and peaked, while the untrimmed data containing more sequencing error generates a “broader” distribution

I also used positional quality plots to evaluate every “trimmed set”, for example Figure 22 shows an equivalent graph to figure 20 but obtained after the trimming procedure. In this figure we can still observe a constant decay in sequencing quality, but conversely to the situation before trimming, we observe that this effect is greatly mitigated as the decrease in proportion of high quality bases corresponds to an increase

of the proportion bases with a still reliable quality score values (between 20 and 30). More importantly as the main effect of the trimming procedure we observe the complete eradication of low quality bases.

Figure 22: Positional quality plot after trimming



As a consequence of the trimming procedure a less marked decrease in high quality scores position is observed (see figure 20). Furthermore to the drop in the positions of higher quality (30<Qscore<40) corresponds and increase of the positions with good quality (20<Qscore<30)

Preliminary assemblies and comparison between different assemblers

To evaluate the practical effects I ran preliminary assembly analyses using a fraction of the data from each trimmed set. I used this empirical procedure to chose the combination of parameters which could enable

us to select high quality data, and possibly evenly sampled across the genome (kmer graph), so that we could achieve a satisfactorily assembly in terms of N50 and possibly with the lowest requirement in computational resources.

Accordingly, for each trim I also performed an explorative assembly of a fraction of the data to measure the requirements in computational space and time, but more importantly the effectiveness of each assembly (N50). To this end we used the Velvet assembler [96]. For each “trim” set we performed 3 alternate assembly based on different Kmer lengths. In each assembly we used only a fraction of the trimmed data (25%) and did not take into account the “pairing” information. Results are summarized in Table 7 and 8.

Overall we observed a slight but constant improvement of the assembly in terms of resources and effectiveness with the increase in stringency of the quality trimming. More important we didn't observe noticeable drops in the N50 even with the adoption of the more stringent parameters. Encouragingly we didn't observe significant changes in the pattern in response to the usage of Kmers of different lengths. These observations convinced me to use the more stringent combination of parameters for the trimming of the data.

As for the first criterion the last 30 bases of each sequence were discarded. For the low quality sequence contexts any sequence following two or more bases with quality lesser or equal to ten, or five or more bases with a quality lesser or equal to 20 were removed. To account for the overall quality of the sequence, a maximum empirical error rate of 5% was used. Finally sequences shorter than 40 bp after

the trimming and having a median quality score lower than 25 were removed as well.

Table 7: Different parameters used for the trimming

	B10	B20	Em Rate	Med Qs	Trimmed bases	Mates after
Ts1	5	-	0.15	15	1.25E+009	41.5 Mil
Ts2	5	10	0.1	15	2.16E+009	38.1 Mil
Ts3	4	8	0.1	15	2.58E+009	36.4 Mil
Ts4	3	5	0.07	20	3.81E+009	32.5 Mil
Ts5	2	5	0.05	25	4.91E+009	28.2 Mil

The different combinations of parameters used to trim the data are shown. B10=remove after N bases of quality < 10, B20= remove after N bases of quality < 10, Em rate=cut off for empircal error probability, Med Qs= minimum
median quality score

Table 8: Performances attained by different assemblers

SOAP denovo									
	K=21			**K=23**			**K=25**		
	N50	**RAM**	**Time**	**N50**	**RAM**	**Time**	**N50**	**RAM**	**Time**
Ts1	31	7G	6.2h	30.5	7G	6h	32	6.9G	6.2h
Ts2	33	7G	6.1h	31.2	7G	5.8h	33	6.8G	6.1h
Ts3	35	6.8G	6h	33.4	6.8G	5.4h	35.5	6.4G	6h
Ts4	35	6.5G	5.9h	33.5	6.5G	4.9h	35.6	6.1G	5.9h
Ts5	36	6.3G	5.4h	34	6.3G	4.4h	36.7	5.4G	5.4h

Abyss									
	K=21			**K=23**			**K=25**		
	N50	**RAM**	**Time**	**N50**	**RAM**	**Time**	**N50**	**RAM**	**Time**
Ts1	32	6.8G	9.1h	34	6.8G	8.1h	35	6.4G	8.1h
Ts2	33	6.7G	8.4h	34	6.7G	7.4h	35	6.1G	7.4h
Ts3	36	6.5G	7.6h	37	6.4G	6.6h	37	6.1G	6.6h
Ts4	36	6.1G	7.3h	37.5	6G	6.3h	38.5	6G	6.3h
Ts5	38	6G	7.1h	39	5.9G	6.1h	41	5.3G	5.1h

Velvet									
	K=21			**K=23**			**K=25**		
	N50	**RAM**	**Time**	**N50**	**RAM**	**Time**	**N50**	**RAM**	**Time**
Ts1	31	4.8G	5.1h	32	4.7G	4.9h	34	4.3G	4.4h
Ts2	33	4.7G	4.4h	35	4.6G	4.4h	37	4.1G	4.1h
Ts3	35	4.5G	3.6h	36	4.3G	3.6h	38	4.1G	3.5h
Ts4	37	4.1G	3.3h	37.5	4G	3.2h	38.5	3.9G	3.1h
Ts5	38	4G	3.1h	38	3.9G	3.1h	40.5	3.87G	2.9h

N50, memory requirements and computational times achieved by SOAP denovo, Abyss and Velvet in the assembly of the different trim set

PE data and difficulties in assembly

Subsequent to the trimming, we run the assembly. To assemble the 2 genomes I adopted 3 of the most widely used NGS assemblers: SOAP denovo, Abyss, and Velvet. All of these programs also implement "internal" scaffolding routines, therefore I decided to take full advantage of the PE reads and required to the programs to perform the scaffolding as well. To perform a more unbiased comparison of the 3 tools, we run for each one 3 different assembly with slightly altered values for the Kmer size parameter. It is acknowledged that this comparison has little statistical relevance, as the assemblers were tested on just a small dataset changing a few parameters. However in our case we just wanted to asses which assembler was the more suited to our dataset. Unsurprisingly all the programs achieved similar results, although the computational resources and time required differed greatly, with SOAPdenovo and Abyss being more greedy than Velvet which turned out to be the fastest and less demanding in our case (Results are reported in Table 8). To economize on computational resources I decided to adopt Velvet as the main assembler in our project.

The results of this first run of assembly were not particularly satisfactory, at least for one of the 2 genomes, as I was expecting (from the published genome contig sizes for comparable projects) to achieve an N50 of at least 50 Kb. further assemblies with Velvet using a broader range of values for the Kmer size did not produce significant improvements. Indeed what we consider our final (best) assembly of these data lead to an N50 of 76 and 27 Kb respectively. Scaffolding,

which at this stage was performed using the velvet scaffolding routines had a low impact overall, but this wasn't particularly surprising as insert size was not much greater than the combined length of the reads (insert size around 380 bp, reads 100 bp x 2)

I was also surprised to observe that the results of the assembly were remarkably different for the 2 genotypes, considering that the genomes are from very closely related specimens. Furthermore, contrarily to any logical expectation the genome that attained the best assembly is that from the hybrid isolate (library 2) which would be expected to be more heterozygous and hence harder to assemble. From now on I will refer to this better assembly, associated with the presumed fujikuroi-proliferatum hybrid as to A2 and the other (*F. fujikuroi*) as A1. Results of the assemblies are reported in Table 9.

Table 9: Results from the first assembly

	N50	**N° contigs**	**Kmer**	**N° scaffolds**
A1	27Kb	1280	21	92
A2	75Kb	636	23	45

Best results achieved by Velvet in the assembly of Fusarium data. The number of N50 of the assembly, number of contigs and scaffolds and the length of the Kmers used to construct the Kmer graph are reported

I reasoned that possible explanations could include either an uneven sampling for the A1 sequencing data, or a consistent expansion of repeat families (more likely direct and simple repeat) in the A1 genome.

Explorative comparison of the assemblies

To gain a deeper insights on the situation I decided to compare the 2 assemblies by aligning their respective contings. As the contigs from the A2 were consistently longer, I decided to use the A2 as the reference assembly to verify the consistency of the cognate A1. To this extent I had first to establish, where possible, an univocal relationship between the contigs in the 2 assemblies.
To assign every conting from the A1 univocally to a corresponding contig in A2 I applied a simple scoring scheme based on the output of the blast program. After performing an initial blast sequence similarity search with strict parameters (list) for any contig in A1, I considered the list of hits (contigs in A2). For each hit I used the complete set of non overlapping HSPs to calculate a simple score by summing linearly the products of the similarity rates and length of each HSP.

I assigned a contig in A1 to a corresponding contig in A2 when the contig in A2 was found to have the best similarity score and the second best score was at least 25% lower than the best one. To avoid confusing situations derived from small and potentially low complexity/highly repetitive contigs, I applied this procedure only to contings longer than15 Kb. In this way, I could assign indisputably 153 contigs from the A1 (corresponding to 4.2 Mb) to its cognate in A2.

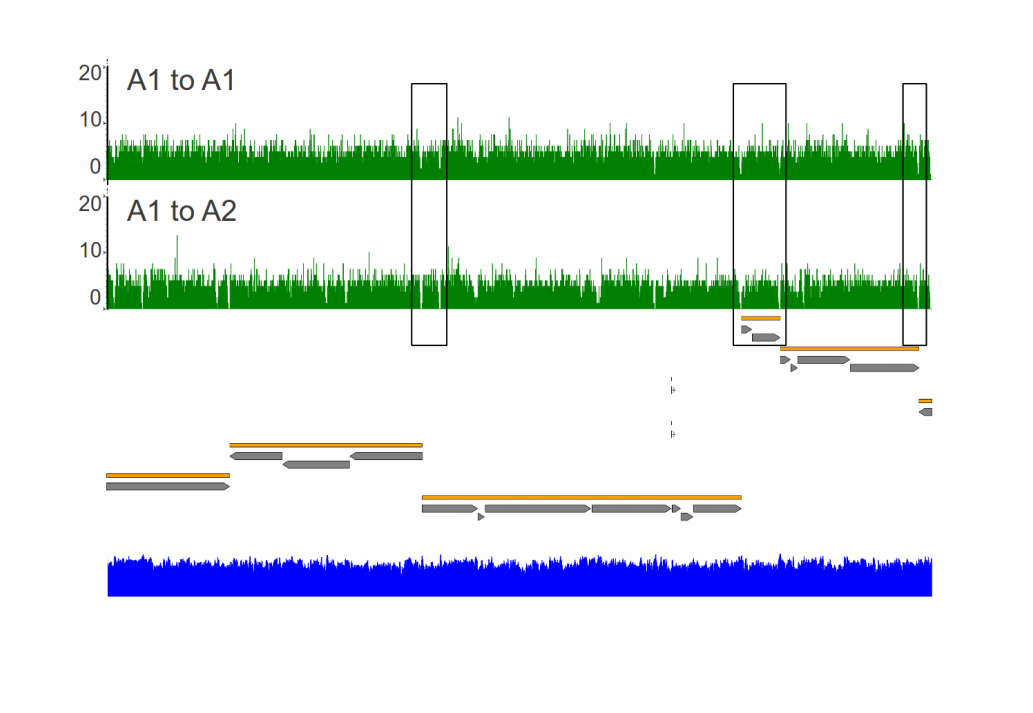
To investigate the hypothesis of a more uneven genomic sampling in A1 respect to A2 I compared the coverage of highly similar contigs from the sequencing reads of the 2 libraries, while to understand the effect of

highly repetitive regions I excluded from this comparison all the reads having more than 10 equally good mapping solutions on the contigs. To generate an easily interpretable “snapshot” of the situation I produced graphical representations of these features, and analysed different figures empirically. An example is reported in Figure 23.

I produced and carefully analysed images like these for more than 150 A2 contigs (corresponding to 4.2 Mb in sequence), there was no particular evidence for uneven sampling of the reads in library I1. The main emerging pattern from the figures is that we constantly observe coverage drops in both libraries corresponding to the ends of the A1 contigs. I reasoned that the most likely explanation for this could either be the expansion of a simple repeat in the A1 genome, or the presence “regions of low sequentiability” at the boundaries of the A1 contigs. By low sequentiability I mean regions whose composition reduces the possibility of being sequenced. For example Illumina technology suffer from severe drop in sequencing quality in the presence of AC rich regions [51].

Thus I was not able to produce a conclusive answer to the question: indeed the only guaranteed way to solve the dilemma would have been to look at the composition of the sequences lying between the presumed gaps in the A1 assembly. However, I regularly noted that where more than one A1 contig mapped into a single A2 contig, the coverage of the junction region was low also in the A2 contig.

Figure 23: Example of plots used to compare the assemblies



The alignment of different Contigs from the A1 assembly (orange) and the corresponding BLAST HSPs (grey) against the most similar A2 contig (not shown) are displayed. Various coverage track are used to assay the differences bewteen the assemblies. A GC % content track calculated on overlapping windows (blue) of 40 bp is used the verify the effect of compositional biases. Two coverage tracks (green) are used to show the mapping of the reads from A1 (lower) and A2 (upper). In the correspondence to the ends of A1 contigs coverage drops are observed for both datase (rectangles). The coverage is calculated as the absolute number of aligned reads “starting” at each position.

MP data and scaffolding

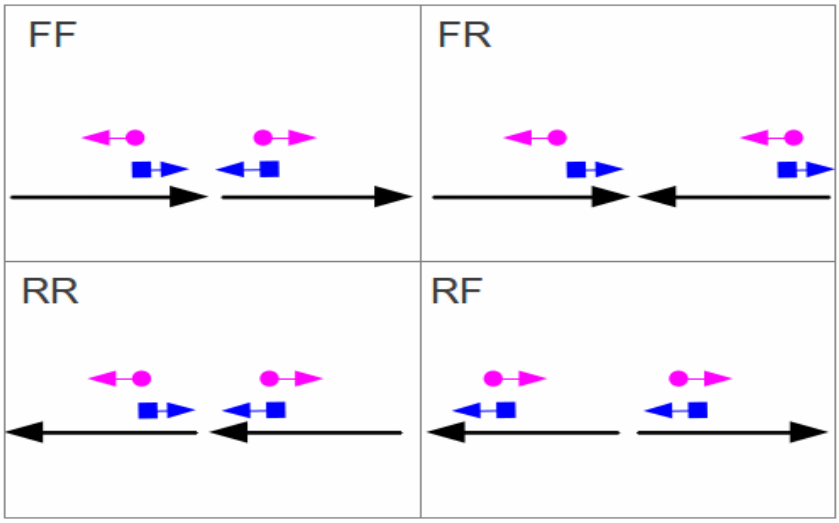
To improve the assemblies 2 additional MP libraries, with a theoretical insert size of 3 Kb for each genome, were produced, with the objective of enhancing the scaffolding step. Again each of the 2 libraries constituted more than a theoretical 100X coverage of the 2 genomes,

as reported in table X. As admitted by the manufacturer (see above) it is expected that due to inconsistency in the biotinylated DNA affinity purification, MP Illumina libraries may contain a fraction of PE reads, in a proportion of in between 10 and 15% according to Illumina.
It is straight-forward to identify such PE contaminants when a reference genome is available. As when mapped to the reference the MP reads will map with opposite orientation and at far longer distance than PE.
To estimate the PE contamination in our libraries I mapped the reads on the A1 and A2 contigs. Surprisingly analysis of the mapping pattern suggested that about 60% of the reads in the libraries were of PE type. Discouraged by the high contamination but reassured by the quantity of data available, I decided to proceed further with the scaffolding. Before starting with the core scaffolding procedure I applied to the MP data the same quality evaluation and trimming procedure described for the PE reads.

The core idea of the scaffolding process is to bridge contigs by the use of mated reads. In our case given the high contamination from PE reads in the Illumina MP libraries, we had to deal with a mixed library with a bimodal insert-size distribution. Therefore to be able to produce a reliable scaffolding first we had to distinguish the nature of the bridging “pairs”. Even though the discrimination of PE from MP is straightforward when both reads can be mapped on the same reference molecule, it is more difficult when the two mates map on different contigs, especially because the contigs may have discordant orientations. As depicted in figure 24, there are four possible combination of relative orientation between 2 contigs (FF,FR,RR and and accordingly 4 possible ways of

bridging them with PE or MP reads.. However whence the contigs have the same orientations, the FF or RR case are indistinguishable from a read oriented point of view, as the PE and MP reads are not directional. Once established the expected mapping patterns and the maximum distance allowed between the 2 mates, it is trivial to verify if a pair of reads is bridging 2 contigs in way amenable with those described in the picture. In our case I estimated the maximum distance empirically from the insert size distribution, as the 99th pencentile of the distribution itself. This value corresponded to 3.61 and 3.58 Kb respectively for the 2 libraries.

Figure 24: Expected orientations of PE and MP



FF(forward forward),RR(reverse reverse),FR(forward reverse),RF (reverse forward), are relative contigs orientations. MP reads are in purple. PE reads are in blue. Arrows are used to show the relative orientations. Reads orientation are displayed assuming MP reads from the Illumina protocols.

At this point I produced a custom script to identify all the possible bridging MP and PE within our libraries. It was no surprise to ascertain that only a few bridging PE reads could be identifies, given that the prior scaffolding attempts based on the PE library did not produce significant results, while I was encouraged by the good number of MP links found.

Development of a new scaffolder

At the time when we performed this analysis the MIP scaffolder wasn't yet available, while SSPACE was universally acknowledged as the best program (subsequent tests show that the performances of the 2 tools are not so dissimilar although MIP performs slightly better).
Therefore I proceeded with the scaffolding using the SSPACE program.

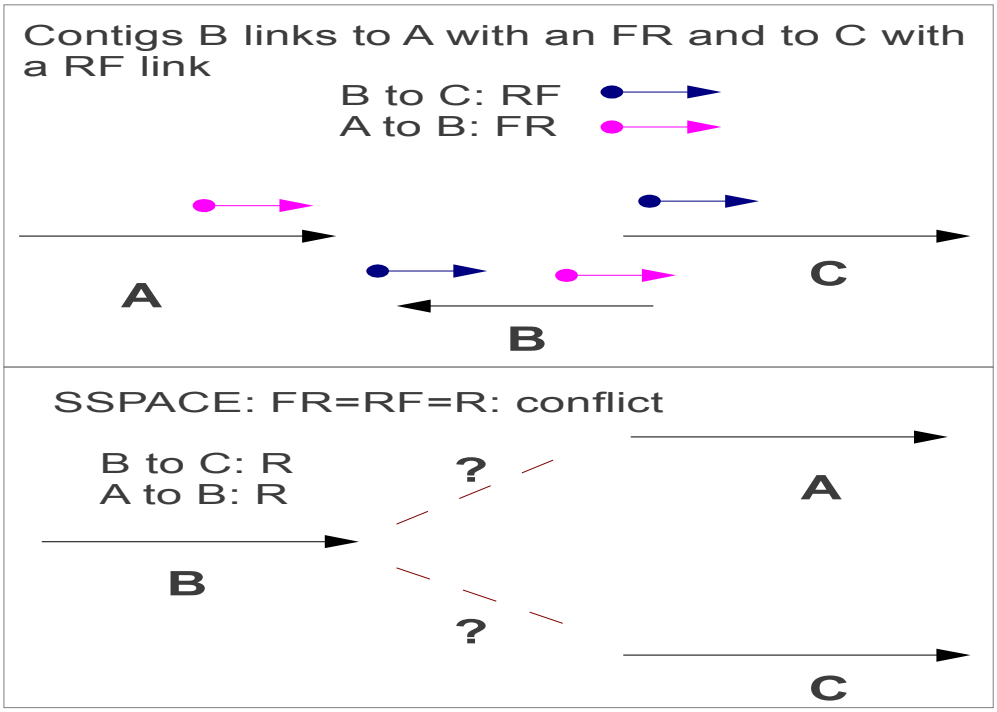
The minimum number of mates supporting a link is a sensible parameters for any scaffolding algorithm. This value is used to discriminate between random links due to chimeric reads of inconsistencies in the mapping and actual links deriving from contiguous sequences. The SSPACE program requires this parameter to be estimated by the user. Again I estimated it empirically. I counted the number of distinct mate couples supporting every link and computed a “link support” distribution. To set the cut off I arbitrarily chose the 95th percentile of the distributions, resulting in 43 and 52 links respectively.

Having established all the possible connections and cut off values I performed the scaffolding with SSPACE. The results seemed relatively

good (see Table 10), however after a careful examination of the ancillary files produced by SSPACE I was puzzled as apparently it was missing a certain number of links and scaffolds which it was expected to find.

Further analysis indicated how the problem was almost exclusively from pairs of contigs with opposite orientations. To my surprise, and through inspection of the SSPACE source code, I discovered that the SSPACE software doesn't discriminate between FR and RF when joining contigs with opposite orientation. This is particularly harmful in cases such as that depicted in Figure 25, i.e when a contig shows 2 “reverse” links (links to contig with reverse orientation). If the 2 links are of the same type (FR or RF) they're evidently conflicting and little can be done especially if they have similar support, however when the reverse links are of the different type the scaffolding is completely legal, and the 3 contigs should be bridged (see again the figure). Obviously without taking into account the nature of the reverse links SSPACE can't resolve such “ambiguities” systematically therefore loosing a good part of the information.

Figure 25: Apparent conflict between RF and FR



In the presence of 2 reverse links, as SSPACE doesn't discriminate between FR and RF orientation (lower panel), the links are recovered as conflicting even if a legal scaffolding solution (upper panel) exists.

To verify further my findings I developed an in house ad-hoc scaffolder, which is completely identical by design to SSPACE apart from the feature that it can resolve “double” FR, RF reverse links.

The algorithm implementation is straight-forward. Initially a procedure analogous to that described above is adopted to estimate empirically proper cut-off values for the maximum length allowed between the RP and minimum support for a link (99th and 95th percentile of the underlying distributions respectively). Subsequent all the RP mapping to different

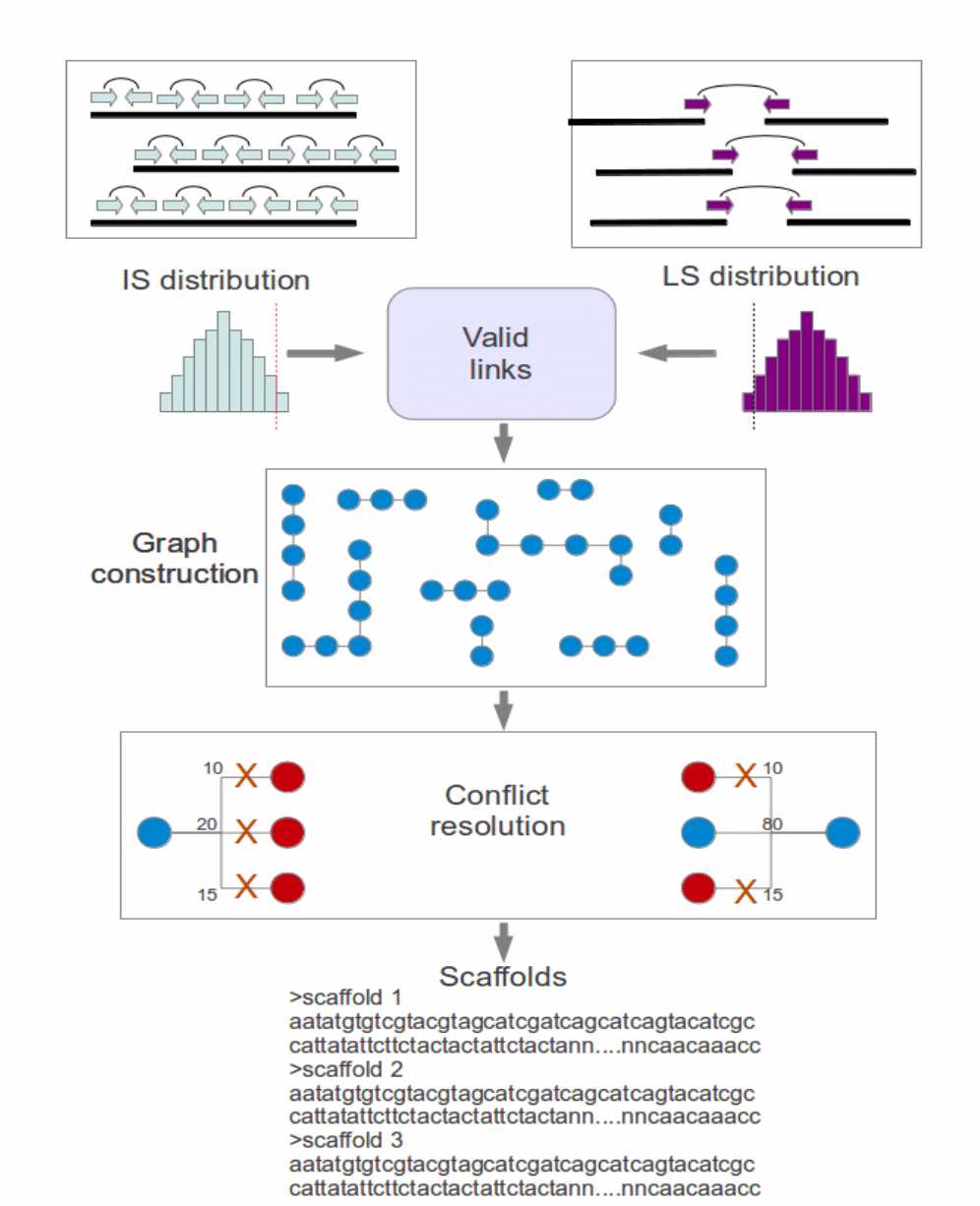
contigs are evaluated and in the light of the cut-off values the valid links between contigs (those sufficiently supported by RP mapping within the maximum distance) are ascertained. A simple direct a-cyclic graph,is then constructed from this valid links. Contigs constituting the nodes of the graph. An exhaustive procedure is then used to explore the whole graph. Each node (contig) is assayed and in the presence of ambiguities (conflicting links of the same type) a simple and conservative conflict resolution procedure is applied: the best supported link of that type has to be 4X or more more supported than the second best link to resolve the conflict. While in the opposite case the conflicting links are removed from the graph.

Once the ambiguities are resolved a simple exhaustive walk in the graph is used to “call” the scaffold which are finally printed to a file in fasta format.

The distances between scaffolded the contigs are estimated as the minimum observed distance between 2 supporting mates.

An overview of this approach is depicted in Figure 26.

Figure 26: In house scaffolder



Reads mapping on contigs are used to calculate the insert size distribution. The 99th percentile of such distribution is used as length cut-off. Analogously a link support distribution is computed from RP bridging the contigs and the 95th percentile of the distribution is used as cuf -off. Valid links in the light of the cut-off criteria are identified and a scaffolding graph is constructed. An exhaustive walk in trough graph is performed and conflicts are removed. Finally scaffolds are printed in fasta format

I applied my simple scaffolder to the fungal contigs and MP libraries and compared its performance to that of SSPACE. The results are summarized in Table 10, and show that my algorithm achieves a considerably better N50. Furthermore a comparison of the scaffolding performed on the A1 assembly, using the A2 as reference showed how there is no difference in specificity between our scaffolder and SSPACE. To this extent we used the same procedure described before to assign each A1 conting to the most similar contig in A2, recorded the spacing and orientation and compared it to the results of the scaffolding. In both cases we found perfect accordance between the scaffolding performed by us and SSPACE and the similarity analysis.

Table 10: Comparison between SSPACE and in house scaffolder

	N50 PE	**N50 MP**	**Longest scaffold**	**N° of scaffold**	**Avg contigs per scaffold**	**Concordant A1-A2**
A1 SSPACE	27 kB	121 Kb	350 Kb	201	2.1	98.20%
A1 In house	27 Kb	180 Kb	553 Kb	180	2.93	98.50%
A2 SSPACE	75 Kb	840 Kb	950 Kb	102	2.81	
A2 In house	75 Kb	1.02 Mb	1.4 Mb	84	3.4	

N50 longest conting and average number of contigs per scaffolded achieved by my in house scaffolder and by SSPACE in scaffolding the fungal data. To evaluate the accuracy of the scaffolders I used the A2 as reference to measure the accordance of the A1 scaffolds to the A2 assembly.

This analysis, by definition could not give a precise estimate of the False Positive Rate, as I could not compare the scaffolds against a reference genome. To further convince myself of the validity of the approach I

compared our scaffolder and SSPACE using publicly available PE resequencing data. I downloaded 24 million PE resequencing reads for the genome of the well characterized bacterium *Pseudomonas syringae* from the NCBI SRA archive (http://www.ncbi.nlm.nih.gov/sra). In this case the reference genome is also available. After assembling the data with Velvet I compared the performances attained by my scaffolder respect to SSPACE. Results are shown in Table 11 which show how my program achieves a notably larger N50. As in this case the reference genome is available good estimations of the false positives rates of the scaffolders can be provided. Again from the table is completely evident how both the tools produce few errors (between 6 and 7 %) and display an almost identical sensitivity. Thus validating the claim that the tools have similar specificity but my scaffolder has better sensitivity (i.e it detects more "valid" links).

Table 11: Analysis of *P. syringae* resequencing data

	N50 assembly	**N° contigs assembly**	**N50 scaffolding**	**Longenst scaffold**	**Avg N° of contigs per scaffold**	**FP scaffolds**
SSPACE	64 kb	2304	89.4 Kb	294 kB	2.5	22 (6%)
In house	64 kb	2304	121 KB	512 kB	4.1	26 (7%)

The N50 and number of contigs in the original assembly, along with the N50, number of scaffolds, longest scaffold, average number of contigs per scaffold and the estimated false positive scaffolds are reported both for my in house scaffolder and SSPACE. Data used in the comparison are from an Illumina PE resequencing library from the bacterium Pseudomonas syringae. False positive rates have been computed by comparing the scaffolds (order and orientation) to the reference genome of the bacterium.

Conclusion and final remarks

During the 3 years of my PhD studies I trained myself into the fine art of bioinformatic. Having a more biological background (master degree in functional genomics and bioinformatics) I did produce considerable efforts to gain an adequate knowledge of various programming language and the of most common algorithms used in to analyze biological data, with a particular focus on the development of algorithms for the NGS technologies.
While training to gain deeper competences on the more informatics side of the discipline I've been involved in 3 main research project during the course of my studies.

First I participated in a NGS “editome” analysis where the editing pattern of the mitochondrial transcripts of the model angiosperm *Vitis vinifera* have been unraveled. This study demonstrated the applicability of the use of NGS transcriptome sequencing data to characterize the phenomenon of RNA editing for the first time. In the course of this study the strong and weak point of 2 competing NGS technologies have also been assayed, showing how the best results where achieved combining the data from the 2 technologies.

Secondly I've been the main developer of a new software package, for the detection of structural variants at intra-specific level. The software SVM^2, designed to overcome the limitation of already available programs based on similar principles, demonstrated an enhanced sensitivity (more than 4X) respect to its competitors, with a similar

(slightly better specificity). SVM^2 is the first software capable of consistently finding ultra short SV (1 to 10 bp indels) from NGS read pair data. The main novelty in the approach adopted by SVM^2 is the use of supervised learning to infer from an ensemble of meaningful features the presence of different types of SV variants. Limitations such as the optimization of the Pvalues cut-off for a specifical statistical test, which are the main weakness of the competitor programs are avoided, as SVM^2 uses different sources of evidence for taking its final decision over the presence/absence of a SV.
Interestingly SVM^2 demonstrated to be able to compete even with split-mapping based approaches, that is dedicated software specifically designed to find very short indels (while it was longly assumed than any read pair based SV detector would have been useless for this particular task). Indeed the algorithm implemented in SVM^2 demonstrated to fully recover a proportion of ultra-short indels falling in low complexity and repetitive regions which were completely missed by split mapping.

Finally I took part in a de-novo genome assembly project, where I have been responsible to assemble and scaffold 2 genomes from closely related fungal specimens. In this project I developed completely customized quality assessment procedures, and in the presence of inconsistencies and inherent difficultiues in the assembly developed a simple but effective explorative procedure to characterize the assembly patterns of the 2 genomes. Finally and more importantly once evident limitations emerged from already available software for the scaffolding of NGS read pair data, I produced a new custom scaffolder which demonstrated to overcome the limitations of those already available.

References

[1] D. R. Bentley et al., “Accurate whole human genome sequencing using reversible terminator chemistry.,” *Nature*, vol. 456, no. 7218, pp. 53-9, Nov. 2008.

[2] M. Margulies et al., “Genome sequencing in microfabricated high-density picolitre reactors.,” *Nature*, vol. 437, no. 7057, pp. 376-80, Sep. 2005.

[3] J. Shendure et al., “Accurate multiplex polony sequencing of an evolved bacterial genome.,” *Science (New York, N.Y.)*, vol. 309, no. 5741, pp. 1728-32, Sep. 2005.

[4] J. Shendure and H. Ji, “Next-generation DNA sequencing.,” *Nat Biotechnol*, vol. 26, no. 10, pp. 1135-1145, 2008.

[5] J. Zhang, R. Chiodini, A. Badr, and G. Zhang, “The impact of next-generation sequencing on genomics.,” *Journal of genetics and genomics = Yi chuan xue bao*, vol. 38, no. 3, pp. 95-109, Mar. 2011.

[6] L. D. Stein, “An introduction to the informatics of ‘next-generation’ sequencing.,” *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, vol. 11, p. Unit 11.1., Dec. 2011.

[7] A. D. Hershey and M. Chase, “Independent functions of viral protein and nucleic acid in growth of bacteriophage.,” *The Journal of general physiology*, vol. 36, no. 1, pp. 39-56, May 1952.

[8] J. D. Watson and F. H. C. Crick, “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid,” *Nature*, vol. 171, no. 4356, pp. 737-738, Apr. 1953.

[9] A. M. Maxam and W. Gilbert, “A new method for sequencing DNA.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 2, pp. 560-4, Feb. 1977.

[10] 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*, vol. 74, no. 12, pp. 5463-7, Dec. 1977.

[11] L. M. Smith, S. Fung, M. W. Hunkapiller, T. J. Hunkapiller, and L. E.

Hood, “The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis,” *Nucleic Acids Research*, vol. 13, no. 7, pp. 2399-2412, Apr. 1985.

[12] B. G. Barrell, G. M. Air, and C. A. Hutchison, “Overlapping genes in bacteriophage phiX174.,” *Nature*, vol. 264, no. 5581, pp. 34-41, Nov. 1976.

[13] W. Fiers et al., “Complete nucleotide sequence of SV40 DNA.,” *Nature*, vol. 273, no. 5658, pp. 113-20, May 1978.

[14] S. Anderson et al., “Sequence and organization of the human mitochondrial genome.,” *Nature*, vol. 290, no. 5806, pp. 457-65, Apr. 1981.

[15] R. Baer et al., “DNA sequence and expression of the B95-8 Epstein-Barr virus genome.,” *Nature*, vol. 310, no. 5974, pp. 207-11.

[16] A. T. Bankier et al., “The DNA sequence of the human cytomegalovirus genome.,” *DNA sequence : the journal of DNA sequencing and mapping*, vol. 2, no. 1, pp. 1-12, Jan. 1991.

[17] R. Staden, “Sequence data handling by computer.,” *Nucleic acids research*, vol. 4, no. 11, pp. 4037-51, Nov. 1977.

[18] R. Staden, K. F. Beal, and J. K. Bonfield, “The Staden package, 1998.,” *Methods in molecular biology (Clifton, N.J.)*, vol. 132, pp. 115-30, Jan. 2000.

[19] H. S. Bilofsky et al., “The GenBank genetic sequence databank.,” *Nucleic acids research*, vol. 14, no. 1, pp. 1-4, Jan. 1986.

[20] W. R. Pearson and D. J. Lipman, “Improved tools for biological sequence comparison.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 8, pp. 2444-8, Apr. 1988.

[21] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool.,” *Journal of molecular biology*, vol. 215, no. 3, pp. 403-10, Oct. 1990.

[22] L. M. Smith et al., “Fluorescence detection in automated DNA sequence analysis.,” *Nature*, vol. 321, no. 6071, pp. 674-9, Jan. 1986.

[23] Connell C, Fung S, Heiner C, Bridgham J, Chakerian V, Heron E,

Jones B, Menchen S, Mprdan W et al “Automated DNA-Sequence analysis,” *Biotechniques*, vol. 5, no. 5, pp. 342-348, 1987.

[24] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, “Base-calling of automated sequencer traces using phred. I. Accuracy assessment.,” *Genome research*, vol. 8, no. 3, pp. 175-85, Mar. 1998.

[25] B. Ewing and P. Green, “Base-calling of automated sequencer traces using phred. II. Error probabilities.,” *Genome research*, vol. 8, no. 3, pp. 186-94, Mar. 1998.

[26] P. Richterich, “Estimation of errors in ‘raw’ DNA sequences: a validation study.,” *Genome research*, vol. 8, no. 3, pp. 251-9, Mar. 1998.

[27] R. D. Fleischmann et al., “Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.,” *Science (New York, N.Y.)*, vol. 269, no. 5223, pp. 496-512, Jul. 1995.

[28] C. M. Fraser et al., “The minimal gene complement of Mycoplasma genitalium.,” *Science (New York, N.Y.)*, vol. 270, no. 5235, pp. 397-403, Oct. 1995.

[29] J. C. Roach, C. Boysen, K. Wang, and L. Hood, “Pairwise end sequencing: a unified approach to genomic mapping and sequencing.,” *Genomics*, vol. 26, no. 2, pp. 345-53, Mar. 1995.

[30] G. G. Sutton, O. White, M. D. Adams, and A. R. Kerlavage, “TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects,” *Genome Science and Technology*, vol. 1, no. 1, pp. 9-19, Jan. 1995.

[31] R. L. Sinsheimer, “To reveal the genomes.,” *American journal of human genetics*, vol. 79, no. 2, pp. 194-6, Aug. 2006.

[32] F. R. Blattner et al., “The complete genome sequence of Escherichia coli K-12.,” *Science (New York, N.Y.)*, vol. 277, no. 5331, pp. 1453-62, Sep. 1997.

[33] R. A. Welch et al., “Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 26, pp. 17020-4, Dec. 2002.

[34] A. Goffeau et al., “Life with 6000 genes.,” *Science (New York, N.Y.)*, vol. 274, no. 5287, pp. 546, 563-7, Oct. 1996.

[35] C.elegans Sequencing Consortium “Genome sequence of the nematode C. elegans: a platform for investigating biology.,” *Science (New York, N.Y.)*, vol. 282, no. 5396, pp. 2012-8, Dec. 1998.

[36] E. W. Myers et al., “A whole-genome assembly of Drosophila.,” *Science (New York, N.Y.)*, vol. 287, no. 5461, pp. 2196-204, Mar. 2000.

[37] M. D. Adams et al., “The genome sequence of Drosophila melanogaster.,” *Science (New York, N.Y.)*, vol. 287, no. 5461, pp. 2185-95, Mar. 2000.

[38] I. Dunham et al., “The DNA sequence of human chromosome 22.,” *Nature*, vol. 402, no. 6761, pp. 489-95, Dec. 1999.

[39] J. C. Venter et al., “The sequence of the human genome.,” *Science (New York, N.Y.)*, vol. 291, no. 5507, pp. 1304-51, Feb. 2001.

[40] E. S. Lander et al., “Initial sequencing and analysis of the human genome.,” *Nature*, vol. 409, no. 6822, pp. 860-921, Feb. 2001.

[41] C. A. Emrich, H. Tian, I. L. Medintz, and R. A. Mathies, “Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis.,” *Analytical chemistry*, vol. 74, no. 19, pp. 5076-83, Oct. 2002.

[42] J. A. Shendure, G. J. Porreca, and G. M. Church, “Overview of DNA sequencing strategies.,” *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, vol. 7, p. Unit 7.1, Jan. 2008.

[43] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén, “Real-time DNA sequencing using detection of pyrophosphate release.,” *Analytical biochemistry*, vol. 242, no. 1, pp. 84-9, Nov. 1996.

[44] A. R. Quinlan, D. A. Stewart, M. P. Strömberg, and G. T. Marth, “Pyrobayes: an improved base caller for SNP discovery in pyrosequences.,” *Nature methods*, vol. 5, no. 2, pp. 179-81, Feb. 2008.

[45] R. E. Green et al., “A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing.,” *Cell*, vol. 134, no. 3, pp. 416-26, Aug. 2008.

[46] T. Wicker, E. Schlagenhauf, A. Graner, T. J. Close, B. Keller, and N.

Stein, “454 sequencing put to the test using the complex genome of barley.,” *BMC genomics*, vol. 7, p. 275, Jan. 2006.

[47] R. E. Green et al., “Analysis of one million base pairs of Neanderthal DNA.,” *Nature*, vol. 444, no. 7117, pp. 330-6, Nov. 2006.

[48] G. Turcatti, A. Romieu, M. Fedurco, and A.-P. Tairi, “A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis.,” *Nucleic acids research*, vol. 36, no. 4, p. e25, Mar. 2008.

[49] M. Fedurco, A. Romieu, S. Williams, I. Lawrence, and G. Turcatti, “BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies.,” *Nucleic acids research*, vol. 34, no. 3, p. e22, Jan. 2006.

[50] M. Kircher, U. Stenzel, and J. Kelso, “Improved base calling for the Illumina Genome Analyzer using machine learning strategies.,” *Genome biology*, vol. 10, no. 8, p. R83, Jan. 2009.

[51] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, “Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.,” *Nucleic acids research*, vol. 36, no. 16, p. e105, Sep. 2008.

[52] J. Rougemont, A. Amzallag, C. Iseli, L. Farinelli, I. Xenarios, and F. Naef, “Probabilistic base calling of Solexa sequencing data.,” *BMC bioinformatics*, vol. 9, p. 431, Jan. 2008.

[53] Y. Erlich, P. P. Mitra, M. delaBastide, W. R. McCombie, and G. J. Hannon, “Alta-Cyclic: a self-optimizing base caller for next-generation sequencing.,” *Nature methods*, vol. 5, no. 8, pp. 679-82, Aug. 2008.

[54] L. M. Smith et al., *A Theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction.*, vol. 321, no. 6071. 2008, pp. 674-9.

[55] et al. Dimalanta ET, Zhang L, Hendrickson CL, “Increased Read Length on the SOLiD™ Sequencing Platform. Poster SOLiD™ System.,” 2009.

[56] K. A. Wetterstrand, “DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program,” 2011. .

[57] E. E. Schadt, S. Turner, and A. Kasarskis, “A window into third-generation sequencing.,” *Human molecular genetics*, vol. 19, no. 2, pp. R227-40, Oct. 2010.

[58] V. Costa, C. Angelini, I. De Feis, and A. Ciccodicola, “Uncovering the complexity of transcriptomes with RNA-Seq.,” *Journal of biomedicine & biotechnology*, vol. 2010, p. 853916, Jan. 2010.

[59] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, “Computational methods for transcriptome annotation and quantification using RNA-seq.,” *Nature methods*, vol. 8, no. 6, pp. 469-77, Jun. 2011.

[60] L. Zhou, X. Li, Q. Liu, F. Zhao, and J. Wu, “Small RNA transcriptome investigation based on next-generation sequencing technology.,” *Journal of genetics and genomics = Yi chuan xue bao*, vol. 38, no. 11, pp. 505-13, Nov. 2011.

[61] C. Addo-Quaye, W. Miller, and M. J. Axtell, “CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets.,” *Bioinformatics (Oxford, England)*, vol. 25, no. 1, pp. 130-1, Jan. 2009.

[62] S. W. Chi, J. B. Zang, A. Mele, and R. B. Darnell, “Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps.,” *Nature*, vol. 460, no. 7254, pp. 479-86, Jul. 2009.

[63] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao, “Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data.,” *Nucleic acids research*, vol. 36, no. 16, pp. 5221-31, Sep. 2008.

[64] C. S. Ku, N. Naidoo, M. Wu, and R. Soong, “Studying the epigenome using next generation sequencing.,” *Journal of medical genetics*, vol. 48, no. 11, pp. 721-30, Nov. 2011.

[65] E. Eisenberg, “Bioinformatic approaches for identification of A-to-I editing sites.,” *Current topics in microbiology and immunology*, vol. 353, pp. 145-62, Jan. 2012.

[66] B. Hu, G. Xie, C.-C. Lo, S. R. Starkenburg, and P. S. G. Chain, “Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics.,” *Briefings in functional genomics*, vol. 10, no. 6, pp. 322-33, Nov. 2011.

[67] W. J. Kent, “BLAT---The BLAST-Like Alignment Tool,” *Genome Research*, vol. 12, no. 4, pp. 656-664, Mar. 2002.

[68] H. Li and N. Homer, “A survey of sequence alignment algorithms for next-generation sequencing.,” *Briefings in bioinformatics*, vol. 11, no. 5, pp. 473-83, Sep. 2010.

[69] A. Bateman and J. Quackenbush, “Bioinformatics for next generation sequencing.,” *Bioinformatics (Oxford, England)*, vol. 25, no. 4, p. 429, Feb. 2009.

[70] S. Shaham, “galign: a tool for rapid genome polymorphism discovery.,” *PloS one*, vol. 4, no. 9, p. e7188, Jan. 2009.

[71] T. D. Wu and S. Nacu, “Fast and SNP-tolerant detection of complex variants and splicing in short reads.,” *Bioinformatics (Oxford, England)*, vol. 26, no. 7, pp. 873-81, Apr. 2010.

[72] D. Weese, A.-K. Emde, T. Rausch, A. Döring, and K. Reinert, “RazerS--fast read mapping with sensitivity control.,” *Genome research*, vol. 19, no. 9, pp. 1646-54, Sep. 2009.

[73] A. D. Smith et al., “Updates to the RMAP short-read mapping software.,” *Bioinformatics (Oxford, England)*, vol. 25, no. 21, pp. 2841-2, Nov. 2009.

[74] F. Vezzi, C. Del Fabbro, A. I. Tomescu, and A. Policriti, “rNA: a fast and accurate short reads numerical aligner.,” *Bioinformatics (Oxford, England)*, vol. 28, no. 1, pp. 123-4, Jan. 2012.

[75] R. Li et al. “SOAP2: an improved ultrafast tool for short read alignment.,” *Bioinformatics (Oxford, England)*, vol. 25, no. 15, pp. 1966-7, Aug. 2009.

[76] M. David, M. Dzamba, D. Lister, L. Ilie, and M. Brudno, “SHRiMP2: sensitive yet practical SHort Read Mapping.,” *Bioinformatics (Oxford, England)*, vol. 27, no. 7, pp. 1011-2, Apr. 2011.

[77] N. L. Clement et al., “The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing.,” *Bioinformatics (Oxford, England)*, vol. 26, no. 1, pp. 38-45, Jan. 2010.

[78] P. Krawitz, C. Rödelsperger, M. Jäger, L. Jostins, S. Bauer, and P. N. Robinson, “Microindel detection in short-read sequence data.,” *Bioinformatics (Oxford, England)*, vol. 26, no. 6, pp. 722-9, Mar. 2010.

[79] S. Misra, A. Agrawal, W.-keng Liao, and A. Choudhary, “Anatomy of a hash-based long read sequence mapping algorithm for next generation DNA sequencing.,” *Bioinformatics (Oxford, England)*, vol. 27, no. 2, pp. 189-95, Jan. 2011.

[80] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform.,” *Bioinformatics (Oxford, England)*, vol. 25, no. 14, pp. 1754-60, Jul. 2009.

[81] B. Langmead, “Aligning short sequencing reads with Bowtie.,” *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, vol. 11, p. Unit 11.7, Dec. 2010.

[82] Z. Ning, A. J. Cox, and J. C. Mullikin, “SSAHA: a fast search method for large DNA databases.,” *Genome research*, vol. 11, no. 10, pp. 1725-9, Oct. 2001.

[83] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform.,” *Bioinformatics (Oxford, England)*, vol. 25, no. 14, pp. 1754-60, Jul. 2009.

[84] K. Paszkiewicz and D. J. Studholme, “De novo assembly of short sequence reads.,” *Briefings in bioinformatics*, vol. 11, no. 5, pp. 457-72, Sep. 2010.

[85] W. Zhang, J. Chen, Y. Yang, Y. Tang, J. Shang, and B. Shen, “A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies.,” *PloS one*, vol. 6, no. 3, p. e17915, Jan. 2011.

[86] de B. NG, “A combinatorial problem,” *Nederlandse Akad v. Wetenschappen*, vol. 49, pp. 4758-65, 1946.

[87] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol, “ABySS: a parallel assembler for short read sequence data.,” *Genome research*, vol. 19, no. 6, pp. 1117-23, Jun. 2009.

[88] R. Li et al., “SOAP2: an improved ultrafast tool for short read alignment.,” *Bioinformatics (Oxford, England)*, vol. 25, no. 15, pp. 1966-7, Aug. 2009.

[89] F. Menges, G. Narzisi, and B. Mishra, “TotalReCaller: improved accuracy and performance via integrated alignment and base-calling.,” *Bioinformatics (Oxford, England)*, vol. 27, no. 17, pp. 2330-7, Sep. 2011.

[90] J. T. Simpson and R. Durbin, “Efficient de novo assembly of large genomes using compressed data structures.,” *Genome research*, vol. 22, no. 3, pp. 549-56, Jan. 2012.

[91] P. N. Ariyaratne and W.-K. Sung, “PE-Assembler: de novo assembler using short paired-end reads.,” *Bioinformatics (Oxford, England)*, vol. 27, no. 2, pp. 167-74, Jan. 2011.

[92] P. A. Pevzner, H. Tang, and M. S. Waterman, “An Eulerian path approach to DNA fragment assembly.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 17, pp. 9748-53, Aug. 2001.

[93] M. J. Chaisson and P. A. Pevzner, “Short read fragment assembly of bacterial genomes.,” *Genome research*, vol. 18, no. 2, pp. 324-30, Feb. 2008.

[94] R. L. Warren, G. G. Sutton, S. J. M. Jones, and R. A. Holt, “Assembling millions of short DNA sequences using SSAKE.,” *Bioinformatics (Oxford, England)*, vol. 23, no. 4, pp. 500-1, Feb. 2007.

[95] D. W. Bryant, W.-K. Wong, and T. C. Mockler, “QSRA: a quality-value guided de novo short read assembler.,” *BMC bioinformatics*, vol. 10, p. 69, Jan. 2009.

[96] D. R. Zerbino and E. Birney, “Velvet: algorithms for de novo short read assembly using de Bruijn graphs.,” *Genome research*, vol. 18, no. 5, pp. 821-9, May 2008.

[97] J. Butler et al., “ALLPATHS: de novo assembly of whole-genome shotgun microreads.,” *Genome research*, vol. 18, no. 5, pp. 810-20, May 2008.

[98] N. Nagarajan and M. Pop, “Sequencing and genome assembly using next-generation technologies.,” *Methods in molecular biology (Clifton, N.J.)*, vol. 673, pp. 1-17, Jan. 2010.

[99] C. Kingsford, M. C. Schatz, and M. Pop, “Assembly complexity of prokaryotic genomes using short reads.,” *BMC bioinformatics*, vol. 11, p. 21, Jan. 2010.

[100] R. Li et al., “De novo assembly of human genomes with massively parallel short read sequencing.,” *Genome research*, vol. 20, no. 2, pp. 265-72, Feb. 2010.

[101] R. Li et al., “The sequence and de novo assembly of the giant

panda genome.," *Nature*, vol. 463, no. 7279, pp. 311-7, Jan. 2010.

[102] M. Imelfort and D. Edwards, "De novo sequencing of plant genomes using second-generation technologies.," *Briefings in bioinformatics*, vol. 10, no. 6, pp. 609-18, Nov. 2009.

[103] A. Dayarian, T. P. Michael, and A. M. Sengupta, "SOPRA: Scaffolding algorithm for paired reads via statistical optimization.," *BMC bioinformatics*, vol. 11, p. 345, Jan. 2010.

[104] M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano, "Scaffolding pre-assembled contigs using SSPACE.," *Bioinformatics (Oxford, England)*, vol. 27, no. 4, pp. 578-9, Feb. 2011.

[105] L. Salmela, V. Mäkinen, N. Välimäki, J. Ylinen, and E. Ukkonen, "Fast scaffolding with small independent mixed integer programs.," *Bioinformatics (Oxford, England)*, vol. 27, no. 23, pp. 3259-65, Dec. 2011.

[106] K. J. McKernan et al., "Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding.," *Genome Res*, vol. 19, no. 9, pp. 1527-1541, 2009.

[107] R. Xi, T.-M. Kim, and P. J. Park, "Detecting structural variations in the human genome using next generation sequencing.," *Brief Funct Genomics*, vol. 9, no. 5-6, pp. 405-415, 2010.

[108] P. Senapathy, A. Bhasi, J. Mattox, P. S. Dhandapany, and S. Sadayappan, "Targeted genome-wide enrichment of functional regions.," *PloS one*, vol. 5, no. 6, p. e11138, Jan. 2010.

[109] U. Amstutz, G. Andrey-Zürcher, D. Suciu, R. Jaggi, J. Häberle, and C. R. Largiadèr, "Sequence capture and next-generation resequencing of multiple tagged nucleic acid samples for mutation screening of urea cycle disorders.," *Clinical chemistry*, vol. 57, no. 1, pp. 102-11, Jan. 2011.

[110] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.," *Genome research*, vol. 18, no. 9, pp. 1509-17, Sep. 2008.

[111] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics.,” *Nature reviews. Genetics*, vol. 10, no. 1, pp. 57-63, Jan. 2009.

[112] B. T. Wilhelm et al., “Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.,” *Nature*, vol. 453, no. 7199, pp. 1239-43, Jun. 2008.

[113] U. Nagalakshmi et al., “The transcriptional landscape of the yeast genome defined by RNA sequencing.,” *Science (New York, N.Y.)*, vol. 320, no. 5881, pp. 1344-9, Jun. 2008.

[114] M. Sultan et al., “A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.,” *Science (New York, N.Y.)*, vol. 321, no. 5891, pp. 956-60, Aug. 2008.

[115] H. Kim et al., “A short survey of computational analysis methods in analysing ChIP-seq data.,” *Human genomics*, vol. 5, no. 2, pp. 117-23, Jan. 2011.

[116] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by RNA-Seq.,” *Nature methods*, vol. 5, no. 7, pp. 621-8, Jul. 2008.

[117] R. Koehler, H. Issac, N. Cloonan, and S. M. Grimmond, “The uniqueome: a mappability resource for short-tag sequencing.,” *Bioinformatics (Oxford, England)*, vol. 27, no. 2, pp. 272-4, Jan. 2011.

[118] A. Oshlack and M. J. Wakefield, “Transcript length bias in RNA-seq data confounds systems biology.,” *Biology direct*, vol. 4, p. 14, Jan. 2009.

[119] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, “Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.,” *BMC bioinformatics*, vol. 11, no. 1, p. 94, Jan. 2010.

[120] F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch, “Optimal spliced alignments of short sequence reads.,” *Bioinformatics (Oxford, England)*, vol. 24, no. 16, pp. i174-80, Aug. 2008.

[121] C. Trapnell, L. Pachter, and S. L. Salzberg, “TopHat: discovering splice junctions with RNA-Seq.,” *Bioinformatics (Oxford, England)*, vol. 25, no. 9, pp. 1105-11, May 2009.

[122] J. A. Martin and Z. Wang, “Next-generation transcriptome assembly.,” *Nature reviews. Genetics*, vol. 12, no. 10, pp. 671-82, Oct. 2011.

[123] P. Collas and J. A. Dahl, “Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation.,” *Frontiers in bioscience : a journal and virtual library*, vol. 13, pp. 929-43, Jan. 2008.

[124] H. Ji, H. Jiang, W. Ma, and W. H. Wong, “Using CisGenome to analyze ChIP-chip and ChIP-seq data.,” *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, vol. 2, p. Unit2.13, Mar. 2011.

[125] A. P. Fejes, G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, and S. J. M. Jones, “FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology.,” *Bioinformatics (Oxford, England)*, vol. 24, no. 15, pp. 1729-30, Aug. 2008.

[126] J. Feng, T. Liu, and Y. Zhang, “Using MACS to identify peaks from ChIP-Seq data.,” *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, vol. 2, p. Unit 2.14, Jun. 2011.

[127] J. Rozowsky et al., “PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.,” *Nature biotechnology*, vol. 27, no. 1, pp. 66-75, Jan. 2009.

[128] S. Jiao, C. P. Bailey, S. Zhang, and I. Ladunga, “Probabilistic peak calling and controlling false discovery rate estimations in transcription factor binding site mapping from ChIP-seq.,” *Methods in molecular biology (Clifton, N.J.)*, vol. 674, pp. 161-77, Jan. 2010.

[129] L. X. Garmire, D. G. Garmire, W. Huang, J. Yao, C. K. Glass, and S. Subramaniam, “A global clustering algorithm to identify long intergenic non-coding RNA--with applications in mouse macrophages.,” *PloS one*, vol. 6, no. 9, p. e24051, Jan. 2011.

[130] P. J. Park, “ChIP-seq: advantages and challenges of a maturing technology.,” *Nature reviews. Genetics*, vol. 10, no. 10, pp. 669-80, Oct. 2009.

[131] J. Zhao et al., “Genome-wide identification of polycomb-associated RNAs by RIP-seq.,” *Molecular cell*, vol. 40, no. 6, pp. 939-53, Dec. 2010.

[132] D. P. Bartel, “MicroRNAs: genomics, biogenesis, mechanism, and function.,” *Cell*, vol. 116, no. 2, pp. 281-97, Jan. 2004.

[133] D. Baulcombe, “RNA silencing in plants.,” *Nature*, vol. 431, no. 7006, pp. 356-63, Sep. 2004.

[134] B. C. Meyers et al., “Criteria for annotation of plant MicroRNAs.,” *The Plant cell*, vol. 20, no. 12, pp. 3186-90, Dec. 2008.

[135] O. Barad et al., “MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues.,” *Genome research*, vol. 14, no. 12, pp. 2486-94, Dec. 2004.

[136] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, “miRBase: microRNA sequences, targets and gene nomenclature.,” *Nucleic acids research*, vol. 34, no. Database issue, pp. D140-4, Jan. 2006.

[137] J.-H. Yang, P. Shao, H. Zhou, Y.-Q. Chen, and L.-H. Qu, “deepBase: a database for deeply annotating and mining deep sequencing data.,” *Nucleic acids research*, vol. 38, no. Database issue, pp. D123-30, Jan. 2010.

[138] Q. Jiang et al., “miR2Disease: a manually curated database for microRNA deregulation in human disease.,” *Nucleic acids research*, vol. 37, no. Database issue, pp. D98-104, Jan. 2009.

[139] A.-K. Emde, M. Grunert, D. Weese, K. Reinert, and S. R. Sperling, “MicroRazerS: rapid alignment of small RNA reads.,” *Bioinformatics (Oxford, England)*, vol. 26, no. 1, pp. 123-4, Jan. 2010.

[140] M. R. Friedländer et al., “Discovering microRNAs from deep sequencing data using miRDeep.,” *Nature biotechnology*, vol. 26, no. 4, pp. 407-15, Apr. 2008.

[141] D. Hendrix, M. Levine, and W. Shi, “miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data.,” *Genome biology*, vol. 11, no. 4, p. R39, Jan. 2010.

[142] E. Zhu et al., “mirTools: microRNA profiling and discovery based

on high-throughput sequencing.," *Nucleic acids research*, vol. 38, no. Web Server issue, pp. W392-7, Jul. 2010.

[143] M. Hackenberg, M. Sturm, D. Langenberger, J. M. Falcón-Pérez, and A. M. Aransay, "miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments.," *Nucleic acids research*, vol. 37, no. Web Server issue, pp. W68-76, Jul. 2009.

[144] S. Moxon, F. Schwach, T. Dalmay, D. Maclean, D. J. Studholme, and V. Moulton, "A toolkit for analysing large-scale plant small RNA datasets.," *Bioinformatics (Oxford, England)*, vol. 24, no. 19, pp. 2252-3, Oct. 2008.

[145] R. Ronen et al., "miRNAkey: a software for microRNA deep sequencing analysis.," *Bioinformatics (Oxford, England)*, vol. 26, no. 20, pp. 2615-6, Oct. 2010.

[146] W.-C. Wang, F.-M. Lin, W.-C. Chang, K.-Y. Lin, H.-D. Huang, and N.-S. Lin, "miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression.," *BMC bioinformatics*, vol. 10, p. 328, Jan. 2009.

[147] M. A. German et al., "Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends.," *Nature biotechnology*, vol. 26, no. 8, pp. 941-6, Aug. 2008.

[148] C. Addo-Quaye, T. W. Eshoo, D. P. Bartel, and M. J. Axtell, "Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome.," *Current biology : CB*, vol. 18, no. 10, pp. 758-62, May 2008.

[149] A. Portela and M. Esteller, "Epigenetic modifications and human disease.," *Nature biotechnology*, vol. 28, no. 10, pp. 1057-68, Oct. 2010.

[150] T. A. Manolio et al., "Finding the missing heritability of complex diseases.," *Nature*, vol. 461, no. 7265, pp. 747-53, Oct. 2009.

[151] M. Esteller, "Epigenetics in cancer.," *The New England journal of medicine*, vol. 358, no. 11, pp. 1148-59, Mar. 2008.

[152] M. Weber and D. Schübeler, "Genomic patterns of DNA methylation: targets and function of an epigenetic mark.," *Current opinion in cell biology*, vol. 19, no. 3, pp. 273-80, Jun. 2007.

[153] R. Lister and J. R. Ecker, "Finding the fifth base: genome-wide

sequencing of cytosine methylation.," *Genome research*, vol. 19, no. 6, pp. 959-66, Jun. 2009.

[154] M. Pellegrini and R. Ferrari, "Epigenetic analysis: ChIP-chip and ChIP-seq.," *Methods in molecular biology (Clifton, N.J.)*, vol. 802, pp. 377-87, Jan. 2012.

[155] T. Kubo and K. J. Newton, "Angiosperm mitochondrial genomes and mutations.," *Mitochondrion*, vol. 8, no. 1, pp. 5-14, Jan. 2008.

[156] K. Adams, "Evolution of mitochondrial gene content: gene loss and transfer to the nucleus," *Molecular Phylogenetics and Evolution*, vol. 29, no. 3, pp. 380-395, Dec. 2003.

[157] T. Kubo, S. Nishizawa, and T. Mikami, "Alterations in organization and transcription of the mitochondrial genome of cytoplasmic male sterile sugar beet (Beta vulgaris L.).," *Molecular & general genetics : MGG*, vol. 262, no. 2, pp. 283-90, Sep. 1999.

[158] M. M. Robison and D. J. Wolyn, "Complex organization of the mitochondrial genome of petaloid CMS carrot.," *Molecular genetics and genomics : MGG*, vol. 268, no. 2, pp. 232-9, Oct. 2002.

[159] B. Kmiec, M. Woloszynska, and H. Janska, "Heteroplasmy as a common state of mitochondrial genetic information in plants and animals.," *Current genetics*, vol. 50, no. 3, pp. 149-59, Sep. 2006.

[160] D. J. Oldenburg and A. J. Bendich, "Size and Structure of Replicating Mitochondrial DNA in Cultured Tobacco Cells.," *The Plant cell*, vol. 8, no. 3, pp. 447-461, Mar. 1996.

[161] S. Backert and T. Börner, "Phage T4-like intermediates of DNA replication and recombination in the mitochondria of the higher plant Chenopodium album (L.)," *Current Genetics*, vol. 37, no. 5, pp. 304-314, May 2000.

[162] Y. Cho, "Explosive invasion of plant mitochondria by a group I intron," *Proceedings of the National Academy of Sciences*, vol. 95, no. 24, pp. 14244-14249, Nov. 1998.

[163] M. Unseld, J. R. Marienfeld, P. Brandt, and A. Brennicke, "The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides.," *Nature genetics*, vol. 15, no. 1, pp. 57-61, Jan. 1997.

[164] Y. Notsu et al., "The complete sequence of the rice (Oryza sativa

L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants.," *Molecular genetics and genomics : MGG*, vol. 268, no. 4, pp. 434-45, Dec. 2002.

[165] H. Handa, "The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (Brassica napus L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana," *Nucleic Acids Research*, vol. 31, no. 20, pp. 5907-5916, Oct. 2003.

[166] R. Velasco et al., "A high quality draft consensus sequence of the genome of a heterozygous grapevine variety.," *PloS one*, vol. 2, no. 12, p. e1326, Jan. 2007.

[167] O. Jaillon et al., "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.," *Nature*, vol. 449, no. 7161, pp. 463-7, Sep. 2007.

[168] V. V. Goremykin, F. Salamini, R. Velasco, and R. Viola, "Mitochondrial DNA of Vitis vinifera and the issue of rampant horizontal gene transfer.," *Molecular biology and evolution*, vol. 26, no. 1, pp. 99-110, Jan. 2009.

[169] M. Takenaka, D. Verbitskiy, J. A. van der Merwe, A. Zehrmann, and A. Brennicke, "The process of RNA editing in plant mitochondria.," *Mitochondrion*, vol. 8, no. 1, pp. 35-46, Jan. 2008.

[170] M. W. Gray, "Diversity and evolution of mitochondrial RNA editing systems.," *IUBMB life*, vol. 55, no. 4-5, pp. 227-33.

[171] F. Kempken, W. Howard, and D. R. Pring, "Mutations at specific atp6 codons which cause human mitochondrial diseases also lead to male sterility in a plant.," *FEBS letters*, vol. 441, no. 2, pp. 159-60, Dec. 1998.

[172] J. P. Mower and J. D. Palmer, "Patterns of partial RNA editing in mitochondrial genes of Beta vulgaris.," *Molecular genetics and genomics : MGG*, vol. 276, no. 3, pp. 285-93, Sep. 2006.

[173] E. W. Sayers et al., "Database resources of the National Center for Biotechnology Information.," *Nucleic acids research*, vol. 40, no. Database issue, pp. D13-25, Jan. 2012.

[174] X. Huang, J. Wang, S. Aluru, S.-P. Yang, and L. Hillier, "PCAP: a whole-genome assembly program.," *Genome research*, vol. 13, no. 9, pp. 2164-70, Sep. 2003.

[175] E. Picardi, T. M. R. Regina, A. Brennicke, and C. Quagliariello, “REDIdb: the RNA editing database.,” *Nucleic acids research*, vol. 35, no. Database issue, pp. D173-7, Jan. 2007.

[176] D. Campagna et al., “PASS: a program to align short sequences.,” *Bioinformatics (Oxford, England)*, vol. 25, no. 7, pp. 967-8, Apr. 2009.

[177] A. J. Iafrate et al., “Detection of large-scale variation in the human genome.,” *Nat Genet*, vol. 36, no. 9, pp. 949-951, 2004.

[178] J. M. Kidd et al., “Mapping and sequencing of structural variation from eight human genomes.,” *Nature*, vol. 453, no. 7191, pp. 56-64, 2008.

[179] E. Tuzun et al., “Fine-scale structural variation of the human genome.,” *Nat Genet*, vol. 37, no. 7, pp. 727-732, 2005.

[180] J. Sebat et al., “Large-scale copy number polymorphism in the human genome.,” *Science*, vol. 305, no. 5683, pp. 525-528, 2004.

[181] L. Feuk, A. R. Carson, and S. W. Scherer, “Structural variation in the human genome.,” *Nature reviews. Genetics*, vol. 7, no. 2, pp. 85-97, Feb. 2006.

[182] 1000 G. P. Consortium, “A map of human genome variation from population-scale sequencing.,” *Nature*, vol. 467, no. 7319, pp. 1061-1073, 2010.

[183] P. J. Campbell et al., “Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.,” *Nat Genet*, vol. 40, no. 6, pp. 722-729, 2008.

[184] J. Sebat et al., “Strong association of de novo copy number mutations with autism.,” *Science*, vol. 316, no. 5823, pp. 445-449, Apr. 2007.

[185] P. Stankiewicz and J. R. Lupski, “Structural variation in the human genome and its role in disease.,” *Annu Rev Med*, vol. 61, pp. 437-455, 2010.

[186] A. Itsara et al., “Population analysis of large copy number variants and hotspots of human genetic disease.,” *American journal of human genetics*, vol. 84, no. 2, pp. 148-61, Feb. 2009.

[187] A. M. Snijders et al., “Assembly of microarrays for genome-wide

measurement of DNA copy number.," *Nature genetics*, vol. 29, no. 3, pp. 263-4, Nov. 2001.

[188] D. T. Miller et al., "Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies.," *American journal of human genetics*, vol. 86, no. 5, pp. 749-64, May 2010.

[189] D. A. Peiffer et al., "High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping.," *Genome research*, vol. 16, no. 9, pp. 1136-48, Sep. 2006.

[190] G. M. Cooper, T. Zerr, J. M. Kidd, E. E. Eichler, and D. A. Nickerson, "Systematic assessment of copy number variant detection via genome-wide SNP genotyping.," *Nature genetics*, vol. 40, no. 10, pp. 1199-203, Oct. 2008.

[191] M. D. Mailman et al., "The NCBI dbGaP database of genotypes and phenotypes.," *Nature genetics*, vol. 39, no. 10, pp. 1181-6, Oct. 2007.

[192] J. M. Kidd et al., "Characterization of missing human genome sequences and copy-number polymorphic insertions," *Nature Methods*, vol. 7, no. 5, pp. 365-371, Apr. 2010.

[193] D. P. Locke, "BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications," *Journal of Medical Genetics*, vol. 41, no. 3, pp. 175-182, Mar. 2004.

[194] S. Volik et al., "End-sequence profiling: sequence-based analysis of aberrant genomes.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 13, pp. 7696-701, Jun. 2003.

[195] P. Medvedev, M. Stanciu, and M. Brudno, "Computational methods for discovering structural variation with next-generation sequencing.," *Nat Methods*, vol. 6, no. 11, p. S13--S20, Nov. 2009.

[196] R. E. Mills et al., "Mapping copy number variation by population-scale genome sequencing.," *Nature*, vol. 470, no. 7332, pp. 59-65, Feb. 2011.

[197] J. O. Korbel et al., “Paired-end mapping reveals extensive structural variation in the human genome.,” *Science*, vol. 318, no. 5849, pp. 420-426, 2007.

[198] J. M. Kidd et al., “A human genome structural variation sequencing resource reveals insights into mutational mechanisms.,” *Cell*, vol. 143, no. 5, pp. 837-47, Nov. 2010.

[199] S. Volik et al., “End-sequence profiling: sequence-based analysis of aberrant genomes.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 13, pp. 7696-701, Jun. 2003.

[200] J. O. Korbel et al., “PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data.,” *Genome Biol*, vol. 10, no. 2, p. R23, 2009.

[201] F. Hormozdiari et al., “Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery.,” *Bioinformatics*, vol. 26, no. 12, p. i350--i357, 2010.

[202] K. Chen et al., “BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.,” *Nat Methods*, vol. 6, no. 9, pp. 677-681, 2009.

[203] S. Lee, F. Hormozdiari, C. Alkan, and M. Brudno, “MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions.,” *Nat Methods*, vol. 6, no. 7, pp. 473-474, 2009.

[204] J. A. Bailey et al., “Recent segmental duplications in the human genome.,” *Science (New York, N.Y.)*, vol. 297, no. 5583, pp. 1003-7, Aug. 2002.

[205] D. Y. Chiang et al., “High-resolution mapping of copy-number alterations with massively parallel sequencing.,” *Nat Methods*, vol. 6, no. 1, pp. 99-103, 2009.

[206] C. Alkan et al., “Personalized copy number and segmental duplication maps using next-generation sequencing.,” *Nat Genet*, vol. 41, no. 10, pp. 1061-1067, 2009.

[207] P. H. Sudmant et al., “Diversity of human copy number variation and multicopy genes.,” *Science (New York, N.Y.)*, vol. 330, no. 6004, pp. 641-6, Oct. 2010.

[208] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, “Sensitive and

accurate detection of copy number variants using read depth of coverage.," *Genome research*, vol. 19, no. 9, pp. 1586-92, Sep. 2009.

[209] A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein, "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing.," *Genome research*, vol. 21, no. 6, pp. 974-84, Jun. 2011.

[210] R. E. Mills et al., "An initial map of insertion and deletion (INDEL) variation in the human genome.," *Genome research*, vol. 16, no. 9, pp. 1182-90, Sep. 2006.

[211] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads.," *Bioinformatics*, vol. 25, no. 21, pp. 2865-2871, Nov. 2009.

[212] S. Levy et al., "The diploid genome sequence of an individual human.," *PLoS biology*, vol. 5, no. 10, p. e254, Sep. 2007.

[213] A. W. Pang et al., "Towards a comprehensive structural variation map of an individual human genome.," *Genome biology*, vol. 11, no. 5, p. R52, Jan. 2010.

[214] J. Xing et al., "Mobile elements create structural variation: analysis of a complete human genome.," *Genome research*, vol. 19, no. 9, pp. 1516-26, Sep. 2009.

[215] X. She et al., "Shotgun sequence assembly and recent segmental duplications within the human genome.," *Nature*, vol. 431, no. 7011, pp. 927-30, Oct. 2004.

[216] C. Alkan, S. Sajjadian, and E. E. Eichler, "Limitations of next-generation genome sequence assembly.," *Nature methods*, vol. 8, no. 1, pp. 61-5, Jan. 2011.

[217] M. C. Schatz, A. L. Delcher, and S. L. Salzberg, "Assembly of large genomes using second-generation sequencing.," *Genome research*, vol. 20, no. 9, pp. 1165-73, Sep. 2010.

[218] C. A. Albers, G. Lunter, D. G. MacArthur, G. McVean, W. H. Ouwehand, and R. Durbin, "Dindel: accurate indel calls from short-read data.," *Genome research*, vol. 21, no. 6, pp. 961-73, Jun. 2011.

[219] R. E. Handsaker, J. M. Korn, J. Nemesh, and S. A. McCarroll,

"Discovery and genotyping of genome structural polymorphism by sequencing on a population scale.," *Nature genetics*, vol. 43, no. 3, pp. 269-76, Mar. 2011.

[220] R. E. Mills et al., "Natural genetic variation caused by small insertions and deletions in the human genome.," *Genome research*, vol. 21, no. 6, pp. 830-9, Jun. 2011.

[221] N. Whiteford et al., "An analysis of the feasibility of short read sequencing.," *Nucleic acids research*, vol. 33, no. 19, p. e171, Jan. 2005.

[222] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol, "ABySS: a parallel assembler for short read sequence data.," *Genome research*, vol. 19, no. 6, pp. 1117-23, Jun. 2009.

[223] X. Ruan and Y. Ruan, "Genome wide full-length transcript analysis using 5' and 3' paired-end-tag next generation sequencing (RNA-PET).," *Methods in molecular biology (Clifton, N.J.)*, vol. 809, pp. 535-62, Jan. 2012.

[224] G. F. Hong, "A method for sequencing single-stranded cloned DNA in both directions.," *Bioscience reports*, vol. 1, no. 3, pp. 243-52, Mar. 1981.

[225] E. JD Kececioglu and Myers, "Combinatorial algorithms for DNA sequence assembly," *Algotithmica*, vol. 13, pp. 7-51, 1993.

[226] M. Pop, D. S. Kosack, and S. L. Salzberg, "Hierarchical scaffolding with Bambus.," *Genome research*, vol. 14, no. 1, pp. 149-59, Jan. 2004.

[227] H. Stockmann-Juvala and K. Savolainen, "A review of the toxic effects and mechanisms of action of fumonisin B1.," *Human & experimental toxicology*, vol. 27, no. 11, pp. 799-809, Nov. 2008.

[228] J. Merhej, F. Richard-Forget, and C. Barreau, "Regulation of trichothecene biosynthesis in Fusarium: recent advances and new insights.," *Applied microbiology and biotechnology*, vol. 91, no. 3, pp. 519-28, Aug. 2011.

[229] E. Bergamini et al., “Fate of Fusarium mycotoxins in the cereal product supply chain: the deoxynivalenol (DON) case within industrial bread-making technology.,” *Food additives & contaminants. Part A, Chemistry, analysis, control, exposure & risk assessment*, vol. 27, no. 5, pp. 677-87, May 2010.

[230] S. Döll and S. Dänicke, “The Fusarium toxins deoxynivalenol (DON) and zearalenone (ZON) in animal feeding.,” *Preventive veterinary medicine*, vol. 102, no. 2, pp. 132-45, Nov. 2011.

[231] L.-J. Ma et al., “Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium.,” *Nature*, vol. 464, no. 7287, pp. 367-73, Mar. 2010.

[232] C. B. Michielse and M. Rep, “Pathogen profile update: Fusarium oxysporum.,” *Molecular plant pathology*, vol. 10, no. 3, pp. 311-24, May 2009.

[233] B. Lievens, M. Rep, and B. P. H. J. Thomma, “Recent developments in the molecular discrimination of formae speciales of Fusarium oxysporum.,” *Pest management science*, vol. 64, no. 8, pp. 781-8, Aug. 2008.

Appendix 1

Published online 12 April 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 14* ***4755–4767***
doi:10.1093/nar/gkq202

Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing

Ernesto Picardi[1], David S. Horner[2], Matteo Chiara[2], Riccardo Schiavon[3], Giorgio Valle[3] and Graziano Pesole[1,4,*]

[1]Dipartimento di Biochimica e Biologia Molecolare 'E. Quagliariello', Università degli Studi di Bari, 70126 Bari, [2]Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli Studi di Milano, 20133 Milano, [3]CRIBI, Università degli Studi di Padova, viale G. Colombo 3, 35121 Padova and [4]Istituto Tecnologie Biomediche del Consiglio Nazionale delle Ricerche, via Amendola 122/D, 70125 Bari, Italy

Received September 7, 2009; Revised and Accepted March 9, 2010

ABSTRACT

RNA editing is a widespread post-transcriptional molecular phenomenon that can increase proteomic diversity, by modifying the sequence of completely or partially non-functional primary transcripts, through a variety of mechanistically and evolutionarily unrelated pathways. Editing by base substitution has been investigated in both animals and plants. However, conventional strategies based on directed Sanger sequencing are time-consuming and effectively preclude genome wide identification of RNA editing and assessment of partial and tissue-specific editing sites. In contrast, the high-throughput RNA-Seq approach allows the generation of a comprehensive landscape of RNA editing at the genome level. Short reads from Solexa/Illumina GA and ABI SOLiD platforms have been used to investigate the editing pattern in mitochondria of *Vitis vinifera* providing significant support for 401 C-to-U conversions in coding regions and an additional 44 modifications in non-coding RNAs. Moreover, 76% of all C-to-U conversions in coding genes represent partial RNA editing events and 28% of them were shown to be significantly tissue specific. Solexa/Illumina and SOLiD platforms showed different characteristics with respect to the specific issue of large-scale editing analysis, and the combined approach presented here reduces the false positive rate of discovery of editing events.

INTRODUCTION

Next-generation sequencing platforms (Solexa/Illumina GA, ABI SOLiD and Roche 454) are radically changing the field of genomics (1,2), allowing both re-sequencing and *de novo* sequencing of whole genomes (3) with notable reductions in time and cost with respect to conventional approaches. These technologies are now routinely applied to a variety of functional genomics problems, including, but not restricted to, global identification of genomic rearrangements, investigation of epigenetic modifications and single nucleotide polymorphism (SNP) discovery (4). RNA-Seq—the application of next generation sequencing to entire transcriptomes—can provide accurate gene expression profiles for coding and non-coding RNAs (5) greatly facilitating genome annotation (6).

RNA editing is a widespread post-transcriptional molecular phenomenon that can increase proteomic diversity (7) by modifying the sequence of completely or partially non-functional primary transcripts (8), through a variety of mechanistically and evolutionarily unrelated pathways. 'Substitution' editing by simple base modification is the most frequent type of editing and is seen both in plant organelles and in the nucleus of higher eukaryotes (8–11) as well as in sequences of viral origin (12). In land plant organelles, RNA editing consists almost exclusively of C-to-U substitutions (rarely reverse U-to-C conversions) mostly at first or second positions of codons (9)—typically leading to conservative amino-acid changes and increasing similarity to non-plant homologs. Some plant organellar RNA editing events create translation initiation or termination codons while several known editing events in tRNA or introns improve the stability of functionally relevant secondary structure motifs (13,14). The systematic identification of RNA editing events thus represents an important objective that could significantly improve our understanding of organellar and nuclear molecular genetics. Moreover, the alteration of the RNA editing pattern in plant mitochondria can lead to male sterility, also known as the CMS phenotype (15).

Classically, RNA editing events were identified experimentally by comparing cloned cDNA sequences with their corresponding genomic templates (16). This procedure

*To whom correspondence should be addressed. Tel: +39 080 544 3588; Fax: +39 080 544 3317; Email: graziano.pesole@biologia.uniba.it

allows the study of a relatively small number of sequences and does not take into account potential cloning artefacts. More recently, large-scale identification of RNA editing sites has been performed using collections of expressed sequence tags (ESTs) and full-length cDNAs mainly stored in public databases (17,18). However, the generally low quality of EST sequences, and the incomplete nature of some editing events markedly hampers such approaches. Indeed, C-to-U editing has been explored at the whole mitochondrial (mt) genome level in only four higher plants, *Arabidopsis thaliana* (19), *Brassica napus* (16), *Beta vulgaris* (20) and *Oryza sativa* (21). High-throughput transcriptome sequencing by next-generation technologies provides deep coverage per reference nucleotide and indications of base call qualities and may overcome existing limitations and improve the large-scale detection of RNA editing sites.

Recently, human RNA editing sites have been identified using massively parallel target capture and DNA sequencing employing computationally predicted A-to-I sites (22). In another approach, Life Science (Roche) 454 Amplicon Sequencing technology has been used to determine global expression of known RNA editing sites during brain development (23).

In the present work, focused on the *de novo* detection of C-to-U editing modifications occurring in coding and non-coding genes of the *Vitis vinifera* mitochondrial genome, we also present a novel strategy to investigate the landscape of RNA editing at the genome level through RNA-Seq. This strategy involves the use of millions of short reads generated by Solexa/Illumina GA and ABI SOLiD systems. Over 6 000 000 short reads (from both platforms) mapping uniquely onto the grapevine mitochondrial genome provided significant support for 401 C-to-U alterations in coding regions. Sixty percent of the identified events occurred at second codon positions. Forty-four additional editing modifications (38 C-to-U and 6 U-to-C) were identified in tRNAs and group II introns, supporting the notion of pervasive RNA editing in grape mitochondria. Interestingly, 76% out of all C-to-U conversions in coding genes represent partial RNA editing, and 28% of them were shown to be significantly tissue specific.

In this study, we prove the effectiveness of RNA-Seq data for the the global identification of RNA editing sites and the relative performances of the Solexa/Illumina GA and ABI SOLiD systems to reliably identify editing sites. The computational strategy presented here can be applied to the discovery of substitution editing events of any type in both nuclear and organellar compartments of different organisms.

MATERIALS AND METHODS

Assembly and annotation of the PN40024 mitochondrial genome

Ad-hoc perl scripts making use of the NCBI Blast URL API were used to automate similarity searches of the PN40024 genome sequencing project trace archive with overlapping 10-kb windows of the Pinot Noir ENTAV115 mitochondrial genome [GenBank: NC_012119]. Only traces showing greater than 95% identity to the ENTAV115 genome were retained. The 'query_tracedb' script provided by NCBI was used to recover sequences and associated quality scores (16 789 putative mitochondrial sequences of which 13 682 were identified as mate pairs). The average read length was 785 bases, implying a hypothetical redundancy of greater than 20 times. The software PCAP (24) was used, without reference to the ENTAV115 sequence, to assemble four contigs of 339 264, 132 252, 202 123 and 76 068 nt. Our assembly represented 96.37% of the reference sequence, with which it showed 99.92% identity. Similarity searches using the ENTAV115 annotation allowed the identification of all of the genes of mitochondrial origin proposed by Goremykin *et al*. (25). In addition, the mitochondrial origin of each coding gene was confirmed comparing grape ORFs to genomic and unedited mitochondrial genes downloaded from the specialized REDIdb database (http://biologia.unical.it/py_script/search.html) (26).

Short read sequencing and mapping

In total, 205 435 765 short reads were obtained by sequencing cDNA obtained from four tissue samples with the Solexa/Illumina technology: leaf (11 lanes), root (9 lanes), callus (9 lanes), stem (14 lanes) (6). The mRNA molecules were purified from total RNA extractions and fragmented before cDNA synthesis. The single-end reads obtained were 33-nt long, except for five lanes in the callus sample, where the reads were 35-nt long. Total RNA from PN40024 grape cultivar was sequenced with the SOLiD-2 technology, resulting in 139 467 080 short reads from leaf and 188 742 647 short reads from root. All SOLiD short reads were 35-nt long. For the construction of the SOLiD libraries we had early access to the Applied Biosystems Whole Transcriptome Shotgun procedure. Poly(A)+ RNA was enzymatically fragmented and directionally ligated to adaptors, essentially as indicated in the AMBION Small RNA Expression Kit (SREK).

Solexa/Illumina and SOLiD short tags, pooled from all tissues, were mapped to the assembled *V. vinifera* mitochondrial genome using version 0.5 of the PASS software (27) with a seed length of 12, a minimum identity of 90% and a minimum alignment length per read of 30 nt. Similar to a BLAST approach, PASS seed sequences (called long word anchors) are extended on the flanking regions using DNA words of predefined length (typically 6 or 7 bases) for which the alignment scores are pre-computed according to Needlelman–Wunch. Significant matches are then refined to improve the global alignment quality. In particular, we used a pre-computed scoring matrix (PST) based on DNA words of 7 bases long (W7M1m0G0X0.pst, downloadable from the PASS web site: http://pass.cribi.unipd.it/), filtering hits having more than 11 discrepancies. Moreover, we filtered out Solexa/Illumina and SOLiD short tags containing more than 5 bases with a quality threshold less than 15. In case of Solexa/Illumina reads, we used the -gff option to print out mapping results in the standard GFF (version 3)

format (see http://www.sequenceontology.org/gff3.shtml for more details about this format).

SOLiD reads, derived from a ligation-mediated sequencing strategy, are not collected as nucleotide sequences, but instead are recorded in color space where each color provides information about two adjacent bases but their identification is not provided (a complete description of 2-base color codes can be found at the ABI web site http://www3.appliedbiosystems.com/AB_Home/). For this reason, SOLiD data need a distinct processing method, including an accurate decoding step in which color reads are converted to sequence reads. However, decoding should not be performed before mapping because sequencing errors may affect the translation to base space leading to significant inaccuracies. Therefore, we mapped SOLiD reads to the known reference within color space, again using PASS, allowing at most four color mismatches using the option -SOLiDCS. Next, resulting query-to-reference alignments in color space were parsed by custom python scripts in order to correct sequencing errors and identify isolated color changes corresponding to valid base space mismatches (main scripts are available upon request). Moreover, we performed a further modification—using SOLiD quality scores per single base to reliably call individual nucleotides. For the SOLiD technology, a quality score is assigned to each color (corresponding to a pair of adjacent nucleotides) and each nucleotide (except the first and the last) is read twice as it is included in two adjacent colors. Consequently, a per-base quality score can be reasonably assigned calculating the average quality between two adjacent colors (i.e. two overlapping dinucleotides). If two neighboring colors have high quality scores, the nucleotide in common between them has a high quality score. If two adjacent colors have very different quality scores we call the base in common between them according to a defined quality threshold. The threshold, set at 15 for both Solexa/Illumina and SOLiD reads was generated considering the distribution of detected quality scores per base and considering the fact that SOLiD quality values are also calculated using a phred-like scale.

SOLiD mapping results, in addition to potential mismatches, were finally saved in GFF format. Solexa/Illumina and SOLiD mapping data in GFF format are available upon request.

Computational identification of RNA editing sites

Solexa/Illumina and SOLiD mapping results in GFF format were used to identify C-to-U changes due to RNA editing in the grape mitochondrial genome of the cultivar PN40024 by means of *ad hoc* custom python scripts.

The main script, in particular, takes as input a GFF file, the reference sequence of the grape mitochondrial genome in FASTA format and a textual file containing protein-coding annotations. It collects all uniquely mapping reads (with at most two mismatches and no indels) falling in annotated genes and for each reference position calls the corresponding read nucleotide if its quality score is above the fixed threshold of 15. Finally, for each reference position, the script calculates the frequency of the modified nucleotide (if any) over the total recorded signal (sum of modified and not modified nucleotides) (Figure 1). Results obtained from Solexa/Illumina and SOLiD data are available as Supplementary Data in tab-formatted text files.

RNA editing sites due to C-to-U changes were detected separately for each platform and tissue. Rates of sequencing errors were estimated for each sample as the total frequency of non-C↔U substitutions. Among the potential editing sites, corresponding to sites where a genomic C was aligned to one or more U from RNA-Seq data, statistically significant editing sites were determined by applying the Fisher's exact test by comparing the observed and expected C and U occurrences in the aligned reads. A confidence level of 0.05 (also with FDR or Bonferroni correction) was used as cut-off.

A putative editing site is classified as 'conserved' if one or more homologous sites in other plants are experimentally known to be edited or if a fully conserved U is observed in all homologous sites, according to the data collected in the REDIdb database (26).

RNA editing sites in non-coding grapevine genes and group II introns were detected according to the same computational strategy. These results are also available as Supplementary Data.

Statistically significant edited sites have been classified fully or partially edited depending on if the observed fraction of RNA-Seq aligned U was above or below 90%.

All statistically significant RNA editing events have been submitted to the specialized REDIdb database (http://biologia.unical.it/py_script/search.html) (26) and can be freely consulted in their gene context under the accessions EDI_000000804–EDI_000000840. Finally, data providing additional editing information per each coding gene, tissue and platform, including short read coverage per gene and single reference position, are supplied as Supplementary Data.

Characterization of grape mitochondrial editing sites

All statistics to characterize detected RNA editing sites in grape mitochondrial protein-coding genes, including affected codon positions and amino acid changes, were calculated by custom python scripts. The effect of RNA editing alterations in tRNA genes was evaluated according to secondary structure predictions by the tRNA-Scan web server (http://lowelab.ucsc.edu/tRNAscan-SE/) (28), whereas the impact of C-to-U modifications in the domain V of the group II intron *nad7i4* was manually checked.

Tissue-specific editing sites were identified by means of a chi-square statistical test comparing for each edited position the observed and expected distributions of Cs and Us in all available tissues. Three degrees of freedom were used for Solexa/Illumina data (four tissues) and one for SOLiD reads (two tissues). Significant sites were detected at 0.05 and 0.01 confidence levels, corrected for false discovery rate according to Benjamini and Hochberg (29). The Bonferroni correction, while highly conservative, was also used.

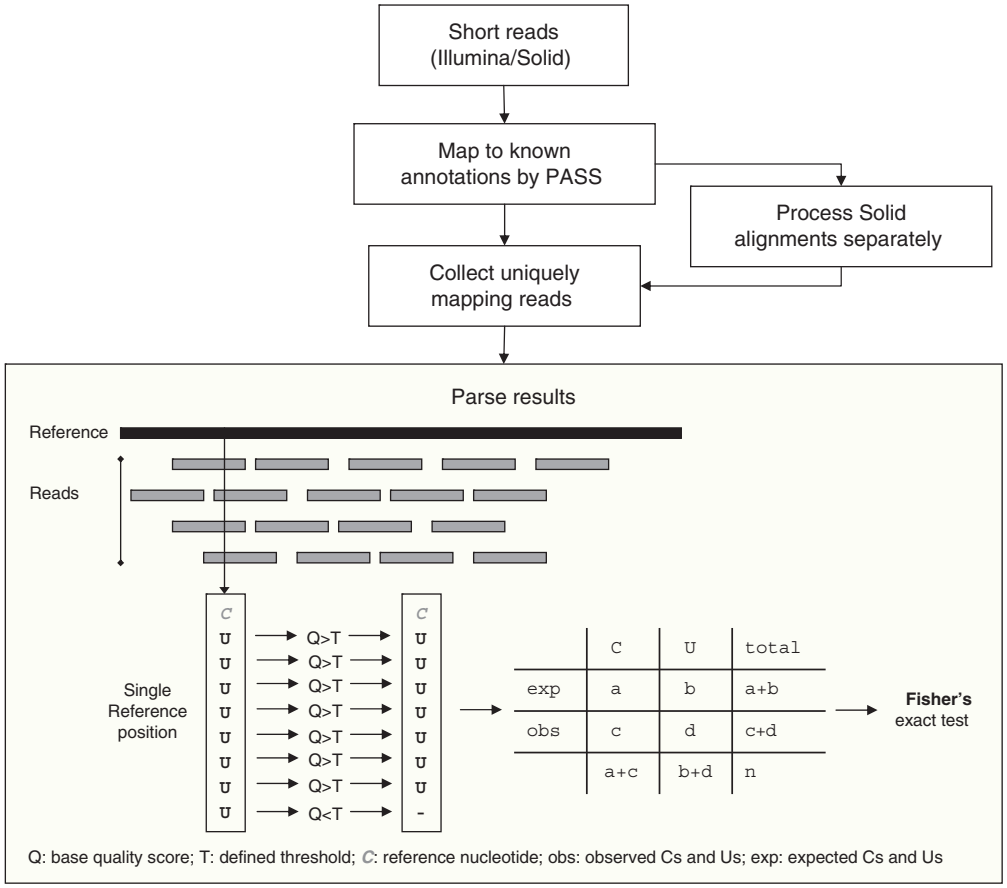


Figure 1. Methodology overview. Graphical overview of the computational methodology used to detect RNA editing sites by short sequencing reads of next generation platforms.

Nucleotide sequences (40-bp long) around RNA editing sites detected in protein-coding genes were examined in terms of relative entropy using windows of 1, 2 or 3 bases, according to the computational methodology described by Mulligan *et al.* (30). Sequence logos were generated by the WebLogo program (version 3) (31).

Domain searches in edited and unedited grapevine mitochondrial genes were performed through the Pfam webserver (http://pfam.sanger.ac.uk/search) using 1.0*e*–05 as E-value cut-off (32).

RNA editing in *A. thaliana* mitochondria

To detect C-to-U changes in mitochondria of *A. thaliana*, we used 63 850 661 Solexa/Illumina short reads (distributed over five runs) from floral tissue of Col-0 ecotype downloaded from NCBI Short Read Archive under the accession SRX002554. All these reads, each of 50 nt in length, were mapped onto the reference *Arabidopsis* mitochondrial genome [GenBank:NC_001284] using PASS with settings as described above. Potential RNA editing sites were identified according to the computational strategy previously explained. Known *Arabidopsis* C-to-U substitutions were downloaded from REDIdb database and used to identify new editing sites.

RESULTS

Grapevine mitochondrial genome assembly and annotation

The complete mitochondrial genome sequence of the Pinot Noir, clone ENTAV115 was recently presented by Goremykin *et al.* (25). The genome is, at over 773 kb in length, the largest sequenced higher plant mitochondrial genome. Notably, Goremykin *et al.* (25) estimate that >42% of the *Vitis* plastid genome has been incorporated into the mitochondrial sequence, and the high similarity of such sequences to their plastidic forbearers (25) indicates that such transfers have occurred recently. While plant mitochondrial-coding regions tend to show extremely high levels of conservation (33), for the purposes of the current study, we wished to compare transcriptome reads to genomic templates derived from identical cultivars (PN40024). Accordingly, we used overlapping windows along the Goremykin *et al.* sequence (25) to perform similarity searches against the PN40024 genome sequencing project trace archive (Sanger sequencing reads) (34). Assembly of 16 789 putatively mitochondrial reads yielded four contigs covering 96.37% of the *Vitis* mitochondrial template. Interestingly, the positions where assembly of contigs was not possible consistently corresponded to regions containing large plastid-like insertions in the Goremykin *et al.* assembly (25), suggesting either that

some such insertions occurred after the divergence of the two cultivars in question or that some such regions have undergone elimination or rearrangement after the divergence of the two clones. Unsurprisingly, similarity searches allowed us to confidently identify all 37 mitochondrial genes (24 components of the respiratory chain and 13 ribosomal proteins) previously annotated (25), in addition, we were able to identify 13 tRNA genes of mitochondrial origin and a number of potentially functional tRNAs of plastidic origin. Protein-coding regions were almost identical to those previously identified by Goremykin *et al.* (25). Indeed within the 37 protein-coding genes of mitochondrial origin studied in the current work, only a single potential synonymous polymorphism was identified between the two clones. A detailed description of patterns of variability between non-coding portions of grapevine mitochondrial genomes will be presented elsewhere. The PN40024 mitochondrial genome contigs are available through Genbank under accessions GQ220323, GQ220324, GQ220325 and GQ220326.

Computational strategy to detect RNA editing sites by short sequencing reads

The strategy proposed here is conceptually simple, computationally tractable, and suitable for Solexa/Illumina and SOLiD short sequencing RNA reads. In the first part of our approach, depicted in Figure 1, we mapped and aligned short reads to the reference genome using the PASS software (27) (see 'Materials and Methods' section for more details). To reduce inconsistent results, we retained only alignments of at least 30 nt in length with a minimum identity of 90% and no indels. In addition, problematic reads were discarded *a priori* by setting PASS (27) quality parameters as described in 'Materials and Methods' section. We recovered only reads mapping once to the reference sequence with at most two mismatches. For each reference position we collected all corresponding reads, scoring hits only if their corresponding quality scores were above a defined threshold (Figure 1). In this way, potential sequencing errors are minimized obtaining a high confidence set of bases per reference position.

RNA editing sites are finally detected by interrogating the reference position by position. A site is considered potentially edited if a C is observed in the reference genome and one or more U in the aligned reads at the same position. The Fisher's exact test has been carried out, as described in 'Material and Methods' section, to assess the statistical significance of each potentially edited site. This statistical assessment was performed separately for every tissue and platform to account for tissue specificity and the different features of Solexa/Illumina and SOLiD systems. Indeed, Solexa/Illumina and SOLiD platforms show different behaviours in terms of base substitution pattern (see below for details) and coverage per base that may affect the identification of genuine editing sites increasing the false discovery rate.

Editing of grapevine mitochondrial RNAs is revealed by Solexa/Illumina and SOLiD RNA-seq

RNA editing in higher plant mitochondria (predominantly C-to-U conversions) represents one of the most investigated types of editing (9), although its molecular mechanism is yet largely unknown (11). Data stored in primary and specialized databases indicate that the mitochondrial genomes of *A. thaliana*, *B. napus*, *B.vulgaris* and *O. sativa* contain 441, 427, 357 and 491 C-to-U edited sites, respectively. We analyzed 205 million reads obtained by Solexa/Illumina technology from four different tissues (stem, root, callus and leaf) as well as 328 million reads produced by SOLiD technology from leaf and root tissues of the highly homozygous PN40024 clone.

We aligned Solexa/Illumina and SOLiD reads to grape PN40024 mitochondrial contigs, recovering 939 554 unique Solexa/Illumina alignments and 5 207 827 unique SOLiD alignments. The different fraction of uniquely aligned reads (0.45 and 1.59% for Solexa/Illumina and SOLiD, respectively) also reflect quite different coverage patterns, which seem much more biased for SOLiD (Supplementary Table S1). We noted that despite the much higher overall fold coverage of SOLiD (158×) than SOLEXA (35×) both platforms provided a similar percentage of covered nucleotides in the coding regions, 96.9 and 96.6%, respectively (see Supplementary Table S1). Furthermore, 16 out of 37 annotated mitochondrial coding genes were fully covered by Solexa/Illumina reads while only 11 were fully supported by SOLiD data (Supplementary Figures S1–S3). Looking at reads distribution along the reference sequence, we also noted local maxima in SOLiD reads in which several mitochondrial regions appeared deeply covered.

While the patterns of coverage seem to indicate a notable bias in the per-site distribution of the coverage depth across coding genes for the SOLiD data, a moderate, but highly significant ($r = 0.25$, $P < 0.0001$) correlation was observed between per base coverage by SOLiD and Solexa/Illumina sequencing for individual positions in the coding sequences of the 37 genes of mitochondrial ancestry—possibly due to a known dependence of recovery of fragmented cDNA (by gel elution) on GC content (35). However, distinct coverage patterns by these different sequencing strategies contribute to a substantially higher coverage when both technologies were combined—complete coverage of 25 genes out of the 37 and an overall coverage of 98.3% of all coding nucleotides (see an example in Figure 2 or extended images in Supplementary Figures S1–S3).

Both *Vitis* mitochondrial assemblies harbor two identical copies of *rps19*, one upstream of the *rps3* and *rpl16* genes and another downstream of a pseudo *atp1* gene. Experimental data suggest that the evolutionarily conserved cluster *rps19*, *rps3* and *rpl16* is transcribed as a polycistronic RNA in land plants (36). When only reads that map uniquely to the genome were considered, the *rps19* gene was, unsurprisingly, not covered. When we allowed the use of reads mapping on at most two genome locations, we found eight C-to-U modifications in the *rps19* coding region, three occurring at the third

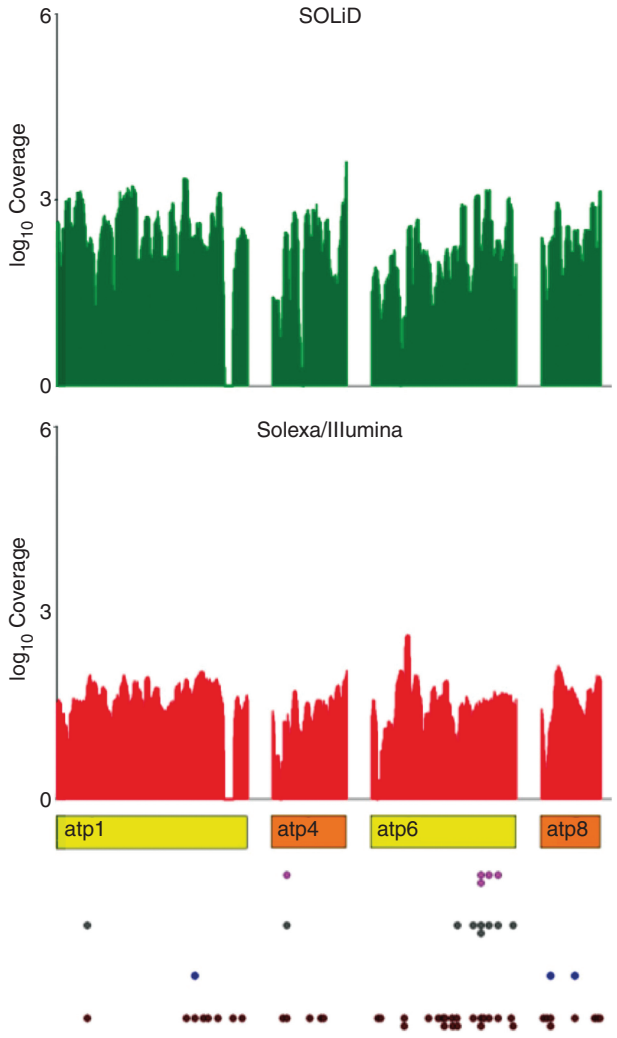


Figure 2. Solexa/Illumina and SOLiD coverage for four mitochondrial atp genes. Plot showing the coverage depth for Solexa/Illumina and SOLiD reads in four mitochondrial genes coding for subunits of the atp synthase. Rectangles in colour indicate protein-coding genes (orange if the strand is forward and yellow if the strand is reverse). RNA editing sites are drawn as brown colored dots below each gene. Blue and grey dots designate the subset of tissue specific sites at 0.05 and 0.01 confidence levels, respectively. Sites passing the Bonferroni test at 1% confidence level are indicated in magenta. Coverage is reported in $\log_{10}$ scale.

codon position and the remaining five in non-synonymous positions that were also conserved across different land plants, except for an event at position 260 that seems to be grapevine specific. We cannot, with confidence, establish if one or both the copies of *rps19* are expressed, although the confirmed expression of *rps3* and *rpl16* genes suggest that at least the *rps19* copy completing the canonical gene cluster should be transcribed.

In total we identified 401 significantly supported editing sites in grapevine mitochondrial coding regions with a 5% confidence level in the Fisher's exact test. To evaluate the effectiveness of the statistical assessment we determined the percentage of conserved edited sites (see 'Materials and Methods' section) of putative editing sites (Supplementary Figure S4). Interestingly, >90% of significantly detected edited sites were conserved, supporting the reliability of the statistical test. Indeed, only a slight increase was observed with more stringent cut-offs (5% confidence level with FDR or Bonferroni correction). To be noted that a remarkable level of conservation was also observed for putative editing sites filtered out by the statistical test. It is highly likely that the read coverage at these positions is not deep enough to provide statistical support. Including all 314 additional putative edited sites with conserved homologous counterparts in other plants, more than 700 sites may be edited in the grapevine mitochondrion (*P*-values for all C residues falling in annotated coding genes are available in Supplementary Data).

All 401 significantly detected editing events were collected in the REDIdb database (26) under accessions EDI_000000804–EDI_000000840. Of these editing events 24.6% were supported by Solexa/Illumina reads and 75.4% were supported by SOLiD data.

A survey of mismatches identified by short reads

In addition to the C-to-U changes, marking editing events in the mitochondrial coding regions, we also analyzed other mismatch types (Table 1). The mismatch distribution, also used for carrying out the statistical tests (see 'Materials and Methods' section), resulted strikingly different between Solexa/Illumina and SOLiD data. In particular, G-to-U, C-to-A substitutions appeared overrepresented by Solexa/Illumina reads with respect to SOLiD data, likely reflecting typical miscalls of Solexa/Illumina reads (37). For the vast majority of G-to-U and C-to-A mismatches at positions covered by both technologies, SOLiD provided no evidence of variation between genomic and transcribed sequences. The established base call quality threshold (>15) likely reduced SOLiD and Solexa/Illumina false mismatches as we observed a slight overrepresentation of mismatches in reads where the corresponding base showed a relatively low quality score (Supplementary Figures S5 and S6). The lower frequencies of non-canonical mismatches recovered by SOLiD data (Table 1) suggest that this sequencing technology shows a higher overall accuracy. However, the combination of SOLiD and Solexa/Illumina data seems particularly suitable for the reliable detection of editing sites.

A survey of nuclear sequences showing more than 95% identity with mitochondrial coding regions revealed, in almost all cases, a cytosine in the detected edited positions. Indeed, rather than resulting from retrotranscription of potentially edited mitochondrial transcripts, mitochondria-like sequences in the nuclear genome derive from mitochondrial genomic fragments. Interestingly, apart from editing sites, differences between mitochondrial genes and their corresponding nuclear pseudogenes were predominantly transitions to A and T in the nuclear compartment [consistent with the high AT content of non-coding regions of the *Vitis* nuclear genome (34)]. Thus, cross matching reads derived from background transcription of nuclear mitochondrial pseudogenes might also account for a proportion of

observed G-to-A and A-to-G mismatches (results not shown).

Overall, we find no compelling evidence for editing events other than the canonical C-to-U.

Characterization of editing sites affecting coding genes in mitochondria of *V. vinifera*

The 401 C-to-U editing modifications detected in coding regions in *Vitis* mitochondria are unevenly distributed across different genes, ranging from 0.8% (*rpl2*) to 18.2% (*rps19*) of total cytosines (Supplementary Table S2) although no significant correlation was observed between sequencing fold-coverage and percentage of edited cytosines (data not shown). Our data also confirm a degree of species specificity of RNA editing. For example, the Vitis *rps3* transcript is edited at 10 sites, whereas the homologs from *B. vulgaris* and *Cycas revoluta* are edited at 8 and 28 positions, respectively (16,36). In grapevine mitochondria, genes coding for subunits of complex I seem to be more edited than genes coding for other subunits. However, the editing extent for each gene of a given mitochondrial complex is quite variable (Supplementary Table S2 and Supplementary Figure S7). The *cob* gene, encoding the cytochrome b of complex III, is the most edited gene, whereas the *sdh3*, a member of the complex II, is the least edited gene (see Supplementary Table S2). Some variability in the extent of editing can be also observed among gene groups belonging to the same complex, with genes of Complex I showing the highest level of edited sites (6.5% of total C) and genes of Complex II showing the lowest level (4.2% of total C) (Supplementary Table S2 and Supplementary Figure S7).

In total, 87% of the 401 editing modifications occurred at the first and second positions of codons, almost invariably resulting in replacement of the encoded amino acid (Figure 3). Indeed, only 1 out of 114 events affecting the first codon position resulted in synonymous changes. All non-synonymous editing conversions could modify the biochemical nature of the affected proteins. As observed in mitochondria of *A. thaliana* (19), the most frequent amino acid changes induced by RNA editing in grapevine were P-to-L (20.0%), S-to-L (19.4%) and S-to-F (13.5%)

Table 1. Base substitution frequencies detected by Solexa/Illumina, SOLiD and both technologies

From	Into				
	A	C	G	U	Any
Solexa/Illumina					
A	–	0.0078	0.0129	0.0037	0.0244
C	0.0177	–	0.0025	0.8768	0.8970
G	0.0187	0.0039	–	0.0273	0.0499
U	0.0057	0.0127	0.0102	–	0.0286
Any	0.0421	0.0244	0.0256	0.9078	
SOLiD					
A	–	0.0022	0.0112	0.0042	0.0176
C	0.0015	–	0.0017	0.9215	0.9247
G	0.0255	0.0029	–	0.0096	0.0380
U	0.0019	0.0151	0.0028	–	0.0198
Any	0.0289	0.0202	0.0157	0.9353	
Both					
A	–	0.0041	0.0118	0.0040	0.0199
C	0.0069	–	0.0020	0.9064	0.9064
G	0.0232	0.0032	–	0.0156	0.0420
U	0.0032	0.0143	0.0053	–	0.0228
Any	0.0333	0.0216	0.0191	0.9260	

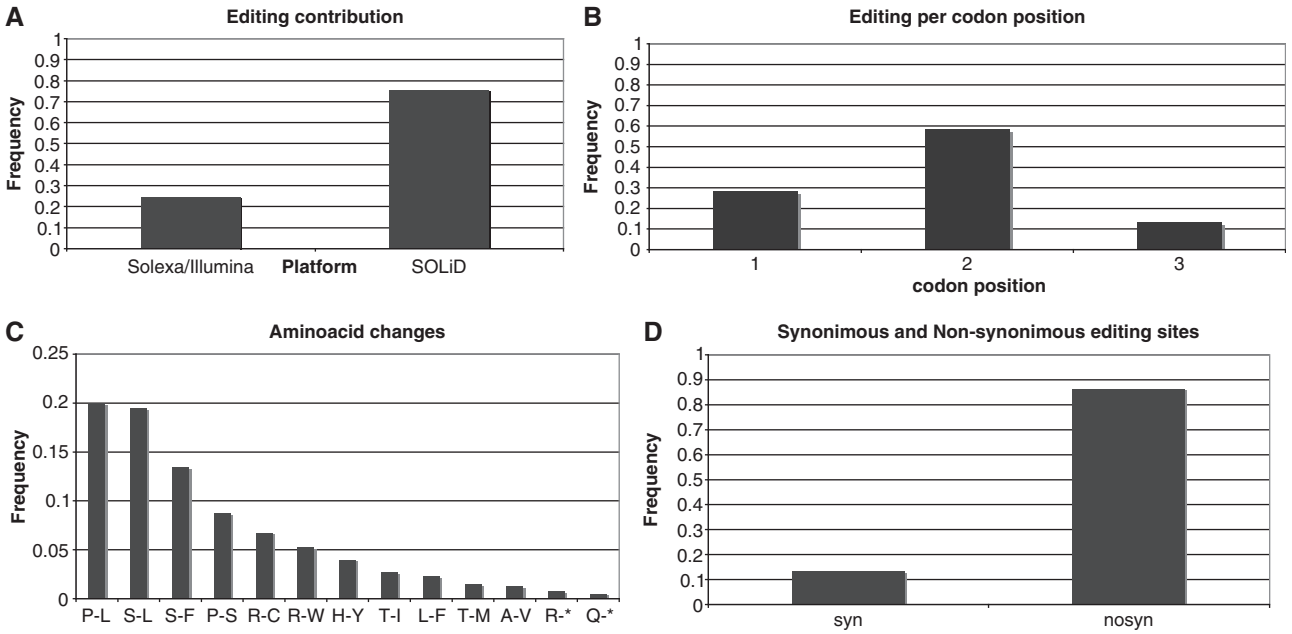


Figure 3. Principal statistics of detected RNA editing sites in *V. vinifera* mitochondria. (**A**) The contribution of each sequencing platform to editing detection; (**B**) distribution of C-to-U editing conversions across codon positions; (**C**) distribution of amino acids changes induced by detected RNA editing; (**D**) frequencies of synonymous and non-synonymous editing changes.

(Figure 3) increasing the proportion of hydrophobic amino acids and suggesting a real functional role for RNA editing through protein modifications in predominantly membrane-localized proteins. Additionally, S-to-L or S-to-F substitutions potentially increase the hydrophobicity of interface residues while P-to-L conversions occurring in secondary structures can contribute to protein functionality by avoiding defects in 3D structures (38).

Besides the non-random distribution of editing with respect to codon positions, we observed a preference of RNA editing towards specific codons. In particular, the three most frequently edited codons were UCA, CCA and UCC, accounting for 32.7% of all edited codons. The only C-containing codons never affected by editing were GGC, AGC and UGC in which editing could only lead to synonymous substitutions. C-to-U variations at specific codons were uncorrelated with codon usage, according to the correlation factor proposed by Giegé and Brennicke (19) (the ratio between the frequency of edited codons and the analogous proportion in the total population of C-containing codons of all investigated grapevine mitochondrial mRNAs). RNA editing in grape mitochondria creates three start codons (for *cox1*, *nad4L* and *rps10* genes), and generates the site of termination of translation in *atp6*, *ccmFC* and *rps10* transcripts. In the *rpl16* mRNA an additional editing event introducing a stop codon in frame with an upstream AUG was found. This suggests that the RPL16 protein is likely translated using a GTG codon just downstream of the edit-generated upstream ORF as initiator. Strikingly, this editing pattern, affecting the protein annotation, is highly conserved across mitochondria of land plants (39).

Although a strict consensus motif for sequences surrounding RNA editing sites has not been identified, bias towards pyrimidines at positions –2 and –1, and a bias towards purines at position +1 have been demonstrated (30). This behavior is also observed in the grapevine mitochondrial genome when the relative entropy in the 40 nt flanking edited and unedited cytidines was calculated. In particular, our data indicate that the relative entropy is extremely high in the immediate vicinity of the editing site (nt from –4 to +1), exceeding the 1% confidence interval calculated by 1000 iterations of random assignment of RNA editing sites. Interestingly, high relative entropy at the 5′-end of edited sites was also evident when it was calculated for 2- and 3-nt windows. Therefore, this region could be directly involved in editing site recognition, especially at position from –5 to –1 and from –18 to –14 as found in computational analyses conducted on four complete plant mitochondrial genomes by Mulligan *et al.* (30). The relative entropy for the 40-nt flanking grapevine editing sites is shown in the Supplementary Figure S8.

RNA editing in coding regions tends to increase cross-species conservation at the protein level and a correlation between amino acids modified by RNA editing and functional residues at protein structure has been shown (38). We performed domain searches of Pfam using either the protein conceptually translated from genomic or edited sequences (32). Interestingly, amino acid changes induced by RNA editing increased the scores of matches to individual Pfam domains from an average of 133.92 to 144.73.

Partial editing and tissue specificity of grape RNA editing sites

Twenty four percent of the 401 C-to-U conversions were classified as fully edited sites while 76% were considered partially edited sites—supporting the hypothesis that partial RNA editing is common in higher plant mitochondria (16,40). A proportion of partial editing might be due to transcripts where editing was not yet complete, while other partial events might derive from tissue-specific edits derived from mixed tissue samples (41).

Our Solexa/Illumina short sequencing reads were generated from total cellular RNA extracted from four different grapevine tissues: stem, leaf, root and callus; while SOLiD short reads were produced from leaf and root RNA (see 'Materials and Methods' section). Therefore, these data offered a unique opportunity to investigate the issue of RNA editing tissue specificity on a large scale. We compared the observed and expected distributions of Cs and Us in all available tissues by means of the chi-square test. 112 editing events were identified as significantly tissue specific at the 5% confidence interval corrected for false discovery rate, whereas 77 of them were selected as significant at 1% corrected confidence level. The Bonferroni correction were also applied at 1% confidence level resulting in a highly conservative estimate of 35 significant tissue specific editing sites (a list of tissue specific editing events is available in the Supplementary Table S4; see also Supplementary Figure S1).

Our findings indicate that tissue specificity accounts for a fraction of the observed partial RNA editing. Tissue specific editing might be required to modulate protein functionality in response to cell-type specific requirements. The high depth of coverage afforded by the SOLiD data resulted in the recovery of the majority of the significantly tissue specific edits by this technology. In summary, using the information from both sequencing technologies we discovered that 71% of all tissue-specific C-to-U changes occurred in leaf, whereas only a small fraction (0.4%) occurred in stem. Tissue specific editing events occurring in root and callus, instead, constituted 21 and 7.6%, of the total, respectively (Supplementary Figure S9).

RNA editing in non-coding regions of grapevine mitochondrial genome

While RNA editing by C-to-U modification occurs mainly in coding regions of land plant mitochondrial transcripts, several alterations to non-coding RNAs have also been described (14). In *Oenothera berteriana* mitochondria, a C-to-U transition at position 4 of the *trnF* gene corrects a mispairing in its acceptor stem improving the corresponding folding (42). Applying our computational strategy to 13 tRNA genes known to be of mitochondrial origin, we identified two C-to-U editing events, one in the

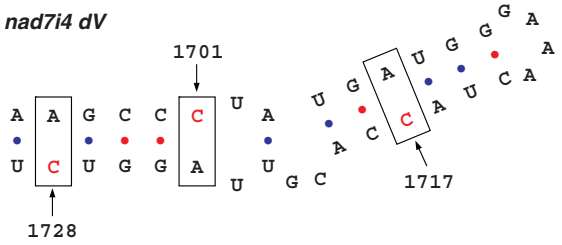


Figure 4. Secondary structure of the domain V of *nad7i4*. In the grapevine *nad7i4* domain V, cytosines subjected to RNA editing are indicated by arrows and included in a rectangle.

anticodon stem of the *trnC* gene altering a C–U mismatch to U–U supported by Solexa/Illumina reads and the other occurring at position 4 of the *trnF* acceptor stem, replacing a C–A mismatch with a conventional U–A Watson–Crick base pair, supported by SOLiD reads from root tissue (Supplementary Figure S10). Although the former editing event does not significantly change the stability of the *trnC* secondary structure, it occurs in the first 3 nt of the acceptor stem, a region that normally provides major identity elements and specific contact points for the cognate aminoacyl–tRNA synthetase. Moreover, this modification has also been described in *trnC* of *Oenothera* mitochondria (42). Notably, Solexa/Illumina reads also identified a reverse U-to-C editing event affecting the *trnP* at position 73. As a consequence, a native G-U match was replaced by a more stable G–C base pair. SOLiD data from leaf and root tissues, instead, supported another U-to-C change located at the first nucleotide 5′ of the anticodon, most likely contributing to codon–anticodon recognition.

RNA editing can also modify C residues in intronic sequences of plant mitochondrial genomes (13). Several C-to-U transitions have been described for group II introns, where they generally stabilize folding (13). Many such modifications occur in intron domains I, V and VI that are important for the excision reaction. We also investigated the extent of RNA editing in grape mitochondrial group II introns (excluding trans-splicing introns). Surprisingly, we observed 36 C-to-U modifications and four potential U-to-C reverse events. Moreover, 10 out of 36 conversions affected the *nad1* intron containing the *matR* gene. Several such editing modifications, as expected, occurred in domains V and VI improving the relative folding and, thus, the intron functionality in terms of self-splicing efficiency. We have analyzed three editing sites occurring in the domain V of the *nad7* intron 4 in detail. These modifications correct three C–A mispairings affecting the folding of this functionally indispensable domain (Figure 4). Several C-to-U events were also conserved across known group II introns of diverse land plants (13). Taken together, our findings indicate that the extent of RNA editing in mitochondrial introns of land plants could be higher than anticipated by previous genome wide studies.

RNA editing in *A. thaliana*

To further confirm the reliability of our computational strategy, we also investigated RNA editing by C-to-U conversions in mitochondria of *A. thaliana*, for which the complete editing landscape has been estimated according to standard experimental procedures based on the Sanger methodology (19). In particular, we used ≈64 million short reads (each of 50 nt in length) generated through the Solexa/Illumina technology from total RNA of floral tissue belonging to Columbia *Arabidopsis* ecotype (43). After the first mapping step, however, we obtained only 241 359 reads uniquely located across *Arabidopsis* mitochondrial protein-coding genes. While the number of mapped reads was limited, we identified 76 C-to-U fully edited sites. Ten of these are new editing sites not previously described by Giegé and Brennicke (19). Three occur at third codon positions and the remaining seven at the first two positions. Several of these changes increase the conservation of the affected protein across land plants. Surprisingly, we found an edited site in position 1277 of the *cox1* transcript, for which no editing sites have been yet observed in *Arabidopsis*. This modification causes the amino acid transition T-to-I for which a hydrophilic residue is replaced by a hydrophobic one. However, the effect of this change on the protein functionality is unknown. A specific protein modulation through RNA editing could also be required in floral tissue. However, this editing position, in addition to another C-to-U change at position 787 of the *rps3* mRNA, are supported by a very limited number of independent reads (<4) and, thus, more investigations are needed to verify the existence of such modifications.

In addition, we also checked for editing sites in non-coding RNAs of *Arabidopsis* mitochondria. According to Giegé and Brennicke, no C-to-U sites were found in tRNAs, whereas new editing conversions where discovered in group II introns. In particular, we detected two new C-to-U changes occurring at the first and third intron of the *nad4* gene other than one additional event in the unique *rpl2* intron. Such editing modifications, however, were again supported by a limited number of short reads (<4).

DISCUSSION

Detecting editing sites by RNA-Seq technology

RNA editing sites are usually identified by direct comparison of transcribed sequences with their related templates (44). Target cDNAs have typically been amplified by gene specific primers or isolated from cDNA libraries and sequenced using the standard Sanger methodology. cDNA sequences are aligned onto their corresponding genomic loci and all detected variations are scored as RNA editing sites (16,44). However, the restricted number of cDNAs per locus, in addition to potential sequencing artefacts, can lead to false positives and prevent the detection of genuine C-to-U editing events. Moreover, poor cDNA sampling can preclude the assessment of tissue specificity of editing modifications and the evaluation of their statistical support. In contrast, deep sequencing can overcome these limitations allowing the characterization of the RNA editing landscape of a given reference annotation. To date, however, no

computational approaches have been developed to this end. To fill this gap and to benefit from RNA-Seq technology for the investigation of editing, we propose a simple strategy that can efficiently handle short reads obtained by massive sequencing of RNAs by using either the Solexa/Illumina GA or ABI SOLiD platforms. Initially, short reads are mapped to a reference sequence using stringent quality criteria and allowing at most two mismatches and no indels. Subsequently we filter mapping results, considering only reads mapping to unique reference locations. This set of alignments is employed to generate a distribution of high quality nucleotides supporting each base of the reference. Unlike previous methodologies based on Sanger sequencing, short reads offer a high coverage depth per reference position and improve the detection of RNA editing sites. We have tested our approach, identifying C-to-U editing modifications occurring in the mitochondrial genome of *V. vinifera*. Plant mitochondrial RNA editing has been extensively studied and many C-to-U substitutions have been characterized in different organisms (9,26). The precise molecular mechanism is unknown but likely depends on nuclear factors belonging to PPR protein family (45). Moreover, the availability of well-annotated mitochondrial editing sites through specialized databases provides a valid benchmark with which to compare grape C-to-U modifications (26).

The availability of genome and RNA-Seq data from the same source, in our study the highly homozygous PN40024 grapevine genotype, is a fundamental requisite for a reliable editing detection. Indeed, nucleotide changes detected by comparing genome and transcript data may be genuine editing events or sequencing errors. In this respect, in addition to the expected C-to-U alterations, the Solexa/Illumina technology identified several potential non-canonical edits that were not supported by SOLiD reads—leading us to believe that for our data at least, the Solexa/Illumina reads are more prone to errors than those generated by the SOLiD technology. The frequencies of base substitutions shown in Table 1 support this hypothesis. The peculiar features of the color-space based SOLiD technology are particularly suitable for a reliable discrimination of real mismatches (two-color changes) from sequencing errors (single-color changes). Coverage depth could also influence the pattern of observed substitutions and contribute to the correction of potential mismatches occurring at low frequency. In our case, SOLiD data provided a mean per-base coverage depth that was three times higher than the Illlumina data (Supplementary Table S1). Indeed, despite the average 3-fold higher coverage, SOLiD data covered a similar number of bases to Solexa/Illumina (Supplementary Table S1).

Furthermore, >99% of SOLiD reads map on the sense strand, while Solexa/Illumina reads are equally distributed between the two strands (Supplementary Table S3 and Supplementary Figures S2 and S3). This is mainly due to the experimental protocol used to generate Solexa/Illumina reads (at the time of this work, the protocol to get strand specific Solexa/Illumina reads was not yet available). Considering the SOLiD data in isolation, we were able to exclude the possibility that the observed partial editing of some sites was a result of noise derived from non-edited antisense transcripts.

However, combining the information from both sequencing technologies we observed a significant increase in coverage and reduction of potential erroneous substitutions (Table 1).

The relatively high frequency for A-to-G and G-to-A mismatches can also be explained by cross mapping of short reads. Sequence similarity searches of the PN40024 nuclear genome revealed a number of regions showing high similarity to genes of mitochondrial origin. Interestingly, these sequences consistently showed higher identity to mitochondrial genome sequences than to edited mitochondrial transcripts. For high scoring segment pairs longer than 100 bases and showing >95% identity with mitochondrial coding regions (~32 000 bases of nuclear DNA), over 400 positions indicated that nuclear insertions were comprised of unedited rather than edited sequences, while only three mismatches with mitochondrial genome sequences suggested the presence of edited sequences. Interestingly, among other mismatches of nuclear to mitochondrial sequences, transitions to A and T were predominant (245/341 of the remaining substitutions). This observation is consistent with the known strong AT bias of intergenic regions of the *Vitis* genome (34) and corroborates our suspicion that some G-to-A changes are due to cross mapping of reads derived from background transcription of nuclear sequences.

Mitochondrial RNA editing in grapevine

The complete RNA editing pattern has been experimentally detected for four higher plant mitochondrial genomes. In total, 441 C-to-U modifications have been found in *Arabidopsis* mitochondria (19) and 427 in *B. napus* (20). Coding genes of *O. sativa* are modified at 491 positions (21), while only 357 editing sites have been found in mitochondria of *B. vulgaris* (16). While we found 401 significantly supported C-to-U editing modifications in 37 mitochondrial protein-coding genes of *V. vinifera*, an additional 314 sites showing non-significant levels of editing corresponded to editing sites in other species. Thus, it is likely that >700 sites are edited in grape mtDNA, and that our test is rather conservative—potentially due to overestimation of sequencing error rates. This implies that editing in *Vitis* is slightly more pervasive than in other plants or that many sites remain undiscovered in other species.

The extremely high level of identity of the PN40024 and ENTAV 115 mitochondrial consensus sequences—particularly those corresponding to coding regions, coupled with the fact that our RNA-Seq data derive from one of these clones (PN40024) lead us to discount the possibility that Single Nucleotide Polymorphisms between the individuals used for genome sequencing and transcriptome analysis should account for a substantial number of inferred editing events.

For the 401 statistically significant events, we found a remarkably conserved pattern of editing: 91% of the grape mtDNA edited sites (366/401) were either edited in the

same position in at least one other species (327/401) or the editing event increased conservation at the genomic level by introducing a uridine/thymine (39/401). For the remaining cases editing was prevalently observed at the third codon position (17/35, 48.6%), a much higher value than the 13.2% observed overall (Figure 3B).

An interesting finding concerns the extent of partially edited sites (in *Vitis* 76% of all detected modifications), and the observation that >85% of edited sites falling at silent (third codon) positions are partially edited. The predominance of partial editing at silent sites could be due to non-specific binding of editing specificity factors rather than an inefficiency of a putative 'editosome' machinery (16).

Partially edited sites may derive from immature transcripts or from differential (and possibly tissue-specific—see below) efficiency of the editing process in different positions. The impact of immature transcripts has been demonstrated by Verbitskiy and colleges (41) who showed that partially edited RNAs are intermediates of RNA editing in plant mitochondria. Moreover, we detected 36 editing sites in grape mitochondrial intervening sequences and all group II introns appeared well supported by short reads, indicating that incompletely processed messages are present in our samples. However, the observed range of variability—from 10 to 90%—of the percentage of unedited reads observed for the subset of deeply covered partially edited sites (>100 reads per site), is suggestive of differential editing efficiency at different sites.

A limited fraction of partially edited sites were shown to be significantly tissue specific. It should be noted that a high per base coverage depth is indispensable for statistical validation of the tissue specificity. Notably, the average per base coverage increases with the level of stringency of the statistical validation (i.e. FDR < 0.05, 165.23 reads per site; FDR < 0.01, 186.02 reads per site; Bonferroni correction, 248.60 reads per site). Therefore, we expect that additional tissue-specific sites would be identified by increasing the sequencing depth.

Considering all detected editing positions, our results are consistent with editing data from other land plants. Ninety percent of all grape RNA editing sites are non-synonymous, occurring with the highest frequency at the second codon position. Moreover, a large proportion of resulting amino acid changes fall in three categories P-to-L, S-to-L and S-to-F. Our results, therefore, validate the proposed computational approach based on next generation of sequencing reads.

Moreover, the detection of RNA editing sites has also been extended to mitochondria of *A. thaliana*. In spite of the restricted number of available short reads (the search for new editing events in *Arabidopsis* mitochondria was limited to fully supported sites in order to avoid potential noise due to false substitutions), ten new C-to-U changes were found in protein-coding genes, in addition to three modifications occurring in group II introns. Such new editing sites could be specific to the floral tissue since previous investigations have been conducted on cell-suspension culture only (19). However, the *Arabidopsis* mitochondrial genome and the Solexa/Illumina data of the accession SRX002554 belong to the same ecotype but not to the same individual and there is evidence that raises the possibility that the ecotype of the accession NC_001284 used by Giegé and Brennicke (19) is not Columbia (46), we can not therefore exclude the possibility that some of the novel *Arabidopsis* editing sites result from genomic polymorphisms.

Finally, we investigated the nucleotide context of edited sites in *Vitis* mitochondria and confirmed previously reported biases towards pyrimidines in nucleotides immediately upstream of edited cytidines and the frequent presence of a purine (generally a G) immediately following edited sites (Supplementary Figures S8 and S9). Thus, our data support the contention that groups of nucleotides in specific locations are important in the recognition of editing sites (30).

CONCLUSIONS

New high-throughput sequencing strategies offer unprecedented opportunities to investigate key molecular mechanisms at the genome level. In particular, RNA-Seq is a powerful tool for high-throughput transcriptome analysis including the investigation of basic post-transcriptional events such as alternative splicing and RNA editing. Editing by base conversion has been extensively studied in animal nuclei and land plant organelles where it seems to be essential for regular gene expression and genome variability maintenance. Indeed, organellar RNA editing may compensate for Muller's ratchet in genomes where nucleotide substitution rates are very low. However, the identification of edited sites is often time-consuming and costly, precluding genome wide investigations.

Recently, high throughput approaches have been used to identify A-to-I sites in human (22) and detect the efficiency of editing for 28 different sites during the development of the mouse brain (23). Such approaches, however, are not based on RNA-Seq and potential editing sites are known from the literature or computational analyses. In this work, we have presented a novel computational strategy that greatly facilitates the discovery of RNA editing sites at the genome level using short sequencing reads. We show that a combined approach including short reads from both Solexa/Illumina and SOLiD technologies may greatly improve the detection of reliable C-to-U editing sites in grapevine mitochondria, significantly reducing the discovery of false substitutions, particularly for editing sites supported by both platforms. However, it should be pointed out that our approach depends on the quality of short reads and should be performed on the same organism and individual. When the last request cannot be satisfied, results must be filtered for known SNPs and the conservation should be taken into account to identify candidate sites.

Although our procedure has been assessed in mitochondria of *V. vinifera* and *A. thaliana*, it can be applied to discover RNA editing events occurring on chloroplast or nuclear genomes, and to investigate the alterations of RNA editing patterns in diverse mammalian diseases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Carmela Gissi for helpful suggestions on the data analysis, Scott Kuersten and Alain Rico (Applied Biosystems) for help and advice with the SOLiD libraries preparations, and Davide Campagna, CRIBI, Padua University for advices in using PASS software.

FUNDING

Funding for open access charge: Ministero dell'Istruzione, dell'Università e della Ricerca (Fondo Italiano Ricerca di Base: 'Laboratorio Internazionale di Bioinformatica' (LIBI); Laboratorio di Bioinformatica per la Biodiversità Molecolare (MBLAB); VIGNA Consortium (Ministero delle Politiche Agricole, Alimentari e Forestali).

Conflict of interest statement. None declared.

REFERENCES

1. Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.
2. Schuster,S.C. (2008) Next-generation sequencing transforms today's biology. *Nat. Methods*, **5**, 16–18.
3. Mardis,E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.*, **24**, 133–141.
4. Morozova,O. and Marra,M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*, **92**, 255–264.
5. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
6. Denoeud,F., Aury,J.M., Da Silva,C., Noel,B., Rogier,O., Delledonne,M., Morgante,M., Valle,G., Wincker,P., Scarpelli,C. *et al.* (2008) Annotating genomes with massive-scale RNA sequencing. *Genome Biol.*, **9**, R175.
7. Gott,J.M. (2003) Expanding genome capacity via RNA editing. *C. R. Biol.*, **326**, 901–908.
8. Gott,J.M. and Emeson,R.B. (2000) Functions and mechanisms of RNA editing. *Annu. Rev. Genet.*, **34**, 499–531.
9. Gray,M.W. (2003) Diversity and evolution of mitochondrial RNA editing systems. *IUBMB Life*, **55**, 227–233.
10. Steinhauser,S., Beckert,S., Capesius,I., Malek,O. and Knoop,V. (1999) Plant mitochondrial RNA editing. *J. Mol. Evol.*, **48**, 303–312.
11. Takenaka,M., Verbitskiy,D., van der Merwe,J.A., Zehrmann,A. and Brennicke,A. (2008) The process of RNA editing in plant mitochondria. *Mitochondrion*, **8**, 35–46.
12. Casey,J.L. (2006) RNA editing in hepatitis delta virus. *Curr. Top. Microbiol. Immunol.*, **307**, 67–89.
13. Carrillo,C., Chapdelaine,Y. and Bonen,L. (2001) Variation in sequence and RNA editing within core domains of mitochondrial group II introns among plants. *Mol. Gen. Genet.*, **264**, 595–603.
14. Brennicke,A., Marchfelder,A. and Binder,S. (1999) RNA editing. *FEMS Microbiol. Rev.*, **23**, 297–316.
15. Kempken,F., Howard,W. and Pring,D.R. (1998) Mutations at specific atp6 codons which cause human mitochondrial diseases also lead to male sterility in a plant. *FEBS Lett.*, **441**, 159–160.
16. Mower,J.P. and Palmer,J.D. (2006) Patterns of partial RNA editing in mitochondrial genes of Beta vulgaris. *Mol. Genet. Genomics*, **276**, 285–293.
17. Blow,M., Futreal,P.A., Wooster,R. and Stratton,M.R. (2004) A survey of RNA editing in human brain. *Genome Res.*, **14**, 2379–2387.
18. Levanon,E.Y., Eisenberg,E., Yelin,R., Nemzer,S., Hallegger,M., Shemesh,R., Fligelman,Z.Y., Shoshan,A., Pollock,S.R., Sztybel,D. *et al.* (2004) Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.*, **22**, 1001–1005.
19. Giege,P. and Brennicke,A. (1999) RNA editing in Arabidopsis mitochondria effects 441 C to U changes in ORFs. *Proc. Natl Acad. Sci. USA*, **96**, 15324–15329.
20. Handa,H. (2003) The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (Brassica napus L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana. *Nucleic Acids Res.*, **31**, 5907–5916.
21. Notsu,Y., Masood,S., Nishikawa,T., Kubo,N., Akiduki,G., Nakazono,M., Hirai,A. and Kadowaki,K. (2002) The complete sequence of the rice (Oryza sativa L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. *Mol. Genet. Genomics*, **268**, 434–445.
22. Li,J.B., Levanon,E.Y., Yoon,J.K., Aach,J., Xie,B., Leproust,E., Zhang,K., Gao,Y. and Church,G.M. (2009) Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science*, **324**, 1210–1213.
23. Wahlstedt,H., Daniel,C., Enstero,M. and Ohman,M. (2009) Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. *Genome Res.*, **19**, 978–986.
24. Huang,X. and Yang,S.P. (2005) Generating a genome assembly with PCAP. *Curr. Protoc. Bioinformatics*, Chapter 11, Units 11 13.
25. Goremykin,V.V., Salamini,F., Velasco,R. and Viola,R. (2009) Mitochondrial DNA of Vitis vinifera and the issue of rampant horizontal gene transfer. *Mol. Biol. Evol.*, **26**, 99–110.
26. Picardi,E., Regina,T.M., Brennicke,A. and Quagliariello,C. (2007) REDIdb: the RNA editing database. *Nucleic Acids Res.*, **35**, D173–D177.
27. Campagna,D., Albiero,A., Bilardi,A., Caniato,E., Forcato,C., Manavski,S., Vitulo,N. and Valle,G. (2009) PASS: a program to align short sequences. *Bioinformatics*, **25**, 967–968.
28. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
29. Hochberg,Y. and Benjamini,Y. (1990) More powerful procedures for multiple significance testing. *Stat. Med.*, **9**, 811–818.
30. Mulligan,R.M., Chang,K.L. and Chou,C.C. (2007) Computational analysis of RNA editing sites in plant mitochondrial genomes reveals similar information content and a sporadic distribution of editing sites. *Mol. Biol. Evol.*, **24**, 1971–1981.
31. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188–1190.
32. Coggill,P., Finn,R.D. and Bateman,A. (2008) Identifying protein domains with the Pfam database. *Curr. Protoc. Bioinformatics*, Chapter 2, Unit 25.
33. Palmer,J.D. and Herbon,L.A. (1988) Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *J. Mol. Evol.*, **28**, 87–97.
34. Jaillon,O., Aury,J.M., Noel,B., Policriti,A., Clepet,C., Casagrande,A., Choisne,N., Aubourg,S., Vitulo,N., Jubin,C. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.
35. Quail,M.A., Kozarewa,I., Smith,F., Scally,A., Stephens,P.J., Durbin,R., Swerdlow,H. and Turner,D.J. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*, **5**, 1005–1010.
36. Regina,T.M., Picardi,E., Lopez,L., Pesole,G. and Quagliariello,C. (2005) A novel additional group II intron distinguishes the mitochondrial rps3 gene in gymnosperms. *J. Mol. Evol.*, **60**, 196–206.
37. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.

38. Yura,K. and Go,M. (2008) Correlation between amino acid residues converted by RNA editing and functional residues in protein three-dimensional structures in plant organelles. *BMC Plant Biol.*, **8**, 79.
39. Bock,H., Brennicke,A. and Schuster,W. (1994) Rps3 and rpl16 genes do not overlap in Oenothera mitochondria: GTG as a potential translation initiation codon in plant mitochondria? *Plant Mol. Biol.*, **24**, 811–818.
40. Zehrmann,A., van der Merwe,J.A., Verbitskiy,D., Brennicke,A. and Takenaka,M. (2008) Seven large variations in the extent of RNA editing in plant mitochondria between three ecotypes of Arabidopsis thaliana. *Mitochondrion*, **8**, 319–327.
41. Verbitskiy,D., Takenaka,M., Neuwirt,J., van der Merwe,J.A. and Brennicke,A. (2006) Partially edited RNAs are intermediates of RNA editing in plant mitochondria. *Plant J.*, **47**, 408–416.
42. Binder,S., Marchfelder,A. and Brennicke,A. (1994) RNA editing of tRNA(Phe) and tRNA(Cys) in mitochondria of Oenothera berteriana is initiated in precursor molecules. *Mol. Gen. Genet.*, **244**, 67–74.
43. Lister,R., O'Malley,R.C., Tonti-Filippini,J., Gregory,B.D., Berry,C.C., Millar,A.H. and Ecker,J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.
44. Takenaka,M. and Brennicke,A. (2007) RNA editing in plant mitochondria: assays and biochemical approaches. *Methods Enzymol.*, **424**, 439–458.
45. Zehrmann,A., Verbitskiy,D., van der Merwe,J.A., Brennicke,A. and Takenaka,M. (2009) A DYW domain-containing pentatricopeptide repeat protein is required for RNA editing at multiple sites in mitochondria of Arabidopsis thaliana. *Plant Cell*, **21**, 558–567.
46. Bentolila,S., Elliott,L.E. and Hanson,M.R. (2008) Genetic architecture of mitochondrial editing in Arabidopsis thaliana. *Genetics*, **178**, 1693–1708.

SVM2: an improved paired-end based tool for the detection of small genomic structural variations using high throughput single genome resequencing data

Journal:	*Nucleic Acids Research*
Manuscript ID:	Draft
Manuscript Type:	3 Methods Manuscript (Online Publication)
Key Words:	Genomic Structural Variation, Genome resequencing, intraspecific variation, next generation sequencing, paired end reads

SVM^2: an improved paired-end based tool for the detection of small genomic structural variations using high throughput single genome resequencing data

Matteo Chiara[1*],Graziano Pesole[2,3] and David S. Horner[1.*]

[1] Department of Biomolecular Sciences and Biotechnology, University of Milan, 20133 Milan, , Italy

[2] Institute of Biomembranes and Bioenergetics, National Research Council, 70125, Bari, Italy

[3] Department of Biosciences, Biotechnology and Pharmacological Sciences, University of Bari, 70125 Bari, Italy

* To whom correspondence should be addressed: tel: +39 02503 14884 email: david.horner@unimi.it Correspondence may also be addressed to: Matteo Chiara: tel: +39 02503 14923, email: matteo.chiara@unimi.it

ABSTRACT

Several bioinformatics methods have been proposed for the detection and characterization of genomic structural variation (SV) from ultra-high throughput genome resequencing data. Recent surveys show that comprehensive detection of SV events of different types between an individual resequenced genome and a reference sequence is best achieved through the combination of methods based on different principles (split mapping, reassembly, read depth, insert size, etc). The improvement of individual predictors is thus an important objective. Here we propose a new a method that combines deviations from expected library insert sizes and additional information from local patterns of read mapping and uses supervised learning to predict the position and nature of structural variants. We show that our approach provides greatly increased sensitivity with respect to other tools based on paired end read mapping at no cost in specificity, and it makes reliable predictions of very short insertions and deletions in repetitive and low complexity genomic contexts that can confound tools based on split-mapping of reads.

INTRODUCTION

The characterization of intra-specific genomic diversity has enormous implications for biomedical sciences and for biology in general and is one of the principal objectives of contemporary genomics. Recently, ultra high-throughput Next Generation Sequencing (NGS [1]) technologies have greatly facilitated ambitious genome resequencing projects and associated studies focused on human health (e.g. http://www.1000genomes.org/ [2]) as well as on generating a wider understanding of genome evolution [3,4].

One of the most interesting general conclusions to emerge from such studies is that, contrary to long-held assumptions, Structural Variations (SV) - genomic rearrangements, including insertions, deletions, copy number variations and inversions - typically explain a very significant proportion of normal intra-specific genetic variation [5,6,7,8,9,10,11]. While the widespread association of SV with hereditary diseases and cancer [12,13,14,15,16,17,18,19] justifies their study, the use of SVs as molecular markers in non human systems, for genome-wide association studies, genetic mapping and marker assisted breeding approaches is also increasing.

Bioinformatics tools to detect SV with high throughput resequencing data tend to be specialized to accommodate specific types of data or rely on different expected patterns of mapping of reads from a resequenced (donor) genome on a reference sequence in the vicinity of structural variations. For example, in the context of the 1000 genomes project [2], mixed samples of genomic DNA from multiple individuals have often been sequenced together as part of an effort to generate a comprehensive catalog of variants and haplotypes in human populations. Dedicated and highly sophisticated tools that use probabilistic methods to identify variations that are not present in all sequenced individuals have been developed and shown to be highly effective [20,21].

Tools developed to detect SVs from high-coverage individual genome resequencing may be categorized as alignment-based or statistics-based. Approaches dependent on the alignment of reads to a reference sequence may include partial de novo assembly of reads [22] or may rely on split mapping of short reads [23]. While such methods should be capable of precisely identifying break-points, difficulties in de-novo assembly, incorporation of sequencing error models and maximum detectable size for insertions in split reads mapping, the impact of repetitive genome sequences and limits in read length mean that they are incapable of identifying all SV events (reviewed in [24,25]).

Statistics-based methods include read density based approaches that exploit the same principle as DNA hybridization arrays. These tools are particularly efficient in detecting copy number variation but cannot easily identify the introduction of novel sequences [26,27]. Paired End (PE) read based approaches are particularly suited for identifying insertion and deletions. Such methods aim to identify genomic loci where donor reads map at inconsistent distances. A number of tools based on this principle have been developed and either detect genomic loci exhibiting statistically significant clustering of PE reads with anomalous mapping distances [28,29] or compare local distributions of mapping distances to an expected distribution in an attempt to identify regions harbouring SVs [28,30]. The first approach is more suited to the identification of long deletions while the second tends to be more computationally intensive but generally applicable. One obvious disadvantage of statistics based methods is that they do not identify precise breakpoints.

While information regarding the mapping of Broken Pairs (BPs) - where only one of two PE reads can be satisfactorily mapped to the reference genome - are routinely used in algorithms designed to detect large genomic rearrangements [22], current mapping-distance based tools use a single metric of insert-size perturbation to predict the presence of SVs. However, different types of genomic rearrangements, even those involving only a few base pairs of DNA, are expected to generate complex and particular signatures in mapping patterns of PE reads. Additionally, each sequencing reaction and reference genome has a series of characteristics which can, in principal, impact upon methods used to identify structural variation. For example, each library has a characteristic insert-size distribution - and each sequencing reaction shows a particular profile and frequency of sequencing errors. Furthermore, individual reference genome sequences show a particular distribution of repetitive sequences. All these factors are relevant to the selection/parameterization of appropriate statistical tests for the identification of SVs from insert size perturbations.

Support Vector Machine is an ensemble of statistics/computational techniques that have been widely employed in biological classification problems including the recognition of miRNA precursors, the discrimination of coding from non-coding sequences, the classification of differential gene expression profiles from microarray data, the recognition of protein secondary structure and the identification of candidate drug targets. SVM uses a series of training data points, each known to belong to one of two classes of origin and described by a number of quantitative features, and, having transformed them into a higher dimensionality than allowed by the number of associated features and through the use of a kernel function, identifies the hyperplane that maximizes their separation by class in a multidimensional space. Once the optimal discriminating function has been established, it is used to classify unknown instances (for an introductory review see [31]). Several software libraries implementing SVM are freely available and the method can be adapted to function in multiple category classification problems.

Here we show that the incorporation of different characteristics of mapping data derived from PE resequencing reads, can improve the sensitivity of detection of relatively small indels (1 – 30bp) that constitute the majority of intra-specific SV events [32]. We use SVM to incorporate these diverse mapping characteristics to address the indel finding/classification problem. The Structural Variation Mapping using Support Vector Machines (SVM^2) software presented here calculates and integrates a combination of features based on statistics and resequencing coverage measures for windows around a given genomic coordinate. The method does not make a-priori assumptions regarding the insert-size distribution of a particular library, or on the optimal pvalue cutoff to be used in any of the statistical tests that it employs, rather, it is trained using a given resequencing dataset and reference genome sequence. In the current work, SVM^2 was trained to discriminate genomic loci flanking four classes of events (deletions, insertions shorter than the library insert size, insertions longer than the

library insert size and hyper-variable regions) from normal genomic regions, although in principal there is no restriction to the number of classes/sizes of events that could be recognized.

SVM^2 attains a similar specificity and a far superior sensitivity than state of the art PE-based methods using the same data and appears to be more robust than split mapping to the confounding effects of some genomic contexts.

Recent surveys confirm that comprehensive detection of SV events of different types between donor and reference sequences is best achieved through the combination, with rigorous filters, of predictions made by methods based on different principles [2]. In this light, the improvement of individual predictors is of course desirable. Indeed, resequencing is becoming ever more accessible and economical, and in some experimental contexts, notably the development of molecular markers for crop and animal positional cloning and marker assisted breeding programs, workers are likely to prefer to use one or two methods to maximize the detection of small to medium insertions and deletions (1-30 bp) without the requirement of implementing and optimizing particularly complex bioinformatics pipelines. We provide evidence that combining our method with split mapping could provide a reasonable starting point for the identification of small to medium sized SV events.

MATERIALS AND METHODS

SVM Features

For each chromosome we store the read mapping data in a sorted (ascending order, by mapping coordinates) doubly linked list. Every node (N) in the list contains the following information:

1. genomic coordinates (start-end) or ID
2. coverage by paired and unpaired reads on each strand (within the coordinates)
3. observed insert size distribution of PE "covering" the node on each strand

consecutive positions with identical coverage statistics are merged into single nodes and positions with no coverage are not incorporated into nodes. Positions and lengths of uncovered regions can be trivially calculated from the difference between the coordinates of 2 consecutive nodes.

For a given node N (which includes a genomic position X) we call M (genomic position Y) the node exactly 1 insert-size downstream in covered bases. The objective here is to identify a site in the reference genome that is beyond the SV event and corresponds to the expected position of mapping of the partners of reads mapping to X. It is acknowledged that bases covered by only redundant mapping reads will lead to errors in the calculation of M as will insertions in the donor genome.

We define the following genomic windows (in ID) X-read length to X, X-10 to X, X to X+10, X to X+read length (and equivalent intervals for Y). The windows of 1 read length correspond to expected positions of peaks of BP reads as X moves within one insert size of an SV event, while the windows of 10 bases were chosen arbitrarily with the objective of accommodating errors in the estimation of position Y and to aid the precise delineation of the sharp peaks of BPs expected to flank junctions of deletions in the donor genome.

For each of these windows, we calculate (for each genomic strand) the mean total coverage per base and normalize these values to the total coverage of X or Y respectively.

For each of these windows we also record (for each genomic strand) the mean proportion of reads mapping to each of these windows that are BPs.

We define an additional window: X-read length to Y+read length (in ID)

We record the length of this long window in genomic bases, the longest interval of consecutive uncovered bases contained within it, and the total number of uncovered bases in the interval

For each visited node in the long window, we perform the following statistical tests: a Z test to compare the observed "upstream" read length distribution to the global insert size distribution, another Z-test to compare the observed "downstream" read length distribution to the global distribution, a Student T test (Welch) and a KS test to compare the "downstream" to the "upstream" distribution to each other. We record the proportion of genomic positions in the window supporting a perturbation of mapping distance according to a particular test with a Pvalue within the following ranges: $\leq 10^{-5}$, 10^{-5} to 10^{-4},10^{-4} to 10^{-3}, 10^{-3} to 10^{-2}

Finally for each node in the long window we compute the broken pairs to total number of reads ratio for each strand and record the proportion of positions on each strand in the window with ratios within the following ranges 0.15-0.25, 0.25-0.50, 0.50-0.75, >0.75

The aforementioned statistics are recorded in an ordered vector and used as a feature set for SVM analysis

Cluster formation and SV calling

Sites classified as non-normal and of the same type by the SVM are merged into clusters when located less than 5 bases apart on the genome. Clusters with a number of "non-normal" positions that exceeds an "indicator cutoff parameter" are promoted to the status of indicators and a comparison of mapping distances for all paired reads mapping to the cluster and pointing towards the putative SV event with the global mean mapping distance is used to estimate where a complementary strand cluster/indicator is expected to fall (2 mean insert sizes plus or minus the estimated size of the SV

event (in the case of deletions and insertions respectively)). If a cluster or an indicator with the same type of predicted event is identified in the expected interval, an event of that type is called at the base falling half way between the outer coordinates of the two supporting clusters. If a cluster or indicator of contradictory type is identified in the expected region, an indeterminate indel (IndIndel) is called. When intervals between two called events overlap by more than 80% of their lengths, the predictions are merged.

Size estimation and detection of heterozygosity

The expected position of the event (or breakpoints) is evidently half way between the two clusters. Once a position has been predicted, a more accurate estimate of the size of the event is obtained by identifying all pairs of reads mapping across the predicted break-point and comparing their mean insert size to the mean global insert size.

To discriminate between homozygous and heterozygous events, we use an EM algorithm and a log-likelihood test similar to that implemented in the software Modil [29] In brief, for any genomic locus where an indel has been predicted, we model the mapping distances of reads covering the predicted event data a single distribution (homozygous) or a pair of distributions (one of which is constrained to the global insert size distribution (heterozygous) and compute the respective likelihoods. At least 30% of reads covering the position must be assigned to each distribution. A log-likelihood test with 1 degree of freedom is used to verify whether the 2 distributions model is significantly more likely (Pvalue $\leq 10^{-3}$).

Coarse filters for the identification of regions potentially containing SVs

The genomic sequence (read map) is traversed in a 5' -3' direction on each strand. Only positions with total coverage above a "minimum coverage" parameter are considered. To avoid unnecessary calculations the SVM is invoked only by sites that satisfy at least one of two "coarse filter" criteria: if the ratio of broken pair reads to mapped pair reads overlapping the position is in the highest "BP proportion parameter" % of genomic sites, or if the mean insert size falls outside of "map distance deviation" standard deviations of the mean of the global insert size distribution.

Training and parameter optimization

Randomly selected genomic regions of at least 15 Kb in length, within which no bases would invoke SVM analysis and where all bases show coverage to expected coverage ratios of between 0.5 and 4 are selected as templates for SVM training and parameter estimation.

To produce the training set for the SVM we use 100 selected regions and randomly introduce a single insertion or deletion of length 1-2500 bp to each. Real sequence reads are then remapped to the in silico mutated genome. The process is repeated to give a total of 1500 indel events. To simulate the effect of hypervariable regions, random windows of length 35-500 were selected and subjected to random mutation at different substitution rates (10% - 25%). A total of 1000 simulations were performed for each combination of size and substitution rate. Each position on the positive strand upstream by less than an insert-size from in-silico break-points (or polymorphic hot-spots) and every position on the negative strand downstream by less than an insert-size are labeled with the relevant type of event (deletion, small insertion, large insertion, hypervariable). For each set of remapped reads the initial coarse filters are re-applied and features calculated around positions which would invoke the SVM. Appropriately labeled feature vectors are used in conjunction with the *libsvm* facilities to train a multi-class SVM and obtain the SVM model file. The polynomial kernel was used in all experiments.

Several parameters required for the analysis must be specified at runtime. The minimum coverage parameter determines the minimum total read coverage of a site for consideration. The “BP proportion parameter” and the “map distance deviation” parameters govern the invocation of the SVM, while the cluster promotion parameter is required in the definition of indicators in the post processing step. These values can be determined by the user or optimized after SVM training using provided scripts. These tools perform simple parameter sweeps and attempt to the select parameter values that minimize the number of overlapping predictions and false positive predictions with the optimized SVM model and a subset of the simulated events that were not used in SVM training.

Data pre-processing and mapping of reads

To evaluate the proposed method, we downloaded 3.5 billion reads (1.75 billion pairs of reads) from the NCBI short read archive: ftp://ftp-private.ncbi.nlm.nih.gov/sra. All reads were 36 bases in length and the libraries contained theoretical insert sizes of around 208 bases. Similarly to Hormozdiari et al. [26] we removed any read (and its mate) where the average phred quality was below 20, and pairs of reads where one read contained more than 2 Ns. This leads to the elimination of 650 million pairs. We aligned the reads to the human genome hg18 reference assembly using SOAP2 [32], allowing only unique mapping reads/pairs with up to 2 mismatches per read.

This generated 1 billion uniquely mapping pairs and 40 million uniquely mapping unpaired reads.

Predictions from other tools, data download and comparison criteria

We ran BreakDancer on our mapping-data using the parameters reported in the original paper. PinDel predictions from the same dataset were downloaded from http://www.ebi.ac.uk/~kye/pindel/ and Variation Hunter predictions from http://compbio.cs.sfu.ca/strvar.htm .

Repeat and gene annotations were downloaded through from the UCSC genome browser (genome-mysql.cse.ucsc.edu).

To compare different validation and prediction set we used the the latest version of the intersectBED program from the BEDtools [33] suite and custom Perl scripts. We used simple overlap (≥ 1bp) between different sets as main criterion of validation/equivalency. As the significant intervals predicted by PE based tools tended to be longer (avg 290 bp), respect to the predictions by Pindel, we extended Pindel predictions by 60 bp upstream and downstream.

RESULTS

Rationale and Description of the Approach

In the vicinity of indels between a donor and a reference genome 3 types of perturbations in the "normal" PE mapping pattern are expected - in different degrees - depending on the type of event (deletion in donor genome, insertion smaller than library insert size, insertion larger than library insert size).

Firstly, PE reads spanning the indel will show a perturbation from the expected mapping distance (increased distance for a deletion, decreased for an insertion in the donor genome provided that the insertion event is smaller than the library insert size. Insertion events larger than the library insert will lead to an absence of PE reads spanning the junction on the donor genome). These phenomena are expected to be observed within one library insert size 5' of junctions of rearrangements.

Secondly, given sufficient sequence coverage and presuming correct and comprehensive mapping of reads, a peak of Broken Pair (BP) mappings is expected to be observed from one library insert size 5' of rearrangement junctions, towards the junctions. In the case of deletions in the donor genome, this peak will be narrow (one read length) as only reads mapping on the rearrangement junction will fail to map, while in the case of an insertion in the donor genome, this peak will extend the length of the insertion towards the rearrangement junction.

Finally, and as a corollary to the previous observation, the rearrangement junctions (and the region deleted in the case of deletions in the donor genome) will show an absence of coverage by any reads (PE or BP). A schematic illustration of these expected patterns is provided in Fig S1.

Existing tools to exploit PE mapping perform a single statistical test, comparing the local insert size distribution to that for all mapped PE reads. The assumption underlying our approach is that avoiding the use of stringent statistical cutoff values by employing a series of *ad-hoc* descriptors of read mapping patterns, supervised learning and searching for concordance between neighboring genomic sites, it might be possible improve sensitivity of SV detection without loss of specificity.

In the current method, for any genomic position we first attempt to identify the expected mapping position for the partners of PE reads covering that position. We then define a series of genomic windows centered on these positions (see methods and Fig S2). For these windows, statistics regarding the aforementioned phenomena are recorded and a multi class SVM classifier is used to assign the site to one of several different categories ("normal", flanking a deletion, flanking a small insertion, flanking a long insertion and flanking a hyper-variable region).

In practice, as the starting position approaches a SV event, the disposition of different types of perturbations along the different windows changes, meaning that a single characteristic pattern of feature value biases cannot be associated linearly with a single type of event. However, the advantage of SVM over hierarchical methods such a decision tree, is that it is not necessarily "looking" for a single combination of feature values to make a classification, rather, it should recognize different patterns that were associated with a class in training.

It is of course expected that multiple sites flanking a single SV event (upstream on each strand) will be recognized by the SVM classifier as indicating a similar type of event, and this expectation is exploited in a post processing step that detects relevant clusters of indicative sites on each strand of the genomic sequence and calls insertion and deletion events between complementary clusters, where such cluster conflicts in their assignment of the nature of their event, we assign an indeterminate indel (IndIndel). Finally, dimensions of called events are estimated by comparing map distances of PE reads spanning the predicted event to the global mean insert size, and a likelihood based method is applied to identify heterozygous SVs. As for other mapping distance methods an inherent weakness of our approach is its relative inability to detect insertion events larger than the PE library insert size. Indeed while it uses BP data and might be expected to detect some such events in regions of high sequence coverage, it is unable to estimate the insertion size.

While the implementation of our approach is efficient and rapid, it is not necessary to apply the SVM to all positions in the reference genome. Our method uses initial filters to identify a subset of genomic positions where either the ratio of mapped unpaired reads to paired reads or the mean insert size of reads on one strand are potentially anomalous with respect the global situation. Given that most SVs are very short (too short to actually perturb the insert size distribution), the net effect is that the SVM is mostly invoked as a consequence of the presence of BP reads.

The SVM itself is trained using the experimental data and genome sequence under study, with simulated insertion and deletion events. Several parameters relevant to the analytical pipeline are also optimized automatically during the training of the system for a particular combination of dataset and genome. The method is implemented in the software SVM^2 - a package written in C++ with accompanying Perl scripts and utilizes the freely available Libsvm package [34]. SVM^2 is rapid and requires only limited RAM after the initial read mapping phase.

Simulation

To estimate the specificity and sensitivity of our method, we artificially implanted 9000 random insertions and deletions of different sizes (1 to 600 bp – note that beyond the library insert size detection of insertion events is not influenced by dimensions of the insertion) into human chromosome 17 (hg18 assembly) and generated artificial reads (theoretical coverage 30X, error rate 1%) from the mutated genome (mate pairs, insert size 208. s.d. 13, theoretical coverage 35X) using the dwgsim program [35] .

Results presented in Table 1 show good overall recall rates (88% and 91% for insertions and deletions respectively) and generally low false positive rates. The column "recall" indicates percentage of simulated events that were correctly classified as insertion or deletion (subdivided by the actual length of the event simulated) while the "recall as any" column shows the total proportion of simulated events that were identified as either insertions or deletions. It is clear that both false positive predictions and misclassification of the nature of events constitute significant issues only with predictions of very short events (less than ten bases). An exception to this trend, is provided by insertions longer than the insert-size which are recovered with a slightly lower recall rate. This is unsurprising given that detection of such events relies exclusively on the presence of broken pairs

Heterozygosity

The identification of heterozygous SV events is an inherently difficult problem for non alignment-based methods. Heterozygosity reduces apparent perturbation in insert sizes and lowers ratios of unpaired to paired reads. Low read depth also raises the probability of unequal sampling of haplotypes, further complicating the issue. However we anticipated that with sufficient depth of coverage our method should be able to recognize longer indel events.

Accordingly, we simulated a set of SVs of different size (1 to 40 bp, 250 events per size per category) with a theoretical 40X depth of coverage, for each SV we generated the heterozygous as well as the homozygous version. The results are summarized in Table 2, for each set of SV we computed the recall rate for the homozygous as well as the heterozygous event, and the fraction of heterozygous SVs that was correctly classified as heterozygous (see Methods). The results confirm that while our method is particularly accurate in detecting homozygous SV of any size, it lacks sensitivity both in the detection and correct classification of heterozygous SVs less than 20 bases in length. Additional simulations showed that, as expected, proportions of heterozygous alleles sampled in resequencing impacts upon detection and classification (Table S1)

Comparison with other tools

To compare the performance of our method to other tools using real PE resequencing data we have taken advantage of publicly available PE resequencing data from an anonymous human donor (Bentley et al [36]) generated with the Illumina technology. The peculiarity of this dataset is that a large and consistent set of SV was previously detected and validated using low coverage (0.3X) longer insert (Sanger + 40Kb fosmids) from the same individual (Kidd et al [37]), thus it has been widely used as a benchmark to compare different SV detection tools. Indeed, the Kidd et al. data was recently subjected to a second analysis [38] and here we consider the union of both sets of predictions as a validated indel set (265264 events).

We compared the performance of our tool with that of BreakDancer [28] (a widely used PE based method) that, in previous studies of the same dataset, exhibited the highest sensitivity and specificity among PE-based tools in detecting relatively small indels (indicatively greater than 10bp) and PinDel [23] , a popular split mapping approach.

The sensitivity (the proportion of indels in the validation set that was recovered by each method, as a function of the validated size of the indel) of each method is shown in Fig1a (and supplementary table 2). Under this criterion, SVM^2 outperforms BreakDancer in all size categories, overall recalling 4.5 times as many events. As expected, the split mapping method (PinDel) is more sensitive in the detection of very small indels (up to 5bp) although SVM^2 recalls a larger proportion of events over this threshold.

The number of predictions and apparent specificity by predicted event size (proportion of predicted indels of coinciding with any indel in the validation set as a function of the predicted size of the indel) for each method is shown in Fig1b (and supplementary table 3). It should be noted that the genome coverage of the Kidd et al. data , 0.3x, represents the maximum theoretical specificity in this benchmark. All of the evaluated methods demonstrate similar overall performance. PinDel in particular shows a marginally better specificity with respect to the smallest events (<10bp) while the size/specificity profile of SVM^2 and BreakDancer are relatively uniform at around 26-27% “validation” for each size bin. Both SVM^2 and BreakDancer suffer an apparent loss in specificity with regard to predicted events greater than 30bp or more. This last observation is likely a stochastic effect due to the fact that larger rearrangements constitute a very small minority of SVs. To partially ameliorate the low genome coverage of the validation set, we compared predictions to all events in dbsnp130 which contains more than 4.2 million known rearrangements derived mostly from Sanger sequencing data [38]. 81.5%, 80.6% and 80.4% of the predictions made by BreakDancer, PinDel and SVM^2 respectively correspond to known human SV events. The specificity by size profile strongly resembles that observed with the Kidd SVs (Supplementary Figure 3a and Supplementary table). Cross referencing the predictions from the various methods with the collection of human genomic SVs provided by the 1000 genomes project, derived from NGS data [2] (1.32 million events) showed that 61% of BreakDancer predictions, 69% of SVM^2 predictions and 80.7% of PinDel predictions were

coincident with events present in that database. 54% of the Kidd/Sanger based validation set events were present in the 1000 genomes database (Supplementary Figure 3b and Supplementary table 4).

While the identification of very large SVs is not a primary objective of our method, we also compared the capacity of several methods to identify 98 long deletions (10 Kb or more) called in the original work of Kidd et al. In this particular task, Variation Hunter, a method developed specifically for the identification of large SVs [30] recovered 65 events, while BreakDancer, and SVM^2 recalled 55 and 51 events respectively. SVM^2 made only 54 predictions of insertions over 200 bases in length.

The Venn diagram in Fig 2 shows the overlap of validated calls made by SVM^2, BreakDancer and PinDel. The union of all methods identified 108158 of the 265264 events recovered from the Sanger data (41%). 24842 (23%) are found by PinDel and SVM^2, 9122 (8.5%) are identified by BreakDancer and SVM^2. Only 1730 (1.5%) are found by BreakDancer only while 49972 (46%) are unique to PinDel and 20974 (19%) are unique to SVM^2. 87% of validated BreakDancer predictions are also made by SVM^2. Taken together, these observations confirm that the incorporation of additional mapping information in SVM^2 allows a great increase in sensitivity over methods that use only mapping distance information. Furthermore, it is evident that a notable proportion of events are recovered by SVM^2 but not other methods. When compared to the sensitivity profile by event size (Fig 1a) it is evident that SVM^2 identifies a significant number of small events not detected by PinDel.

Accuracy of classification and genomic context of predictions

The analysis of simulated data suggested that, for small events, SVM^2 may lack precision in classification. furthermore genomic clustering of SV in variation "hot spots" may additionally complicate classification of real events. As for the simulations, SVM^2 showed a tendency to misclassify only small events (≤5bp). Table 3 summarizes the classification patterns for such events, while Fig S4 illustrates profiles of SV size prediction accuracy for SVM^2. SVM^2 shows a tendency misclassify small (≤5bp) deletions rather than insertions. Consistent with the difficulty of classifying small events, our hyper variable and indindel predictions almost exclusively contain small indels at a similar validation rate to other categories.

Next, we asked whether, for a series of size range bins the sensitivity by genomic context showed obvious differences between methods. Fig 3 confirms that for the smallest events (≤5bp), PinDel outperformed the other methods in most genomic contexts. However, the sensitivity of SVM^2 in SINEs and low complexity regions was comparable to that of PinDel, while in simple repeats SVM^2 outperformed PinDel). As expected – given the small number of predictions by BreakDancer in this size range, the sensitivity was low. For events of between 6 and 10bp in size, SVM^2 was the most sensitive method dramatically outperforming BreakDancer in all genomic contexts. PinDel was almost as sensitive as SVM^2 in DNA transposons and non-repetitive DNA. As event size increases, PinDel shows decreasing sensitivity particularly in low complexity regions and simple repeats (an inevitable

property of split mapping methods). Even for larger (>20bp) events, which BreakDancer was designed to detect, SVM^2 is more sensitive in all genomic contexts. It is notable that, overall, SVM^2 and BreakDancer seem to show much less dependence on genomic context than PinDel.

We were intrigued by the difference of apparent specificities between methods previously observed when using the 1000 genomes SV catalog (but not when using dbSNP or the Kidd et al. data) as a validation set and by the relatively large proportion of the small (<10bp) events found by SVM^2 but not PinDel that fall in low complexity and simple repeat regions (10037/19274, 52%). We reasoned that these observations might be linked by the fact that the 1000 genomes catalog used split mapping to identify small events, and showed that a notable proportion (>97%) of the part of the genome deemed "inaccessible" by their low coverage data, fell in regions annotated as "high copy repeats or segmental duplications" [2]. Accordingly, we investigated the genomic distribution of predictions validated by Sanger sequencing but not by the 1000 genomes catalog by event size and method. We observed that a relatively small proportion of the small events (<10bp) validated by the Kidd et al. data and predicted by PinDel but not supported by the 1000 genomes dataset, fall in low complexity regions and simple repeats (1483/16081, 9.25%), while the equivalent numbers for SVM^2 were (5991/18450, 32%), suggesting that SVM^2, or similar methods, might effectively complement existing tools and pipelines in the detection of very short SVs, particularly in repetitive and low sequence complexity areas of the genome.

Finally, we compared the frequency of predictions by SVM^2 in genic regions with the rest of the genome, reasoning that SV events should occur at lower frequency in the former. 1.2% and 0.27% of predictions fell in genic and CDS regions respectively (using refseq genes). We estimated the significance of the difference between expected and observed frequencies using the Poisson distribution. The departure from the null model that predictions are distributed randomly along the genome was $<10^{-20}$ for both categories.

DISCUSSION

With simulated data, both sensitivity and specificity attained by our method were exceptionally high, although it should be emphasized that other methods have generated similarly impressive results in similar benchmarks but show, in particular, lower sensitivity with real data [28,29]. This is unsurprising as the effects of repetitive sequences and inherent biases in sequence coverage tend to be minimized in simulations. However, for the study of heterozygous events, simulation for now provides the only realistic possibility owing to a fundamental lack of large scale validated heterozygosity catalogs associated with individual genomes. SVM^2 showed relatively poor accuracy in the detection and classification of very short heterozygous SV. All mapping-distance based methods are expected to

suffer from this limitation as distance perturbations are diluted at heterozygous loci. In addition, our current approach employs measures of coverage and in the case of heterozygous deletions, a reduction, rather than an absence of reads in the deleted region would be expected. Conversely, reduced perturbations of BP mapping patterns are expected upstream of heterozygous insertions. These limitations might be partially addressed by some of the potential developments in the strategy that are envisaged (see below). However, in simulation at least, we note a satisfactory performance by SVM2 in the identification and classification of larger heterozygous events.

In the work presented here, SVM^2 was trained to recognize hypervariable regions as distinct from SV events. In practice, few predictions of this type were made. Indeed, an examination of these predictions suggested that they showed a similar specificity in detection of SVs as the other categories of prediction – although all validated predictions in this category corresponded to events of 4 bases or less. This is likely a function of the read mapping strategy employed. Allowing up to 2 mismatches in 35 base reads tends to allow correct mapping of the majority of reads in intraspecific comparisons, and in any case, perturbations of read mapping caused by hypervariable genomic regions are expected to be extremely subtle.

The Bentley/Kidd data represents one of the few cases where extensive Sanger resequencing and SV calling has been performed on an individual for which PE NGS data is also available - providing an “independent” validation set. For this reason, the dataset has been widely used in other studies [22,28,29] and allows immediate comparisons between methods. These considerations notwithstanding, the dataset has several relevant limitations that complicate interpretation of results and merit discussion. Firstly the coverage by Sanger sequencing is rather limited (theoretical coverage 0.3X), suggesting that, even if we make the –optimistic - assumption that all reads were mapped correctly and uniquely, at most less than a third of the SV events between this individual and the hg18 reference could be detected. Secondly, the low coverage implies that the accurate annotation of heterozygous events should be, at best, extremely limited. Finally, the original study of Kidd et al. only attempted to identify events of less than 100bp in length, and while a second evaluation of these data [38] was more comprehensive, the detection of large insertions is limited by the properties of split-mapping methods. Several studies suggest that the vast majority of intra-specific SVs are small [32], and while this generalization is almost certainly correct, our knowledge of the frequency of medium to large events remains rather limited. Our method made few predictions of insertions larger than the insert size of the library. However, this is an inherently difficult category of events to detect by any current approach and it is equally difficult to perform statistical analysis of sensitivity and specificity of tools with respect to detection of such events with the available data

Taken together, these observations render the objective assessment of the overall specificity of methods, with respect to both homozygous and heterozygous SV, extremely difficult. Additionally, the probability that a proportion of the Kidd and Mills predictions are heterozygous complicates estimates of sensitivity with respect to homozygous events. In this context, we believe that while limited in precision, apparent sensitivity and specificity are the best available metrics for comparison of the performance of different methods. By all metrics and validation sets employed, SVM^2 outperformed BreakDancer in terms of sensitivity over a range of SV event sizes, attaining at least the same apparent specificity. This is perhaps not surprising given that additional mapping information, not used by BreakDancer, is employed by SVM^2. Perhaps more significant is the observation that SVM^2 identified a large number of small SVs that were not detected by a contemporary split mapping method.

One alternative to the use of individual genome Sanger resequencing as a biological validation set would be to estimate specificity by comparing genome wide predictions to collections of validated population level SVs (dbSNP [39], 1000 genomes project [2]) making the assumption that coincidence of predictions with an annotated SV implied the presence of the same SV in the donor genome. However, a recent study demonstrated a relatively low overlap between the two aforementioned databases, implying that a significant fraction of human SVs remain undetected [38] It is also worth noting that the 1000 genomes set of SV events was generated from NGS data. Given that our objective was to explore the potential of this very type of data to uncover additional, previously undetected events, we consider that the use of “independent” data from the individual genome under study as our principal validation set to be a justified strategy. Nevertheless, comparisons of apparent specificity of different methods when “validated” by Sanger or NGS based datasets showed interesting patterns, particularly with respect to the genomic context of indel events.

The “elephant in the room” of all methods to determine locations of SV from resequencing data, be they based on split mapping or on statistical approaches is the abundance of repeated sequences in complex genomes. Sequence reads (from any technology) that fall within perfectly repeated regions cannot be unambiguously mapped. PE approaches (dependent on library insert size and repeat length) can ameliorate this problem to some extent, as can probabilistic mapping strategies [30], but the fundamental problem remains. For example, SVs within recent segmental duplications present an almost insurmountable problem for all approaches apart from read-depth methods – and even these will not be able to specify the location of the event. For now, the most promising way to address the problem of repeats may be the maximization of read length and the use of different insert size libraries. The use of larger insert size libraries will aid the detection of larger SV events by insert size-based methods (and contribute to an additional loss of accuracy in the identification of small indels by such methods). Conversely, as the production of longer resequencing reads using NGS technologies becomes more commonplace, the sensitivity of split mapping methods is expected to increase for small to medium size events and to reduce the impact of repetitive sequences on the performance of

all methods. Despite these problems, we note that our analysis of genomic context of predictions and validated predictions suggest that in simple repeats and low complexity regions, SVM^2 attained higher sensitivity than other methods tested, even for small SV events. The observations that a large number of small SV events detected by Sanger resequencing, but not by PinDel (or 1000 genomes) fall in simple repeat and low complexity regions, and that a larger proportion of validated SVM^2 than PinDel predictions fall in such regions are interesting. In this light, the similarity of overall "specificity" between methods when evaluated with the Kidd et al. data or with dbSNP, and the differences in this metric with respect to the 1000 genomes database is intriguing, particularly given the types of data used to construct these catalogs. Simple repeat/low complexity regions represent a notable proportion of the "inaccessible" genome described by the 1000 genomes consortium [2]. We suggest that our method, or others based on similar principles, might be of particular use in addressing SV in such regions.

We can envisage several potential developments to the approach presented here, some of which might be expected to improve the performance with respect to heterozygous SV. Firstly, sequence coverage might be improved by using split mapping in the initial generation of read maps (here we have used only gapless alignment). Secondly, additional features, for example the gapless and gapped alignment coverage for each genomic site could be incorporated into the SVM analysis. Another possible step would be to use positional constraints (based on SVM^2 predictions) in split mapping of reads as a post processing step in establishing additional support for events and in fine mapping positions of SV.

In conclusion, we have shown that inclusion of more detailed information on the local patterns of read mapping can notably enhance the sensitivity of detection of SV events by non split-mapping methodologies.

Furthermore, we showed that insert size-based SV detectors such as SVM^2 can complement split mapping approaches in the localization of ultra short SV events, particularly those in repetitive and low complexity regions of the genome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online: Supplementary tables 1-5, Supplementary figures 1-4. The SVM^2 software, documentation and example files are available via anonymous ftp from ftp:159.149.109.10/pub/svm2.

References

[1] Shendure J and Ji H. (2008) Next-generation dna sequencing. *Nature Biotechnology*, **26**, 1135–1145.

[2] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct 2010.

[3] Iafrate AJ, Feuk L, Rivera MN,Listewnik ML,Donahoe PK,Qi Y, Scherer SW, and Lee C. (2004) Detection of large-scale variation in the human genome. *Nature Genetics*, **36**, 949–951.

[4] Sebat J, Lakshmi B, Troge J Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M et al. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.

[5] Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D et al. (2005) Fine-scale structural variation of the human genome. *Nature Genetics*, **37**, 727–732.

[6] Buchanan JA and Scherer SW. (2008) Contemplating effects of genomic structural variation. *Genetics in Medicine*, **10**, 639–647.

[7] McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL et al. (2008) Integrated detection and population-genetic analysis of snps and copy number variation. *Nature Genetics*, **40**, 1166–1174.

[8] Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P et al. (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

[9] Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK et al. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

[10] Braude I,Vukovic B, Prasad M, Marrano P,Turley S, Barber D,Zielenska M and Squire, JA. Large scale copy number variation (cnv) at 14q12 is associated with the presence of genomic abnormalities in neoplasia. (2006) *BMC Genomics*, **7**.

[11] Bijlsma EK, Gijsbers AC, Schuurs-Hoeijmakers JH, van Haeringen A, Fransen van de Putte DE, Anderlid BM, Lundin J, Lapunzina P,Pérez Jurado LA, Delle Chiaie B et al. (2009) Extending the phenotype of recurrent rearrangements of 16p11.2: deletions in mentally retarded patients without autism and in normal individuals. *European Journal of Medical Genetics*, **52**, 77–87.

[12] McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, Perkins DO, Dickel DE, Kusenda M et al. (2009) Microduplications of 16p11.2 are associated with schizophrenia. *Nature Genetics*, **41**, 1223–1227.

[13] Tam GWC,Redon R, Carter NP, and Grant SGN. (2009) The role of dna copy number variation in schizophrenia. *Biological Psychiatry*, **66**, 1005–1012.

[14] Ballif BC, Theisen A, Rosenfeld JA, Traylor RN, Gastier-Foster J, Thrush DL, Astbury C, Bartholomew D, McBride KL et al. (2010) Identification of a recurrent microdeletion at 17q23.1q23.2 flanked by segmental duplications associated with heart defects and limb abnormalities. *American Journal of Human Genetics*, **86**, 454–461.

[15] Clayton-Smith J, Giblin C,Smith RA, Dunn C, and Willatt L. (2010) Familial 3q29 microdeletion syndrome providing further evidence of involvement of the 3q29 region in bipolar disorder. *Clinical Dysmorphology*, **19**, 128–132.

[16] Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**, 368–372.

[17] Stankiewicz P and Lupski JR. (2010) Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, **61**, 437–45.

[18] Ken Chen et al. (2009) Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, **6**, 677–681.

[19] Hormozdiari F, Alkan C, Eichler EE, and Sahinalp SC. (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research*, **19**, 1270–1278.

[20] Albers CA, Lunter G, MacArthur DG, McVean G, Ouweh and WH, Durbin R. (2011) Dindel: Accurate indel calls from short-read data. *Genome research,* **21,** 961-73

[21] Handsaker RE, Korn JM, Nemesh J, McCarroll SA. (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genetics,* **43**,269-76

[22] Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, and Sahinalp SC. (2010) Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics*, **26**, 1277–1283.

[23] Kai Ye, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.

[24] Medvedev P,Stanciu M, and Brudno M. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, **6**(11 Suppl), S13–S20.

[25] Alkan C, Koe BP, Eichler EE. (2011) Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12**, 363-376

[26] Yoon, S., Xuan, Z., Makarov, V., Ye, K. Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* **19**, 1586-1592

[27] Abyzov, A., Urban, A. E., Snyder, M. Gerstein, (2011) M. CNVnator: an approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974-84

[28] Chen K et al. (2009) Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*, **6**, 677–681

[29] Lee S, Hormozdiari F,Alkan C, and Brudno M. (2009) Modil: detecting small indels from clone-end sequencing with mixtures of distributions. *Nature Methods*, **6**, 473–474.

[30] Hormozdiari F, Hajirasouliha I,Dao P,Hach F, Yorukoglu D, Alkan C, Eichler EE, and Sahinalp SC. (2010) Next-generation variation hunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**, i350–i357.

[31] Noble WS, What is a support vector machine? (2006). *Nature Biotechnology* **24**, 1565 – 1567

[32] Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, **16,** 1182-1190

[33] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. (2010) *Bioinfomatics* **26**, 841-2

[34] Ching C. Chang and Chin T. Lin. (2011) Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 1–27.

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

[36] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

[37] Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E et al. (2008) Mapping and sequencing of structural variation from eight human. genomes. *Nature*, **453**, 56–64.

[38] Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, Kemeza DM, Strassler DS, Ponting CP, Webber C, Devine SE (2011) Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Research* **6**, 830-39

[39] Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID:130}). Available from: http://www.ncbi.nlm.nih.gov/SNP/

TABLE AND FIGURES LEGENDS

Table 1 Simulation

Deletions:

Size	Recall*	Recall as any**	FP rate
1 to 5	69%	83%	8.5%
6 to 10	82%	89%	6.3%
11 to 20	91%	92%	2%
21 to 40	94%	94%	0
41 to 60	97%	97%	0
61 to 100	95%	95%	0
101 to 200	97%	97%	0
>200	97%	97%	0

Insertions:

Size	Recall*	Recall as any**	FP rate
1 to 5	70%	86%	9%
6 to 10	82%	88%	6%
11 to 20	94%	94%	2%
21 to 40	92%	92%	0.5%
41 to 60	93%	93%	0
61 to 100	91%	91%	0
101 to 200	89%	89%	0
>200	86%	86.00%	0

Table 1:Simulation. *correctly classified as insertion or deletion **correctly identified locus, includes indindel and hypervariable predictions

Table 2 Simulations of heterozygous events

Deletions

Size*	Recall rate**	Correctly classified***	Recall rate if homozygous****
1	10%	0	83%
3	10%	0	87%
5	13%	15%	94%
10	40%	20%	98%
15	53%	29%	99%
20	63%	45%	99%
30	85%	87.5%	99%
40	87.5%	93.5%	99%

Insertions

Size*	Recall rate**	Correctly classified***	Recall rate if homozygous****
1	10%	0	80%
3	10%	0	86%
5	17%	3%	94%
10	28%	14%	99%
15	48%	32%	99%
20	57%	47%	99%
30	81%	89%	99%
40	88%	96%	99%

Table 2:Simulation of heterozygous events:

*size of the event, **recall rate for the heterozygous case , *** proportion of recalled indels classified as heterozygous, ****recall rates for equivalent (same locus) Homozygous indels

Table 3: Classification accuracy of short indels predicted by SVM[2]

		Kidd***				
SVM[2]*	**Total ** **	**Insertions**	**Deletions**	**Validation rate **** **		**Misclassification rate ***** **
insertions	50688	11068	2288	13356	26.3%	17.1
deletions	46102	3991	8111	12102	26.2%	32
indIndels	9118	1308	1049	2357	25.8%	
Hyper-variable	8503	1268	982	2250	26.4%	

Table3: Classification accuracy of short indels predicted by SVM[2]. *class predicted by SVM[2]. **number of predictions by SVM[2] by category, *** class of the validating event in the dataset,, **** validation rate for each category of SVM[2] predictions (applies for small only) . ***** Misclassification rate for validated insertions and deletions

Figure 1: Sensitivity and specificity of different methods with the Kidd et al. dataset

Fig 1A Number of indels from the Kidd dataset (binned by size of event in bp) recalled by each method.

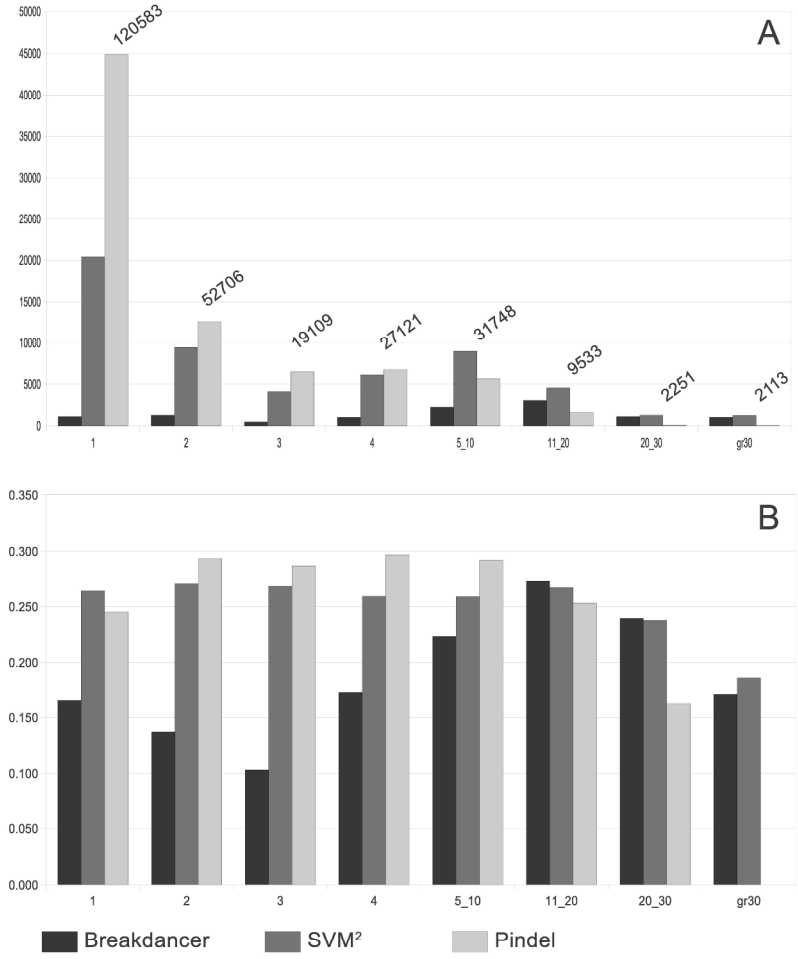
Fig 1B Proportion of predicted indels (binned by predicted sizes) that are validated by an indel in the kidd et al. validation set.

size bins: size≤1, size≤2,size≤3,size≤4,5≤size≤10,10<size≤20,20<size≤30,size>30

Figure 2: Venn diagram showing intersection between validated (by kidd) predictions by each method.

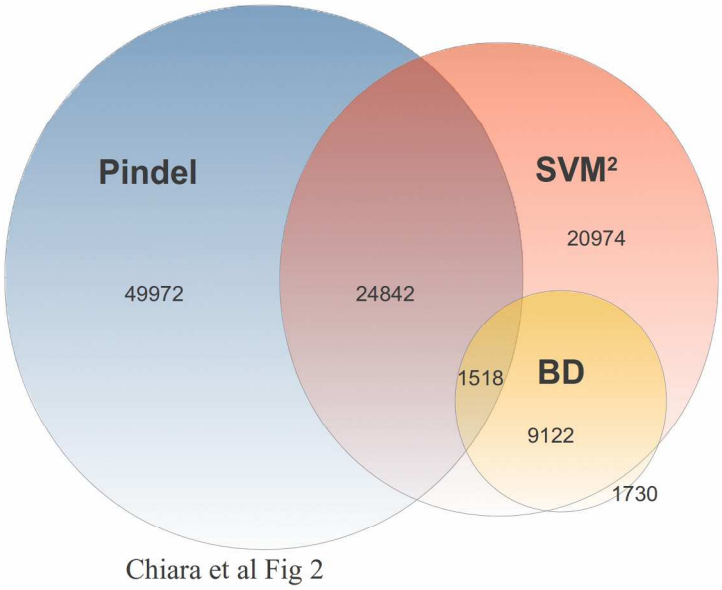
Figure 3: Sensitivity by size and genomic context.

Numbers of events in the Kidd dataset, in different genomic contexts (tDNA=DNA transposon, LTR = long terminal repeats, NR= non repetitive), recalled at different size ranges (size≤5, 5<size≤10, 10<size≤20, size>20) by different methods.



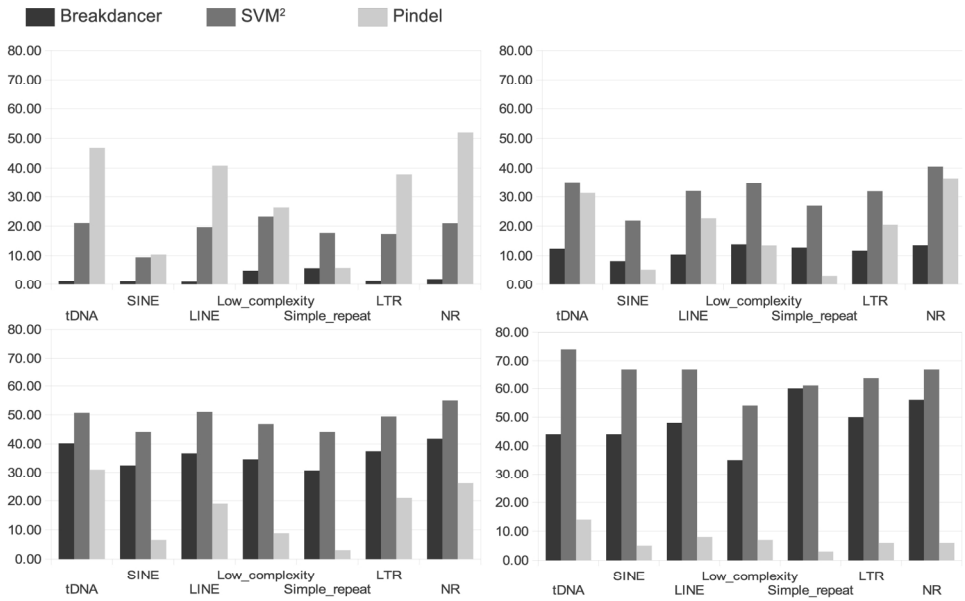
Chiara et al Fig 1

297x420mm (300 x 300 DPI)



Chiara et al Fig 2

729x516mm (72 x 72 DPI)



Chiara et al figure 3

190x130mm (300 x 300 DPI)

Part III Supplementary material enclosed with SVM^2

Supplementary tables

Supplementary table 1 Complete Heterozigosity table.

Deletions

Size*	Proportion**	Detected %***	Het %****	Det Hom %****
1	0.3	0.6	0	83
	0.4	2	0	
	0.5	10	0	
	0.6	15	0	
3	0.3	2.6	0	87
	0.4	4.6	0	
	0.5	10	0	
	0.6	21	3	
5	0.3	3	0	94
	0.4	4	0	
	0.5	13	15	
	0.6	27	15	
10	0.3	4	0	98
	0.4	14	14	
	0.5	40	20	
	0.6	68	20	
15	0.3	9	7	99
	0.4	22	41	
	0.5	66	29	
	0.6	88	30	
20	0.3	12	33	99
	0.4	38	39.7	
	0.5	69	40	
	0.6	90	45	
30	0.3	16	41.6	99
	0.4	52	89	
	0.5	85	87.5	
	0.6	97	83	
40	0.3	21	92	99
	0.4	61	89	
	0.5	87.5	93	
	0.6	98.75	92	

Insertions

Size*	Proportion**	Detected %***	Het %****	Det Hom %****
1	0.3	0.7	0	80
	0.4	7.3	0	0
	0.5	10.7	0	0
	0.6	17.3	0	0
3	0.3	3.3		86
	0.4	7.3	0	0
	0.5	9.3	0	0
	0.6	22.7	2.9	0
5	0.3	3.3	0	94
	0.4	5.3	0	0
	0.5	17.3	3.8	0
	0.6	31.3	6.4	0
10	0.3	3.3	0	99.3
	0.4	16.0	12.5	0
	0.5	28.0	14.3	0
	0.6	58.7	17.0	0
15	0.3	4.0	0	99.3

	0.4	22.0	30.3	0
	0.5	43.3	32.3	0
	0.6	71.3	34.6	0
20	0.3	6.0	44.4	99.3
	0.4	29.3	47.7	0
	0.5	57.3	47.7	0
	0.6	84.0	48.4	0
30	0.3	21.0	63.6	99.3
	0.4	71.0	91.1	0
	0.5	81.0	89.9	0
	0.6	88.7	91.0	0
40	0.3	30	88.2	99.3
	0.4	59.0	93.5	0
	0.5	88.0	96.5	0
	0.6	97.0	96.6	0

*size of the event, **proportion of sampling taken from the mutated haplotype
***recall rate for the heterozygous case ,
**** proportion of recalled indels classified as heterozygous,
*****recall rates for equivalent (same locus) Homozygous indels

Supplementary table 2: Size distributions of call by different methods and apparent specificity on Kidd dataset

size	**BreakDancer**			**SVM²**			**Pindel**			**Valid Set**
	Total	**Valid**	**Valid %**	**Total**	**Valid**	**Valid %**	**Total**	**Valid**	**Valid %**	
1	109	18	16.51	39688	10474	26.39	180760	44286	24.50	120583
2	73	10	13.70	24120	6522	27.04	40668	11936	29.35	52706
3	97	13	10.31	20777	5572	26.82	22303	6396	28.68	19109
4	127	22	17.32	16484	4271	25.91	21854	6486	29.68	27121
5_10	5968	1331	22.30	42236	10935	25.89	18818	5493	29.19	31748
11_20	18310	4989	27.25	20470	5463	26.69	7044	1782	25.30	9533
21_30	6332	1516	23.94	5360	1272	23.73	111	18	16.22	2251
gr30	8575	1470	17.14	7396	1376	18.60				2213
										265264

Supplementary table 3: Number of calls and recall rate (Sensitivity) respect to the Kidd validation dataset (Fig 1A)

SIZE	BreakDancer*	SVM²**	PinDel***	BreakDancer %****	SVM² %*****	PinDel %******	Valid Set *******
1	1133	20409	44897	0.94	16.93	37.23	120583
2	1287	9444	12619	2.44	17.92	23.94	52706
3	470	4101	6504	2.46	21.46	34.04	19109
4	1032	6128	6754	3.81	22.6	24.9	27121
5_10	2246	8978	5652	7.07	28.28	17.8	31748
11_20	3070	4550	1606	32.2	47.73	16.85	9533
21_30	1131	1290	88	50.24	57.31	3.91	2251
gr30	1046	1277	54	47.27	57.7	2.44	2213
TOTAL	11415	56177	78174	4.3	21.18	29.47	265264

ST3: Validation rate (Specificity) per method per predicted size (Fig 1A) on Kidd dataset
*,**** Absolute number and proportion of indels of different size recalled by Breakdancer
,*** Absolute number and proportion of indels of different size recalled by svm2
,** Absolute number and proportion of indels of different size recalled by Breakdancer
******* Number of events in the validation set by size

Supplementary table 4: apparent specificity for each method on dbsnp and 1000 genomes data collections

Predicted_size*	BreakDancer **		SVM²***		PinDel***	
	dbsnp	1000	dbsnp	1000	dbsnp	1000
1	51	17	79.3	70	78.2	75.5
2	58	9	79.6	69.7	83	81.6
3	54	11	79.9	69.6	81.2	87.7
4	59	20	79.3	67.9	82.8	92.3
5_10	71.2	59.5	81.1	66.8	81.9	90.9
11_20	88.6	65.5	88	65.9	79	85
21_30	87.3	62.6	86	63	68	77
gr30	72	45.8	73	46		

Proportions are respect to the numbers reported in supplementary table 2
* predicted size
** apparent specificity for BreakDancer
*** apparent specificity for SVM²
**** apparent specificity for PinDel

Supplementary figures.

Suppl. Figure 1: Expected pattern of read mapping in the presence of different SVs

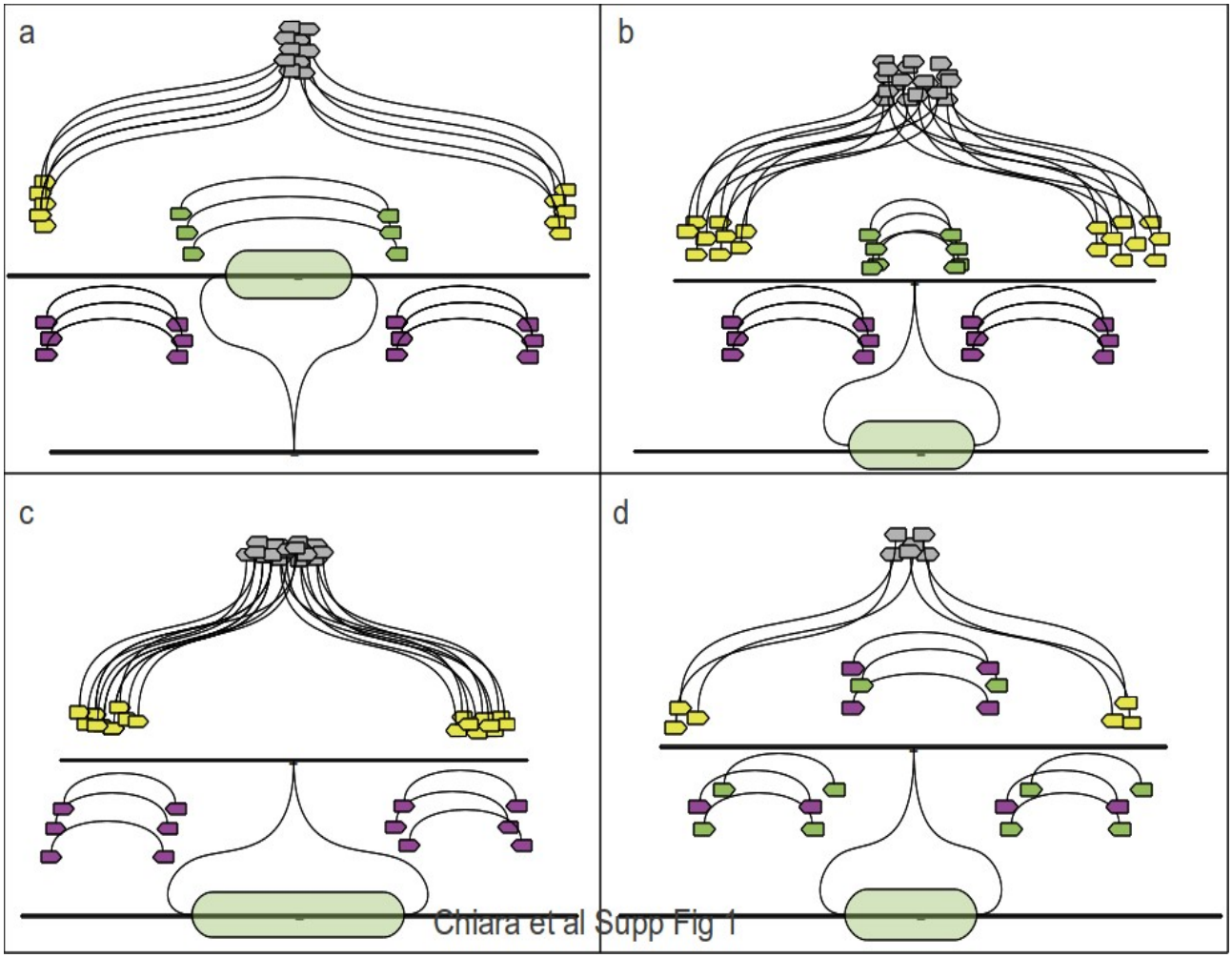


Figure S1 schematically represents the expected pattern of mapping of reads on a reference genomic sequence in the case of a deletion (a) an insertion shorter than the insert-size (b) an insertion longer than the insert-size (c) and in the presence of a particularly variable region (d). The classic approaches based on PE reads to detect indels in this scenario are: 1 define a cut-off to identify aberrant mapping reads and individuate indels as genomic clusters of aberrant mapping mates, this strategy is particularly successful for the detection of long indels; or (2) to detect smaller indels, use a particular statistic to assess whether the local insert size distribution of a particular genomic locus is significantly different from the expectations (global distribution of insert-size).

It is clear from figure 1 that in this scenario there is some additional information which could be useful to integrate in the process, as in each case an SV creates a new genomic junction, which implies that some read from the donor can't map on the reference any-more, thus generating the so called “broken-pairs”.

The difference lies in the fact that each SV event generates such broken-pairs in a specific fashion: in the case of a deletion (a) we expect a sharp peak, while for short insertion (b) we expected a broader one and eventually whence the insertion becomes too long, all we can see is a peak of broken pairs as broad as the insert-size. Furthermore, by looking at their orientation, we can distinguish between PE mapping upstream or downstream respect to an hypothetical breakpoint; this information could be used to broaden the spectrum of statistical tests used for assessing significant insert-size perturbations: indeed instead of just comparing the local distribution to the global (like others do) we could run additional test(cross-checks) by comparing upstream vs downstream, downstream vs global and upstream vs global.

Finally in figure 1 (d) illustrate show there can be some misleading signals in the case of particularly variable and localized regions, which can also lead both to the generation of peaks of broken pairs and to subtle shifts in apparent insert size distributions (although without the directional specificity observed for indels).

Suppl. Figure 2: Features used by SVM²

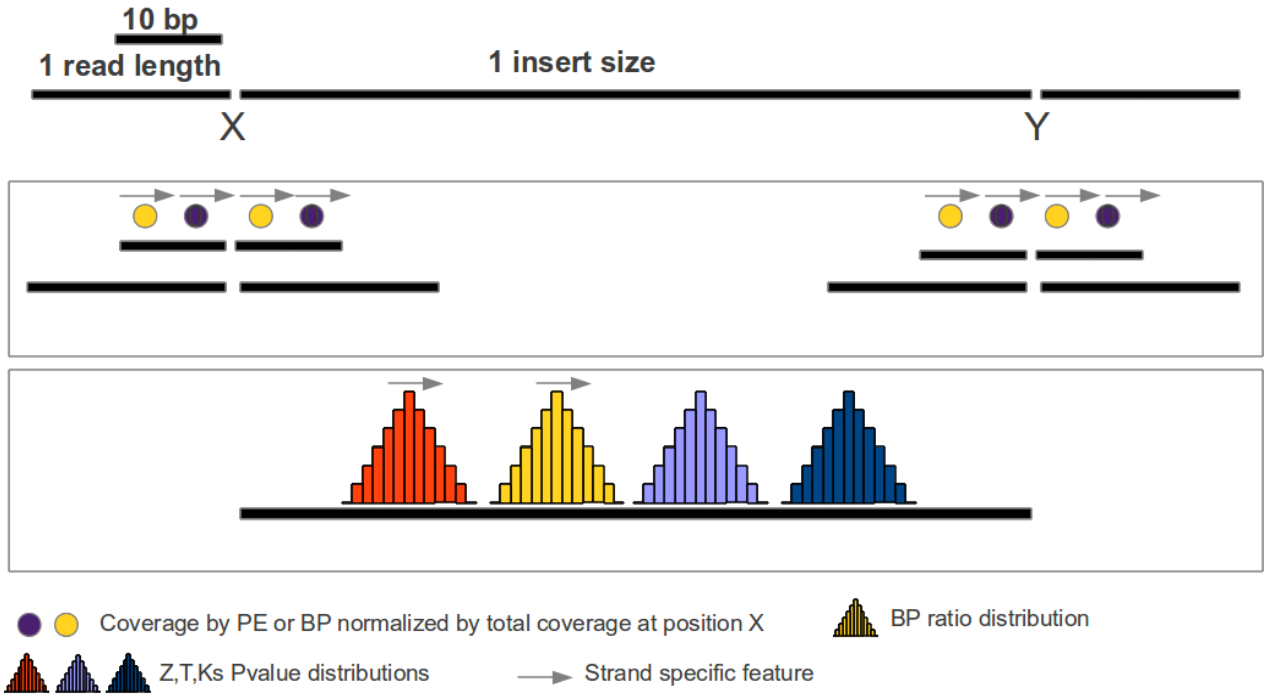


Figure S2: shows the localization and strandness (arrow) of the features used by our SVM. X is the position invoking the SVM, while Y is the genomic position at which mates of X are expected to be found (see methods) PE= paired end, BP=broken pairs Z= Z test, T=T-Welch test KS= Kolmogoroff Smirnov test. Features with an arrow on top are calculated on both strand All the distances are expressed in ID (see methods)

Suppl. Figure 3: Specificity by size using (3a) dbsnp130 or (3b) 1000genomes as validation set

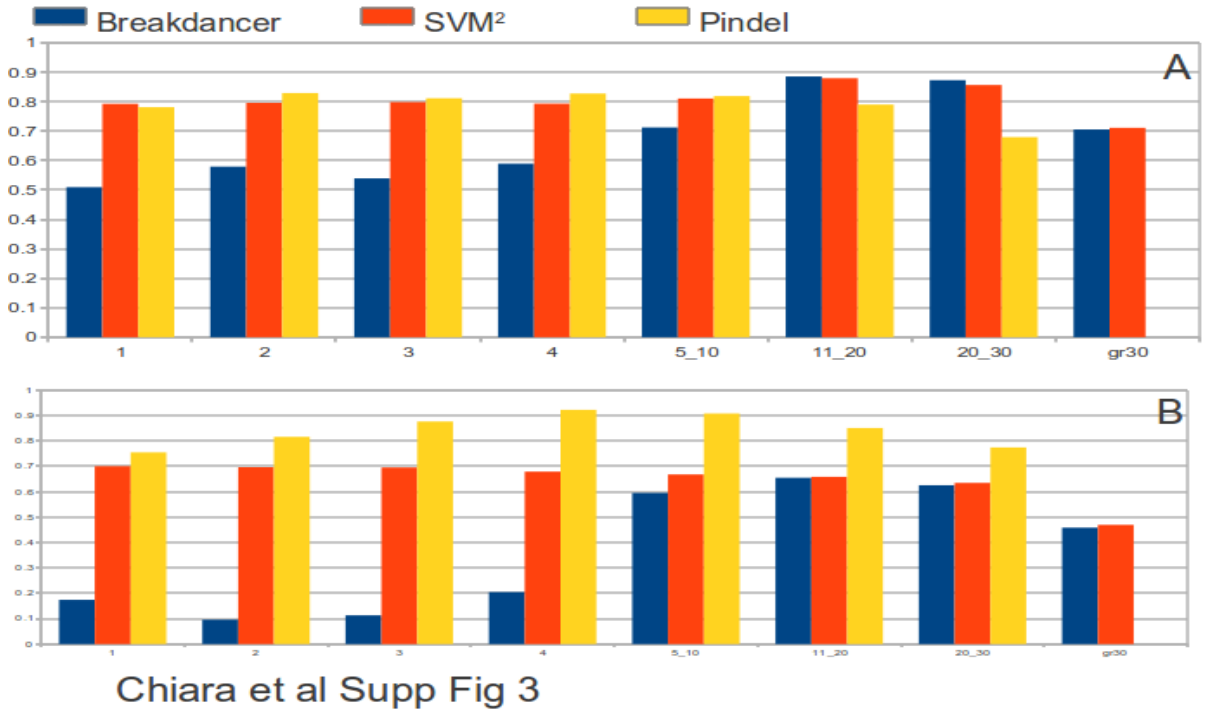
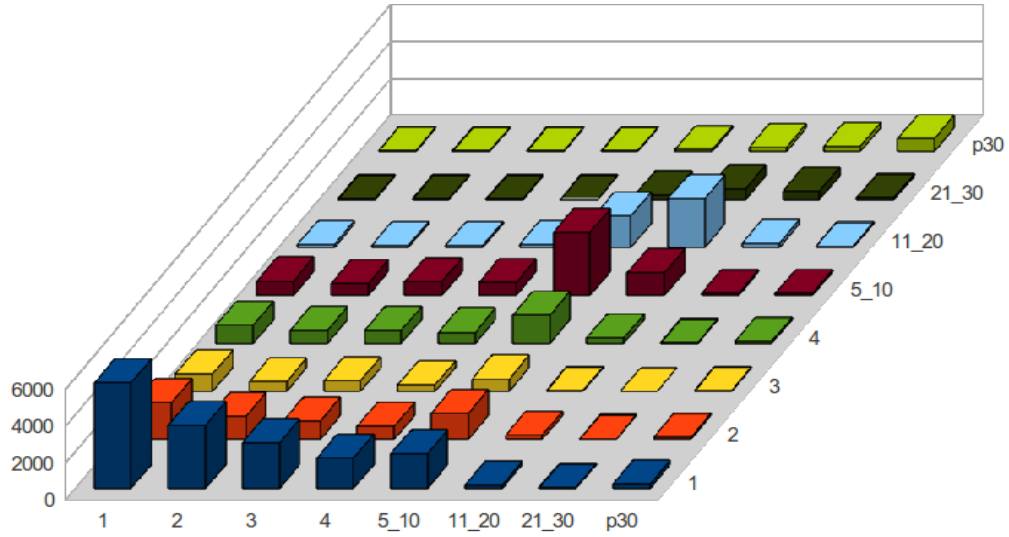


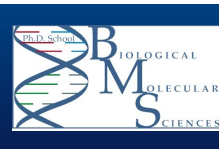
Figure S3a Equivalent to figure 1B but calculated using the whole dbsnp130 [39] as validation set
Figure S3b Equivalent to figure 1B but calculated using the entire 1000genomes SV catalog [2] as validation set

Suppl. Figure 4: 3D size distribution of predicted indels by SVM² validated by Kidd dataset



Chiara et al Supp. Fig 4

The X axis indicates the predicted sizes of events predicted by SVM² while the Z axis shows the real dimensions of the corresponding validated events from the Kidd et al. dataset. Numbers of events are shown on the Y axis.



www.unimi.it http://users.unimi.it/dottscbiolemol/