UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Head Office: Università degli Studi di Padova

Department of Biology

Ph.D. COURSE IN: Biosciences

CURRICULUM: Genetics, Genomics and Bioinformatics

SERIES XXX

Bioinformatics for personal genomics: development and application of bioinformatic procedures for the analysis of genomic data

Coordinator: Prof. Ildikò Szabò

Supervisor: Prof. Franca Anglani

Co-Supervisor: Prof. Giorgio Valle

Ph.D. student: Loris Bertoldi

Abstract

In the last decade, the huge decreasing of sequencing cost due to the development of high-throughput technologies completely changed the way for approaching the genetic problems. In particular, whole exome and whole genome sequencing are contributing to the extraordinary progress in the study of human variants opening up new perspectives in personalized medicine. Being a relatively new and fast developing field, appropriate tools and specialized knowledge are required for an efficient data production and analysis.

In line with the times, in 2014, the University of Padua funded the BioInfoGen Strategic Project with the goal of developing technology and expertise in bioinformatics and molecular biology applied to personal genomics. The aim of my PhD was to contribute to this challenge by implementing a series of innovative tools and by applying them for investigating and possibly solving the case studies included into the project.

I firstly developed an automated pipeline for dealing with Illumina data, able to sequentially perform each step necessary for passing from raw reads to somatic or germline variant detection. The system performance has been tested by means of internal controls and by its application on a cohort of patients affected by gastric cancer, obtaining interesting results.

Once variants are called, they have to be annotated in order to define their properties such as the position at transcript and protein level, the impact on protein sequence, the pathogenicity and more. As most of the publicly available annotators were affected by systematic errors causing a low consistency in the final annotation, I implemented VarPred, a new tool for variant annotation, which guarantees the best accuracy (>99%) compared to the state-of-the-art programs, showing also good processing times. To make easy the use of VarPred, I equipped it with an intuitive web interface, that allows not only a graphical result evaluation, but also a simple filtration strategy.

Furthermore, for a valuable user-driven prioritization of human genetic variations, I developed QueryOR, a web platform suitable for searching among known candidate genes as well as for finding novel gene-disease associations. QueryOR combines several innovative features that make it comprehensive, flexible and easy to use. The prioritization is achieved by a global positive selection process that promotes the emergence of the most reliable variants, rather than filtering out those not satisfying the applied criteria.

QueryOR has been used to analyze the two case studies framed within the BioInfoGen project. In particular, it allowed to detect causative variants in patients affected by lysosomal storage diseases, highlighting also the efficacy of the designed sequencing panel. On the other hand, QueryOR simplified the recognition of LRP2 gene as possible candidate to explain such subjects with a Dent disease-like phenotype, but with no mutation in the previously identified disease-associated genes, CLCN5 and OCRL.

As final corollary, an extensive analysis over recurrent exome variants was performed, showing that their origin can be mainly explained by inaccuracies in the reference genome, including misassembled regions and uncorrected bases, rather than by platform specific errors.

Sommario

Nell'ultimo decennio, l'enorme diminuzione del costo del sequenziamento dovuto allo sviluppo di tecnologie ad alto rendimento ha completamente rivoluzionato il modo di approcciare i problemi genetici. In particolare, il sequenziamento dell'intero esoma e dell'intero genoma stanno contribuendo ad un progresso straordinario nello studio delle varianti genetiche umane, aprendo nuove prospettive nella medicina personalizzata. Essendo un campo relativamente nuovo e in rapido sviluppo, strumenti appropriati e conoscenze specializzate sono richieste per un'efficiente produzione e analisi dei dati.

Per rimanere al passo con i tempi, nel 2014, l'Università degli Studi di Padova ha finanziato il progetto strategico BioInfoGen con l'obiettivo di sviluppare tecnologie e competenze nella bioinformatica e nella biologia molecolare applicate alla genomica personalizzata. Lo scopo del mio dottorato è stato quello di contribuire a questa sfida, implementando una serie di strumenti innovativi, al fine di applicarli per investigare e possibilmente risolvere i casi studio inclusi all'interno del progetto.

Inizialmente ho sviluppato una pipeline per analizzare i dati Illumina, capace di eseguire in sequenza tutti i processi necessari per passare dai dati grezzi alla scoperta delle varianti sia germinali che somatiche. Le prestazioni del sistema sono state testate mediante controlli interni e tramite la sua applicazione su un gruppo di pazienti affetti da tumore gastrico, ottenendo risultati interessanti.

Dopo essere state chiamate, le varianti devono essere annotate al fine di definire alcune loro proprietà come la posizione a livello del trascritto e della proteina, l'impatto sulla sequenza proteica, la patogenicità, ecc. Poiché la maggior parte degli annotatori disponibili presentavano errori sistematici che causavano una bassa coerenza nell'annotazione finale, ho implementato VarPred, un nuovo strumento per l'annotazione delle varianti, che garantisce la migliore accuratezza (>99%) comparato con lo stato dell'arte, mostrando allo stesso tempo buoni tempi di esecuzione. Per facilitare l'utilizzo di VarPred, ho sviluppato un'interfaccia web molto intuitiva, che permette non solo la visualizzazione grafica dei risultati, ma anche una semplice strategia di filtraggio.

Inoltre, per un'efficace prioritizzazione mediata dall'utente delle varianti umane, ho sviluppato QueryOR, una piattaforma web adatta alla ricerca all'interno dei geni causativi,

ma utile anche per trovare nuove associazioni gene-malattia. QueryOR combina svariate caratteristiche innovative che lo rendono comprensivo, flessibile e facile da usare. La prioritizzazione è raggiunta tramite un processo di selezione positiva che fa emergere le varianti maggiormente significative, piuttosto che filtrare quelle che non soddisfano i criteri imposti.

QueryOR è stato usato per analizzare i due casi studio inclusi all'interno del progetto BioInfoGen. In particolare, ha permesso di scoprire le varianti causative dei pazienti affetti da malattie da accumulo lisosomiale, evidenziando inoltre l'efficacia del pannello di sequenziamento sviluppato. Dall'altro lato invece QueryOR ha semplificato l'individuazione del gene LRP2 come possibile candidato per spiegare i soggetti con un fenotipo simile alla malattia di Dent, ma senza alcuna mutazione nei due geni precedentemente descritti come causativi, CLCN5 e OCRL.

Come corollario finale, è stata effettuata un'analisi estensiva su varianti esomiche ricorrenti, mostrando come la loro origine possa essere principalmente spiegata da imprecisioni nel genoma di riferimento, tra cui regioni mal assemblate e basi non corrette, piuttosto che da errori piattaforma-specifici.

Contents

List of Figures

List of Tables

List of Boxes

List of Abbreviations

AD	*Allele Depth*
AO	*Alternate Allele Observations*
BAM	*Binary Alignment Map*
BED	*Binary Extended Data*
CCD	*Charge-Coupled Device*
CMOS	*Complementary Metal-Oxide-Semiconductor*
CRIBI	*Interdepartmental Research Center of Innovative Biotechnology*
DB/FOAR	*Donnai-Barrow/Facio-oculo-acustico-renal*
DD	*Dent Disease*
ddNTP	*di-deoxynucleotidetriphosphate*
DO	*Disease Ontology*
DP	*Filtered Depth*
dsDNA	*double-stranded DNA*
ESP	*Exome Sequencing Project*
ExAC	*Exome Aggregation Consortium*
FGS	*First Generation Sequencing*
GFF	*General Feature Format*
GMAF	*Global Minor Allele Frequency*
GO	*Gene Ontology*
GT	*Genotype Codes*
GTF	*General Transfer Format*
HGP	*Human Genome Project*
HGVS	*Human Genome Variation Society*
HPO	*Human Phenotype Ontology*
LSD	*Lysosomal Storage Diseases*
MAiR	*Minor Allele in Reference*
MAF	*Minor Allele Frequency*
MNP	*Multiple Nucleotide Polymorphism*
NGS	*Next Generation Sequencing*

NHGRI *National Human Genome Research Institute*

PGM *Personal Genome Machine*

RO *Reference Allele Observation Count*

ROH *Runs Of Homozygosity*

RPKM *Reads Per Kilobase per Million mapped reads*

SAM *Sequence Alignment/Map*

SGS *Second Generation Sequencing*

SMRT *Single-Molecule Real-Time*

SMS *Single Molecule Sequencing*

SNP *Single Nucleotide Polymorphism*

SNV *Short Nucleotide Variant*

SOLiD *Sequencing by Oligo Ligation Detection*

SQL *Structured Query Language*

ssDNA *single-stranded DNA*

TGS *Third Generation Sequencing*

uBAM *unmapped BAM*

VC *Variant Calling*

VCF *Variant Call Format*

WES *Whole Exome Sequencing*

WGS *Whole Genome Sequencing*

XML *eXtensible Markup Language*

ZMW *Zero-Mode Waveguide*

1. Introduction

1.1. A survey on sequencing: 40 years of development

1.1.1. First Generation Sequencing

Nucleic acid sequencing is a powerful technique developed for solving the specific order of nucleotides within a DNA or RNA molecule, which have met the interest of many branches of medicine and biology. Although several methodologies had already been proposed at that time [1–3], the advent of sequencing can be found in the middle of the 1970s, when Maxam and Gilbert (1977) [4] and Sanger and Coulson (1975) [5] proposed in parallel two different sequencing approaches, called the "chemical sequencing method" and the "chain-termination method", respectively.

Briefly, in the first system, DNA fragments were labelled with a radioactive compound (usually γ-^{32}P) at one 5' end and then purified. The following chemical treatments triggered to a partial modification of the bases causing specific pattern of cleavage, depending on the chosen reaction (G, A+C, C, C+T). The obtained marked DNA chunks were separated by electrophoresis and revealed using autoradiography. The sequence could be deduced from presence and absence of specific fragments [6].

Instead, the chain-termination approach, also called Sanger sequencing, was based on DNA elongation, mediated by DNA-polymerase, which was blocked when a modified di-deoxynucleotidetriphosphates (ddNTPs) was incorporated. In this way, four parallel reactions, in which only one of the four radiolabeled-ddNTPs (ddATP, ddCTP, ddGTP and ddTTP) was added, were enough to easily obtain all fragments, necessary to inferred the sequence, after their electrophoresis on a thin acrylamide gel and corresponding bands revelation by autoradiography [7, 8].

Although at the beginning chemical sequencing had become more popular, since purified DNA could be directly handled, it was soon replaced by the Sanger sequencing, as it was less hazardous, less complex and more scalable, so resulting more prone to be automatized and scaled up [9]. In particular, the advent of capillary electrophoresis for DNA separation [10, 11] and the introduction of fluorescent ddNTPs (dye-terminator sequencing) [12] allowed to

sequence DNA in a single reaction, boosting the speed and simultaneously dropping the costs.

In this way, the automated Sanger sequencing became the gold standard of the first generation sequencing (FGS) methods, dominating the scenario for more than two decades and leading to extraordinary progresses in life sciences, including the publication of an initial draft [13, 14] in 2001 and then the complete sequence of the human genome in 2004 [15]. Nevertheless, the Human Genome Project (HGP), which at that time was the largest collaborative biological project, had also elucidated the limitations of FGS, highlighting in particular two huge aspects: the time spent (~13 years), directly linked to the low throughput, and the costs (US$3 billion) [16]. These problems were widely discussed by scientific community, identifying the urgency of decreasing the cost of DNA sequencing in order to reach the goal of $1000 for a genome. For this purpose, in 2004 the National Human Genome Research Institute (NHGRI) began to fund projects focused on the development of new technologies capable of reducing by four to five orders of magnitude the sequencing expenditure, committing more than $100 million to 50 research teams [17].

1.1.2. Second Generation Sequencing

The request of faster and cheaper sequencing approaches triggered the development of second-generation sequencing (SGS) methods, referred also as next-generation sequencing (NGS). NGS has overcome Sanger sequencing through at least three substantial improvements: first, the huge increasing of the throughput, thanks to the parallelization of sequencing reactions which allow the concurrent reading of millions of DNA fragments belonging to a single sample; second, the dramatic reduction of the sequencing cost, directly linked to the massive parallelization; and third, the drop of the required sequencing time, as bacterial cloning was replaced by library preparation and the output detection, performed cyclically and in parallel, has become direct, thus avoiding the electrophoresis step. On the other hand, the big number of produced reads and their relatively short length raised novel issues mainly regarding data analysis and data interpretation, constituting the pitfalls of NGS technologies [16–18].

The NGS revolution began in 2005 with the commercialization of Roche 454's pyrosequencing method, directly followed by the appearing on the market of the

Solexa/Illumina Genome Analyzer platform based on the sequencing-by-synthesis approach in 2006 and the Sequencing by Oligo Ligation Detection (SOLiD) system released by Applied Biosystems (now Life Technologies) in 2007. The last SGS method was proposed in 2010 by the Ion Torrent (now Life Technologies) through the sale of the Personal Genome Machine (PGM) based on the semiconductor sequencing [16–19].

The different kinds of platform have shown several innovations not only in the way in which the sample is sequenced, but also in how the templates should be prepared. The specific aggregation of the various protocols allows to discriminate one technology from another and it directly influences also the type of data produced. All the previously mentioned SGS platforms require an initial step of template preparation, where DNA is randomly broken into small sizes and common adaptor sequences are added to generate either mate-pair templates or fragment templates. This cell-free system has the advantage of avoiding the arbitrary loss of genomics portions, typical of cloning-based procedures [20, 21]. The templates are usually fixed or blocked on a support in order to be spatially separated, allowing in this way the simultaneous execution of a huge number, from thousands to billions, of sequencing reactions [21]. The sequencing is preceded by an amplification step where several copies of each template are generated, forming clusters: this passage is necessary as the majority of detectors have been designed to collect multiple fluorescent signals, but at the same time it permits a high signal magnification. Solid-phase amplification (i.e. bridge PCR, solid-phase PCR, asymmetric solid phase PCR) (Figure 1B) [22] and emulsion PCR (emPCR) (Figure 1A) [23] have been the two most chosen protocols for clonally amplified template preparation [21], even if other methods such as in situ polonies [24], in situ rolling circle amplification (RCA) [25, 26], and picotiter PCR [27] have been proposed [28].

Bridge PCR (Figure 1B) is the solid-phase amplification protocol integrated in its sequencers by Illumina, which is applied to create on a slide a series of randomly spread, clonally amplified clusters derived from both fragment and mate-pair templates. The slide is coated with a certain density of forward and reverse primers attached to the surface at their 5’ ends through a flexible linker. When a ssDNA template is added, it binds one kind of adaptor blocked on the surface, which is then elongated by polymerase creating a double-stranded DNA (dsDNA). The dsDNA is denatured and the original template is washed away. The remaining strand, which now contains on the top the second type of adapter, bends over

and hybridizes to a complementary oligonucleotide on the flow cell. The synthesis of the complementary strand creates a dsDNA bridge which results, after denaturation, in two single stranded copies of the molecule, that are tethered to the slide. The process is then repeated several times and it occurs concurrently for millions of clusters, triggering to the clonal amplification of all the fragments. Each cluster contains an average of 1000 copies of a single member of the library. To avoid overcrowding and to maximize the cluster density, it is necessary to accurately measure the concentration of the starting template library [16, 19–21, 29–31].

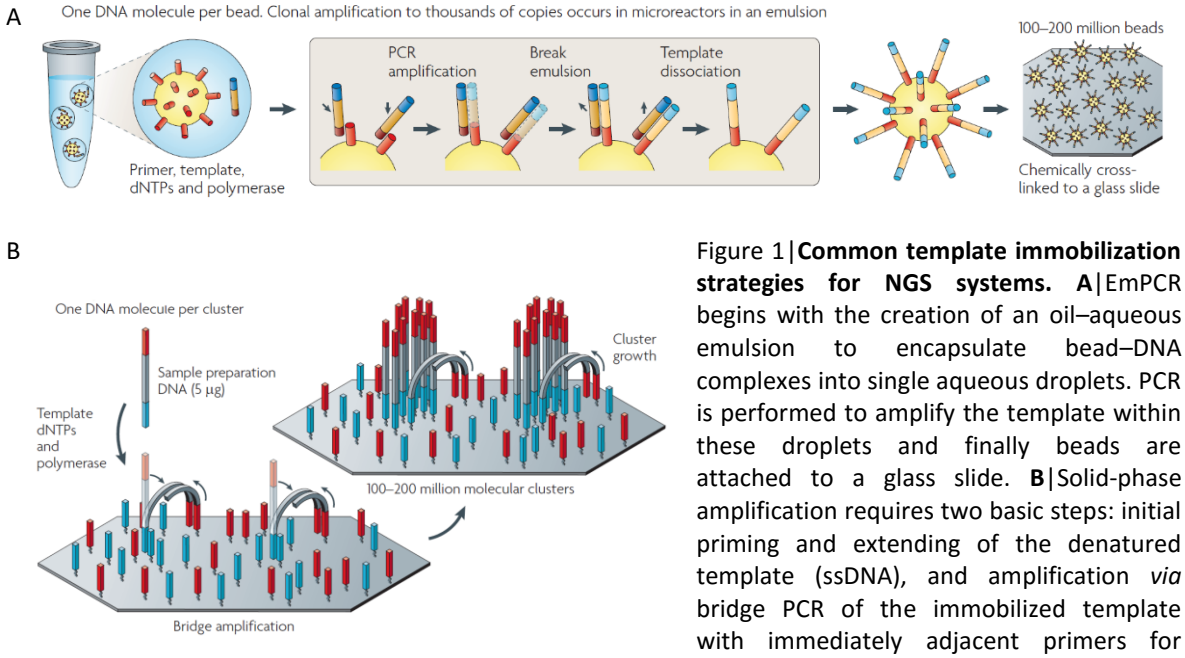


Figure 1|**Common template immobilization strategies for NGS systems. A**|EmPCR begins with the creation of an oil–aqueous emulsion to encapsulate bead–DNA complexes into single aqueous droplets. PCR is performed to amplify the template within these droplets and finally beads are attached to a glass slide. **B**|Solid-phase amplification requires two basic steps: initial priming and extending of the denatured template (ssDNA), and amplification *via* bridge PCR of the immobilized template with immediately adjacent primers for forming clusters. Figures adapted from [22].

On the other hand, emPCR (Figure 1A) is adopted by almost all other platforms, including SOLiD, Ion Torrent and Roche/454. In this method, NGS library is captured on micron-scale beads, tethering on the surface one of the PCR primers linked via the 5' end [20, 21, 32]. Theoretically, each bead should host only one fragment if the template concentration is correctly established. However, due to the usual low template concentration chosen, it is more frequent having the formation of unbounded beads, which then will be unproductive PCR reactors, rather than obtaining multi-bounded particles, associated with by-products production [33]. A water-in-oil emulsion including PCR reagents and one bead per droplet is generated to amplify each template individually, producing up to 10^7 copies per bead [31, 32]. The dsDNAs, formed after PCR, are denatured and the emulsion is then broken, in order to distribute each bead in a single well of a fiber-optic slide (Roche/454, Ion Torrent) or on a

glass surface (SOLiD). Each bead carries on its surface the amplification products deriving from only one fragment of the library [16, 20, 32, 34].

The NGS library preparation has surely overcome the cloning methods applied in FGS. However, it is not free from biases, which are mainly derived from the step involving the PCR process. To improve the yield of PCR, various adjustments have been performed including a deep assessment of the less noisy DNA polymerases and the best conditions at which they work [35]. Instead, to reduce the losing of sample, some specific protocol which integrates DNA fragmentation, end-polishing and adaptor-ligation in a single reaction, as in the Nextera technology [36], have been developed. These progresses directly influence the amount of DNA required: currently, few nanograms of starting material are enough for completing the whole sequencing process [37].

Although different approaches for sample preparation have been proposed, the main differences among the SGS platforms can be appreciated at the sequencing level. Here, I will describe only the four methods (Roche/454 pyrosequencing, Ion Torrent semiconductor sequencing, SOLiD sequencing-by-ligation and Illumina sequencing-by-synthesis) implement in the most widely spread sequencers (Figure 2); however, other methodologies such as the combinatorial probe anchor ligation (cPAL) sequencing, exploited by the Polonator G.007 of Complete Genomics [26], and the single molecule sequencing (SMS), adopted by the HeliScope of Helicos BioSciences [38], have found a remarkable interest not only in the scientific community, but also on the market [30].

Pyrosequencing technology (Figure 2A) is the methodology used by Roche/454 sequencers [32]. It is a non-electrophoretic, bioluminescence method. The DNA polymerase derived from *Bacillus stearothermophilus* and a single-stranded binding protein are preincubated together with the beads coated with the amplified template, produced after the emPCR. The solution is then deposited on a PicoTiterPlate (PTP) in order to have only a single bead within each well. Wells are also loaded with smaller beads bearing the other enzymes necessary for the reactions (ATP sulfurylase and luciferase) and with the remaining reagents, including primer, luciferin and adenosine 5’ phosphosulfate (APS). The PTP wells are exposed on the top to the flow of 2’-deoxyribonucleoside-triphosphate (dNTPs), while on the bottom they are linked to a fiber-optic bundle which is directly bound to a high-resolution charge-coupled device (CCD) camera.

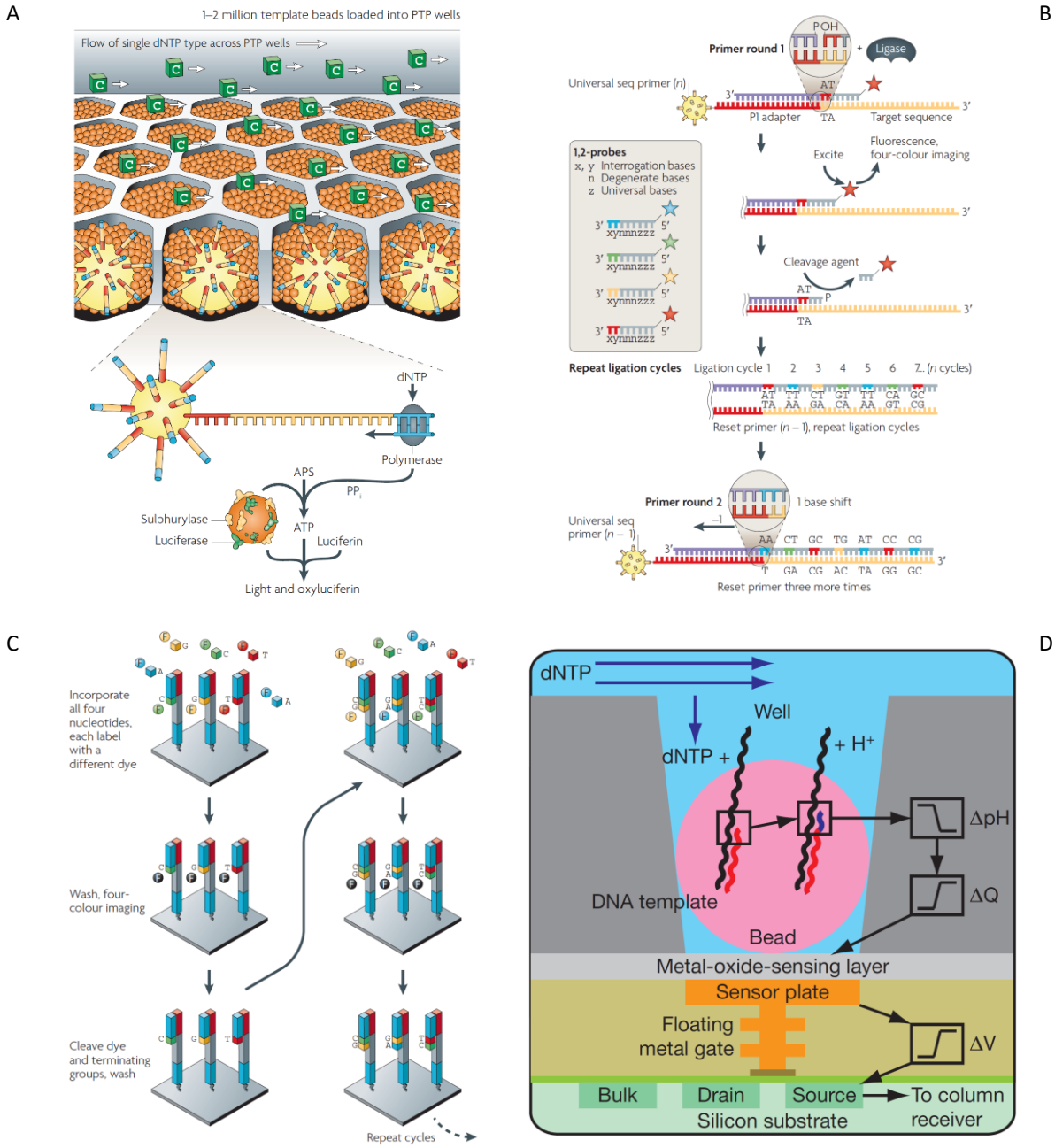


Figure 2|**NGS sequencing approaches. A**|Pyrosequencing used by Roche/454 platforms. The incorporation of a known dNTP causes the release of a PPi which is converted by sulfurylase into ATP, using APS. Luciferin consume ATP, producing a burst of light, detected by a CCD camera. The signal is proportional to number of dNTPs incorporated. As the dNTP added at each cycle is known, the template sequence can be established. **B**|Sequencing by ligation used by SOLiD platforms. Labeled octamers, whose fluorochrome changes depending on an internal dinucleotide (interrogation bases), are ligated to the primed-template by a DNA-ligase. Signal is detected and the last three nucleotides chemically cleaved. Other cycles are performed until the end of the sequence. The extended primer is then stripped and four more ligation rounds are performed, changing the initial reading position (n-1). **C**|Sequencing by synthesis used by Illumina platforms. Primers annealed to the immobilized templates are elongated using four different fluorescent reversible terminators. The propagation of polymerization is inhibited by the blocking element, which is removed and washed away together with the dye after the image acquisition, allowing the incorporation of a further nucleotide. **D**|Semiconductor sequencing used by Ion Torrent platforms. The method is really similar to pyrosequencing, but a changing in pH is detected by a transistor (pHFET), when a nucleotide is incorporated. Figures A, B, C adapted from [21], D from [39].

This apparatus permits the detection of the light emitted by the PTP wells after the pyrosequencing reaction [20, 21, 30]. The latter begins when a dNTP is incorporated on the complementary strand of the template, causing the release of an inorganic pyrophosphate (PPi). PPi is combined with APS by sulfurylase producing ATP, which is used by luciferase to convert luciferin in oxyluciferin, triggering to the emission of a burst of light [40]. The flash is detected by the CCD, recording the coordinate of the specific well and the intensity of the signal: the peak height will be proportional to the number of nucleotides incorporated [41]. The free dNTPs are degraded by apyrase and the by-products washed away. Then, another dNTP is added into the flow cell repeating the pyrosequencing reaction. As the nucleotide added at each cycle is known, the sequence of the template can be established [19, 40].

The semiconductor sequencing (Figure 2D) can be seen as an improvement of pyrosequencing, because the chemistry of the sequencing reaction is very similar, but the signal detected regards the release of a proton (H^+) instead of a pyrophosphate when a nucleotide is incorporated [16]. Nevertheless, semiconductor sequencing, proposed by Ion Torrent division of Life Technologies, differs from the other SGS platform as it has replaced the use of the complex optical systems based on CCD cameras with an ion-sensitive field effect transistor (ISFET), working as pH field effect transistor (pHFET), produced using standard complementary metal-oxide-semiconductor (CMOS) processes [34]. The introduction of such type of chip putting sequencing definitely into the framework of Moore's Law, to the extent that the Ion Torrent chips have been classified by the International Technology Roadmap for Semiconductors (ITRS) as "More-than-Moore", because they have proposed a "functional diversification" to the original applications of CMOS chips [39]. To have a briefly focus on the chemistry, the amplicon-bearing beads are distributed on 3.5 µm diameter wells, specifically designed to host at maximum a single bead. dsDNA is denatured, then primed and finally loaded with the DNA polymerase [34]. A first trial solution containing one type of dNTP flows over the wells. If the dNTP is complementary to the template sequence, it will be incorporated causing a release of a pyrophosphate and a H^+, which triggers to a pH changing ($\Delta pH = 0.02$). Such variation is detected by pHFET, converted to a voltage and digitized [39]. Otherwise, no pH modifications are collected when the nucleotide is not incorporated. After each trial flow, unbound dNTPs and by-products are washed away. A different dNTP is then made flowing and the process is repeated until all the templates are sequenced. When homopolymeric

(HP) tracts are present, more than one nucleotide is incorporated by polymerase, causing an higher release of H^+ ions. The associated amplitude of the incorporation signal is applied to evaluate the length of HP region [34]. As for the pyrosequencing, this kind of estimation constitutes the greater source of error, even if other biases related to the sequencing of AT-rich genomes have been reported [42]. The magnitude of these problems seems to be reduced after the introduction of new chips and the improvements of PCR chemistry and base-calling software.

Another method preceded by emPCR is the sequencing-by-ligation (Figure 2B) developed for the SOLiD platforms [28, 43]. It is driven by a DNA-ligase, rather than the polymerase [44]. The first step of the ligation-based sequencing corresponds to the hybridization of a universal primer, exposing a free 5'-PO_4, to the SOLID-specific adaptors, which are captured on the surface of paramagnetic beads and linked to the template molecules [20, 45]. Each cycle of the process involves the ligation of a degenerate octamers population, which contains a ligation site (first nucleotide), a cleavage site (phosphorothiolate linkage between fifth and sixth nucleotides), a fluorescent dye attached to the eighth nucleotide, and inosines in the last three positions to reduce the probe complexity [21]. The color of the dye depends on the sequence of a specific dinucleotides (interrogation bases) which can correspond to the first two nucleotides or nucleotides in position 4 and 5, depending on the cycle number [45]. After the ligation of the proper probe (interrogation bases must be complementary to template) and the signal detection, a cleavage step mediated by silver ions [43, 45] removes the last three nucleotides and consequently also the fluorophore, enabling a subsequent round of ligation. The process of probe ligation is repeated more times in order to reach the end of the template and it triggers to the reading of a certain combination of dinucleotides. The synthesized reads are removed by denaturation and washed away. A further lap of sequencing starts with the hybridization of a second universal primer in position n-1, followed by another cycle of probes ligation. Five ligation rounds allow to complete the reading of the whole template. Although the method is quite slow, the accuracy is really high, as each base of the template is read twice (*2 base encoding*), allowing an easier identification of miscalling [20, 45].

The last SGS technology discussed in this section is the sequencing-by-synthesis (Figure 2C) approach implemented by Illumina, which is based on the four-color cyclic reversible terminator (CRT) chemistry [29]. After bridge PCR, the several million clusters, each

containing ~1000 copies of the same fragment, represent both the forward and the reverse strand of the template. To prevent issues derived from steric hindrance or from unintended complementary base pairing, one of the two strands (usually reverse) is removed exploiting the different adaptor by which the templates are kept on the slide [30]. Linearized amplicons are primed exploiting the adaptor located at the top of the template and then loaded with a mutant DNA polymerase, able to incorporate modified nucleotides. The latter are called 3'-blocked reversible terminators, because their 3'-OH end is protected by a blocking element [46], as for example the 3'-O-allyl [47] or the 3'-O-azydomethyl [29, 48], linked via an etheric bond. This group inhibits the propagation of the polymerization, but at the same time it should be easily cleaved by chemical agents, after the signal detection. Reversible terminators present also a fluorescence dye connected to the base. Such modification, in addition to the possibility to be removed, does not have to hugely modify the base structure, in order to keep its capability to be recognized by DNA polymerase [46]. Sequencing process begins when a mixture of four labelled nucleotides flows over the slide. Primers bound to the template are extended by only a single-base, thanks to the presence of termination group, while the remaining unincorporated nucleotides are washed away. The image acquisition is performed using a CCD camera and two lasers, which are able to excite couples of nucleotides (A/C and G/T), allowing the correct base recognition through a set of optical filters, working on emission spectra. After imaging, both groups, the terminator and the dye, are chemically cleaved. Finally, a further step of washing is carried out, before starting another CRT cycle. The number of cycles depends on the desired read length, which depends on the way in which the library was built [20, 21, 29, 31]. When the first strand is completed, few bridge PCR cycles are performed for second strand synthesis (usually reverse). The first strand is then removed and the second strand is sequenced. This approach is called "paired-end" sequencing and it is widely used also by other technologies, even if it is extensively applied by Illumina sequencers [30].

1.1.3. Third Generation Sequencing

Nowadays, the SGS technologies are still dominating the sequencing market, even if in the last decade a variety of new systems overcoming in many aspects the NGS methods have been proposed [49]. This new category of platforms has been classified as third generation

sequencing (TGS) because they promised higher throughput, shorter run time (real time detection), smaller quantities of starting material (theoretically a single molecule), longer reads (several Kb), higher consensus accuracy for rare variant discovery and, finally, lower cost, reaching the goal of a high coverage genome for less than $100 [50].

Currently, none method has achieved all these targets. In addition, for a couple of platforms, it is even difficult to understand if they should be included into SGS or into TGS group. Among these transitioning technologies, Ion Torrent and HeliScope devices are the two main examples [50]. In fact, the first one has removed the need of scanning system thanks to the introduction of the CMOS pHFET chip; otherwise it still requires PCR for the initial amplification and the washing of byproducts between one cycle and the following [34]. The HeliScope instead has been the first system based on SMS, but the use of virtual terminators nucleotides, whose dyes have to be cleaved after the signal detection, heavily influences the speed of sequencing [38].

Nevertheless, in the TGS technologies, all such platforms able to perform fast SMS in real-time, could be included. They can be divided into three approximate categories [50]: 1) direct observation of single DNA molecules through cutting edge microscopy systems, as for example the aberration-corrected scanning transmission electron microscopy (TEM) [51] or the scanning tunneling microscope (STM) [52]; 2) single molecule real-time SBS, where the processing of DNA polymerase is detected, exploiting the zero-mode waveguide (ZMW) technology [53] or the fluorescence resonance energy transfer (FRET) [21, 54]; 3) nanopore sequencing [49].

The two most successful TGS methods can be found in the last two groups, corresponding to the single-molecule real-time (SMRT) sequencing, developed by Pacific Biosciences (PacBio) and the nanopore sequencing, proposed by Oxford Nanopore Technologies (ONT). Both systems still require a library preparation which is usually quite simple to prepare and not time-consuming. The PacBio template is called SMRTbell (Figure 3A) and it consists in a single stranded circular DNA, produced by ligating hairpin adaptors, which provide the primer binding site, to both ends of a dsDNA fragment [55]. The nanopore template preparation (Figure 3B), instead, requires a common genomic DNA fragmentation, followed by end-repairing and dA-tailing to add an adenosine to 3’ end of the fragment. Then two adaptors are ligated: the leader adaptor, also referred as “Y adaptor” for its “Y” shape, where the sequencing process begins, and the “HP adaptor” with a hairpin-like structure,

binding site of the hairpin protein. Finally, a purification step using His-beads is performed for removing nucleotides and enzymes. To allow a 2D base calling, before loading the library is briefly incubates (30 min) with the motor protein and the HP [56].

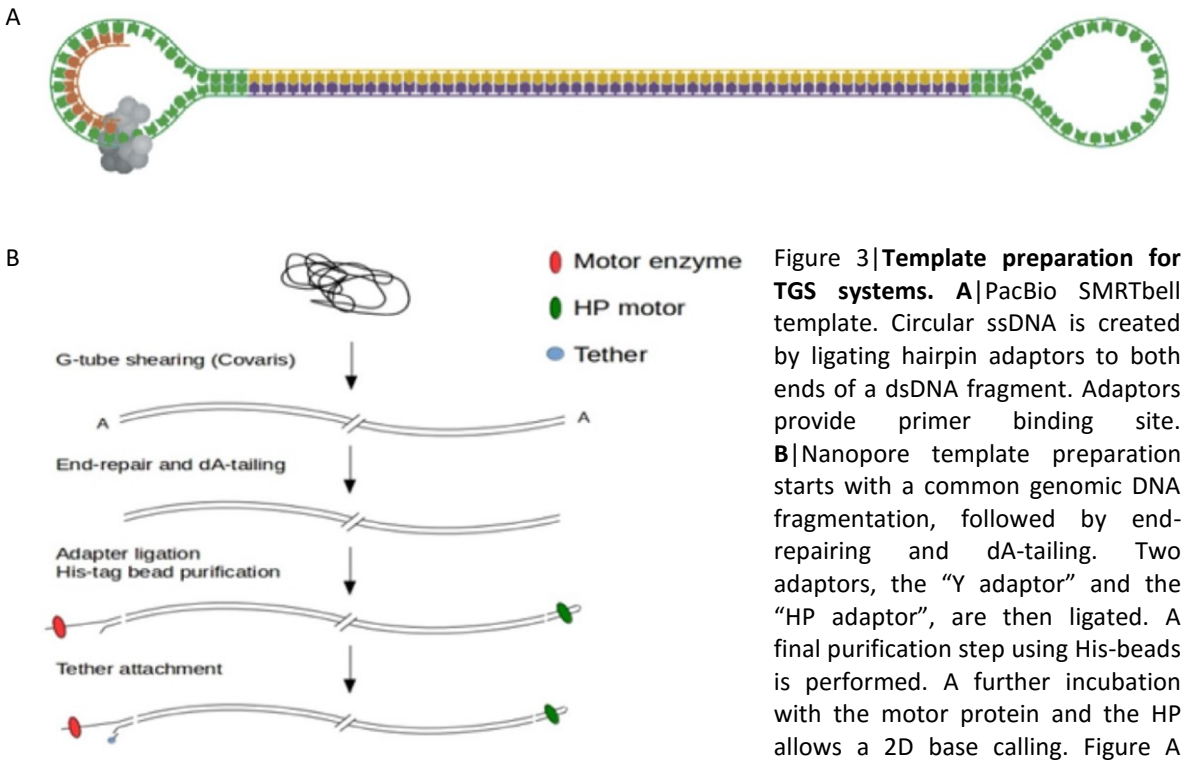


Figure 3|**Template preparation for TGS systems. A**|PacBio SMRTbell template. Circular ssDNA is created by ligating hairpin adaptors to both ends of a dsDNA fragment. Adaptors provide primer binding site. **B**|Nanopore template preparation starts with a common genomic DNA fragmentation, followed by end-repairing and dA-tailing. Two adaptors, the "Y adaptor" and the "HP adaptor", are then ligated. A final purification step using His-beads is performed. A further incubation with the motor protein and the HP allows a 2D base calling. Figure A adapted from [57], B from [56].

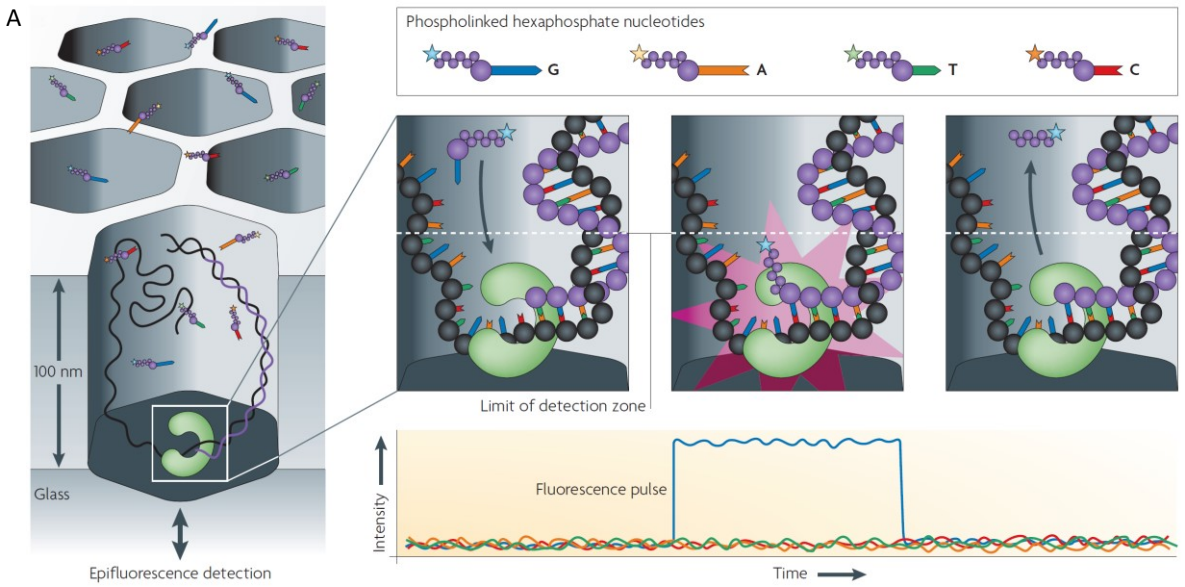
In the PacBio SMRT sequencing (Figure 4A), when the SMRTbell library is ready, it is loaded into a chip, named SMRT cell. Each SMRTbell diffuses into a nanophotonic structure, the zero-mode waveguide (ZMW), which constitutes the smallest available volume for light detection [57]. On the bottom of each ZMW, a single molecule of engineered Φ29 DNA polymerase, chosen for its favorable properties [58], is bound using streptavidin/biotin interaction. When all polymerases are loaded with primed SMRTbell, a mixture of four different colored γ-labeled phospho-nucleotides, generating distinctive spectra, are flowing within the SMRT cell [53]. The structure of ZMW does not allow laser light penetration along the hole, avoiding the excitation of labeled nucleotides during their migration, unless it diffuses through the bottom 30 nm of the ZMW. Nevertheless, as the diffusion is really fast (μs), the emitted fluorescence results in a signal background [21]. Instead, when the right nucleotide is incorporated by DNA polymerase, it is blocked over the laser light for a longer time (ms), creating a high *signal-to-noise* ratio, which is distinctly detected as a pulse. The released labeled byproduct rapidly diffuses far [21, 50]. As soon as the next cognate

nucleotide reaches the polymerase binding site, the subsequent pulse is recorded. Thanks to the SMRTbell circular structure and the high processivity of polymerase, both strands of the fragment can be sequenced several times, producing a continuous long read (CLR), which is finally splitted in subreads after the adaptors recognition [57].

Differently from PacBio, Oxford Nanopore technology does not apply any imaging systems to detect nucleotide incorporation. This process is revealed by measuring a change in the ion current when DNA passes through a pore (Figure 4B) [59]. The system is constituted by a lipid double layer, necessary to separate the different concentrations of KCl between the two sides of the membrane, where an engineered staphylococcal α-hemolysin protein pore is also inserted. The protein pore modifications not only regard some amino acidic substitutions in the heptamers $(M113R/N139Q)_6(M113R/N139Q/L135C)_1$, but also the covalent ligation within the barrel structure of a complex β-cyclodextrin ($am_6amPDP_1\beta CD$), working as sensor for base detection [60]. The sequencing process starts when the motor enzyme, usually a polymerase or a helicase, loaded with the DNA is attracted towards the pore by the application of a voltage on the two Ag/AgCl electrodes [61]. The motor enzyme, Φ29 DNA polymerase, provides multiple functions, as in addition to the DNA carrier activity it is also able to separate the two strands of the loaded DNA, regulating in this way the speed of DNA translocation through the pore. The latter feature is fundamental, because if the zip opening is too fast, the nucleotide reading could not be performed [62]. Once reached the top of the pore, the motor enzyme begins to unzip the dsDNA, causing the DNA translocation into the pore. When the nucleotide approaches the sensor, a perturbation in the current is measured, whose pattern and magnitude are used to decodify the base. Thus, the data streams are elapsed to a microchip named the application-specific integrated circuit (ASIC) and finally processed by the MinKNOW software [56, 63]. Thanks to the introduction of HP into the protocol, the complementary strand can also be sequenced, triggering to a 2D base calling. In this way, the information of both strands is included allowing a higher base quality [56].

Since the beginning the nanopore technology raised enthusiastic interest in the scientific community. For this reason, a lot of work has been recently done in order to exploit new kind of pores, not only biological (α-hemolysin [60], MspA [64], Φ29 [65]), but also solid-state nanopores or nanogap electrodes [66]. In particular, the solid-state pores can be manufactured in a wide range of shape and size, offering also the compatibility with CMOS

technology and thus a great cost reduction, due to the industrial scalability [67]. So, the significant improvements achieved have greatly simplified the experimental process, allowing the diffusion of these new nanopore technologies in many fields of DNA sequencing, creating a perspective for a rapid and low-cost fourth-generation DNA sequencing methods [66].



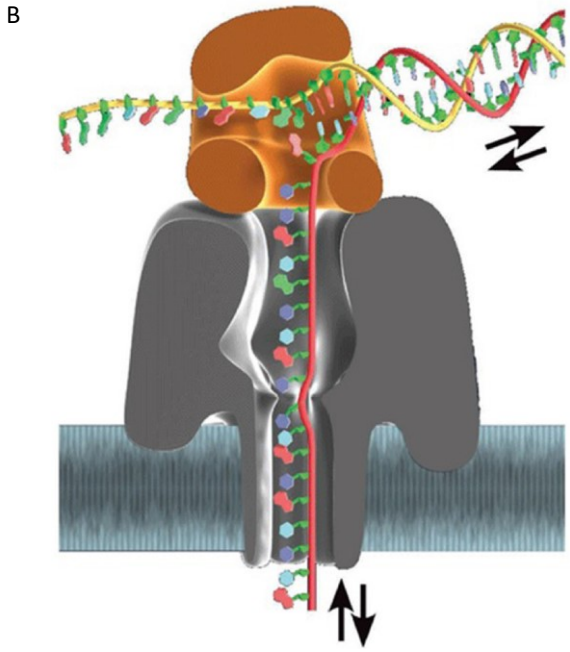


Figure 4|**TGS sequencing approaches. A**|SMRT sequencing used by PacBio platforms. The reduced observation volume of ZMW allows the entrance of a small number of stray fluorescently labelled molecules, limiting the background signal. When a modified nucleotide is incorporated by the anchored DNA polymerase, the dye remains near the bottom of the well for a longer time, causing the detection of a fluorescence pulse. The signal type is used for distinguishing the various nucleotides. **B**|Nanopore sequencing used by Oxford Nanopore platforms. When the motor enzyme reclines on the top of the pore, it starts to denature the dsDNA, causing the entrance of a strand into the hole. The perturbation of the ion current, whose pattern and magnitude are used to decodify the base, is detected by a sensor within the pore and is transmitted to the ASIC microchip. Figure A adapted from [21], B from [49].

Platform	Generation	Library	Template preparation	Chemistry	Time/run	Read Length	Read Number	Output (Gb)	Output Type	Pros	Cons	Biological Application	Release Date
Sanger 3730xl	1	Frag	PCR, cloning	Dideoxy chain terminator	20 m - 3 h	400 bp - 900 bp	-	1.9 - 84 Kb	Chromatogram	High quality, long reads, low cost for very small studies	Low throughput, very high cost for large data amount	Mutation detection, NGS validation	2002
Roche/ 454's GS FLX Titanium	2	Frag, MP	emulsion PCR	Pyrosequencing	24 h	400 bp up to 1000 bp	1 M	0.7	Standard Flowgram Format (SFF)	Longer reads than other SGS	High error rate, high cost of reagents, low throughput, no support since 2016	Small genomes assembly	2009
SOLiD 5500xl	2	Frag, MP, PE	emulsion PCR	Sequencing by Ligation	1 w for frag, 2-3 w for MP and PE	1 x 75 bp Frag, 2 x 50 bp for MP & PE	1.2-1.4 G	~160	eXtensible SeQuence (XSQ)	High accuracy due to double reading of each base, high throughput	Short reads, long run times, almost retired	small WGS, WES,	2010
Illumina MiSeq	2	PE	solid-phase bridge-PCR	Sequencing by synthesis	4-55 h	2 x 300 bp	25 M	15	Binary Base Call (BCL)	Moderate cost instrument and per Mb, fast, longest Illumina read lengths	Relatively few reads and Higher cost per Mb compared to NextSEq or HiSeq	Targeted sequencing, 16S metagenomics, small WGS	2011
Illumina NextSeq	2	PE	solid-phase bridge-PCR	Sequencing by synthesis	12-30 h	2 x 150 bp	400 M	120	Binary Base Call (BCL)	Easy to use, moderate instrument and run costs	Version 2 of new chemistry not yet as good as older chemistry	small WGS, WES, WTS	2014
Illumina HiSeq 4000	2	PE	solid-phase bridge-PCR	Sequencing by synthesis	24 – 84 h	2 x 150 bp	5 G	1500	Binary Base Call (BCL)	Low cost per read and per MB	High instrument cost, high cost per run, requires highly trained personnel	WES, WTS	2015

Illumina NovaSeq 6000	2	PE	solid-phase bridge-PCR	Sequencing by synthesis	19 – 40 h	2 x 150 bp	20 G	6000	Binary Base Call (BCL)	Most powerful Illumina sequencer, very high throughput, current lowest cost per read or per MB	High instrument cost, high cost per run, requires highly trained personnel, high data storage	Large WGS, WES, WTS, methylation sequencing	2017
Ion Torrent PGM	2	Frag	emulsion PCR	Semiconductor sequencing	4.4 h / 7.3 h	200 bp / 400 bp	4-5.5 M	0.6 - 1 / 1.2 - 2	Binary sequence Alignment Map (BAM)	Low cost instrument, three chips available, very simple machine	More hands-on time and fewer reads at higher cost per Mb relative to MiSeq	Targeted sequencing,16S metagenomics, small WGS	2010
Ion Proton	2	Frag	emulsion PCR	Semiconductor sequencing	2 - 4 h	200 bp	60-80 M	10	Binary sequence Alignment Map (BAM)	Moderate cost instrument for medium throughput applications, fast run times	More hands-on time and fewer overall bases of data than Illumina, smaller user community	small WGS, WES, WTS	2012
Ion S5	2	Frag	emulsion PCR	Semiconductor sequencing	2.5 h / 4 h / 4 h	200 bp / 400 bp / 600 bp /	60-80 M	3 - 4 / 6 - 8 / 1.5 - 4.5	Binary sequence Alignment Map (BAM)	Cost-competitive vs. MiSeq and potentially NextSeq, fast run times	More hands-on time and fewer reads at higher cost per Mb relative to MiSeq	small WGS, WES, WTS	2015
Oxford MinIon	3	Frag	Single molecule	Real Time Sequencing	1 m - 48 h	10 kb up to 1 Mb	4.4 M	40	FAST5	Portable instrument, USB device, extremely low-cost instrument, extremely long reads	High cost per read, quite high error rate	*de novo* genome assemblies, WTS, structural variations discovery	2015

Oxford PromethION (48 Flow Cells)	3	Frag	Single molecule	Real Time Sequencing	1 m - 48 h	10 kb up to 1 Mb	1.25 G	11000	FAST5	Very high throughput, extremely long reads	Similar to MinION, available only joining the PromethION Early Access Programme	Large WGS, WES, WTS, epigenetics, CNV, structural variations discovery	2017
PacBio RS II (16 SMRT cells)	3	Frag	Single molecule	Real Time Sequencing	8 h - 96 h	> 20 kb	0.9 M	16	HDF5	Extremely long reads, ability to detect base modifications, short run time, random error profile	High error rates at high coverage, low total reads number, high cost per Mb, high capital cost	de novo assemblies, WGS, WTS, haplotype detection, methylation profiles	2013
PacBio Sequel (16 SMRT cells)	3	Frag	Single molecule	Real Time Sequencing	8 h - 160 h	> 20 kb	5.84 M	120	HDF5	Similar to RSII but with lower cost	High cost per Mb and per read relative to other platforms	de novo assemblies, WGS, WTS, epigenetic detection	2015

Table 1|**Sequencing platforms comparison.** Technical information, advantages, drawbacks and biological applications of the most diffused sequencing platforms are reported. Frag: fragment; MP: mate-pair; PE: paired-end.

1.1.4. Application and comparison of high-throughput sequencing technologies

Clear differences among the various series of the sequencing technologies are appreciable, in particular regarding costs, limitations and advantages [20]. Considering these three features is necessary for designing the experimental flow in order to choose the best sequencer for achieving the planned objectives. For this purpose, a detailed comparison among the several platforms now available on the market or which have covered an important role in the NGS revolution is reported in Table 1. For example, the Sanger sequencing is yet confined to small projects, where few kilobases have to be analyzed, or thanks to its high accuracy it can be applied to mutation detection on a specific gene or more frequently to validate a variation found using NGS methodologies. The latter can be used for a various range of applications, although the heavy differences regarding throughput and read number between small benchtop platforms (Illumina MiSeq or Ion Torrent PGM) and the most powerful sequencers (Illumina NovaSeq and Illumina HiSeq) directly influence the main usage of the machinery. The NGS platforms have found use in variant discovery, metagenomics, transcriptomics and small RNA analysis, epigenetics and chromatin immunoprecipitation sequencing (ChIP-Seq) and finally small genome assembly [21]. In particular, the genome resequencing for the comprehensive polymorphisms and variant discovery in human genomes has been probably the most successful NGS application, due to its direct implication on the study of genetic diseases. At the same time in these projects, also the human DNA variability was evaluated, triggering to the development of population allele frequency databases, such as ExAC [68], ESP6500 [69] or dbSNP [70], exploited by clinicians and geneticists to understand if mutations under analysis are rare or not.

Remaining in the field of human genetic variation detection, it is important to notice that although the sequencing cost has fallen very fast in the last decade, when high coverages are required, the whole genome sequencing (WGS) could still be too expensive [71]. Thus, as the majority of Mendelian diseases are associated with mutations in the coding regions [72], in many cases WGS can be replaced by the whole exome sequencing (WES), where only exons are analyzed. This method is cheaper than WGS, as the region covered is near to 1.5% of the entire genome [18], but it requires an intermediate step where exons are

extracted from the whole DNA sequence, referred as target enrichment. A lot of work has been performed to optimize the enrichment process, for achieving the best values of sensitivity, specificity, uniformity, reproducibility and design coverage [73]. The nature of the mechanism involved in the process is used to separate the current techniques, dividing them into three groups [74]: I) *hybrid capture*, where the sample DNA is hybridized with specific probes, complementary to the target region, either in solution or on a solid support, allowing the physical capturing and the following purification of the sequences of interest [75]; II) *molecular inversion probes* (MIP) or *selective circularization*, where universal ssDNA molecules incorporating on their edges segments complementary to the boundaries of the target regions are firstly hybridized against sample DNA, then gap-filled and finally ligated, resulting in circularized DNA molecules containing the target sequence in addition to other features useful for downstream analyses [76]; III) *PCR amplification*, where a huge number of couple of primers (up to 20000) is designed to amplify specific target regions through a highly multiplexed PCR [34].

The same enrichment techniques can be applied for targeted sequencing, where a limited set of genes (gene panel) or genomic regions are taken into account. This approach guarantees very high coverages (>1000X) at affordable costs and should be the first choice for studying disorders in which a middle heterogeneity is recognizable [77]. Gene panels for analyzing variants in patients affected by cardiomyopathy [78, 79], lysosomal storage disease [80] and epilepsy [81, 82] are only a few examples of the medical fields in which targeted sequencing has been employed [83].

The platforms eligible for resequencing projects depend on the amount of data required. Surely, for targeted sequencing or WES, all Life Technologies platforms, the Illumina benchtop apparatus (MiSeq and NextSeq) and the SOLiD system can be used, while the Illumina HiSeq or NovaSeq sequencers are mostly prone for WGS [19]. Nevertheless, the performance in variant discovery among the various technologies is quite comparable [42], with a more precise single nucleotide polymorphism (SNP) calling for Ion Torrent platforms, at the expense of a higher rate of error in homopolymeric segment [34], where SOLiD or Illumina systems seems to keep a very high accuracy [84]. On the other hand, due to their high error rate, TGS methods are not prone for discovering short nucleotide variants (SNVs). However, exploiting the very long reads produced, they are really suitable for detecting full-length gene isoforms, which instead are scarcely found by SGS, and structural variations

[57], such as copy-number variations (CNVs), inversion, translocation and segmental uniparental disomy [85], whose characterization is really crucial, because they cover an average of 18.4 Mbp in a diploid genome [86].

1.2. NGS data analysis: from raw reads to variant prioritization

Since 2005, the introduction of the high throughput technologies caused a huge decreasing of sequencing cost leading to the breaking of the $1000 threshold for an entire genome. Anyway, if it is now relatively easy to sequence the genome of a patient, it is instead really troublesome giving the correct meaning to the obtained data. Other challenges derived from the needing of working with large datasets, as terabytes of data are produced every day from big sequencing centers. In this way, also biologists have now joined the big-data club, encountering issues with handling, processing and transferring all such information [87]. For this reason, if during the FGS years the main bottleneck was constituted by sequencing, now it is moved downstream in the pipeline, as the most time wasteful and expensive step has surely become the data analysis [88].

As described in the previous paragraphs, the applications of NGS are very widespread, and although some parts are shared between the various data analysis pipelines, it is extremely long giving a plenty information for each of them. For this reason, as the whole thesis in centered on different types of variant analysis and manipulation, this section of introduction will be dedicated to the description of the main steps involved in this process (read preprocessing, alignment, variant calling, variant annotation and prioritization) and the associated data formats.

1.2.1. Read preprocessing and FASTQ format

Table 1 shows the various data format used by sequencers to provide their outputs. Although quite different, almost all these formats can be converted to a simple extension of FASTA, called FASTQ, that has become the widest used interchange read file format, thanks to its simplicity [89]. An exception is constituted by Ion Torrent which directly returns the reads aligned on the reference genome, rather than raw reads. The FASTQ format, as frequently happened in biology, has suffered of a lack of formal definition, starting from a Sanger FASTQ, passing through different kinds of Solexa/Illumina format, and reaching the

current solution, which is almost the same of the first proposed. Thus, the quality of each base called is encoded by a PHRED quality score from 0 to 93 using ASCII characters from 33 to 126 [89]. The PHRED quality score is defined as the estimated probability of error:

$$Q_{PHRED} = 10 \times \log_{10}(P_e)$$

FASTQ format describes each read using four rows (Box 1). The first line begins with the @ character followed by the read identifier and an optional description. The second row displays the raw sequence of the read. The third line starts with *+* character and can optionally contain again the read information. The last row contains the quality values of the bases and must have the same length of the sequence expressed in line 2.

Box 1|**Example of FASTQ format.**	
1) Identifier	@Lactobacillus_493302_495389_1_1_0_0_0:3:0_0:3:1_12cb59/1
2) Raw Sequence	ATAGCAGTTTAAGCACGGTTCTTGTTCTGAAAAAGTTTGCGATAG
3) Optional Info	+
4) Quality	?:8755433221110000////.....--------,,,,,,,,,

Before being processed, FASTQ data are usually briefly analyzed to understand if reads can be used for the following analysis. For this reason, several tools for viewing, manipulating and summarizing FASTQ data before the pre-processing step have been developed over the years. Among them, the most widely used are fastx-toolkit [90], fastq-tools [91], seqtk [92] and the newer fqtools [93], efficiently designed for working also with the very long TGS reads.

Once FASTQ files have passed this sort of validation, they undergo to a set of processes which are commonly referred as the preprocessing step [94]. In this phase, read adaptors or primer portions kept within the sequence are recognized and thus removed [95, 96]. Then low-quality bases are usually trimmed out depending on the threshold chosen by the user [97, 98] and so the resulting reads can be discarded if too short [99]. Other programs based on various approaches, including maximum expected error (MEE) [100] or overlapping analysis [101], have also been implemented for a complete preprocessing step. Moreover, at the end of the preprocessing, a summary of read statistics (mean length, mean quality, GC content, *etc.*) is usually calculated using software like FastQC [102]. Concluding, some protocols such as the best practices of Broad Institute [103] suggest converting the final

FASTQ files into an unmapped BAM (uBAM) as it can store more information including all metadata.

1.2.2. Read alignment and SAM format

The read alignment is probably the most complex process at computational level in a common pipeline for variant detection. Considering a read of length *m* and a genome of length n, a common dynamic programming algorithm solves the approximate string matching with indels using a quantity of memory and time that grows proportionally to $m \times n$. Filtering and indexing are two different ways proposed to address the problem of large input sizes, regarding both the huge number of reads and the length of complex genome (3.2 Gbp for a human) [104].

In the filtering methods, wide reference regions, where no approximate match can be found, are rapidly excluded. Several approaches exploiting filtration have been developed, but almost all of them are based on a hash table exploiting the *q*-gram index (or *k*-mers index). These seeding-based methods identify short portions of the reference, called *k*-mers, which constitute the keys of the hash table, sharing a small piece of the read with no error (seed). The seed is then extended allowing or not the presence of gap (spaced seed) [105]. The remaining regions, where no association is found, are filtered out. To choose which regions of the reference can be rejected as not presenting any approximate match, the applied filters are usually based on two main principles: the pigeonhole [106] or the *q*-gram [107]. Both establish a minimal length or number of perfect *q*-grams that a read with a definite count of mismatches shares with the reference sequence [104].

In the easiest form of the pigeonhole lemma, assuming that we want to find all approximate matches of a read with at most k errors, if the read is divided in $k + 1$ seeds, at least one piece appears without error [108]. Thus, for each read, $k + 1$ non-overlapping seeds of almost the same length are obtained applying a pigeonhole filter. Their matches are then found in parallel by scanning the hash table of the genome. Widely spread read mappers are based on different variants of the pigeonhole lemma, such as SeqMap [109], SOAP [110] and MAQ [111], where reads are broken into more seeds (e.g. $k + 2$) and at least two matches seeds are searched, or on the other hand splitting reads into fewer pieces, thus allowing a fews errors, including Bowtie2 [112] and Masai [113].

A potential issue associated with these seed-based approaches regards lack of gaps within the seeds. To overcome this problem, other software including SHRiMP [114] and RazerS [115] adopted the q-gram filter which supports an index that natively permits gaps. The *q*-gram principle takes into account all overlapping *q*-grams (substrings of length *q*) of a read and provides a bottom limit for the number of substrings that a *k*-error match in the reference shares with the read. With these premises, a *q*-gram counting filter looks for genomic regions where at least a lower bound number of *q*-grams of a read can be detected, returning them as a possible match. The remaining regions can be securely eliminated [104].

Moreover, in some new aligners as mrsFAST [116], mrsFAST-ultra [117] and Masai [113], the indices are made on the genome and the reads, which are then scanned in parallel for finding matches with the same *q*-gram.

More recent software, instead of using filters and the q-gram index, implement a powerful data structure, referred as suffix tree, which straight searches reads with gap or mismatches. This approach is advantageous as multiple identical copies of a substring collapse into a single path in the tree and so they are aligned only once, while each copy should be singularly aligned if a common hash table is used [105]. The suffix tree of a certain string is made by all suffixes of such text and a linear time is required to build it. All occurrences of a read can be looked for in optimal time. In the native definition proposed by Weiner in 1973 [118], the suffix tree is a rooted tree, where from each internal node depart at least two edges, labeled with different substrings, deriving from the original text. The external leaves contain the starting index of the associated suffix. As the suffix tree requires a lot of space and it does not allow an efficient use of the cache memory, it has been firstly replaced by the suffix array [119], which still suffered of low performance in searching time, and then by two other methods, the enhanced suffix array [120] and the Ferragina-Manzini index (FM-index) [121].

All these approaches have been implemented into published read mappers. In particular, the suffix tree is used in MUMmer [122] and OASIS [123], Segemehl [124] and Vmatch [125] exploit the enhanced suffix array, while BWA [126], SOAP2 [127] and Bowtie [128] implement the FM-index.

Column	Field	Type	Example	Description
1	QNAME	Str	Lactobacillus_3072417	Query template NAME
2	FLAG	Int	4, 67, 163	Bitwise FLAG
3	RNAME	Str	chr1, 1	Reference sequence NAME
4	POS	Int	3070108, 1274308	1-based leftmost mapping POSition
5	MAPQ	Int	0, 22, 40	MAPping Quality [$-10 \log_{10} P_e$*]
6	CIGAR	Str	75M2D25M, 100M	CIGAR string
7	RNEXT	Str	=, *	Ref. name of the mate/next read
8	PNEXT	Int	3072347, 1334817	Position of the mate/next read
9	TLEN	Int	-2240, 1941	Observed Template LENgth
10	SEQ	Str	GCCGTTTATTTCTCA	Segment SEQuence
11	QUAL	Str	?:BA87554332211	ASCII of Phred-scaled base QUALity+33

Table 2|**Overview of the mandatory fields in the SAM format.** Each column is associated with an explanatory example. * P_e indicates the probability that mapping position is wrong.

Although the read aligners are based on different methods, the final output follows the specifications indicated by the Sequence Alignment/Map (SAM) format [129]. SAM is a general tab-delimited alignment format capable of storing information regarding aligned and not aligned reads with a maximum length of 128 Mbp independently by the sequencing platform used. It usually contains a header block before the alignments, where each line starts with a @ character, followed by two letters representing the record types, such as the header line (*HD*), the reference sequence dictionary (*SQ*), the read group (*RG*), the program used (*PG*) and the one-line text comment (*CO*). Each record has also predefined tags: for example, the format version (*VN*), the sorting order of alignments (*SO*) and the grouping of alignments (*GO*) for the HD type. Below the header section, the various alignments are listed, each of them contains 11 mandatory columns, as explained in Table 2. Almost all fields are self-explanatory. However, FLAG and CIGAR are quite complex and need a more exhaustive description. For the CIGAR, I remand to the SAM format documentation [130], while the FLAG is briefly explained in Box 2 using information collected in Table 3.

Bit	Binary	Meaning
1	000000000001	Template having multiple segments in sequencing → it is a mate
2	000000000010	Each segment properly aligned according to the aligner
4	000000000100	Segment unmapped
8	000000001000	Next segment in the template unmapped
16	000000010000	SEQ being reverse complemented
32	000000100000	SEQ of the next segment in the template being reverse complemented
64	000001000000	The first segment in the template
128	000010000000	The last segment in the template
256	000100000000	Secondary alignment
512	001000000000	Not passing filters, such as platform/vendor quality controls
1024	010000000000	PCR or optical duplicate
2048	100000000000	Supplementary alignment

Table 3|**Explanation of bits used for composing the FLAG field.**

Box 2|**Example of decodification of the FLAG meaning.**

163 = 000010100011

- 000000000001 → it is a mate (1)
- 000000000010 → both mates are properly aligned (2)
- 000000100000 → its mate is reverse (32)
- 000010000000 → it is the second mate (128)

The SAM format can accept also optional fields, where the mapper shall provide more information regarding the alignment. Such column can often store also read features, when a uBAM is produced. Nevertheless, independently from what is contained, optional columns must follow the TAG:TYPE:VALUE format, where TAG is a two character string, while TYPE is a single case-sensitive letter (A, Z, H, B, i, f) which defines the format of VALUE.

The compressed form of SAM format is called BAM format, where B means binary. In fact, due to the large dimensions of the SAM files, they are almost always converted into BAM format, which promises a size reduction of 4-5 times, keeping the compatibility with the majority of post-mapping programs. Differently from SAM, BAM format contains 0-based coordinates and the letters used in CIGAR string are converted into number. Surely, other differences can be identified, but for the complete description of BAM format I suggest

referring to its online specification [130]. However, another important feature of BAM is the possibility of indexing. In this way, the alignments overlapping a determined region can be fastly retrieved without sliding the whole file. Before indexing, BAM must be sorted by reference ID and by the leftmost coordinate. The basic indexing algorithm is based on the UCSC binning scheme [131].

After mapping, the obtained BAM files are further processed in order to improve the quality of the information contained. In particular, optical and PCR duplicates are marked or removed, then initial alignments are refined by local realignment around known indels and finally base quality scores are recalibrated exploiting an empirically accurate per-base error model [132].

Although all these processes are becoming standard for variant analysis, the correct parametrization of each single step is still difficult to perform as it requires a lot of testing. Indeed, the parameters chosen for a specific problem could not be the best for a different type of project, because for example the read length or the mean coverage can change.

1.2.3 Variant calling and VCF format

After the read alignment and the post-processing, the final BAM is ready to be handled in order to find all the differences between the base called and the reference. This process is known as variant calling (VC) and it is usually followed by genotype calling, where it is determined the genotype of the called variants [133].

Essentially, two different types of variant can be discovered through the VC: germline or somatic. Although the software to detect variants, referred as variant callers, are often capable of performing both kind of calls, the basic processes are quite different. For defining germline variations, the comparison is made between the sequenced sample and the reference genome, which is a digital assembled sequence providing the best representation of the sequence of a particular specie [134]. On the other hand, instead, the detection of somatic variants is usually achieved finding the difference between the genome of a tumoral specimen and its normal counterpart, which works as a reference [135].

More than 40 tools for VC can be found in literature, able to determine both single nucleotide polymorphisms (SNPs) and short insertion deletion (indels). These tools are based on various algorithms, allowing the reaching of different values of sensitivity and

positive predictive value (PPV) [136]. The first methods for VC were based on the direct counting of alleles at each site, using simple base quality cutoff, typically Q_{PHRED} equal to 20, for determining either the variant or the genotype [137]. Further improvements regard the introduction of probabilistic approaches and local realignment for variant detection. In particular, GATK UnifiedGenotyper [132], FreeBayes [138], Strelka [135] and SAMtools [129] are based on Bayesian statistical approaches, while VarScan2 [139] detects variants applying an heuristic method and a statistical test. Moreover, LoFreq [140] relies on Poisson–binomial distribution, instead SNVer [141] employs a binomial-binomial model for testing how significant is the difference between the observed allele frequency against sequencing error. The most advanced variant callers integrate more complex methods based on local realignment or *de novo* assembly for the haplotypes reconstruction. For example, GATK HaplotypeCaller [142] is designed for calling together SNPs and indels, exploiting a local *de novo* assembly of haplotypes in a region which displays sign of variation. On the other side, VarDict [143] carries out supervised and unsupervised realignment, for calling short and long indels, respectively. Last, Platypus [144] relies on a local *de novo* assembly using de bruijn graph, followed by a local realignment and by probabilistic haplotype estimation. Almost all tools give a set of parameters for a comprehensive description of the reported variants and also some suggestions for a proper filtration.

Moreover, it has been proved that the use of multiple individuals triggers to an improvement in the genotype calling accuracy over using single samples, that is even increased if linkage disequilibrium (LD) information are used. Otherwise the gain in accuracy mediated by LD is great for variants with high-allele frequencies, while is not so significant in the calling of rare allele [133].

Nevertheless, several issues still afflict the VC process, as currently no software or pipeline succeeded in calling all mutations. In particular, it was pointed that sensitivity and precision seems to be inversely proportional, in sense that high sensitivity is always accompanied by the increasing of the false positives number [136]. Consequently, as the variant callers use various approaches to optimize sensitivity and PPV, the resulting variants concordance is usually quite poor, in particular when indels [145] or data from low coverage regions [146] are compared. For these reasons, some strategies using a consensus of data derived from different software [147, 148], in addition to a specific attention on variants falling in low

complexity regions allowed an increasing of accuracy [149], even if a lot of work is needed for getting reproducible results.

The information obtained during the VC step are collected into a standardized format called variant call format (VCF). It is capable of storing the prevailing types of sequence variation, like SNPs, structural variants and indels, in addition to other kinds of annotation [150]. VCF is a tab-delimited text file format, containing at the beginning several meta-information lines preceded by two '#' characters, a header line starting with only '#' char, and then data rows each including information about a position in the genome [151]. The meta-data rows provide information regarding file format or associated with INFO, FILTER, FORMAT or sample IDs fields of the VCF file. The header line contains eight mandatory columns (#CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO), followed by a FORMAT field when genotype is present. The latter describes which features are shown into the following arbitrary number of sample IDs. The body of VCF, usually constituting by the variants even if in a slightly modified version of this format, referred as genomic VCF (gVCF), also non-variant blocks are included, presents the same number of header columns, whose meaning is explained in Table 4.

Col N	Field	Type	Example	Description
1	CHROM	Str	chr1, 1	Chromosome name
2	POS	Int	307058, 1274308	1-based position sorted numerically
3	ID	Str	rs21589, COSM15, .	Semi-colon separated list of unique identifier
4	REF	Str	A, AT, AGN	Reference base(s) included in A, C, G, T, N
5	ALT	Str	AG, AGGG, *	Comma separated list of alternative alleles
6	QUAL	Int	0, 22, 40, 365	Phred quality score for ALT [$-10 \log_{10} P_e$*]
7	FILTER	Str	PASS, q10	Filter status; PASS → all filters passed
8	INFO	Str	NS=3;AF=0.01	Additional information, semicolon-separated
9	FORMAT	Str	GT:AD:DP	Specification of the data types and order
10	SampleID1	Str	0/1:28,32:60	Specific values of FORMAT keys in Sample1
...	...	...	...	...
N	SampleIDN	Str	1/1:0,44:44	Specific values of FORMAT keys in SampleN

Table 4|**Overview of the fields in the VCF format.** Each column is associated with an explanatory example. Columns from 1 to 8 are mandatory, while the others are optional, even if they contain the information linked to the analyzed samples. *P_e indicates the probability that call in ALT is wrong.

The variability in the keywords included in the various fields, in particular INFO, FORMAT and consequently Sample ID, is wider than what present in the other described formats and it mainly depends on the software used for making the VC process and on the type of mutation is going to be detected.

1.2.4. Variant annotation and prioritization

Once variants are called, researchers need to give them a meaning in order to understand their functional impact. This process of adding useful information on variants is called variant annotation. Two main level of annotations can be found: variant level and gene level.

The first type of annotation includes features directly derived from the VC step, such as the variant and the genotype quality scores, or calculated after the localization of the mutation at genomic level or found in various external sources [152]. Among the latter, the predominantly used are the allele frequencies databases, including the 1000 genome project [153], dbSNP [70], the Exome Variant Server (EVS) [154] and the Exome Aggregation Consortium (ExAC) [68], the pathogenicity predictors, as for example SIFT [155], CADD [156] and DANN [157], and the conservation scores, such as PhyloP [158], PhastCons [159] and GERP++ [160]. Almost all these information, in addition to predictions of mutation affecting splicing sites, are collected into dbNSFP database [161], which is often exploited for adding such annotations on VCF files by the most famous variant annotators, including VEP [162] and ANNOVAR [163]. The most complex type of annotation seems to be related to the assign of the impact of the DNA mutation at the transcript and protein levels. In fact, several types and versions of genes and transcripts, which differ each other, can be used, even if the main ones are provided by the Ensembl [164], the UCSC [165], the NCBI RefSeq [166] and the GENCODE [167]. In this way, a variant can affect a coding region in a specific gene set while it can be intergenic in another one. For this reason, it is usually recommended to employ the consensus coding sequence (CCDS) in the case that protein coding genes should be annotated, as CCDS are normally characterized by all gene sets [168]. Moreover, another problematic point regards the proper using of the HGVS nomenclature [169]. The main discordances between variant annotators are found in the indels as they are not always correctly realigned to the most 3’ possible position of the reference sequence, thus causing

multiple depiction of the same variant. Furthermore, taking into account the annotation of a frameshift variant, which theoretically causes the creation of a premature stop codon and therefore a shorter protein, it has to be pointed that this is only a prediction, probably wrong, because the process is dominated by the nonsense-mediated RNA decay phenomenon [170].

When variations lack of their intrinsic annotation, it is important to access data at the gene level. The Gene Ontology (GO) [171] is surely one of the most widely used, as it allows a characterization of the gene considering the cellular component, the biological process and the molecular function. Moreover, further information can be obtained accessing to the Disease Ontology (DO) [172], the Phenotype Ontology (PO) [173], the Genotype Tissue Expression (GTEx) [174] or the Gene Expression Atlas [175] data. Even pathway sources, such as KEGG [176] and Reactome [177], can be employed at this purpose. All these features are not usually integrated during a standard protocol of variant annotation, unless specific plugins have been used, constituting one of the main problem of the gene level annotation.

The information added during the annotation step are handled for finding all the variant which could be associated to the disease or condition under study. This process can be performed through a subsequent application of filters, which eliminate variants no overtaking the imposed thresholds, or through a prioritization system, which allows to bring out the most feasible variants, ordering them for relevance exploiting a certain number of criteria. The latter approach is more powerful, as it does not suffer of the main problem of filtration which regards the risk of removing something that is just below the threshold for one of the criteria, while being well above the threshold for the other criteria [178]. The main typologies of criteria work directly on variants features, including the mode of inheritance (e.g. autosomal recessive), the genomic localization (e.g. CDS or UTR), the type (e.g. missense or nonsense), the frequency in the population (e.g. ExAC or dbSNP), the pathogenicity predictions (e.g. CADD) and the previous description in the databases (e.g. LOVD [179] or ClinVar [180]).

Although several methods for annotating and prioritizing variants have been published, currently there is no gold standard able to clarify all the possible situations. Moreover, with the near switching from WES to WGS, new software will have to be implemented, allowing the interpretation of more kinds of variants, including those that nowadays are considered almost neutral such as the synonymous or the intronic ones [152].

2. General aims and thesis structure

The typical bioinformatic workflow that is currently used for the identification of genetic variants is affected by several problems that make these analyses difficult and prone to errors. As extensively explained in the introductory paragraphs, there are some crucial steps which are particularly prone to errors due to their complexity or to intrinsic features of NGS data [20]. For example, read mappers fail to align short NGS reads when they match to duplicated regions, while variant callers need an extensive parametrization. Moreover, an accurate annotation of genes and variants is often difficult to achieve because there are no tools that can provide this information, in a direct and comprehensive way, using a standard format.

With these premises, to keep up with the progress, in 2014 the University of Padua funded the BioInfoGen Strategic Project with the goal of developing technology and expertise in the area of bioinformatics and molecular biology applied to personal genomics.

Since this field of research requires multidisciplinary expertise, five different groups of the University of Padua have been involved, including four departments and the Interdepartmental Research Center of Innovative Biotechnology (CRIBI) (Figure 5). Each unit owns specific expertise, which are shared with the other groups to reach common goals. In particular, the Department of Biology guarantees the support in molecular biology, genomics and bioinformatics, while the Department of Mathematics, through its top-level knowledge in the machine-learning approaches, including neural networks and kernel-based methods, supplies the informatics expertise, particularly suited to be applied to the information collected in the genomics research. Moreover, the two medical units (Department of Women's and Children's Health and Department of Medicine) supply excellent skills in the genetics of the lysosomal storage diseases (LSD) and the renal proximal tubulopathies respectively. The case-studies considered in this research, besides their inherent medical interest, are also very useful for the setting up of a robust bioinformatics platform for personal genome analysis. Last, the CRIBI center contributes with the DNA sequencing facility and the computing resources, accompanied by the decennial experience in genomics and bioinformatics fields.

However, although each research units needs to achieve specific objectives within their respective fields, the general aim of the project is even more important. In fact, the establishment of an advanced platform for personal genomic analysis can emerge only from the joint effort of the different units, as the cooperation and the transversal knowledge diffusion are two strongholds for reaching ambitious scientific target.

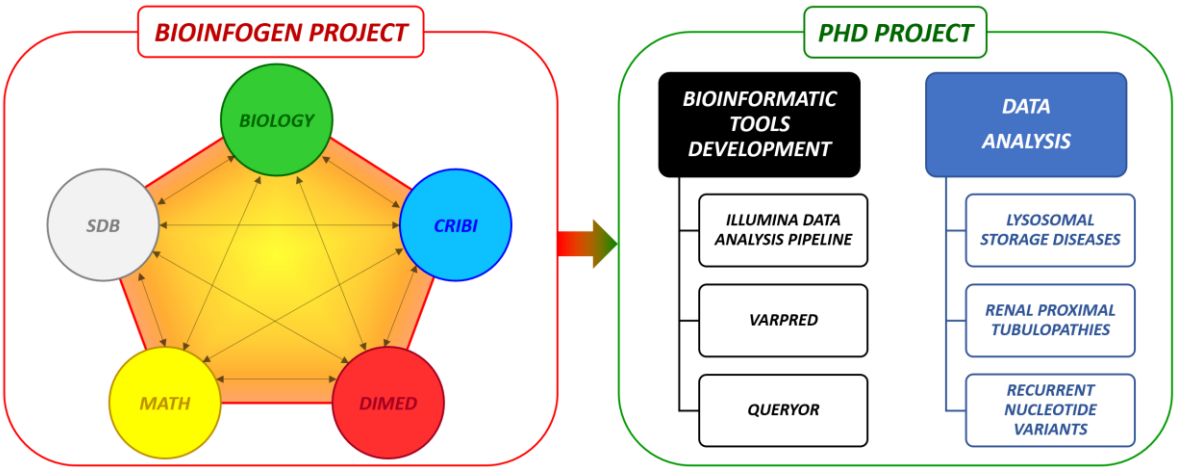


Figure 5|**Schematic representation of the two projects debated within the thesis**. On the left, the five units involved in the BioInfoGen are depicted. The extensive interconnection among the various research groups is highlighted using the arrows. On the right, the two main branches of my PhD project are summarized, proposing at the same time a further division into the subsections discussed within the thesis. SDB: Department of Women's and Children's Health.

My PhD has been funded by the BioInfoGen project (Figure 5), and many of the objectives of my research activity follow the general aims of the project. In particular, I was involved in the setting up and in the development of suitable software and solutions to overcome many of the problems associated with the NGS data analysis.

To reach the purpose, I developed three main tools: I) a comprehensive pipeline for Illumina data analysis (§3.1), which grants a cutting-edge detection of nucleotide variants; II) a variant annotator, named VarPred (§3.2), to provide a quick and exhaustive transcript-level annotation; III) a web-platform for variant prioritization, referred to QueryOR (§3.3), which allows an easy investigation of mutations affecting patients with rare genetic diseases.

Although the various programs have been specifically implemented for solving distinct problems, such as read alignment, variant calling, variant annotation and variant prioritization, these tools can be considered as part of a whole bioinformatics platform for high-throughput data analysis which should allow to get closer to personalized medicine. In fact, the proposed procedure can be applied not only to the two case studies involved in the

BioInfoGen, the lysosomal storage diseases (§4.1) and the renal proximal tubulopathies (§4.2) respectively, but also to other exome sequencing projects, where rare monogenic or oligogenic diseases need to be investigated.

To provide an easy comprehension of the proposed thesis, I decided to split it into two main parts, each of them is further divided into subsections. In particular, the first part describes the three major bioinformatic tools that I developed, while the second one shows the application of such methods to investigate the two clinical cases and to perform a deep study of recurrent variants in human exomes (§4.3).

I preferred to present the various subsections using a paper style format rather than organizing all the information in the usual material and methods (M&M), results and discussion (R&D), and conclusion blocks. In this way, I think the reader is facilitated in the interpretation of the discussed arguments. Nevertheless, each part is however subdivided using the common scheme based on introduction, M&M, R&D and conclusion paragraphs. A final paragraph (§5) summarizing all the obtained results is also proposed as conclusion of the manuscript.

3. Development of new bioinformatic tools

3.1. Illumina data analysis pipeline

3.1.1. Introduction

The purchasing of a new Illumina sequencer at the CRIBI sequencing service triggered the need of developing new systems for data analysis, because quite different in terms of software from what applied for Ion Proton data. Considering that the Illumina systems are the most widely spread sequencing technologies and for this reason such company is the market leader, several data analysis pipeline have already been published in literature [94, 181], in addition to specific suggestions proposed by the Broad Institute of Cambridge [103]. Commonly, after the base calling, reads are converted into FASTQ files, which are then aligned against the reference genome. A further processing involves the detection of duplicates derived from PCR or clusters misreading, followed by indel realignments and/or base quality score recalibration [182]. The last step requires the application of a variant caller in order to get somatic or germline variations.

Although the choosing of the software to implement within a pipeline is quite easy when the Broad Institute best practices are followed, the selection of the best parameters is instead challenging as it depends on the type of variants to detect, the read length, the coverage and many other factors. Moreover, as often some tools have been designed to be precise and sensitive at the expense of the performances, another important issue to consider is the computation time.

3.1.2. Materials and Methods

The following pipeline (Figure 6) is a bash script able to create and launch job files, derived from specific templates suitably designed for the various steps employed into the analysis. The majority of the applied programs are included in the Picard tools (version 2.3.0) [183] or in the GATK suite [181], both developed by the Broad Institute, with the exception of the read aligner (BWA [126]), the somatic variant caller (VarDict [143]) and a binary to calculate statistics (SAMtools [129]).

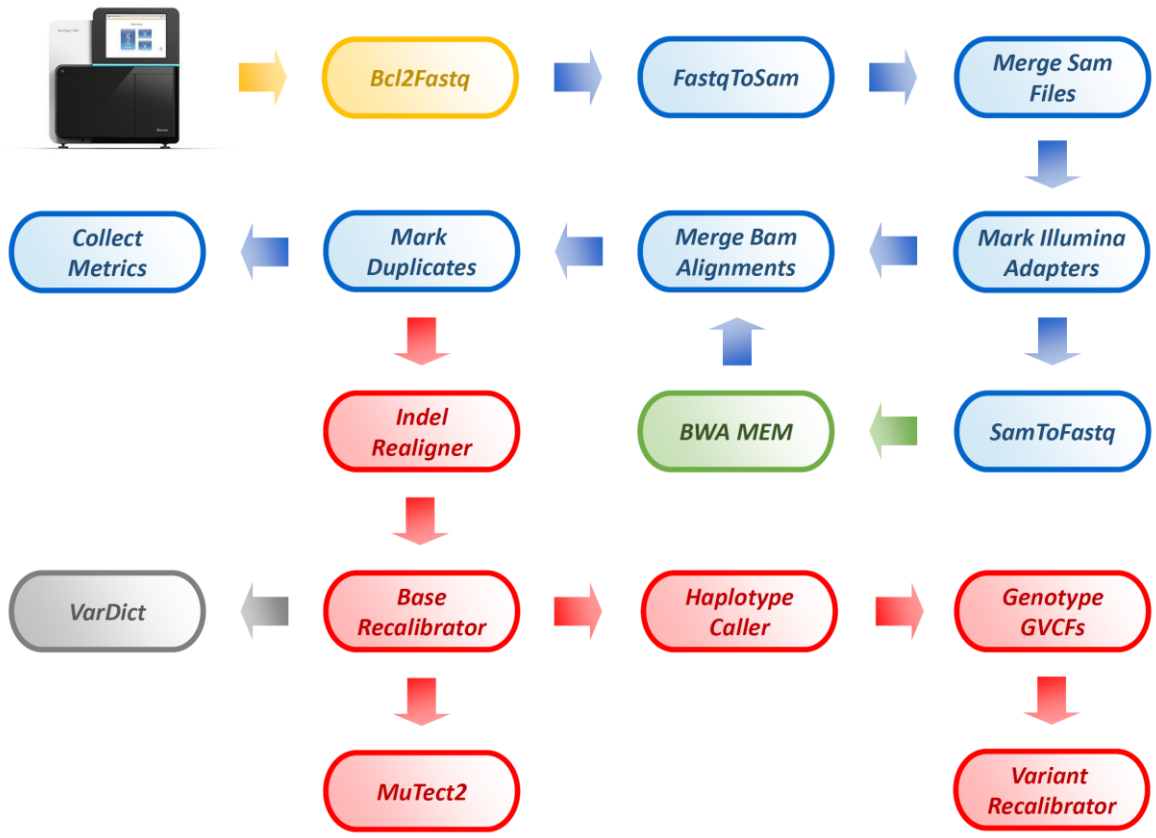


Figure 6|**Illumina data analysis pipeline.** All the steps included within the pipeline, from raw reads up to variant detection, are summarized. The different backgrounds are used to distinguish the packages which the various tools refer to (yellow: Illumina, blue: Picard, green: BWA, red: GATK, grey: VarDict)

The first step, called demultiplexing, is performed using the Illumina software *bcl2fastq.* It allows the separation of read clusters, depending on the information provided in the sample sheet, and the removal of residual portions of primers and/or indexes erroneously sequenced. The various lanes containing the sequenced reads are converted into the uBAM format using *FastqToSam*, and subsequently merged together through *MergeSamFiles*, for creating a unique uBAM file for each sample. A further step of sequence checking is performed applying *MarkIlluminaAdapters*. At this point, the resulting file is reconverted into an interleaved FASTQ file though *SamToFastq*, as the *BWA mem* aligner (version 0.7.15) cannot accept uBAM format. The obtained BWA output (SAM file) is then merged together with the previously processed uBAM using *MergeBamAlignment*, achieving the final aligned file with all the necessary information for the following manipulation. Optical and PCR duplicates are detected using *MarkDuplicatesWithMateCigar* and some statistics are then computed thanks to both *CollectHSmetrics* and *SAMtools stats* (version 1.3). The subsequent step is carried out applying a software included into the GATK suite (version 3.6-

0), named *IndelRealignment*, which performs a local read realignment around known indels taken from dbSNP v144. The final stage of read manipulation is accomplished with GATK *BaseRecalibrator*, which performs the base quality score recalibration (BQSR); such method helps to achieve a really accurate variant calling process.

When the read analysis is completed, a second block of instructions for variant calling is executed. It depends on the planned type of research, as somatic variants detection requires different parameters and tools from those used for the germline calling.

Thus, for common exome studies, where germline variations have to be found, only GATK tools are essentially used. First of all, *HaplotypeCaller* is run against all the samples, obtaining a set of VCF files. These latter are then processed all together by the *GenotypeGVCFs* in order to get a joint genotype for the called variants. Concluding, to obtain a refined set of calls, the variant quality score recalibration (VQSR) is performed, leading to the exclusion of several false positives. Unfortunately, this sophisticated machine learning approach carries to high quality data only when a large batch of samples (at least 30 exomes) and highly curated sets of known variants are available. Otherwise, the improvements are poor.

On the other hand, for somatic variants detection I chose to integrate two different programs, *VarDict* (version February 2017) and *MuTect2* (version 3.6-0) respectively. Depending on the conditions required for the analysis, it is possible to select one of the two callers, which allow various kinds of parametrization.

3.1.3. Results and Discussion

The automatization of the various software for Illumina data manipulation into a high-performance pipeline allows an easy and rapid analysis of the NextSeq 500 output. In particular, the system has been thought to be handled not only by bioinformaticians, but also by biologists with a minimal knowledge of the Linux shell, allowing a wide usability among the people of the CRIBI sequencing center. Indeed, a single command with few required inputs including the sequencing run name, the folder on which FASTQ data are stored, the interval file with the coordinates of the enriched regions, the reference genome which should be used and the applied sequencing type (paired-end or single-end), is enough for performing the whole analysis. The last two options are really useful as they improve the

overall versatility of the system: in fact, both kinds of reads can be accepted and, more interestingly, almost all species with a known reference sequence can be processed and analyzed. Regarding the parameterization, an intensive critical reading accompanied with a lot of simulations using testing samples has been performed, in order to obtain reliable and reproducible results.

However, the most time-expensive part of the pipeline implementation regarded the performance optimization. Indeed, special attention has been dedicated to the reduction of the pipeline wall time and to the saving of disk-space, as in the NGS field data storage has become a tricky problem to tackle [184].

To face up the latter issue, I tried to use, where possible, compressed BAM files, rather than their extended SAM form, diminishing at the same time the number of intermediates. For example, the read alignment using BWA required a first step of read reconversion into FASTQ format, followed by the merging of the obtained BAM file with the previously prepared uBAM file. Thus, many temporary intermediates will be created and saved if this triptych of processes is performed separately. Instead, in the described system, the three commands have been concatenated into a unique instruction, where the output of a program is directly redirected into the input of the following one. This method allows to store only one file for each sample, simultaneously diminishing the wall time. The drawback of such approach is related to the need of repeating all the steps if an error arises when the whole block is not finished yet.

On the other hand, the issue regarding the long wall time has been solved applying a wide parallelization, not only associated with the software used, which is not dependent on the pipeline implementation, but mainly over the various samples which are usually sequenced together into a single run. In fact, each specimen, from FASTQ conversion to the final recalibration, is processed independently and parallelly to the others thanks to a series of jobs which are automatically created and executed into the thirty blades of the CRIBI cluster. Furthermore, the step of somatic variant calling, which supports both *VarDict* and *MuTect2* callers, has been found to constitute the most time-wasteful stage, as it can take more than one week to process a single matched pair of tumor-normal specimens. Surely *VarDict* offers a better performance than *MuTect2*, but the software parallelization is not so efficient to significantly reduce the time required for the analysis. Instead, for the system speeding up, the Broad Institute suggests the application of its pipeline, called *Queue* [185],

that is based on scatter-gather parallelism (Figure 7). Exploiting such method, the *Queue* engine will execute the same GATK command on separate portions of the input data (scatter step), storing results in temporary files. Then once all the runs will be completed, the engine will combine all the results into the final output files, simulating an event comparable to a single running command (gather step). The great advantage of such form of parallelism consists in the possibility to extend the programming at cluster level, allowing at the same time the multithreading parallelism within each single machine. However, as integrating a pipeline within another one is a difficult task, not always associated with such a great improvement to justify the effort, I decided to directly integrate an approach similar to the scatter-gather parallelism into my program.

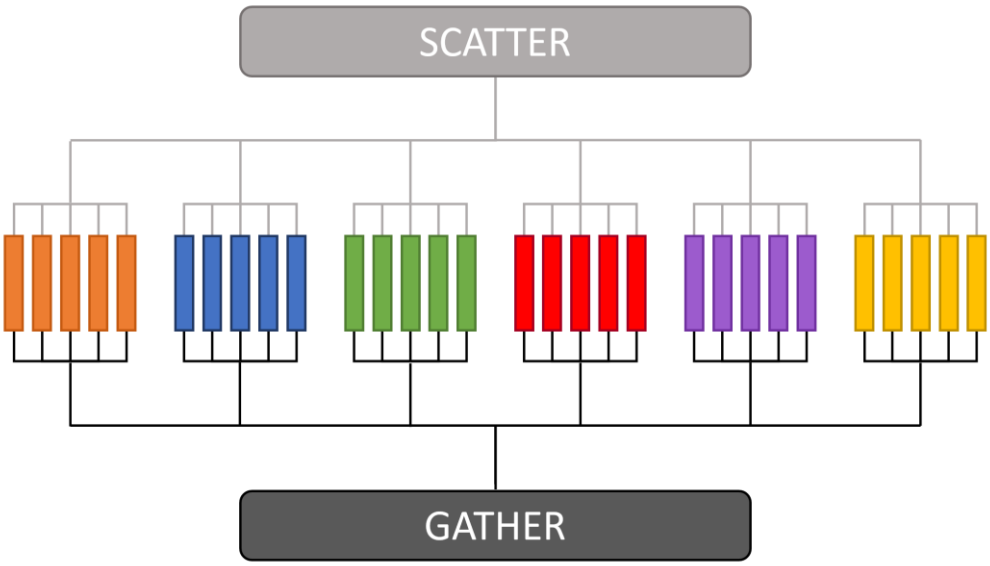


Figure 7|**Scatter-gather parallelism.** After the initial split of the input data, the same command is run on each of the previously created junk (scatter step). When the processing (single-threading or multi-threading) is completed, the partial outputs are merged together to obtain the final result (gather step).

For this purpose, BAM files are splitted per chromosome, then the variant calling is linearly executed (tests on multithreading did not evidence substantial improvements) comparing case with control on each couple of chromosomes and finally partial VCF files are joined together producing the final output. This kind of implementation reduced the wall time of somatic calling by about twenty times, allowing to carry out the process in less than one day. A possible limitation of such method could be associated with the reduction of extension of the sample background, used by callers to determine variants. Nevertheless, the comparison of the results produced by *MuTect2* applying or not the scatter-gather

parallelism on the same couple of samples did not show significant differences, taking into account also the heuristic component of the software. For this reason, such approach seems to be a very efficient form of parallelization, that can be moved and applied in many other bioinformatic tasks, as for example the variant annotation. Considering the whole pipeline, it can completely process a high coverage (~100X) case-control couple of exomes in nearly one day using the resources available for a single user in the CRIBI cluster.

At the time of the thesis writing, the pipeline was already used to successfully analyze dozens of normal exomes for germline variants detection, in addition to more than thirty exomes of matched pairs of tumor-normal specimens derived from patients affected by various types of tumor, including gastric cancer [186].

3.1.4. Conclusion

The proposed pipeline is a good compromise to process and analyze Illumina data, in order to detect either germline or somatic variants. The easy command line should guarantee the usability also to non-bioinformatician personnel, while the application of a job scheduler system and the scatter-gather parallelization permit to widely reduce the analysis time, without losing the consistency of the obtained results. Indeed, the overall quality of the final output has already been proved by Di Bartolomeo and colleagues [186], but also in other internal controls, allowing thus to retain concluded this part of the work, even if further ameliorations are always possible following the development of the technology.

3.2. VarPred: a flexible tool for genetic variant annotation

3.2.1. Introduction

As extensively explained in the previous paragraphs, the development of NGS sequencers has quickly moved the main bottleneck of the sequencing studies from the sequencing process linked to Sanger technology to the data analysis, which has become almost always the most cost-effective step of each project [88]. In fact, the big amount of the produced data raised novel tasks linked to the downstream steps, including the calling of the variants and lastly their annotation and prioritization. Several bioinformatic approaches have been proposed to solve the problems associated to each specific step, even if a lot of work is still ongoing to ameliorate the existing software or to implement new one. In particular, getting

back on specific issues related to short nucleotide variants (SNVs), once they are called and reported into a VCF file, they have to be annotated using various information including gene features, HGVS descriptions [187], allele frequencies, variant database identification codes and ontologies. The majority of variant annotators, including the most used VEP [162] and SnpEff [188], have chosen the Sequence Ontology (SO) [189] for describing the specific traits of biological sequence, while other tools such as VariOtator [190] adopt the Variation Ontology [191] to depict nature, outcomes, consequences and processes of variation. A comprehensive annotation is also necessary to carry out a deep prioritization process, which can exploit sequential filtering steps, as in wANNOVAR server [163], or more complex criteria selection, integrated for example in QueryOR platform [178].

Although a lot of work has been recently made to standardize the VCF annotation, various standard formats have been proposed, including the *clinical sequencing nomenclature* (CSN) [170] which is mainly based on the principles of HGVS guidelines and the *Variant annotations in VCF format* [192] which instead takes into account both HGVS notation and SO terms, slightly modified in order to eliminate inconsistencies with VCF format. Surely, a standard annotation format is fundamental to facilitate benchmarking, to allow an easy integration of each annotator into common NGS data analysis pipelines, and to solve issues regarding particular cases, including indel realignment.

Another task often unconsidered regards the various representations by which a single variant can be described. In fact, the lack of variant normalization can trigger misleading variant annotation as the same variant has not a unique VCF depiction. In particular it is possible to encounter this problem when multiple alleles, usually classified as indels or multiple nucleotide variants (MNVs), are detected in a specific position by variant callers. To prevent this ambiguity and to allow a consistent representation, a variant needs to be *left normalized* and *parsimonious*, which respectively mean that the position of the VCF entry is the smallest and the allele is the shortest between all alleles representing the same variant [193].

Moreover, even if the introduction of recent variant callers, based on haplotype reconstruction, has partially solved such issue, in many cases adjacent variants, falling in the same reads, should be fused into a single feature, constituting a MNV. This event has to be treated with special attention when these variations are located in a single codon, as often their separate annotation leads to an incorrect interpretation of the real variant. To my

knowledge, none of the publicly available annotators integrates a system to perform such step of neighboring variants aggregation, exploiting only the information contained in the VCF file. Indeed, the main challenge of the methodology resides in the haplotype phasing, which requires the BAM file with the aligned reads. MAC software [194] has been specially implemented to correct potentially misannotated MNVs among the reported SNVs, getting in input both files, the VCF and the BAM, but then the real annotation is performed using ANNOVAR, SnpEff or VEP.

It has been proved that also the gene-set employed for the annotation has a great impact on the variant prediction even if the main differences among GENCODE [167] and RefSeq [166] can be found in non-coding transcripts and UTR sequences [195]. This issue is directly linked to the need of annotating all the gene transcripts, not only the canonical isoform, as the same variant could have different impacts depending on the transcript in which it falls.

Considering all these issues, together with the will to give the user the possibility to highly customize the version of the input databases required for the annotation, I implemented VarPred, a flexible and efficient variant annotator able to characterize all types of SNVs, including indels and complex rearrangements. Although its development is not yet completed, VarPred is available as stand-alone software, allowing its integration into variant discovery pipelines, and as web-platform, ensuring the opportunity to explore results in a simple, but at the same time accurate manner also to non-bioinformaticians.

3.2.2. Materials and Methods

VarPred is implemented in Python 3 and it requires the installation of Pandas library and its dependencies, tabix [196] and python3-tables module, which provides an efficient data storage, based on HDF5 and pickle formats. Pandas library has been chosen as it allows a simple and efficient management of tabular data, in which almost all genomic data can be converted, promising at the same time great computational performance, thanks to its compiled C core. Differently from the other variant annotators, which require an initial step of precomputed data downloading, VarPred directly works on the original annotation files, allowing an easy knowledge of the version of data used. The required inputs are very few: the VCF file to be annotated, a set of transcripts in GTF or GFF3 format and the indexed reference genome (Figure 8). Moreover, optional annotation data can be added to the

software, including three databases of allele frequency (ExAC [197], ESP6500 [69] and dbSNP [70]) and one linked to somatic mutations as the Catalogue Of Somatic Mutations In Cancer (COSMIC) [198]. Other options can be set to annotate not common organisms, mainly when the nuclear or the mitochondrial genetic codes are not the standard ones. Also variations in plants are supported thanks to a specific module which handles the plastid genetic code. Finally, to manage the memory required by the software, the number of variants which have to be annotated for each run can be selected, in addition to the possibility to declare how many processors can be employed.

GENCODE, Ensembl and RefSeq transcripts can be used by VarPred as genomic scaffold for the annotation. During the first running on a specific gene set, the software carries out several computations in order to obtain the highest number of information on CDSs, exons, introns and UTRs, but also on whole transcripts. Moreover, the presence of common splicing sites within multi-exons transcript is checked and possible inconsistencies due to superimposed features are removed. For each gene, the various transcripts are sorted depending on the rules for the canonical transcript definition provided by Ensembl [199]. The tables obtained during the previously described processing are saved into a HDF5 file, while the computed searching structures are included into a pickle file. Both files allow a quick access to the stored data, which are employed to determine the annotation at gene, transcript and eventually protein level.

During the annotation step, variants are firstly normalized following the method proposed by Tan *et al* [193] with slight modifications. Then, the information included into the selected databases is added, considering that in many cases such features need to be rearranged, and not simply copied and pasted. It happens, for example, with minor allele in reference (MAiR) variants where minor allele frequency (MAF) value, usually corresponding to the variant frequency, is not related to the variant itself, but to the frequency of the reference allele. Thus, the variant frequency has to be somehow calculated. In the following step, variants are divided depending on their type and each subgroup is annotated exploiting a dedicated function. The annotation format chosen is based on suggestions proposed by Cingolani and colleagues [192] and it includes the HGVS predictions, the sequence ontology annotation, in addition to the various IDs associated with the adopted gene-set and the position at transcript and protein level. The final output of the program is an annotated VCF, to which also a TSV file can be added.

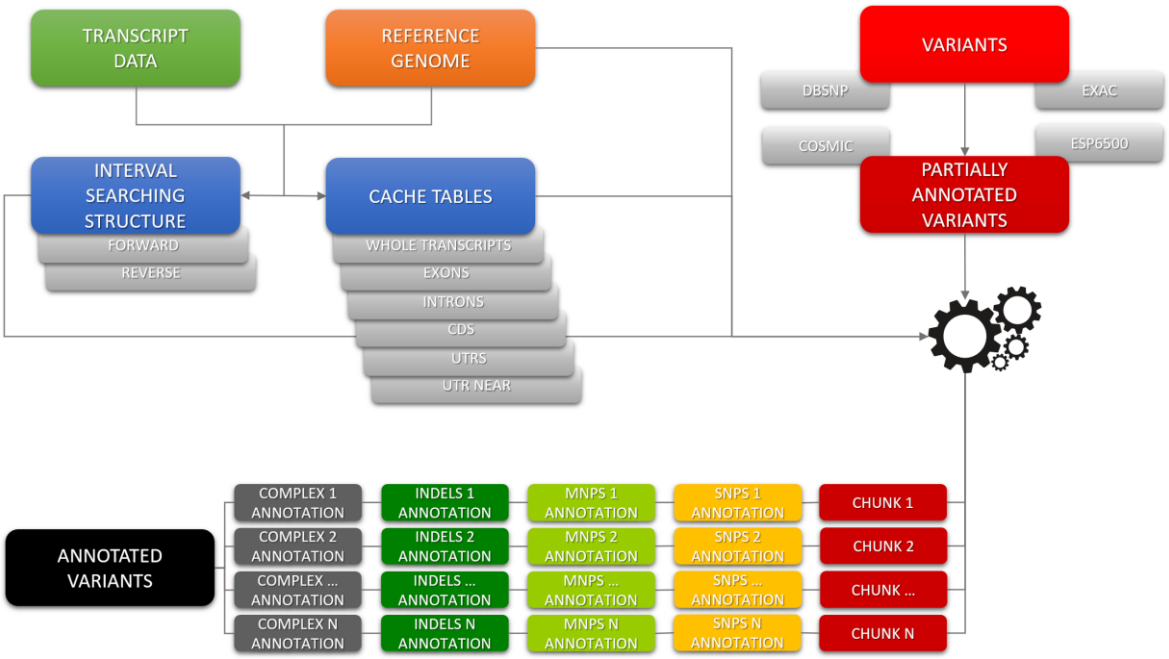


Figure 8|**VarPred annotation process.** During the first run, the software employs the transcript data in GTF or GFF format and the reference genome to build some cache tables, containing genomic features, and a searching structure to provide a fast detection of the regions in which the variants fall. These files will be then used to complete the annotation of the input VCF file, after the association of the database related features (allele frequencies and IDs) to such variants. Before to undergo the main annotation procedure, variants are also splitted into a number of chunks equal to the declared CPUs. In this way, the parallelization process can be enabled. Then, special functions are applied to annotate SNPs, MNPs, indels and complex rearrangements contained in each chunk. Finally, the annotated variants are joined together into the final file.

Even if not yet completed, VarPred owns also a web graphic interface, built using Django, a high-level Python web framework. The user can select one of the implemented species, which are the widest used in the scientific area, including *Homo sapiens*, *Rattus norvegicus*, *Danio rerio* and *Mus musculus*. Subsequently, one or more VCF files can be uploaded and the parameters for annotation, as for example the frequency databases, can be selected. With the collected knowledge, the system creates a job file containing the command line and all the accessory information for correctly executing the instruction. The job is submitted to the CRIBI cluster, while a progress bar shows the status of the processing. When it is completed, a notification email, containing the link to the results, should be sent to the address indicated during the file uploading (this part is not yet implemented). The managing of the results is performed using a plug-in for the JQuery JavaScript library, called DataTable, which allows easy pagination, instant search and multi-column ordering of the VCF table calculated during the annotation step. No database is employed to store the

results, which instead are saved into pickle files and then processed using Pandas before to be treated by the JavaScript. A basic filtering strategy is also integrated, similar to what proposed by wANNOVAR or the online Ensembl VEP, where user defined cut-offs are applied to reduce the searching space of the analysis.

To evaluate the performance of VarPred, it has been compared with the state-of-art variant annotators, including ANNOVAR (release June 2017), SnpEff (v4.2) and VEP (v89). Various conditions have been tested on Intel(R) Xeon(R) CPU E5-2687W v2 @ 3.40GHz exploiting all the 16 CPUs, changing in particular the dimension of the input VCF (~45000 variants for a typical human exome and ~2900000 variants corresponding to a human genome-size dataset, even if in this case coding mutations of COSMIC v74 have been used) or adding common parameters, such as the dbSNP annotation. Moreover, to assess the accuracy of VarPred, always considering the other tools, a subset of near 2 million variants selected from coding mutations of COSMIC v74 have been analyzed. In particular, protein predictions contained in COSMIC and those computed by the annotators were compared, dividing them into SNPs (1964578), insertions (18015), deletions (47002), MNPs (10130) and complex rearrangements (819), similarly to what reported in other studies [200].

3.2.3. Results and Discussion

Software implementation

VarPred is an easy-to-use and flexible software guaranteeing a comprehensive variant annotation of almost all species for which an assembled genome and a complete gene-set exist.

Two different data formats for transcript data, GFF3 and GTF, are accepted by the annotator raising the possibility to choose among GENCODE, Ensembl and RefSeq gene annotations. This feature is really important as it has been proved in other studies that the selection of the reference transcripts is challenging for the functional annotation of the variants, as wide effects can be found on the outcome [195]. For solving this issue, it was developed also a consensus sequence among the various transcript sets which unfortunately are limited to protein coding regions [168]. Moreover, still considering the genomic features, VarPred has also been developed to annotate all the transcripts found within the provided gene-set, as different impacts could be linked to a certain variant depending on the transcript in which it

falls. In fact, the same variation could be intronic in a particular transcript, while it could affect the coding region in another one, constituting for example the disease-causing mutation, as such transcript is the most expressed isoform in the considered tissue. Nevertheless, as often researchers focus their attention on the canonical isoform, VarPred sorts transcripts following the rules supplied by Ensembl [199], based on the presence of CCDS, the length of the coding regions and the whole length of the transcript. Thus, the possibility to extend the analysis to different annotations, considering at the same time the whole plethora of transcripts, can give a wider comprehension of the genomic context associated with the variants under consideration.

This possibility is supported also by the other annotators, which however require the downloading of several cache data related to the chosen gene set. Instead, VarPred calculates such files during the first running of the software, facilitating in this way also the using of user-created custom gene set, deriving for example from non-model organisms or from a re-annotation of common species.

The reproducibility of the annotation is guaranteed by the normalization system based on the algorithm proposed by Tan *et al* [193]. Thus, a specific variant, even if depicted in various ways within the VCF, as it can happen for particular indels or MNPs, will be always associated with the same prediction, facilitating the evaluation of the results and avoiding possible misleading in variant interpretation. Moreover, such normalization system is also applied to correctly realigned variants to the most three prime position of the transcript, in order to compute an accurate HGVS prediction at transcript and protein level. Indeed, the exact characterization of HGVS notation was one of the main reason which promoted the development of VarPred, because in the late 2014 none of the available annotators had completely fixed the problem. Obviously, nowadays such issue has been predominantly solved by almost all the previously mentioned tools, even if some discrepancies can still be found when a variant falls near the boundaries of uncompleted protein coding transcripts, where the coordinates of the first or the last codon are not properly assigned.

Regarding the annotation format, I chose what suggested by Cingolani and colleagues [192], as the proposed rules are based on the HGVS recommendations and on the SO terms, allowing a compact but exhaustive representation of all the information computed during the annotation process. The knowledge included into the VarPred annotation ranges from the HGVS notations at genome, transcript and eventually protein level, to the consequence

prediction using SO terms and a simple estimation of the putative impact, in addition to the various IDs associated with the adopted gene-set, the position at transcript and protein level and the transcript biotype. Moreover, errors, warnings or informative messages, regarding possible issues raised during the annotation, are reported. Although other formats, including CSN, have been proposed, such scheme has already been adopted by default by SnpEff, and selecting a specific option by VEP. This indicates the willingness of the developers to make a common effort to achieve a standard and easily handled format, favoring also the developing of new VCF parser which will facilitate the final data prioritization.

Web-interface implementation

The process of variant prioritization is not directly integrated in the stand-alone version of VarPred, as it was established since the beginning to develop a web interface of the software when its implementation was almost finished. Even if the platform is not still ready to be publicly released, its skeleton is essentially completed. Indeed, not only for testing, but also for convenience, the web version of VarPred was frequently used to annotate various type of VCF and to explore the obtained results. For using the platform, implemented through the Django Python framework, no registration is required, but only a valid email address where a message with a private URL for the results visualization will be sent. Firstly, the user has to choose among one of the integrated species (Figure 9A) which are essentially the most used in the biosciences including some model organisms, as the zebrafish and the mouse, and obviously the human. This initial selection is necessary for the following steps because the different species are associated to peculiar parameters and features (Figure 9C), such as which databases or genes sets are available. These latter are set up in the second step, where also several VCF files can be uploaded (Figure 9B). Once variants and the other information are submitted, a job file is properly modified and sent into a specific designed queue in the CRIBI cluster. The processing can be followed in the specific web page through a loading bar (Figure 9D). When the annotation is concluded, the user can begin to explore the results, which are presented as an interactive table, managed by the JQuery DataTable library. If more VCF files have been uploaded, the same number of pads will be clickable on the top of the web page. On the left, instead, it is located the menu

for the filtering, where the type of variant, the allele frequencies, the genomic context and the putative impact on protein can be selected (Figure 9E).

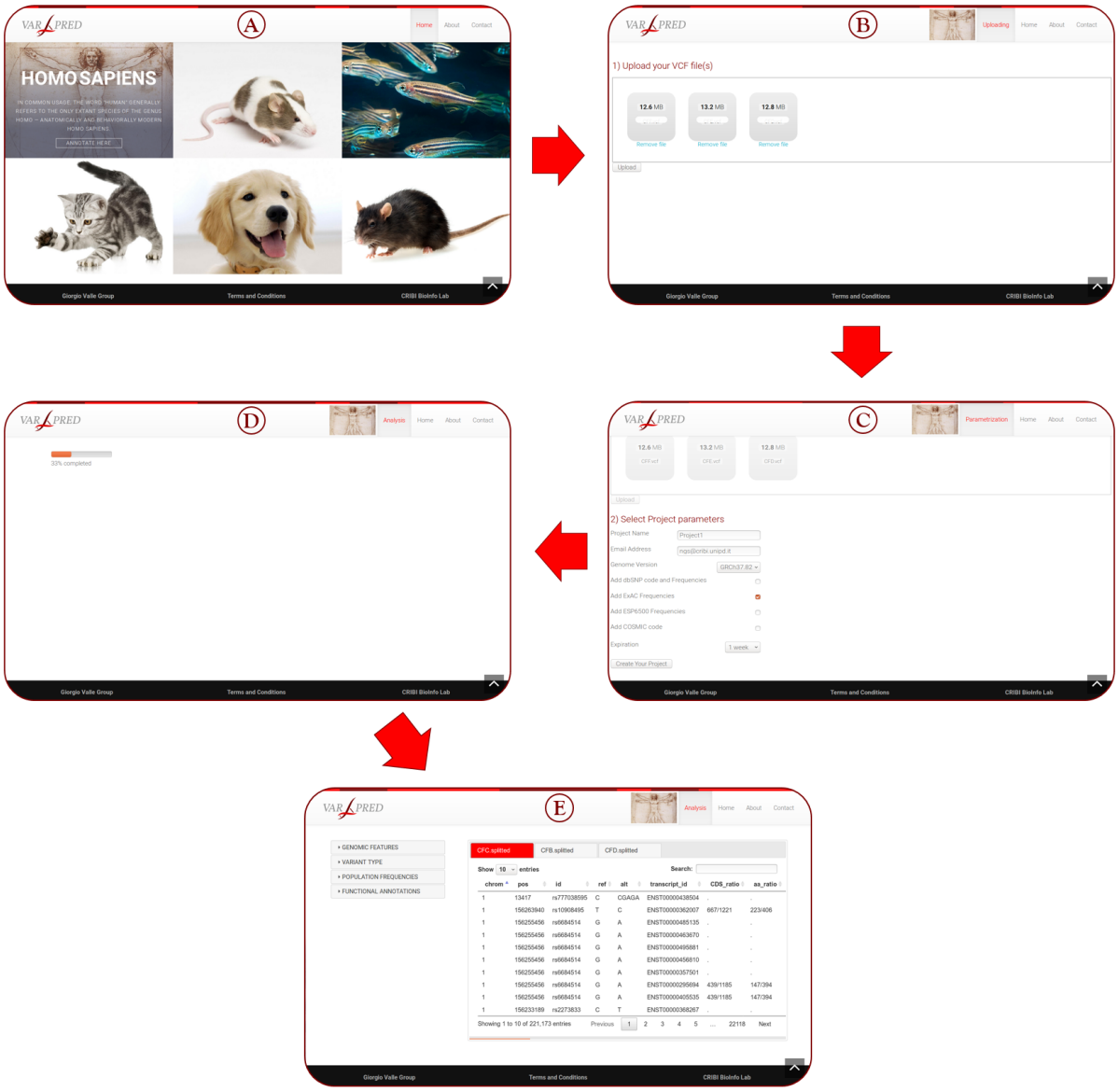


Figure 9|**Web implementation of VarPred. A**|The typical workflow, using the web implementation of VarPred, starts with the selection of the species to be annotated, among those implemented. **B**|Then, several VCF files can be easily uploaded in few seconds through a '*drag and drop*' window. **C**|The analysis parameters including the project name, a valid email address and the expiration time, are set up in the following step, **D**|which triggers to the beginning of annotation process, notified by the loading bar appearance. **E**|Finally, the user can filter the annotated results exploiting the interactive menu on the left side of the page and refine the output using the controllers integrated into the table.

Contrary to the majority of similar web tools, which save data into relational databases, the web interface of VarPred exploits pickle files, where annotated variants are stored. Once the user's request for filtration or sorting is sent to the server, a Python Pandas script collects the information, loads the specific pickle file, executes the request and returns the final

results. The whole operation usually takes few seconds, ensuring an almost real-time visualization of the final outcomes. Although not yet implemented, the results will be accompanied by several interactive charts, which will help the user in the comprehension of the managed data.

As it is possible to understand, the VarPred web server is currently a work in progress, and for this reason it is probably the area in which more improvements could be introduced in the near future. In fact, in addition to the features which will be included into the stand-alone software, the platform will be ameliorated in order to become widely usable, thanks to the development of a fully interactive user-friendly interface, but at the same time flexible and easily maintainable, to be kept always updated.

Tool comparison

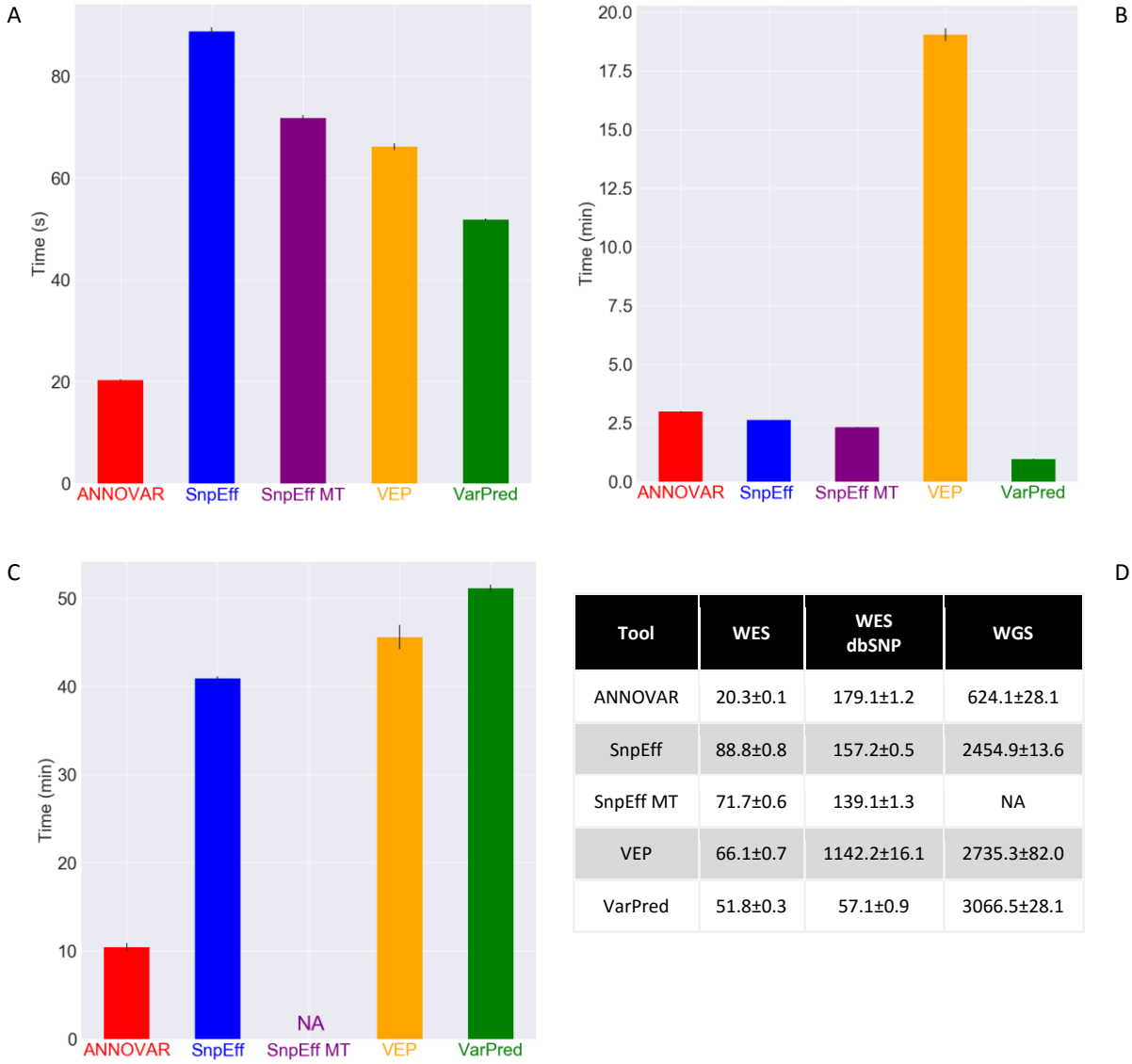
To assay the performance in terms of speed and quality of annotation provided by VarPred, the software has been compared with the most widely spread variant annotators, including ANNOVAR, SnpEff and VEP, which collect respectively 3280, 1746 and 1160 citations in literature. The speed tests are quite simple to perform as it is enough to run multiple times all the tools on the same machine with identical or at least similar parameters in various conditions. The quality of annotation is instead really difficult to establish for several issues. First of all, no benchmark is available for such kind of testing, excluding the COSMIC database whose annotation is homemade, but almost all predictions do not follow the HGVS nomenclature. Moreover, the output of the different tools is quite various even if a great effort has been done to standardize it. Indeed, VarPred, SnpEff and VEP follow the rules provided by Cingolani and colleagues, but in many cases a rearrangement of protein predictions is necessary to efficiently compare them. Finally, also the choice of the gene-set selected for the annotation has a great impact on the comparison; unfortunately, it is not always possible to trace the version or the type of transcripts used, as it happens for COSMIC.

Nowadays the annotation time is not so relevant, if compared with the whole time taken by a common analysis pipeline from raw data conversion to variant detection. Nevertheless, the time performance of the previously mentioned variant annotators has been tested in different conditions, changing in particular the number of variants to process (WES or WGS) or setting up a common parameter, such as the addition of rs codes and allele frequencies

obtained from dbSNP. For each test, five replicates have been performed and the final averages have been plotted with their matched standard deviations to facilitate the comparison. Moreover, SnpEff has been assayed twice, as the multithreading configuration has been considered separately from the common implementation, as previously reported in the work of McLaren and colleagues [162].

VarPred shows an overall good performance. In particular, considering the annotation of a VCF derived from a human WES project (~45000 variants), VarPred is really competitive, showing a higher computation time only when compared with ANNOVAR (Figure 10A). Moreover, VarPred is the quickest software for annotating the same VCF using also the dbSNP database. In such condition, all programs have significantly increased their running time from 2 to 15 times, with the exception of VarPred (Figure 10B). This evidence could be explained by the integration of tabix algorithm, which guarantees a high-level computing, within the function for associating the dbSNP information to the variants. At genome level, instead, VarPred seems to perform a little bit worse than the others even if its performance is quite similar to the Ensembl VEP (Figure 10C). A possible explanation could be associated to the fact that VarPred is the only software among the considered ones to display almost a linearity between the run-time and the variants number, while SnpEff is the tool whose processing is less affected by the increasing of variant number. Surely, the best performance of SnpEff, when the cohort to be annotated is wider, depends on the programming language used for the software implementation. In fact, Java thanks to its compiled nature is typically faster than the interpreted languages, such as Python or Perl, in which VarPred and VEP are respectively written. On the other hand, Java employs much more time to begin the processing, explaining in this way the low efficiency of SnpEff in annotating a strict group of variants. Instead, the multi-threading version of SnpEff has not be included in the comparison at the genome level as the run time reproducibility was really low, due to several random errors, including an out of memory in a 48gb RAM server. Nevertheless, the fastest tool in almost all the conditions is ANNOVAR (Figure 10D), that is implemented in Perl like VEP. Similarly to what reported by McLaren and colleagues [162], the shorter processing time of ANNOVAR is due to the reduced annotation depth of such tool. Indeed, the other applications perform the annotation at transcript level, while ANNOVAR reports a prediction at gene level, usually considering only the canonical transcript or at most a few number of them.

The performance of VarPred has also been tested using various species with different complexities such as the *Mus musculus* or the *Saccharomyces cerevisiae*. It was seen that the number and the complexity of genomic features used during the annotation highly impacts on the run time of the software. This behavior results in faster analyses for species scarcely characterized than those associated with a rich annotation such as the human and the mouse.



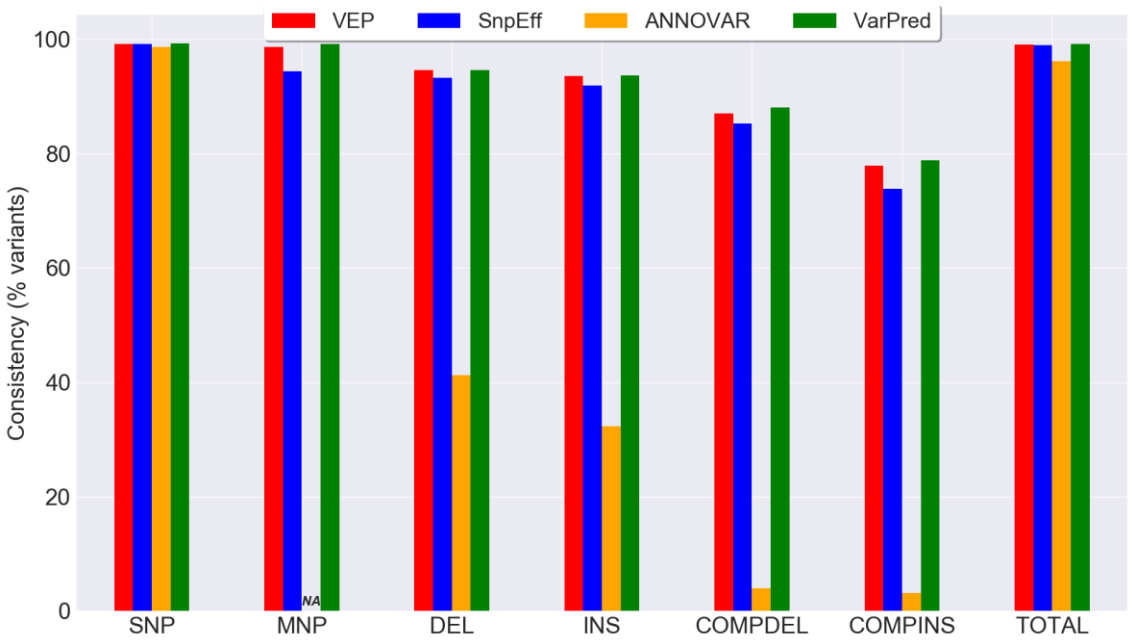
D

Tool	WES	WES dbSNP	WGS
ANNOVAR	20.3±0.1	179.1±1.2	624.1±28.1
SnpEff	88.8±0.8	157.2±0.5	2454.9±13.6
SnpEff MT	71.7±0.6	139.1±1.3	NA
VEP	66.1±0.7	1142.2±16.1	2735.3±82.0
VarPred	51.8±0.3	57.1±0.9	3066.5±28.1

Figure 10|**Time comparison among variant annotators using bar plots.** ANNOVAR June 2017, SnpEff v4.2 (common and multi-threading), VEP v89 and VarPred have been assayed. Tests on human WES VCF file (45K) **A**|using common annotation features or **B**|adding the dbSNP annotation. **C**|Test on COSMIC Coding Mutations v74 (2.9M) using common annotation features. **D**|To summarize all the results reported in the bar plots, the average running times, expressed in seconds, with their standard deviation are also numerically detailed. Common annotation features include all HGVS predictions, gene, transcript and protein IDs, SO terms and position at transcript and protein level. The NA value for SnpEff MT is due to several errors raised during the software execution, which inhibited the possibility to get reliable results.

The COSMIC v74 dataset, previously filtered to eliminate variants falling out of the coding regions, has been chosen for testing the reliability of annotation provided by the software included in the comparison. The obtained group, including about two million of variant, constitutes a good benchmark for such kind of evaluation, even if, as explained above, the predictions have to be modified in order to be compared with those calculated by the annotators. Although most of the dataset is composed by SNPs, it shows also a wide variability in the incorporated data, including complex rearrangements, thus almost completely covering the whole landscape of the possible types of annotation.

A



B

	VEP	SnpEff	ANNOVAR	VarPred
SNP	99.191	99.194	98.599	99.283
MNP	98.667	94.413	NA	99.191
DEL	94.583	93.279	41.215	94.583
INS	93.589	91.918	32.312	93.667
COMPDEL	87.052	85.259	3.984	88.048
COMPINS	77.918	73.817	3.155	78.864
TOTAL	99.027	98.962	96.165	99.118

Figure 11|**Consistency of annotation among variant annotators.** VEP v89, SnpEff v4.2, ANNOVAR June 2017 and VarPred have been tested. Consistency is displayed by the percentage of variants matching protein annotations in a filtered version of COSMIC Coding Mutations v74 based on 1964578 SNPs, 10130 MNSs, 18015 INSs, 47002 DELs, 502 Complex DELs and 317 Complex INSs, for a total of 2040544 variants. NA, protein level annotations not available. **A**|Graphical representation of the results using bar plots. **B**|Tabular format of the same values expressed in seconds.

VarPred essentially outperforms all other tools for the consistency of annotations (Figure 11A), displaying however similar results to what obtained with the Ensembl VEP, as the differences are not significative (data not shown). Indeed, VEP is the software showing the best outcomes among the published tools, followed quite closely by SnpEff, while ANNOVAR is really far from giving an accurate annotation (Figure 11B). In fact, ANNOVAR has a very high error rate on insertions and deletions, mainly due to the lack of realignment, it performs even worse with the complex variations, and finally it is not able to analyze MNPs. This is surprising as ANNOVAR is the most widely used variant annotator counting more than twice the number of citations of VEP or SnpEff.

VarPred is really competitive and precise, as it is able to correctly annotate more than 99% of variants, showing the lowest performance with the complex insertions which is however near to the 80% (Figure 11). It should be pointed that the overall annotation quality could be even better, as not always the COSMIC predictions seem to be corrected: in fact, in such situations VarPred and VEP often show concordance among their computed values, which instead differ from what reported in the benchmark. Another source of error is related to the different transcript set used for annotating COSMIC: in fact, a certain number of variations are defined as coding, when instead with the new release it has been shown their non-coding origin. Finally, the COSMIC annotations in some occasions are hardly convertible into the right HGVS codes, causing a slight decreasing of the performance rate, even if such difference is not significant.

Taking into account all the issues discussed in the introduction, VarPred solved all of them with the exception of the correction of adjacent SNPs in order to create a MNP, without using the information contained into the BAM file. Theoretically the only case of difficult interpretation is when both variations are heterozygous, while in the other two cases, both homozygous and one heterozygous and one homozygous, are quite easy to analyze. The above problem can be assessed starting from the analysis of the genotypes and the coverage values of the two variants. In fact, unless the two positions are exactly covered by the same number of mutated and normal reads, through the analysis of linkage disequilibrium between variants coverage, it should be possible to assign a confidence for defining if the two variations fall in the same read or not. Another problem of such method derives from intrinsic difficult to extract the coverage information from VCF files: indeed, the fields for including the coverage values are not standardized among the VCF files

released by the various software, and furthermore some variant callers report only genotypes and the associated reliability of the computed values, without including the necessary data. In these last cases, it is not possible to fuse the SNPs into a unique MNP. However, the integration of this functionality within VarPred should allow to ensure the best level of annotation among the publicly available tools, although its consistency is already the highest one.

3.2.4. Conclusion

Thanks to its implementation, VarPred is a valuable tool for a fast and really accurate annotation of VCF files derived from NGS project. It can be employed not only in human or model organisms, but essentially in all species for which a reference genome and a gene annotation are available. This is possible as VarPred, in contrast with the other published tools, does not apply predigested files, produced by the developers, to perform the annotation, but it starts from the source files directly downloaded from the archives. Moreover, it is able to extend the annotation to all the transcripts included into the adopted gene-set, accepting data in both GTF and GFF3 format and from various repositories, including Ensembl, GENCODE and RefSeq. The reliability of annotation is guaranteed by the normalization system integrated into the software and by the choice of a standard output format. Also the web interface is almost completed, although some improvements and an extensive testing should be performed. Nevertheless, VarPred shows the highest consistency of annotation (>99%) among the published tools. Taking into account all these features, VarPred can be considered ready to be released to the scientific community.

3.3. QueryOR: a comprehensive web platform for genetic variant analysis and prioritization

This section contains a slightly modified version of the article that I published in April 2017 on the BMC Bioinformatics journal [178]. The paper is accessible in its final format following the link https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1654-4. The figure numbers have been changed to be sequential with the enumeration of the thesis.

3.3.1. Introduction

Over the past few years, the advances in DNA sequencing technology have opened new perspectives in many fields of Life Sciences. In particular, Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) are contributing to the extraordinary progress in the study of genetic variants, improving the understanding of causative genes in human disorders.

While “Next Generation Sequencing” (NGS) is making the production of sequencing data progressively easier, bioinformatic analysis is still a problem when dealing with genes and pathologies not well characterized at the molecular level.

The initial bioinformatic steps for variant analysis are quite standard: the NGS reads are firstly aligned on the human reference genome [201], then the resulting SAM file [202] is parsed for the identification of genomic variants. As a result, a Variant Call Format (VCF) file with the list of variants is generated [182].

The selection of candidate variants responsible for the phenotype or disease under study remains a challenging task. Firstly, we need to functionally characterize and annotate the large number of variants that are typically detected: tens of thousands for WES and millions for WGS. Several approaches have been developed to accomplish this task. Programs like SIFT [203] and PolyPhen-2 [204] evaluate variants by focusing on the impact of amino acid changes on protein function, while ANNOVAR [163] extends the functional annotation considering other features such as phylogenetically conserved regions and allele frequency in populations.

Once the variants have been annotated further action is required to choose the most effective criteria for “prioritizing” candidate causative variants. It is unfeasible to conceive an all-purpose protocol as the type of problems and the available data may be very disparate. Moreover, field-specific expertise may be essential both in the definition of the criteria and in the interpretation of the data.

If the genetic disease is well characterized at the molecular level, then the obvious action to take is to focus on the variants occurring on known causative genes. Unfortunately, our knowledge is still limited as ~50% of Mendelian monogenic diseases have not yet been associated with causative genes [205], while most polygenic disorders remain uncharacterized at the molecular level.

Taking into consideration that the function of many genes is still unknown, bioinformatic approaches such as Endeavour [206] prioritize candidate genes on features shared with other genes that are involved in the same biological process or disease under study. Several phenotype-driven approaches have been implemented in programs like eXtasy [207], PhenIX [208], Phenolyzer [209], PHIVE [210], Exomiser [211] and Phevor [212], taking advantage of resources such as Gene Ontology (GO) [213], Human Phenotype Ontology (HPO) [214] and Disease Ontology (DO) [215].

As previously mentioned, the prioritization process usually requires the integration of a wide range of functional information about variants, genes and diseases as well as mode of inheritance when multiple individuals are considered. Currently, the standard strategy involves the application of filters with arbitrary thresholds that progressively remove variants not satisfying the criteria. As a result, there is the risk of removing something that is just below the threshold for one of the criteria, while being well above the threshold for the other criteria.

Prioritization is not only confined to the problem of merging information on variants, genes and phenotypes. An issue that is often disregarded is that the clear majority of genes undergo alternative splicing [216]. As a result, the same variant may have very different functional outcomes, for instance it may generate a stop codon in a transcript and a silent variant in another isoform of the same gene. For this reason, the annotation of variants should refer to each alternative transcript rather than the putative major isoform.

Recently, some web-servers [217] have been developed to analyze exome data, but they do not satisfy most of the above requirements, thus limiting the spectrum of possible analyses. Stand-alone programs such as VariantMaster are available [218], but they are driven by line-commands that make their usage cumbersome and difficult for most users. An additional problem is that our knowledge on human genomics is changing very rapidly at all levels, needing continuous updates, implementations and integration of data, tools and ideas. Therefore, a platform for prioritization that combines usability and comprehensiveness has become a priority.

With these premises in mind, QueryOR has been engineered as a user-friendly web-platform that integrates the most advanced prioritization criteria. Furthermore, QueryOR is built on a robust set of XML-defined rules that allows an easy implementation of new criteria without modifying the program code. Currently, 70 different criteria of prioritization have been

implemented in the platform and can be selected by users to build dynamic tailor-made queries and to facilitate expert-driven variant and gene prioritization.

QueryOR is freely available for academic institutions at http://queryor.cribi.unipd.it/.

3.3.2. Materials and Methods

Web-interface implementation

QueryOR has been implemented in CGI/Perl combined with Apache web-server. JavaScript, Jquery, AJAX and CSS properties are used to dynamically render some parts of HTML pages and to define their structures and layouts. The pages for criteria selection and transcript report are built on dedicated XML-files. For this reason, we have developed a XML-language that describes standard database queries and their web representation (layout, form elements, hyperlinks, highlighted columns). Thus, any selection criterion or transcript data table is completely specified in a XML node, making the system flexible and scalable. The XML language also allows the user to integrate custom databases into the QueryOR platform. This integration is easily obtained loading multicolumn files with information related to genes (one column must contain the ENSEMBL gene ID) or variants (four columns are mandatory: chromosome, position, reference allele and alternative allele). Once the file is loaded, the user can select the fields on which one or more filters have to be created. Then, the system automatically fills a new database associated with the project and builds specific XML-files containing the new queries, which will be available with all other criteria.

Data processing implementation

The data processing step is based on in-house scripts developed in Perl, Python and Bash; it runs on a blade cluster, managed by a PBS job resource manager (TORQUE). ANNOVAR software and dbNSFP database (v2.9) [219] are used for the annotation of variants, in addition to a homemade script. All project data are stored in a local database using MariaDB, a MySQL open-source fork, with the TokuDB® engine. The database is designed to contain both annotation tables and user data tables. The former host human gene annotations and known SNP information (global minor allele frequency, clinical significance, etc.) and are regularly updated every 6 months. The latter stores the data uploaded by the user and the associated meta-data produced during the “Data processing” step.

ENSEMBL data and variant annotation integration

The hg19 release 81 of human gene and transcript data has been downloaded from ENSEMBL (http://grch37.ensembl.org/info/data/ftp/) [220]. Two different databases of known mutations have been integrated in the platform: dbSNP [71] version 144 (http://www.ncbi.nlm.nih.gov/SNP/) [221], modified to recover old variants excluded from this last release but present in the online version, and Exome Variant Server version ESP6500SI-V2 (http://evs.gs.washington.edu/EVS/) [154] have been chosen to annotate allelic frequencies in the population. Disease information has been obtained from OMIM (http://www.omim.org/) [154, 222] and associated to gene and transcript data. Regarding somatic mutations, QueryOR incorporates COSMIC database [223] version 74, whose SQL table has been created starting from VCF files containing both coding and non-coding mutations and the complete export file of COSMIC. In case of new releases of gene annotations, dbSNP files or OMIM data, a custom set of Perl/Python scripts have been developed for the automatic update of all SQL tables.

Integration of functional and phenotypic annotations

QueryOR integrates several gene annotations derived from different public databases, which have been directly obtained from their respective websites or through ENSEMBL BioMart [243]. Within these annotations, QueryOR embeds Gene Ontology [171] and InterPro [224] data, as well as two different pathway repositories, KEGG (Kyoto Encyclopedia of Genes and Genomes) [225] and Reactome [177], which have been collected using the Graphite package [226] of Bioconductor [227]. QueryOR also makes available gene expression data derived from the GTEx portal (version 6) [174]. The information contained in this atlas has been processed to link Ensembl ID to tissues and sub-tissues in which the gene is expressed. The level of expression is measured in RPKM (Reads Per Kilobase per Million mapped reads) [228]. Moreover, regarding the phenotype annotation, the platform accommodates two main databases: DisGeNET version 3.0 [229] and Human Phenotype Ontology (HPO) release 98, whose entries have been further processed to be associated to ENSEMBL-ID. The updating of these functional annotations has been automatized through a set of Perl/Python scripts as described in the previous section.

Chromosome map tool implementation

The "runs of homozygosity" (ROHs) are calculated by comparing the user-uploaded variants and the high-polymorphic dbSNP variants (GMAF higher than 0.3) falling into the target regions. The algorithm extracts those positions where only dbSNP variants, and no custom variants, are mapped. The resulting locations are those with a homozygous genotype for the reference allele (0/0) in the analyzed sample.

Using these spots, the script finds all the ROHs, computes the length distribution and selects the stretches whose length exceeds the 95th percentile of the distribution. Then, the algorithm tries to extend all the ROH seeds in both directions as long as the homozygosity ratio (number of positions with 0/0 genotype divided by the sum of homozygous and heterozygous positions in the considered region) remains above 0.9. ROHs are used to build the "chromosome map" chart in association with the genes selected during the prioritization process.

Case study dataset

The exome data from the "Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability" (study EGAS00001000287, https://www.ebi.ac.uk/ega/studies/EGAS00001000287) [230, 231] were obtained from the European Genome-Phenome Archive (EGA) website.

3.3.3. Results

We have implemented QueryOR dividing the process into three main steps as shown in Figure 12. Each step is further divided into different sub-steps and procedures, as detailed below. Users will spend most of their time at step 3, querying and browsing the system in the search of possible causative variants. To test the potential and features of the querying step, several sets of data have been made openly available on the platform, including some trio data from de Ligt *et al.* [230], as well as data produced by our own group.

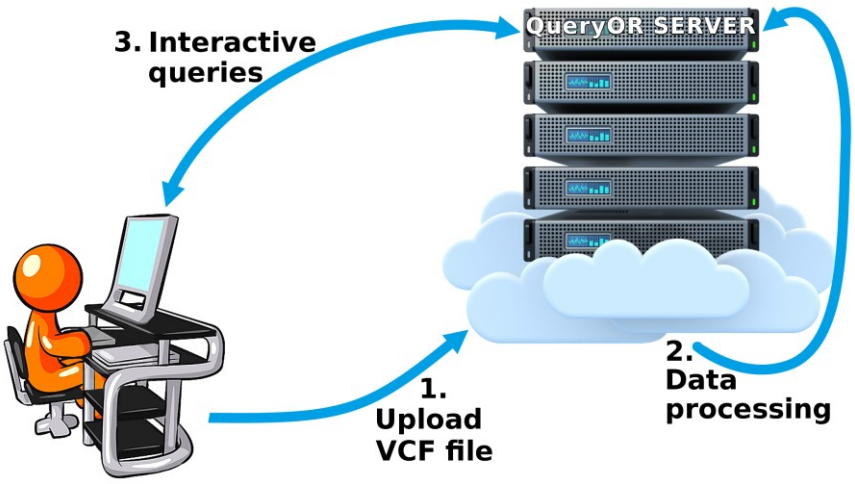


Figure 12|**The three main steps of QueryOR analysis.** Step 1 and step 3 require interaction with the user, whereas step 2, data processing, is automatically performed by the system after uploading VCF files.

Uploading and updating VCF files

All QueryOR's activities are centered on projects that the users can create and possibly share with their collaborators. Projects can be related to single individuals, trios or families, as well as population or cohorts. Starting a project is very simple, but users must first register, both for privacy reasons and for permitting the retrieval of their data.

The creation of a project requires the uploading of VCF files that must satisfy several requirements. Firstly, each individual sample should be labeled with a unique name that will be used as identifier in the subsequent steps. Secondly, the information about genotype, allelic depth and total read depth, which are usually found in the GT, AD and DP fields, must be available. Although VCF is a well-established format, not all variant callers implement the VCF fields in the same way; for instance, the Torrent Variant Caller does not fill the AD and DP fields. Therefore, we have developed specific scripts that calculate the allelic and total read depth from other parameters, such as Alternate allele Observations (AO) and Reference allele Observation count (RO). As a result, the platform accepts VCF files produced by all the commonly used variant callers.

In the upload/update step the user can also upload BED files containing regions of interest. BED files should have four columns for each row: chromosome number, starting position, ending position and sample ID; the latter is used to associate the genomic coordinates to the right individual. These custom-defined regions will be shown in the graphical synopsis of variants and transcripts (Fig. 13-Q3) as yellow boxes. We usually exploit this feature to mark on each sample the regions with low coverage.

Once the files are uploaded, QueryOR takes some time, from minutes to hours, to process data, depending on the number of uploaded samples and variants. The user can check the job status while the processing is running. The beginning and the end of the process are notified by automatic emails to the user's registered address.

Data processing

Data are processed by an automatic back-end procedure that provides a comprehensive annotation of the variants, linking them to genes, transcripts, encoded proteins and biological ontologies. QueryOR takes into consideration that alternative splicing may generate multiple transcripts from the same gene. As a result, a variant may have different effects depending on the transcript isoform. With this premise, we thought that the common practice of limiting variant annotation to the major transcript isoform is a coarse approximation. Therefore, to manage this problem QueryOR annotates variants on all the predicted ENSEMBL transcripts derived from alternative splicing events. Furthermore, the distribution of variants on the different splicing isoforms can be displayed and examined by the user as a part of the interactive result analysis described in the next paragraph.

Besides QueryOR's own procedures, a further double annotation is performed using both ANNOVAR [163] and dbNSFP [219], thus obtaining a wide set of measures, scores and constraints related to each variant, that among others include SIFT [155], PolyPhen [204], MutationAssessor [232] and GERP++ [160].

Data processing involves many other steps, including the association of variants to the available information in dbSNP, such as the allelic frequency in the global population and in ethnic groups, as well as the presence in the 1000 Human Genome Project [233]. Moreover, we discovered several thousand SNPs in the reference genome (both in GRCh37 and GRCh38) that do not correspond to the major allele in the population and as a consequence are found as "false positive" in most individuals. To overcome this problem, the reference positions characterized by a dbSNP frequency lower than 0.1 are annotated as MAiRs (Minor Allele in Reference).

When a project involves the analysis of multiple patients such as trios and families, the platform runs a specific module that automatically computes how variants are shared between individuals. Moreover, possible Runs of Homozygosity are calculated for each sample, as explained in the Methods section.

All the retrieved and computed information obtained by the data processing step is stored in the QueryOR database.

The overall time required for loading and processing data is approximately proportional to the number of variants. Typically, for ~100,000 unique variants (6-8 exomes) the time required is less than 20 minutes. A more detailed analysis of the loading time is given in Figure S1.

Interactive queries and results analysis

After the completion of data processing, the user can explore the information that has been associated to the project, following the general procedure shown in Figure 13. Queries can be formulated very easily and the resulting answers are typically delivered in a few seconds that can extend to minutes for very complex queries. Thus, it is possible to experiment different criteria and parameters, to perform a comprehensive investigation and to get progressively closer to possible causative genes. A detailed analysis of the querying time, as a function of the number of criteria and variants can be found in Figure S2.

The complete route from query to variant takes five progressive steps that correspond to pages appearing on the web browser, labelled Q1 to Q5. At each step, some decisions must be taken: Q1 is for the query, Q2 is for choosing a gene from the resulting list, Q3 is for the selection of a specific transcript among the different isoforms, Q4 corresponds to the transcript report where a certain variant can be chosen and Q5 is the description of the variant. Like being in a maze, you may explore some paths and you can go back if the route leads to a dead end. In the web browser, Q1 to Q5 will open as independent pages making it easy to return to any of the previous steps. Some integrated QueryOR tools are associated to different points of this route, to make decisions easier. The main features of this process are described in the following paragraphs.

Figure 13|**Common analysis in QueryOR.** A typical route for a QueryOR investigation starts with the selection of criteria (Q1); a ranked list of genes and variants is returned in Q2. The selection of a gene, for instance FUCA2, leads to page Q3 where variants and affected transcripts in their genomic context are shown. The black track at the top of Q3 shows the target regions of exome capturing. The blue tracks just below show that the analysis was done on two samples named APN and APM, that share a heterozygous variant (white circles). The yellow boxes report the positions specified in the optionally uploaded BED file, indicating for instance low coverage regions. The bottom part of Q3 shows two alternative transcripts where the same variant in one case is located in an exon, generating a missense substitution (dark blue circle) while in the other case is located in an intron (gray circle). By clicking on a transcript of Q3, the system returns Q4, where several transcript features are directly linked to external resources, as well as to the variant overview page (Q5). For a full list of symbols used in Q3, see Figure 14. A more detailed description of the entire process is given in the main text.

Query procedure (Q1) - Page Q1 allows the user to select the criteria for prioritization that are grouped into seven main sections. Three sections (ENSEMBL Features, Functional Annotation and Phenotype Annotation) are related to genes, pathways and phenotypes. In these sections, it is possible to select for specific lists of genes and transcripts as well as features like gene ontology, gene expression and associations to pathways, diseases or phenotypes. The remaining four sections are related to variants. These include Variants Annotation (for instance genomic context and functional prediction scores), Variants Databases (for instance dbSNP, EVS and COSMIC), Variants Sharing and Segregation (variants in homozygosity and/or heterozygosity present or absent in different individuals) and VCF User Data (for instance variant coverage, genotype and quality calls).

Each section can be exploded to visualize sub-sections that can be further expanded to see the selectable criteria. Figure 13-Q1 shows a query page where the section Variants

Databases shows four sub-sections and where the last sub-section (dbSNP) shows six selectable criteria. The selected criteria are shown on the right side of frame Q1 where GMAF is under definition, while other 7 defined criteria are shown in their "collapsed" view.
By default, all criteria have the same relevance in the ranking process, but this can be modified by assigning different weights to each criterion. There are no restrictions in the number of selected criteria, but very complex queries may take a longer processing time.

Engine (Q2) - When a query is submitted, the system performs an independent search for each of the selected criteria; then, the score of each variant is calculated as the sum of the weights of the satisfied criteria. Finally, genes are ranked according to their highest-score variant. The results from the query are summarized in a score table (right part of Figure 13-Q2) that shows the number of genes and variants associated to each score. The two top-scores shown in the right side of Figure 13-Q2 were selected and expanded to produce the results matrix on the left, where each row reports a single gene combined with the number of variants satisfying the prioritization criteria.
By clicking on a gene name in the results matrix, more details show up. For instance, the image in Figure 13-Q2 was taken after expanding FUCA2 and BPIFB3. This feature is useful to better understand the results. In fact, although the first six genes have positive variants in every column, as shown by the blue background, only 2 genes satisfy all the 8 selected criteria, resulting in an associated score of 8. This apparent incongruence can be explained by looking at the expanded data of BPIFB3, showing that although the gene has some variants satisfying all the criteria, the two best variants satisfy only 7 criteria.
From the bottom line of Q2 (Total Number of Variants) it is possible to appreciate the depth and the stringency of each filter and to make a general evaluation of the prioritization. Thus, the user can reconsider some of the criteria and go back to Q1 to redefine the query.

Gene overview (Q3) - This page is shown after a gene is selected by clicking on the Gene-ID, in the results matrix. The page displays a compact graphical representation of alternative transcripts associated to the selected gene, together with the position and type of each variant across all samples. In Figure 13-Q3, two samples named APM and APN are shown at the top of the frame. Both samples share a heterozygous variant, represented by the white dots. The bottom part of the Q3 frame shows two alternative transcripts in which the same

variant acts as a missense mutation (dark blue dot) in one transcript and as an intronic mutation (gray dot) in the other.

In the case of trio studies, samples are differently tracked to highlight parental heritage of allelic variants (haplotype configuration), as shown in Figure 14.

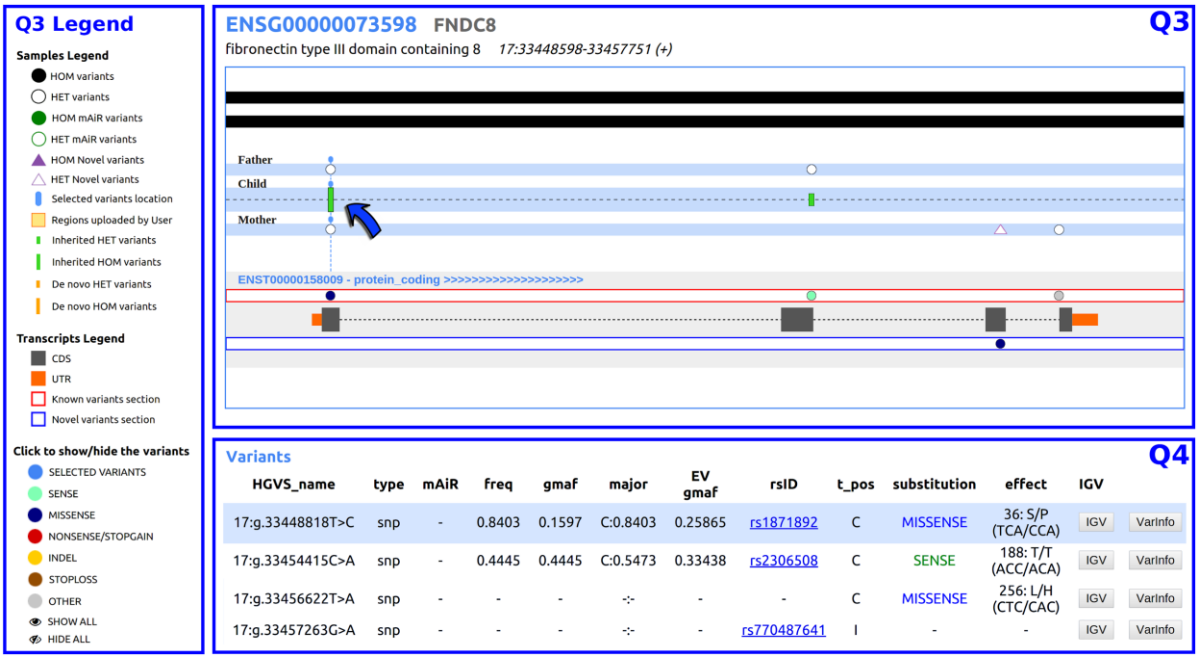


Figure 14|**Trio analysis.** In the Q3 section, the arrow points to a variant that is heterozygous in both parents and homozygous in the child (full green bar). At the end of the next exon, the child displays a heterozygous variant, shown as a small green bar, which was directly inherited from the father. A detailed description of the variants is given in the Q4 section where the user can also find a link to the IGV viewer, that will be conveniently opened on the appropriate genomic position.

Transcript report (Q4) - Detailed information about the transcript selected in Q3 is shown in Q4 (Figure 13 and Figure 14), where various contents are briefly described and directly linked to their primary source on the web. The variants that emerged from the prioritization process are highlighted with a blue background. If the BAM file is available on the client side, the user can consider to launch IGV [234] that will automatically point to the position of the variant under analysis to view the alignment of the reads on the genome. By the "Varinfo" button the user can move to Q5.

Variant overview (Q5) - This page allows the evaluation of the specific features of the candidate variant (Figure 13-Q5) where several pathogenicity scores are accessible, including the above mentioned PolyPhen and SIFT, as well as Mutation Taster [235], CADD

[156] and DANN [157]. Although these features are sometimes discordant, it is useful to have a global view to estimate the possible pathogenicity of the variant under analysis.

Advanced analyses - From page Q2 it is possible to access other QueryOR's tools such as the "Variants Report" that is a printable table summarizing the information on variants, genes and pathogenicity. Another link builds a "Chromosome map" reporting possible Runs Of Homozygosity, that can be important in the analysis of human disorders, as they represent a good clue for the presence of deleterious variants responsible for recessive diseases [236]. A further link takes the user to the "Gene Analysis tool" that allows the identification of genes carrying different mutations among a group of patients. With this tool, it is possible to investigate unrelated patients or to investigate diseases caused by de novo mutations, where it is more important to know if the same gene is mutated in different patients rather than if they share the same variant. This information comes as a summary table flanked by a distribution chart (data not shown). Each group of genes can be further investigated searching for shared biological terms, using DAVID [237], or for common pathways within Reactome [177] and KEGG [176].

Case study

To evaluate the performance of the platform we re-analyzed some of the data published by de Ligt et al. [230], (EGA study EGAS00001000287), concerning patients affected by recessive forms of cognitive impairment and mental retardation. Our prioritization strategy was achieved by applying several criteria on trio number 4 (VCF files EGAZ00001004509, EGAZ00001004510, EGAZ00001004511). In particular: 1) we selected high confidence variants with coverage level >60 and 2) with alternative allele coverage >30; 3) we only considered variants that changed the amino acid sequence; 4) as the disease is rare, we imposed a low frequency threshold with MAF<0.05; 5) the results were further fine-tuned by considering the "intellectual disability" Phenotype Ontology keyword; 6) taking into consideration the pattern of inheritance, we selected variants that are homozygous only in the child. QueryOR identified only two variants that could satisfy these six criteria. Interestingly, one of the two is a missense variant placed in the PDHA1 gene, in the X chromosome, corresponding to that proposed in the aforementioned work [230]. It is interesting to point out that with only six criteria it was possible to achieve a very effective

prioritization. The above case is fully explained in a tutorial available at http://queryor.cribi.unipd.it/cgi-bin/queryor/tutorial.pl. To prevent any incidental findings and to preserve patient's privacy, the tutorial is based on the exome of a healthy patient, manually edited to insert the above variant.

3.3.4. Discussion

It is normal that when a new technology starts to produce novel types of data, the development of software analysis runs a little behind and eventually catches up. In the case of Whole Genome and Exome Sequencing this problem is particularly relevant because the scope of the prioritization process is not limited to the variants as such, but it extends also to a wide variety of data and information that is continuously updated and is often superseded by new discoveries.

When we started the development of QueryOR, this context of generalized "work in progress" was one of our main concerns. Prioritization is essentially a process of data integration and to develop it using unstable datasets would be a vain effort. On the other hand, we thought that a user-friendly variant-prioritization platform, suitable for a wide range of analyses, could be of great utility. To overcome the problem of sustainability, QueryOR has been designed on a general schema rather than on predefined databases. A dedicated XML language permits the declaration of the datasets to be implemented in the platform. Each dataset is defined for its content, for the possible queries and for their web representation (layout, form elements, hyperlinks, highlighted columns), thus making the system flexible and scalable.

Thanks to this flexibility many datasets are available in the platform while more will be added in the future. Although a query could be potentially made by selecting different features from all the available datasets, in a normal session only some of the data will be interrogated. Thus, there is a double level in which the information is organized: at the basal level, there are all the available datasets implemented by the QueryOR manager, while at the top there is the information emerging from the queries performed by the end-users.

		Features	QueryOR	SeattleSeq	wANNOVAR	VEP	BierApp	PhenIX	OVA
Data uploading		vcf support	✔	✔	✔	✔	✔	✔	✔
		multisample vcf	✔	✔	✔	✔	✔	✔	✔
		multiple vcf	✔	✘	✘	✘	✘	✘	✘
		custom features	✔	✘	✔	✘	✘	✘	✔
Support for filtering and prioritization		variant annotation	✔	✔	✔	✔	✔	✔	✔
		progressive filtering	✔	✔	✔	✔	✔	✘	✘
		overall prioritization	✔	✘	✘	✘	✘	✔	✔
		no. of available criteria	70	5	23	54	29	3	15
		system preset query	✔	✘	✔	✔	✘	✔	✔
		custom preset query	✔	✘	✘	✘	✘	✘	✘
		alternative transcripts effect	✔	✔	✔	✔	✔	✔	✔
		links to external resources	✔	✔	✔	✔	✔	✔	✔
		users' provided database	✔	✘	✘	✘	✘	✘	✔
		filtering on sample subsets	✔	✘	✘	✘	✔	✘	✘
Main prioritization criteria	Gene level	gene ID, symbol, description	✔	✘	✔	✔	✔	✘	✘
		transcript ID, symbol, class	✔	✘	✘	✔	✘	✘	✘
		functional annotation	✔	✘	✘	✘	✘	✘	✔
		phenotype annotation	✔	✔	✔	✔	✘	✔	✔
	Variant level	codon impact	✔	✔	✔	✔	✔	✘	✔
		allelic frequency	✔	✘	✔	✔	✔	✔	✘
		minor allele in reference (MAIR)	✔	✘	✘	✘	✘	✘	✘
		coverage	✔	✘	✘	✘	✘	✘	✘
		genotype	✔	✘	✘	✘	✔	✘	✘
		mendelian inheritance	✔	✘	✔	✘	✔	✔	✔
		homozigosity sharing	✔	✘	✘	✘	✔	✘	✘
		variants sharing	✔	✘	✘	✘	✘	✘	✘
Result output		interactive analysis	✔	✘	✔	✔	✔	✘	✘
		hypertextual html report	✔	✔	✔	✔	✔	✔	✔
		plain text file	✔	✔	✔	✔	✘	✘	✔
		homozygosity map	✔	✘	✘	✘	✘	✘	✘
Support info		tutorial	✔	✘	✘	✔	✔	✘	✘
		manual	✔	✔	✔	✔	✔	✔	✔
		trial data/results	✔	✔	✔	✔	✔	✘	✘

Figure 15|**Comparison of QueryOR with other platforms for variant prioritization.** The platforms were tested using a VCF input file. The indicated number of available criteria is approximate due to different ways of implementation.

In literature, several bioinformatic tools for whole exome analysis are reported, but only a few of them are suitable for a comprehensive and efficient exome investigation. In fact, while some platforms center their analyses on gene features found in biological ontologies, others focus primarily on variant annotations, disregarding gene function. In QueryOR we

combined the most useful features found in other tools, gathering and expanding them within a single platform. Moreover, to enhance the potential of the analyses, we implemented some important features such as the annotation of minor alleles in the reference genome, several prioritization criteria based on VCF information such as coverage, genotype and quality score, as well as criteria based on sharing variants and homozygosity in different individuals. Furthermore, we introduced the possibility to implement customized prioritization criteria based on databases supplied by the user. A detailed description of the procedure for submitting custom tables is given in the User Manual, available in the "Info" section of the web site. Figure 15 compares the main features of QueryOR with other available tools, including SeattleSeq [238], wANNOVAR [163], VEP [162], BierApp [217], PhenIX [208] and OVA [239].

To our knowledge, QueryOR is the open web tool with the widest spectrum of applicable criteria (currently 70) for exome data prioritization, spanning from gene and variant annotations, to intrinsic features of the VCF file. Another interesting peculiarity of QueryOR regards the opportunity to select a subset of samples within a multisample project, allowing focusing on attributes found only in the chosen group of samples.

A major effort has been made to simplify the formulation of complex queries. To perform a query the user can select any combination of criteria and associated parameters. For instance, one of the criteria could be the minimal coverage of the locus where a SNP occurs and the associated parameter could be "30". Criteria can be classified in three main categories. The first group is based on the knowledge of genes and diseases, exploiting the functional and phenotypical annotation integrated in QueryOR as well as lists of candidate disease genes when available. The second category discriminates variants on the basis of information contained in the VCF file including coverage, genotype and quality of calling. The third category is related to variant features, such as pathogenicity scores, effect on protein, population frequency and distribution among the project samples. In particular, it is possible to impose a specific inheritance model in trios as schematized in Figure 16, or families and cohorts, allowing for instance the selection of variants shared or not shared among different patients or that are homozygous in some patients and heterozygous in others.

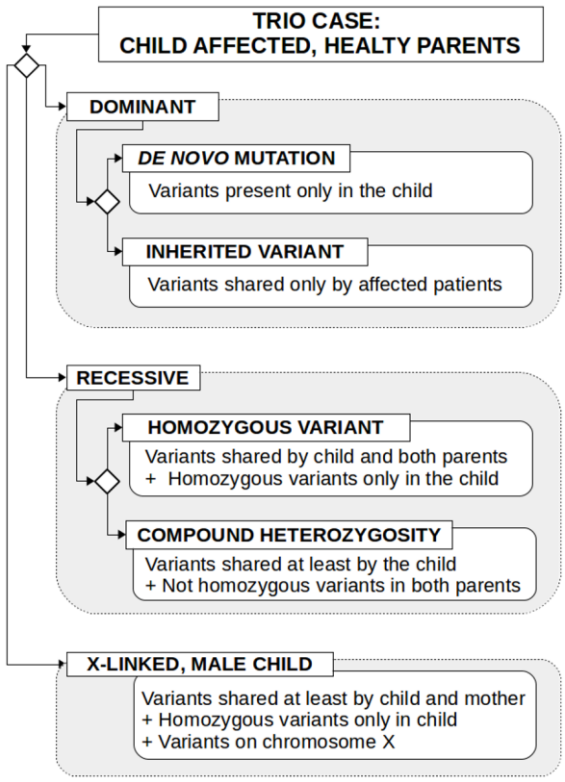


Figure 16|**Usage of the criteria for "shared" and "homozygous" variants in a trio case.** Diamonds indicate different hypotheses that can be made. For instance, if we hypothesize a recessive homozygous variant in the child we should set two criteria: 1) shared variants by child and both parents and 2) homozygous variants only in the child. Whereas, for a compound heterozygosity we would expect that the child shares the variants, but we do not know which variant is in which parent; furthermore, the variant should not be homozygous in the parents. Compound heterozygosities are generally difficult to find and criteria based only on sharing and homozygosity would not be selective enough. In this case the "Gene Analysis tool" described in the text could help in the selection of genes carrying different mutations. Sometimes it may be useful to set criteria that may appear useless, like homozygosity on a X chromosome; however, this may help to reduce false positives.

In the development of the graphical user interface, we dedicated a particular attention to user friendliness, both for setting the criteria and for interpreting the results. As an example, Figure 17 shows how de novo mutations can be searched and visualized in a trio of mother, father and child.

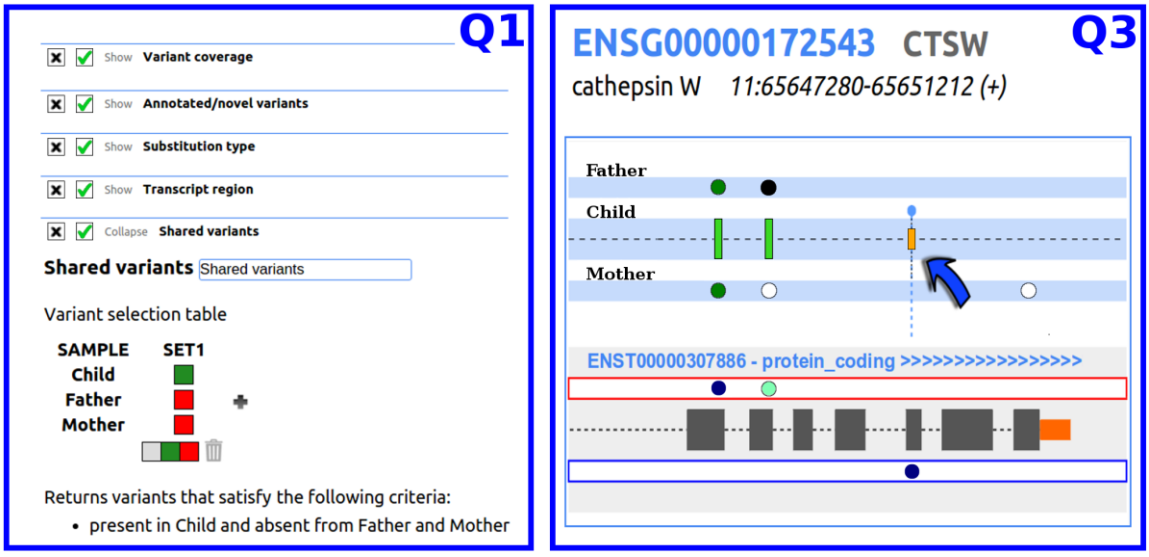


Figure 17|**Searching for de novo mutations in a trio.** Q1: to set the criteria the user should select "Shared variants" and click the box beside each patient, selecting green, red or gray respectively for present, absent and ignore, while the sentence underneath will report in plain English the meaning of the settings; more criteria can be set by clicking the + symbol. Q3: the results include haplotype phasing. The yellow bar indicated by the arrow is a possible de novo mutation. For the meaning of other symbols see Figure 14.

In contrast with other similar tools that return only the items that simultaneously satisfy all the query specifications, QueryOR sorts the results on the number and weight of satisfied criteria; thus, the user can have a global view of which criteria are or are not met for every gene and can decide whether to continue the investigation or modify the query. The integration of a wide range of heterogeneous information and the automated annotation procedure provides the end user with the ability to evaluate the information at various levels in order to establish the relationships between different data and to discriminate between causal and neutral variants.

Several other innovative features of QueryOR make the process of prioritization thorough and at the same time easy. For instance, an important issue is that we annotated all the variants that in the reference genome are represented by rare alleles, that we named MAiRs (minor Allele in Reference). These variants can either be filtered off by the query specification or alternatively they will be automatically labelled as MAiR when seen on the selected genes.

3.3.5. Conclusion

Currently, QueryOR is primarily used to analyze exomes and gene panels, however it has been successfully employed also for whole genomes. In this respect, the main problem is the lack of functional information that can be associated to variants belonging to non-coding sequences. As this information will become available we will take advantage of the flexibility of QueryOR to implement datasets that may facilitate the prioritization of variants in whole genome analyses.

In conclusion, the comprehensiveness of the implemented criteria and the aptness to add new features together with a user-friendly environment make QueryOR very suitable to support researchers, clinicians and geneticists engaged in variant analyses.

3.3.6. Supplementary Materials

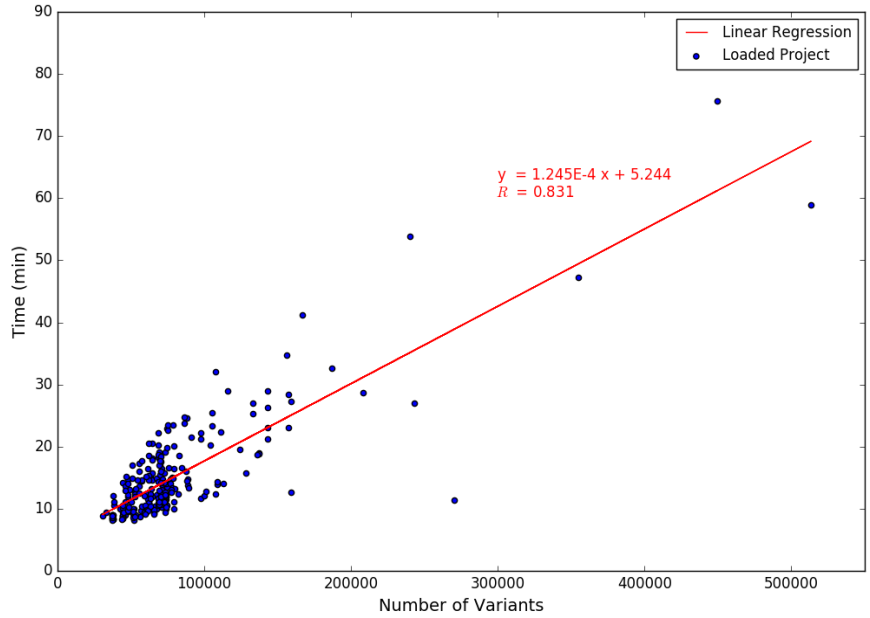


Figure S1|**Time required for uploading and processing a project.** The Figure shows the loading and processing time of about 200 projects, as a function of their number of unique variants. It can be seen that the required time is roughly proportional to the number of unique variants; however some discordance may be due to different ratios of novel/known variants, as well as to resource availability on the computer cluster.

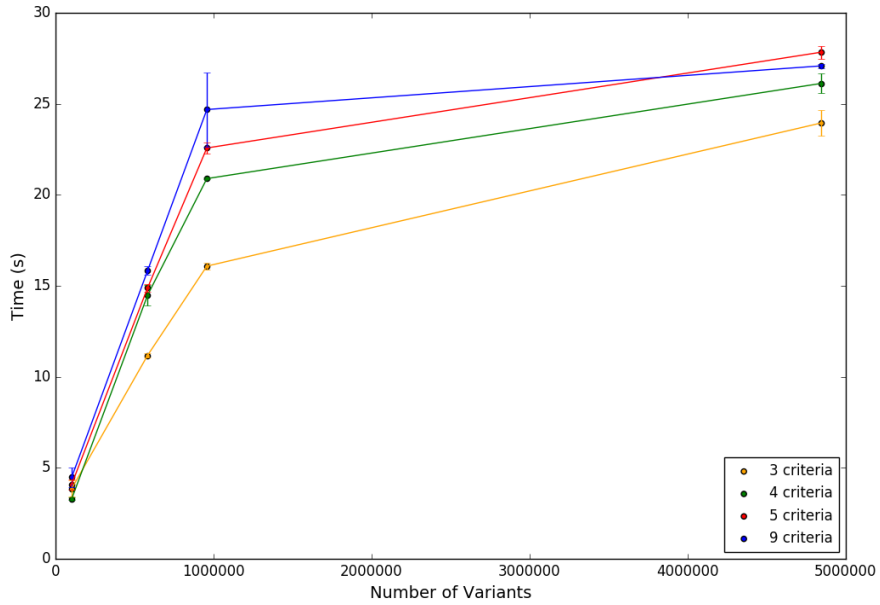


Figure S2|**Time required for the processing of a query.** The Figure shows the time required for the processing of queries with different number of criteria and with increasing number of variants. All the tests were repeated three times and the figure reports the mean and standard deviation of each point.

4. Data analysis: from clinical cases to recurrent nucleotide variants in WES studies

4.1. A targeted sequencing panel for the analysis of the exons and the conserved intronic sequences of 50 lysosomal storage disease genes

4.1.1. Introduction

Lysosomal storage diseases (LSD) are a group of monogenic metabolic disorders, each one leading to the accumulation of specific substrates due to the deficit of a lysosomal hydrolase. Although individually rare, overall incidence of LSD is estimated around 1:5000-1:8000 [240]. Affected children generally appear normal at birth and the first signs and symptoms develop during the first few months of life and progressively worsen; however, LSD can occur also as late-onset juvenile and adult forms. The diagnosis of LSD requires clinical expertise as most clinical features are not specific and could be shared by different LSD; in some cases, the diagnosis is very difficult and may take several years. The first diagnostic assessments performed are biochemical assays, useful to evaluate the accumulation of specific substrates and/or the enzymatic activity of one or more enzymes. Then molecular analysis of the suspected gene is performed to reveal the disease-causing genetic variants. This diagnostic route could be potentially reversed given the accessibility to NGS technologies which allow the simultaneous sequencing of several genes in a short time. An approach of targeting sequencing could be the primary screening tool in the diagnosis of LSD thus drastically shortening the time from the onset of first symptoms to the diagnosis formulation. In this study, we evaluated a LSD-genes targeted sequencing panel as a potential diagnostic tool for these disorders. The peculiarity of such panel derives from the inclusion of conserved intronic fragments (CIFs) within the design in order to detect mutations playing their pathogenetic role without directly affecting the coding region.

4.1.2. Materials and Methods

Genes selection and panel design

For genes selection we evaluated the Orphanet list of LSD, the SSIEM LSD list and the genes list reported by Fernandez-Marmiesse and colleagues for their panel design [80]. Genes associated with extremely rare disorders or disorders presenting with a very peculiar phenotype were removed from the list. The Ion AmpliSeq™ platform (Thermo Fisher Scientific®) [241] was used for the design of a custom panel including the selected genes. For each protein-coding transcript the exons, a 50 bp flanking sequence on each side and both UTRs were given to the Ion Ampliseq™ Designer software as target sequence. Moreover, we included the Conserved Intronic Fragments (CIF) obtained using a homemade pipeline developed *ad hoc*. To extract such regions, we used several kinds of data: a base-wise conservation score (phastCons) calculated on the basis of multiple alignments of 33 placental mammal genomes to the human genome [242], a list of common gene names and two gene annotations, RefSeq (version January 2015) and Ensembl GRCh37.75 [243].

Samples selection

A total of 80 samples were collected from different European Clinical and Diagnostic Centers and from the Telethon "Cell Line and DNA Biobank from Patients Affected by Genetic Diseases" [244]. 59 were positive controls (PC); 12 of them belonged to patients who were diagnosed only through enzymatic analysis (biochemically diagnosed: BD). 9 out of 80 samples came from suspected LSD patients for which a diagnosis had not been formulated yet (UD).

Library construction, enrichment and sequencing

DNA library preparation was performed according to the Ion AmpliSeq™ DNA Library Preparation protocol (Thermo Fisher Scientific) in combination with the Ion AmpliSeq™ Library kit, version 2.0. For the Ion Proton sequencing, libraries were loaded into a Ion PI chip and sequenced using the Ion PI HiQ Sequencing 200 kit. The produced reads were mapped using the Torrent Mapper (TMAP) and variants called by the Torrent Suite Variant Caller (TSVC). On a subset of BAM files (n=30), corresponding to the initial two sequencing runs (8+22), a coverage study has also been carried out to understand inter-runs and inter-samples coverage variations, using principal component analysis (PCA) and density plots.

Variant analysis and prioritization

Variant prioritization was performed by using QueryOR [178]. First an accurate selection of the most suitable criteria capable to choose in the positive controls the variants known to be associated with the pathology was done. Hence for each sample the analysis was splitted in three distinct queries aiming to prioritize: missense, nonsense and sense variants (query 1), frameshift, inframe, stoploss and stopgain variants (query 2) and splicing-affecting variants (query 3). In case of no appreciable results obtained with the first three queries, an additional optional forth query was performed with the aim to prioritize the 5' UTR, 3' UTR and intronic variants (Table 5). Finally, in case of no results, the same queries were re-launched removing the coverage filters to detect poorly covered variants. When the presence of long indels was suspected, a copy number variation (CNV) study using a in-house software, specially designed to detect amplicons whose coverage significantly differs from what found in the same amplicons in the other samples of the same run, was carried out. All the variants identified were verified for coverage and chromosomal position through IGV [234] and annotated using the HGVS nomenclature through Ensembl VEP [162] or VarPred (§3.2).

Filter Type	Query 1	Query2	Query 3	Query 4 (optional)
Allele frequency	<0.01	<0.01	<0.01	<0.01
MAiR	no	no	no	no
Substitution type	Missense, nonsense, sense	Frameshift, inframe, stoploss, stopgain	-	-
Transcript region	-	-	Acceptor site, donor site	5'UTR, 3'UTR, intron
CADD phred score	>10	-	-	-
DANN score	>0.8	-	-	-
Alternative variant coverage	>10	>10	>10	>10

Table 5|**Prioritization strategy for LSD samples.** Description of the four queries and the relative criteria used for variants prioritization through QueryOR platform.

Analysis of Conserved Intronic Fragments (CIF)

The intronic variants located in the CIF included in the panel were filtered for frequency <0.01 by QueryOR and analyzed using different tools. SPANR (Splicing-based Analysis of

Variants) [245] was used to predict both intronic and exonic variants affecting RNA splicing. For each variant, which may be up to 300 nucleotides into an intron, the tool computes a score for how strongly genetic variant affects RNA splicing. Variants falling in regulatory regions and predicted to have a deleterious impact were obtained through Ensembl VEP.

Variants validation

The sequence variants identified in the BD patients and in the UD patients were checked in dbSNP [70], 1000 genomes [246] and ExAC [197] for frequency. Moreover, the variants were confirmed by Sanger sequencing in both directions duplicate PCR products. Obtained sequences were compared to the genomic reference sequence through BLAST.

4.1.3. Results and Discussion

Implementation of pipeline for CIF detection

The pipeline for Conserved Intronic Fragments (CIF) detection is characterized by two steps, each of them implemented in Python. The first script allows to select from the phastCons scores the bases presenting a conservation value equal or higher than the threshold chosen by the user. Several tests have been performed to define an optimal threshold value which has been set up to 0.85. The output of the program is used as input of the second software, which contains the real core for designing the CIF; the other inputs are represented by two different gene annotations (Ensembl and RefSeq) and a simple gene list. First, all exons of the various transcripts assigned to the genes of interest included into the two annotations are extracted and merged together in order to collect all possible coding bases. This process allows to expand the region covered by common gene panels provided by the companies as they are usually based on the coordinates of the exons of the canonical transcript. In this way, not only exons, but also the intronic regions are identified. Within these latter, the conserved fragments are recognized and eventually fused into a unique feature when the gap among two of them is equal or lower than 2 bp. Then, CIF shorter than 20 bp are eliminated, while the rest are checked for the possible overlapping with exons of unconsidered genes, and in case discarded. Boundaries of the extended exons, the CIF and the UTRs are saved using a BED format with an extra column containing the gene name that the feature is associated with.

The file obtained running such pipeline over the selected LSD genes was submitted to the Ion AmpliSeq Designer™ platform to design specific enrichment primers.

Panel design

The total target sequence length is 202.59 kb and includes 50 LSD genes (Box 3) and 230 CIF with an average length of 40 bp. The panel design output is a 187.42 kb sequence covered by 1561 amplicons; the average amplicon length is 240 bp with 93% of the whole target sequence covered. Considering only exons, their flanking sequences and UTR, the target sequence coverage is 92.4%. The less covered genes result DNAJC5, CLN8, IDUA, NPC2, HYAL 1 whose sequence is covered for a percentage between 55% to 80%. Considering only the coding sequence the most affected gene is IDUA with 8 exons being partially or totally uncovered.

Box 3|**List of lysosomal storage diseases genes included into the panel.**

AGA, ARSA, ARSB, ASAH1, CLN3, CLN5, CLN6, CLN8, CTNS, CTSA, CTSD, CTSK, DNAJC5, FUCA1, GAA, GALC, GALNS, GBA, GLA, GLB1, GM2A, GNE, GNPTAB, GNPTG, GNS, GUSB, HEXA, HEXB, HGSNAT, HYAL1, IDS, IDUA, LAMP2, LIPA, MAN2B1, MANBA, MCOLN1, MSFD8, NAGA, NAGLU, NEU1, NPC1, NPC2, PPT1, PSAP, SGSH, SLC17A5, SMPD1, SUMF1, TPP1.

Sequencing

Samples sequencing was performed in 4 separate runs having the following quality control metrics: 99% enrichment percentage, 22% polyclonal beads; 4% low-quality reads and 72% usable reads. Each library included about 300.000 reads with a mean length of 181 bp. The average coverage was at least 100X.

On the first two runs, containing respectively 8 and 22 samples, a coverage study applying the principal component analysis (PCA) and analyzing the corresponding density plots was performed. The mean coverage of the two studies was 222.1±24.6 and 128.5±14.0 respectively.

Focusing on the results of the coverage study, the first run shows a quite uniform coverage distribution for 7 samples (Figure 18A), even if the corresponding PCA (Figure 18B), which represents the variation of the coverage distribution among the various amplicons, seems to identify three different groups of samples. Indeed, analyzing the mean coverage of the various target regions (data not reported), the clusterization reported in the PCA is quite easy to detect, as samples AWT and AWU display completely different behavior from the 5 samples clustered on the top-left of figure 18B and from sample AXA (top-right, Figure 18B).

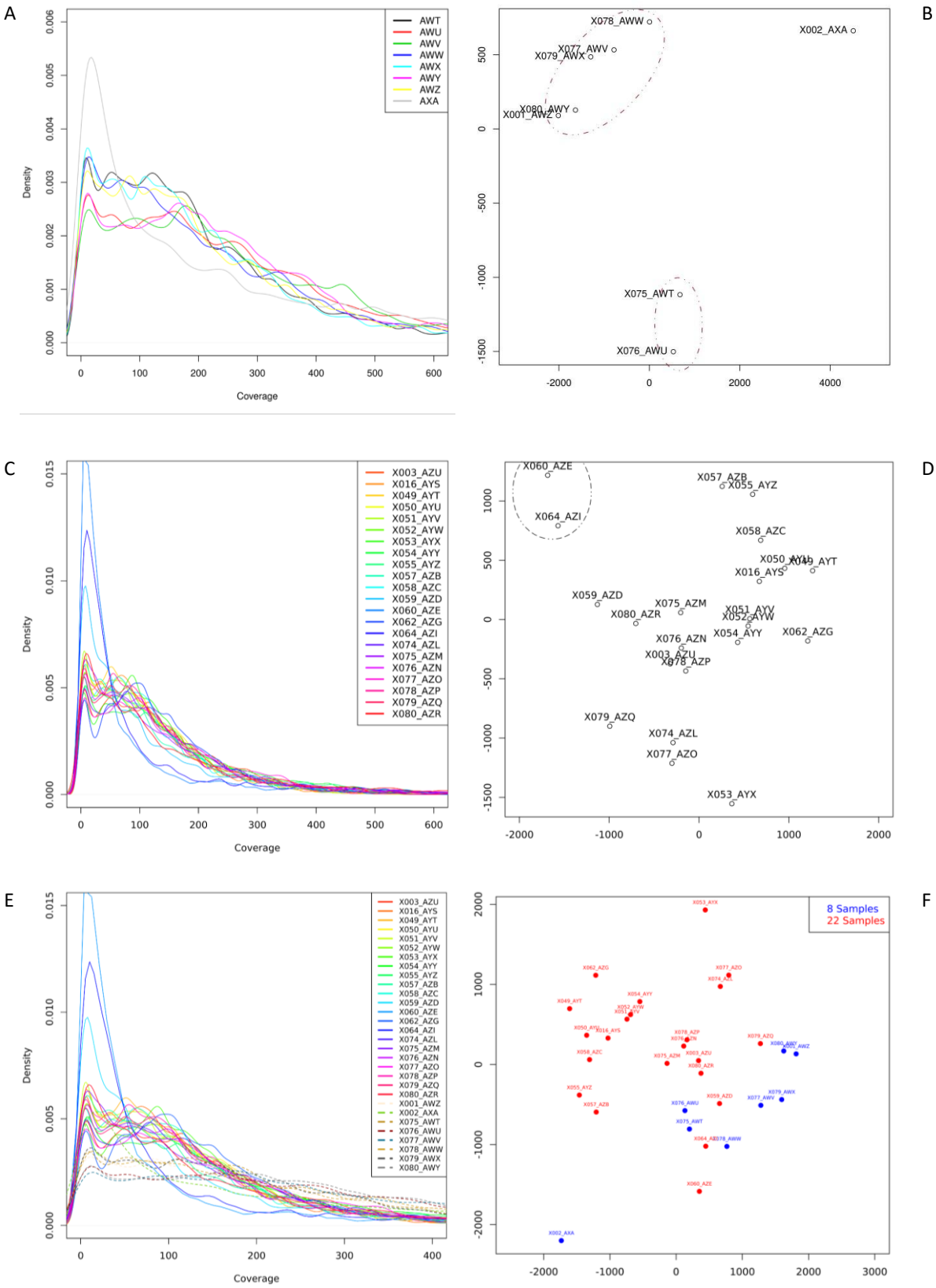


Figure 18|**Coverage study performed using density plots and PCA. A**|Density plot and **B**|PCA on the first sequencing run including 8 samples. **C**|Density plot and **D**|PCA on the second sequencing run including 22 samples. **E**|Density plot and **F**|PCA on all the samples of the two runs (30 samples). The comparison allows to establish intra-run and inter-runs differences.

Furthermore, this latter is the only specimen associated to a different coverage distribution (gray line, Figure 18A), indicating a higher number of bases with low coverage (pick at the beginning), but at the same time a slightly higher fraction of bases with really high coverage (>500X): in fact, its mean coverage (213.4X) is not significantly different from the others.

Also the coverage distribution of the second run is almost homogeneous, with the exception of three samples (AZE, AZI and AZD) (Figure 18C), whose behaviors are similar to what found for AXA in the first experiment, even if in this case the coverage of the most unbalanced sample (AZE) is also the lowest of the whole run (105.1X). Moreover, for AZE and AZI, also a disequilibrium in the amplicon coverages can be observed. In fact, both PCA (Figure 18D) and CNV analysis (data not shown) allow to easily separate these two samples from all the others.

Considering samples all together, with the exception of the previously discussed specimens, the coverage distributions have quite similar shapes, even if the first run shows a flattened trend due to the higher mean coverage (Figure 18E). The PCA confirms the results obtained using the density plots, as only AXA and AYX can be isolated from the main block of samples (Figure 18F). The behavior of AYX is difficulty explainable, while AXA was already widely disconnected in the first analysis.

Thus, although the significative difference of the average coverage among the samples included in the two experiments (*p-value*=1.786×10^{-13}), the reproducibility of amplicon coverage distribution between the different runs seems to be quite good. Indeed, the presence of outliers can be due to anomalous samples, which have met problems during library preparation or during sequencing or whose starting DNA quality was really low.

Variant analysis

Variant analysis has been carried out through a prioritization process performed by QueryOR platform. The total number of known variants per sample ranges from 63 to 359 with an average of 253 variants; the average number of novel variants per sample is 7. To analyze the three groups of samples *positive controls* (PC), *biochemically diagnosed* (BD) and *undiagnosed* (UD), we used the same flowchart which consists in performing a set of 4 queries, each capable of detecting a specific type of variant or group of variants (Table 5). If the output of a single query is greater than one variant, priority is given to alleles with the

lowest frequency or to not annotated alleles and to those presenting the most pathogenic scores.

All 'pathogenic' variants but few have been detected applying the first 3 queries, being most variants located in the coding regions or in the nearest intron-exon boundaries. The analysis leads to the identification of pathogenic variants in 64% of the PC's alleles, if applying the coverage filters (>10X). Failed variants detection may be caused by: I) variants not covered by the amplicons due to panel design; II) low-covered variants due to poor amplification of specific amplicons; III) large deletions not detected by variant calling process. For analyzing this latter type of variations, we implemented a software able to discover amplicons with unbalanced coverage comparing to what found in the other similar samples, better if sequenced in the same run. This should allow to identify various types of CNVs. The program takes as input a table containing the mean coverage of each amplicon for the analyzed samples. Such values are normalized for the median coverage of the sample and then the average and the standard deviation of coverage within the amplicons are calculated. Two different approaches are thus applied for defining problematic amplicons: the ratio between normalized values and the median coverage over the amplicons, and the deviation from the mean amplicon coverage. The parametrization is left to the user, but by default the ratio is considered interesting if lower than ½ or higher than 2. Similarly, a variation from the mean of more than two standard deviations is worthy to be account. The output is a html table with a color code for an easy visualization of the obtained results.

If the program detects a possible CNV, its trustworthiness is checked using IGV. For homozygous deletions, the evaluation of reliability is easier than the corresponding heterozygous case, as the amplicon is completely uncovered. Also for tandem repeats, the coverage should be twice or more times higher than the mean coverage in the other samples, facilitating the final interpretation. Surely, the efficiency of the software should be improved integrating more complex statistical methods or implementing new strategies based, for example, on Hidden Markov Models (HMM), as it happens in several tools for performing this kind of analysis, including CONDEX [247], XHMM [248] or CNAseg [249]. However, the detection of CNVs from genes panel data is still difficult and the using of one of the latest published tools, as HMZDelFinder [250], could help in such challenging task. Nevertheless, for some pathologies it has been proved that the increased benefit of exon

level deletion/duplication analysis is poor [251], leading to an improvement of the final efficacy that is not enough to completely justify the effort.

Confirmation of biochemical diagnosis

The panel analysis brought to confirmation of previous enzymatic diagnosis for 6 out of 12 subjects in which we found both mutations. In two and three samples respectively we found only one mutation and no mutations.

In one of the latter case, we found two mutations in GNPTAB gene in a patient biochemically reported as affected by galactosialidosis (CTSA). The patient carries a previously described mutation (c.3503_3504delTC) and a new variant presumably affecting a splicing donor site and potentially causing an exon skipping according to Human Splicing Finder [252] and SpanR [245]. Further cDNA analysis and the re-analysis of the clinical features of this patient for a better definition of his clinical picture should help in the confirmation of this hypothesis.

New diagnoses achieved

Two new diagnoses were achieved among the 9 undiagnosed patient (UD) analyzed. In a child suspected of mucopolysaccharidosis a novel hemizygous variant was found in the IDS gene (mucopolysaccharidosis type II); the same variant was detected in heterozygosis in the mother. In a girl suspected to be affected by mucolipidosis, we found two previously described mutations in GNPTAB gene (mucolipidosis II α/β, III α/β); the same mutations were detected in her parents. For the remaining 7 patients, a deeper analysis of the biochemical and clinical data is still undergoing in order to focus the panel analysis on specific genes whose gaps, if necessary, will be filled by Sanger sequencing.

Analysis of Conserved Intronic Fragments (CIF)

The analysis of CIF is performed to identify potentially dangerous variants located in intronic regions and is focused on those samples from undiagnosed patients in which no variants have been found through the previous analysis. 345 intronic variants with frequency <0.01 or with no frequency (not annotated variants) filtered by QueryOR have been uploaded in SpanR and in Ensembl VEP. Nine variants have been selected by SPANR as potentially deleterious, but unfortunately none of them was carried by undiagnosed patients. The VEP

analysis instead revealed that 61 variants fall in regulatory regions, of which 6 can be found in promoters, enhancers or flanking regions of undiagnosed samples. The potential pathogenicity of these variants is currently under analysis, even if a final validation is possible only using “wet” approaches.

Variant validation

Sanger validations till now performed on BD samples confirmed the panel results with exception of one case in which a poor covered missense mutation was not confirmed.

4.1.4. Conclusion

Targeted sequencing is an appealing approach to implement routine diagnostic strategy, given its low sequencing costs and short sequencing time. However, a good coverage must be assured and validation by Sanger sequencing have to be carried out on the proband and on his parents as final step, also to exclude the presence of deletions in case of homozygous variants finding. Moreover, the possibility to fill the gaps in the panel design have to be guaranteed, especially in case of strong suspicion for a specific disease.

4.2. LRP2 gene variants in Dent disease patients with no detectable mutation in CLCN5 and OCRL genes

4.2.1. Introduction

The second case study has been investigated in collaboration with the group of Prof. Franca Anglani of the Department of Medicine. It regarded the evaluation of a small cohort of patients showing many signs recallable to Dent disease (DD) [253], but with no mutation in the previously identified disease-associated genes, CLCN5 [254] and OCRL [255]. These seven patients were classified as DD3. Globally, on the total cohort of DD patients, DD3 constitute approximately the 25% of the cases. Considering our previously published work [256], we hypothesized that such group may represent unknown or unrecognized phenotypes of already known nephropathies, including the Donnai-Barrow (DB) syndrome, where mutations in LRP2 gene are typically associated with the onset of the disease [257]. The most interesting patient (BDA) affected by common symptoms of DD, but with pathogenic variations in LRP2 gene, has been further analyzed and described within a paper

submitted to Clinical Genetics, together with a second similar subject investigated by the group of Prof. Mauro Longoni of the Massachusetts General Hospital of Boston. A revisiting of such article will be presented as a final extended and deepen corollary (section §4.2.5) of the general analysis performed on the DD3 cases which constitutes the main topic of such part of BioInfoGen project.

4.2.2. Materials and Methods

To investigate the possible involvement of known disease-associated genes in the pathogenesis of DD3 patients, whole exome sequencing (WES) with the Ion Proton sequencer was conducted in all the seven collected subjects. The average read coverage for each sample was near 80X. The data analysis has been performed following the suggestions provided by the manufacturer: read alignment using TMAP and variant calling with TSVC. Variant prioritization has been achieved through QueryOR platform [178]. In particular, for missense variants, the prioritization strategy was based on query coverage, low MAF values and a possibly-probably damaging/deleterious prediction. Variants were finally validated by Sanger sequencing.

4.2.3. Results and Discussion

We identified in 4 patients 8 different variants of LRP2 gene which encodes for Megalin (Table 6). Two of them are novel predicted pathogenic variants.

As explained in the article reported in a following section (§4.2.5), by deepening phenotypic features of BDA patient, we highlighted mild characteristics of Donnai-Barrow (DB)/FOAR syndrome, due to two non-conservative mutations in LRP2. A known LRP2 disease-causing mutation was detected in AMV, associated with a novel pathogenic mutation of CUBN gene encoding for Cubilin. This patient has no signs of DB syndrome, but presents an Alport-like glomerulopathy. No mutation in Alport disease genes was found by WES. A very rare pathogenic LRP2 missense variant was found in AMQ, in association with an OCRL disease-causing mutation. Four different LRP2 missense variants were detected in AMT, two of them are common polymorphisms, while the remaining ones are low-frequency variants predicted to be pathogenic by MutationTaster [235]. No sign of DB syndrome was highlighted also in this patient, however an Alport-like glomerulopathy was reported. No

mutation in the Alport-causing genes was detected by WES. Notably, also in this patient an uncommon CUBN missense variant was observed.

Patient	Codon Substitution	dbSNP code	ExAC Frequency	Pathogenicity
BDA	p.Arg2243Ter	NA	novel	CV: NA MT: disease_causing (1.000)
BDA	p.Ile81Asn	NA	novel	CV: NA MT: disease_causing (1.000)
AMT	p.Gly259Arg	rs34693334	TOT: 0.0632 EUR: 0.0880	CV: probable-non-pathogenic MT polymorphism (1.000)
AMT	p.Asn2632Asp	rs17848169	TOT: 0.0295 EUR: 0.0426	CV: probable-non-pathogenic MT disease_causing (1.000)
AMT	p.Gly669Asp	rs34291900	TOT: 0.0285 EUR: 0.0430	CV: probable-non-pathogenic MT: disease_causing (1.000)
AMT	p.Val3999Gly	rs79723119	TOT: 0.0089 EUR: 0.0130	CV: probable-non-pathogenic MT: polymorphism (1.000)
AMQ	p.Thr2086Ser	rs146149181	TOT: 0.0015 EUR: 0.0010	CV: NA MT: disease_causing (1.000)
AMV	p.Asp2054Asn	rs138269726	TOT: 0.0011 EUR: 0.0016	CV: pathogenic MT: disease_causing (1.000)

Table 6|**Interesting variants found into LRP2 gene of four patients** (BDA, AMT, AMQ, AMV). Codon substitution is reported using HGVS protein code. Pathogenicity is expressed using information included into a public archive of clinically relevant variants (ClinVar, CV) [180], while the prediction scores are obtained from MutationTaster (MT) [235]. NA: Not Available.

LRP2 and CUBN genes may be excellent candidates for DD3. They work in the same pathway of CLCN5 gene regulating the tubular reuptake of albumin and LMW proteins. However, mutations in these two genes are known to cause different monogenic diseases i.e. DB [257] and Imerslund-Gräsbeck [258] syndromes respectively. Both disorders show a DD-like renal phenotype.

4.2.4. Conclusion

Given our results, we can consider LRP2 variants as 1) causative of DB syndrome in BDA patient, 2) modifier of DD2 phenotype in AMQ, 3) possible disease-causing in association with CUBN mutation in AMV, and 4) difficult to interpret with unknown significance in AMT. Interestingly, in all patients glomerulopathy was present.

4.2.5. Hypercalciuria and nephrolithiasis: expanding the renal phenotype of Donnai-Barrow syndrome

Introduction

Donnai-Barrow/Facio-oculo-acustico-renal (DB/FOAR) syndrome (MIM #222448) is a rare inherited condition characterized by typical craniofacial features, agenesis/hypogenesis of the corpus callosum, high-grade myopia, sensorineural hearing loss, and low-molecular-weight proteinuria (LMWP) (Figure 19). Congenital diaphragmatic hernia and omphalocele are frequent additional findings. Mutations in the LRP2 gene, encoding for Megalin, cause DB/FOAR syndrome. Proteinuria is a defining feature of this condition, as Megalin is expressed in the renal proximal tubule where it accounts for the uptake of retinol binding protein, vitamin D binding protein, and lipoproteins, among other ligands [259].

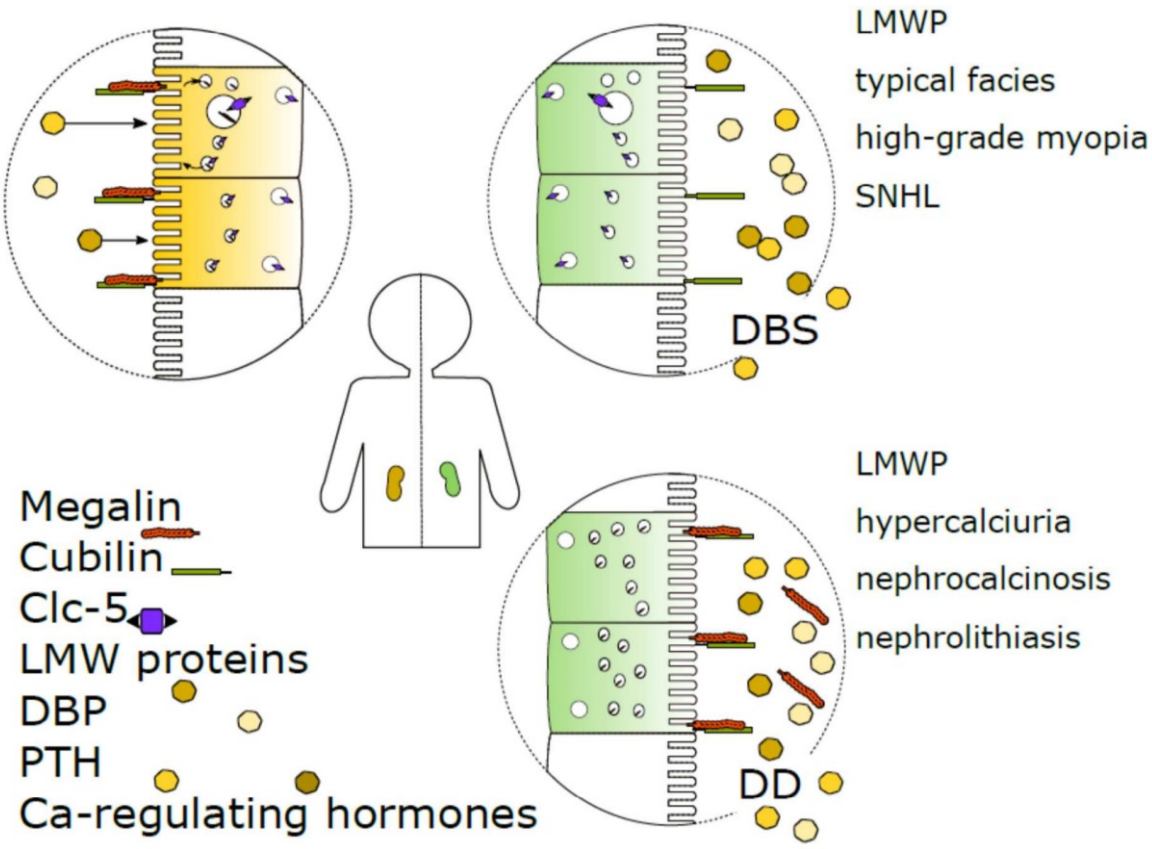


Figure 19|**Graphical portrayal of different conditions in the tubular cells**. Normal homeostasis is figured on the top-left, while Donnai-Barrow syndrome (DBS) (lack of megalin) and Dent disease (DD) (lack of ClC-5 antiporter) are depicted on the right, at the top and at the bottom, respectively. LMWP: Low Molecular Weight Proteinuria; SNHL: SensoriNeural Hearing Loss; PTH: parathormone.

Dent disease (DD) identifies a group of X-linked renal tubulopathies characterized by the triad of symptoms LMWP, hypercalciuria, and nephrocalcinosis and/or nephrolithiasis (Figure 19) [260, 261]. Rickets and osteomalacia are also relevant features of DD phenotype. DD usually presents in childhood or early adult life. The most common genetic cause of DD is a mutation in the CLCN5 gene encoding the Cl-/H+ antiporter ClC-5, (Dent disease 1; MIM#300009) [254, 255, 262]. Mutations in the OCRL gene, encoding the phosphatidylinositol 4,5-bisphosphate-5-phosphatase OCRL1, usually associated with Lowe syndrome (MIM #309000), have been identified in about 10-15% of DD patients (Dent disease 2; MIM#300555) [254]. Whereas DD1 only affects the kidney, the spectrum of symptoms in DD2 can range from apparent exclusive kidney manifestations to the involvement of other organs, notably brain, muscles, and eyes in overlap with Lowe syndrome [255]. It remains an open question whether a third gene is responsible for DD in patients without identifiable mutations, or whether they represent atypical disease phenotypes of already known hereditary nephropathies [256].

In this study, we investigated two patients, an adult and a child with LMWP, hypercalciuria and nephrocalcinosis/nephrolithiasis without CLCN5 and OCRL variants. Exome sequencing revealed novel likely pathogenic variants in the LRP2 gene in both individuals.

Materials and Methods

Clinical cases

Case 1 - The patient is a 69 years old male with a family history of increased urinary proteins (mother and sister) and nephrolithiasis (father). At the age of 14, he developed progressive bilateral hearing loss and left eye blindness. Moreover, at age 40 he developed glaucoma of his right eye. When he was 25 years old, he developed an acute kidney injury (AKI) with microhematuria, glomerular and LMW proteinuria, and granular casts in his urinary sediment. Urinary findings were presumed secondary to chronic tonsillitis and documented anti-streptolysin O (ASLO) positivity. AKI remitted completely after tonsillectomy but mixed proteinuria persisted (500 mg/day), with hyaline and granular casts, uric acid crystalluria, and incomplete Fanconi syndrome. This prompted a kidney biopsy which showed 1/23 hyaline glomeruli, while in the other glomeruli there were mild PAS-positive mesangial hyperplasia, focal sclerosis of Bowman's capsule, tubular cells with granular cytoplasm and few hyaline intraluminal casts. No specific therapy was administered. Bone biopsy was

performed when he was 30, and showed osteomalacia, consistent with a diagnosis of “renal rickets”. At age 40, he developed calcium oxalate kidney stones. By the age of 50, he developed hypertension and slow progressive chronic kidney disease (CKD) (creatinine 1.6-1.7 mg/dL) with mixed proteinuria of 1-2 g/24h. At age 58, he developed also renal glycosuria. Since hypercalciuria (400 mg/day) was observed with normal calcemia and parathormone levels, therapy with amiloride-hydrochlorothiazide was started, reducing calcium excretion below normal levels. However, after 10 years he developed a new episode of nephrolithiasis.

Case 2 - The patient is a boy of Senegalese origin with a negative family history for nephropathy. He was born weighing 3000 g at full term. The perinatal period was uneventful. Omphalocele repair surgery was performed at age of 6 months. Growth failure and a history of chronic constipation were reported in childhood. At the age of 5, he was hospitalized for clinical tests which documented marked growth failure (height -3/-4 DS, weight -2 DS), facial dysmorphisms (hypertelorism, flat and enlarged nasal bridge, broad forehead and prominent parietal bossing), and bilateral cryptorchidism. He displayed mild psychomotor retardation. Laboratory testing of the patient’s urine showed LMWP, hypercalciuria, and microhematuria, but no acidosis. Hypophosphatemia and hypovitaminoses D and A were also present. Ultrasound examination showed tiny calcifications in his left kidney compatible with the presence of nephrocalcinosis. After few months, the child was hospitalized for retinal detachment in his left eye, further complicated by ocular listeriosis and vitritis. His right eye demonstrated severe myopia with peripapillary atrophy. Shortly thereafter, the child underwent cataract surgery in the left eye. At age 12, he was hospitalized for an acute bowel obstruction with detection of hypokalemia (K 2.9 mmol/L); this was treated with intravenous potassium supplementation for few days, and then with oral supplementation. Currently, the patient is 16 years old and wears hearing aids.

Whole Exome Sequencing

Whole Exome Sequencing (WES) was performed using the Ion Proton System (Thermo Fisher Scientific, MA USA) in Case 1. The patient gave informed consent. We obtained an average read coverage near to 70X (Table S1). Alignment and variant calling were performed

using the software suggested by the company for Ion Proton data analysis. To annotate and prioritize the short nucleotide variants, we used QueryOR (http://queryor.cribi.unipd.it) [178].

WES of patient 2 was performed at the Broad Institute (Cambridge, MA) on the Illumina HiSeq 2000 after enrichment with Agilent SureSelect v.1.1. Mean coverage was approximately 100X (Table S1) with paired-end 76 bp reads. Data analysis was performed as previously described [263]. Briefly, sequence data was preprocessed with the Burrows–Wheeler Aligner (BWA 0.7.5a) and SAMtools version 0.1.19. Variant calling was performed with the Genome Analysis Toolkit (GATK) (https://software.broadinstitute.org/gatk/) [264] according to GATK Best Practices [103], with minor adjustments. Mutation analysis was performed with Ingenuity Variant Analysis™ (Qiagen, MA USA) and seqr (https://seqr.broadinstitute.org/) [265]. Informed consent was obtained by the legal representatives of patient 2 (Partners Human Research Committee, Protocol Number: 2009P001589). Variants that passed the in silico prioritization strategy (Table S2) were submitted for molecular validation by Sanger sequencing according to a previously published protocol [263].

Results

The two patients received a provisional diagnosis Dent disease because their renal phenotypes were consistent with the triad of classical symptoms: LMWP, hypercalciuria, and nephrocalcinosis/nephrolithiasis. Both patients, however, also presented with extrarenal symptoms involving vision and hearing. Facial dysmorphisms were absent in the adult patient (case 1), but was present in case 2. Clinical DNA sequencing did not detect mutations in CLCN5 or OCRL; accordingly, WES was performed.

Two alleles c.[242T>A];[6727C>T] of LRP2 (gene ID: ENSG00000081479, transcript ID ENST00000263816, reference sequence NM_004525.2) were detected in case 1 (https://databases.lovd.nl/shared/individuals/00131892) [266]. One allele was predicted to cause the non-conservative amino acid substitution p.(Ile81Asn), while the other to introduce a premature stop codon at position 2243 of the aminoacidic chain p.(Arg2243Ter). Neither variant is reported in dbSNP, 1000 Genome, EVS, or ExAC. Parental DNA was not available for testing.

A homozygous missense variant c.7624C>T p.(Arg2542Cys) was identified in Patient 2 (https://databases.lovd.nl/shared/individuals/00131950) [267]. Both parents were unaffected carriers. The variant is not reported in any of the SNV databases listed above.

Discussion

Pathogenic loss-of-function variants in LRP2 are associated with Donnai-Barrow/Facio-oculo-acusticorenal (DB/FOAR, MIM#222448) syndrome [257, 268, 269], characterized by typical craniofacial features, and by high-grade myopia, congenital diaphragmatic hernia or omphalocele, sensorineural hearing loss, LMWP, brain anomalies, and intellectual disability [270]. The LRP2 gene encodes Megalin, a large single-spanning transmembrane multiligand endocytic receptor with a small intracellular region, expressed in the proximal tubule of the adult kidney where it works in the same pathway as the Dent disease-associated protein ClC-5 [271]. The extracellular domain has a modular structure with alternations of LDL-receptor class A and class B domains, interspersed with EGF-like repeats and YWTD spacer regions [272].

Patient 1 is a compound heterozygote for a likely null nonsense mutation and the p.(I81N) missense variant at the N-terminus of the protein. The other missense Megalin variants reported in DB patients are instead localized to the C-terminal half of the protein (Figure 20) [257, 268, 273, 274], including one possibly pathogenic variant present in a patient with intellectual disability [275]. While we speculate that the p.(I81N) could act as a hypomorph, thus explaining the mild phenotype of patient 1, sequencing of larger cohorts will be needed to confirm this observation.

In our patients, nephropathy was the presenting phenotype, characterized by incomplete Fanconi syndrome with LMWP, hypercalciuria, nephrocalcinosis/nephrolithiasis and rickets. The last three features are typical of DD phenotype but were never described before in DB/FOAR, although disturbances in systemic calcium homeostasis and bone metabolism can be observed in a mouse model with conditional inactivation of the LRP2 gene in the kidney [276]. Megalin is absent in the proximal tubular cells of patients with null LRP2 variants [277], while faint staining was detected in the cytosol of a DB/FOAR patient with two missense variants [268]. The latter had a LMWP pattern similar to Dent disease, yet without hypercalciuria and nephrocalcinosis/nephrolithiasis.

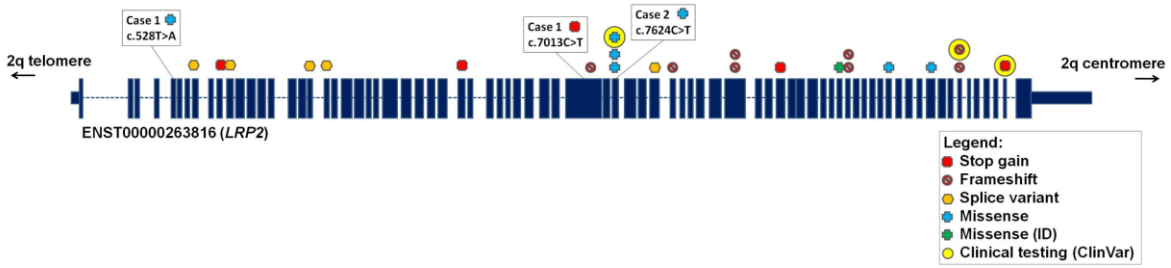


Figure 20|**Diagram of LRP2 exons.** Sequence variants identified in probands. p.(I81N) (pt1) is in a LDL-receptor class A. p.(R2542C) (pt2) is in a LDL-receptor class B. Missense (ID): homozygous missense variant in a patient with intellectual disability without signs of DBS/FOAR. Clinical testing: variants identified in clinical testing of patients and deemed pathogenic or likely pathogenic (ClinVar variant IDs 9450, 211391, 374076); in these cases, the available clinical information is insufficient for phenotypic classification. References to the relevant papers are in the main text. Introns not in scale.

In conclusion, we identified LRP2 mutation-positive patients with LMWP, hypercalciuria, and nephrocalcinosis/nephrolithiasis. We propose that a subset of patients presenting as DD may represent unrecognized cases or mild forms of DB/FOAR, or be on the phenotypic continuum between the two conditions.

Supplementary materials

Feature	Case 1	Case 2
Total number of reads	33051726	103331000
On target reads	31708617	99817562
On target bases	3887903103	3220721127
Mean target coverage	67.33X	98.89X
Percentage of target bases > 20X	73.78%	89.8%

Table S1|**Sequencing metrics of the two analyzed case studies.**

Criteria for filtering	Case 1	Case 2
Total number of variants	49995	51858
Variants with coverage > 30	39206	38876
Variants with quality score > 50	38265	38591
Number of variants with minor allele frequency ≤ 0.01	3967	7171
Exonic and splicing variants	1856	2385
Exclusion of synonymous variants	1092	1361
Pathogenic predicted variants (DANN > 0.99)	256	155
Genes associated to kidney diseases*	9	8
Clinically relevant variants	LRP2: NM_004525.2, Exon 3, c.242T>A (p.Ile81Asn) LRP2: NM_004525.2, Exon 39, c.6727C>T (p.Arg2243Ter)	LRP2: NM_004525.2, Exon 41, c.7624C>T (p.Arg2542Cys)

Table S2|**Prioritization strategy applied on the two analyzed case studies.** *Genes associated to kidney diseases for conditions with matching DisGeNET term: Kidney Diseases (C0022658).

4.3. Analysis of recurrent nucleotide variants reveals inconsistencies in the human reference genome

4.3.1. Introduction

Since its first draft release in 2001 [13, 14], the reference sequence of the human genome underwent several updates and improvements. Notably, in 2009 the Human Genome Reference Consortium made available the GRCh37 release (also known as hg19), that was followed by the GRCh38 release in 2013 and further updated in the form of “patches” in the following years [278].

Interestingly, many users are still adopting GRCh37 for their studies [134]. This is due to the many difficulties in updating tools and pipelines when a new version of the genome becomes available. Indeed, most commercially available exome kits, for instance the “Ion AmpliSeq Exome RDY Kit” from Thermo Fisher Scientific or the “Nextera Rapid Capture

Exome" from Illumina are still based on the old GRCh37 release. As a consequence, GRCh37 is also recommended for bioinformatic analyses.

This reluctance to update to the new release of the genome is unfortunate because GRCh38 contains several important improvements [134]. It was derived from many donors instead of a few and led to the correction of 8248 bases; furthermore, GRCh38 supports the representation of complex haplotypes with the introduction of alternate loci, and includes many regions that were missing in the previous release, such as segmental duplications, centromeres and telomeres [134].

The problems derived from using the old reference genome for next generation data analysis have already been widely discussed in the literature. Two different studies demonstrated that the poor representation of repeated sequences in GRCh37 produces read misalignments and false-positive variants [279, 280]. To solve this problem, the authors propose the integration in standard pipelines of 'decoy' sequences [279] or 'sponge' databases [280] representing a collection of sequences omitted from the GRCh37 assembly that are supposed to result in mismapped reads. This integration allows an improvement in read mapping and in the resolution of false heterozygous calls [279].

More recent studies confirmed that the new sequences introduced in the updating of GRCh37 to GRCh38 improved the read mappability and lowered the number of false-positive single-nucleotide variants [281, 282]. However, despite the above improvements, we observed that the problem of false-positive variants remains also in GRCh38.

A large number of unexpected false-positive variants is certainly due to the inclusion of MAiRs, minor alleles in the reference genome, and can be easily filtered out with appropriate tools [178]. However, even after this filtering process, several thousand variants still remain. This is quite surprising because they are not reported as common variants in the databases and at the same time they are found in most exomes. Interestingly, exomes obtained with different technologies such as Ion-Proton, Illumina and SOLiD, exhibit a largely overlapping set of these false-positive variants, therefore they are not due to artifacts of a particular chemistry or sequencing platform.

The study presented in this chapter was carried out in collaboration with other colleagues and has two main aims: firstly, we wanted to evaluate and classify the recurrent short-nucleotide variants, both in GRCh37 and GRCh38. We believe that a clear repertoire of the

recurrent miscalls will help geneticists in analyzing exome data, facilitating the process of variant prioritization.

A second, but not less important scope is to better understand the nature of this problem and to verify the hypothesis that sometimes these unexpected variants may originate from duplicated regions that are not reported in the reference genome. This can be experimentally verified because the "collapsed" repeated sequence of the reference genome would be the target for the alignment of reads derived from two or more real genomic regions, resulting in a disproportion between frequency, heterozygosity and homozygosity of the corresponding allele.

With this premise, we analyzed a few hundred exomes from different platforms, using both GRCh37 and GRCh38. We found that the problem of collapsed repeats is indeed responsible for the call of many false-positive variants, several of which are still remaining in GRCh38. Furthermore, we suggest a few positions of the reference genome that require a revision in future updates.

4.3.2. Materials and Methods

Datasets

In this study, we used three different datasets: one case study dataset, which has been extensively analyzed, and two control datasets. The study dataset was composed by all variants collected from 222 different exomes sequenced at the CRIBI facility of the University of Padua. These exomes were enriched with the Ion AmpliSeq Exome panel and sequenced with the Ion Proton system (Thermo Fisher scientific). Samples came from the most different research projects ranging from cohorts of individuals to trios (Table 7).

The two other datasets are technological controls chosen to appraise platform-specific errors. One was composed by all variants found in 22 exomes belonging to two different projects and enriched with Illumina TruSeq Exome panel and then sequenced with Illumina NextSeq 550 platform at CRIBI. The other was composed by variants identified in 300 exomes belonging to the study published by J. de Ligt *et al* [230]. It should be pointed out that these samples, enriched with SOLiD-optimized target enrichment and sequenced with SOLiD 4 System (Life Technologies), belong to 100 trios composed by patients with

unexplained severe intellectual disability and their unaffected parents (data deposited into The European Genome-phenome Archive under submission EGAS00001000287, [231]).

Project	Exomes number
1	47
2	45
3	29
4	22
5	18
6	17
7	10
8	9
9	9
10	6
11	5
12	2
13	2
14	1
Total exomes	222

Table 7|**Number of exomes for each project.**

Alignment and variant calling on GRCh37

All samples of study dataset and control datasets were aligned against the release GRCh37 of the human reference genome.

Study dataset - Each exome included into the study dataset was sequenced using the CRIBI Ion Proton system to reach a final mean coverage of 80x and a target uniformity higher than 90%. Alignment and variant calling were carried out according to the Torrent Suite 5.0 exome analysis pipeline, as suggested by the manufacturer. Variants were merged into a unique file using CombineVariants of Genome Analysis Toolkit (GATK v. 3.6) and then normalized applying the method proposed by Tan and colleagues [193] in order to eliminate different representations of the same variant. Variant annotation, based on GRCh37.82 version of Ensembl transcripts, was performed using VarPred (§3.2).

Illumina control dataset - Each sample was sequenced with 76 bp paired-end reads by using the CRIBI Illumina NextSeq 500 to reach a final mean coverage of at least 40x with an average of 103x. Alignment and variant calling were performed applying the Illumina data analysis pipeline discussed above (§3.1), following the suggestions reported in the GATK Best Practices [103]. The obtained variants were then normalized as previously described.

SOLiD control dataset - VCF files were downloaded from The European Genome-phenome Archive (EGA). For alignment and variant calling procedures please refer to de Ligt *et al* paper [230]. Variant normalization was performed as indicated above.

Alignment and variant calling of study dataset on GRCh38

The 222 exomes of the study dataset were also aligned against GRCh38.p10 downloaded from Ensembl [283]. Alignment and variant calling were performed according to the Torrent Suite 5.0 exome analysis pipeline as described in the Alignment and Variant Calling on GRCh37 section. Variants from all samples were merged and processed as described above. Applying CrossMap [284] the coordinates of the resulting variants were converted to GRCh37 coordinates, allowing a comparison between such variants and those obtained using the release GRCh37 of the reference genome.

Identification of Minor Allele in Reference positions

We called Minor Allele in Reference (MAiR) those positions in the human reference genome that present an allele that is not the most frequent in the population. To identify if in our study dataset some variants fall in such positions, the allele frequencies in the total population reported in 3 databases including I) dbSNP [70] version 144 [221], modified to recover old variants excluded from this release but present in the online version, II) NHLBI ESP version ESP6500SI-V2 [154], III) ExAC version 0.3.1 [68], were analyzed. In particular, the reference allele frequencies were compared with the alternative allele frequencies. Thus, a genomic position was marked as MAiR if the reference allele frequency was lower than any alternative allele frequency in all databases.

Confirmation of variants in MAiR positions at protein level

Variants in GRCh37 MAiR positions confirmed in GRCh38 genome were annotated using both SnpEff [188] v4.2 and VEP [162] v84, employing respectively UCSC and RefSeq

transcripts. Two different annotations were chosen to avoid transcript-dependent biases. Missense variants were selected from the two annotated VCF files as associated to protein changing. These protein variations were independently compared with the Human polymorphisms and disease mutations release 2017_05 [285] of UniProt, in order to understand if these high frequency exomic variants have a known correspondence at protein level. Moreover, the presence of the mutated amino acid in the primary protein sequence was also evaluated. This analysis was performed using the reviewed Swiss-Prot human sequences [286]. The comparisons have been done using a in-house python script.

Statistical test on heterozygous genotype frequencies

Variants with a heterozygous genotype frequency significantly higher than the expected were investigated. This analysis was performed only for biallelic variants, defined as loci that have two observed alleles: the reference and one alternative allele. For each variant, the observed allele frequency was calculated as the number of times the specific allele was found divided by the total allele number (444 alleles). Then we calculated the observed frequency for the three possible genotypes of each variant as the number of times we found that genotype divided by the total number of exomes (222 exomes). Expected genotype frequencies were computed using the formula $(p+q)^2=1$, where p and q are the observed frequencies of reference and variant allele respectively. We then performed a one-tailed binomial test for the heterozygous genotype. P-values were corrected for false discovery rate using the Benjamini-Hochberg procedure [287]. Observed genotype frequencies were considered significantly higher than the expected if the corrected p-value was lower than 0.01.

Realignment of reads containing interesting variants

The analysis was focused on those reads belonging to amplicons containing more than one variant having an unbalanced heterozygous genotype and confirmed by Illumina and SOLiD. From three randomly chosen samples, we extracted for each amplicon those reads containing all the variants and reads with none of them. These two groups of reads were aligned against the GRCh37 and GRCh38 toplevel human reference genomes using BLAST. Toplevel genomes are defined as those assemblies containing chromosomes, regions not assembled into chromosomes and N padded haplotype and patch regions. Files were

downloaded from Ensembl ftp website [288, 289]. In this analysis, the identity percentage cutoff was set to 90%.

Variant database creation

All information collected on variants during the various analyses has been gathered into a unique tabular file using an in-house python script. The table contains one row for each variant and one or more columns for each analysis.

4.3.3. Results

Preliminary analysis of the study dataset

After the normalization process, the obtained study dataset was composed by 264303 variants, including 239255 SNPs and 25048 small INDELs (14075 deletions and 10973 insertions). Among the total variants, 245088 were defined as biallelic while the remaining 19215 variants occurred in positions in which more than one alternative allele was present.

It was also observed that 9313 (3.52%) variants were shared by at least the 90% of exomes and even 2349 were identified in all the samples. Surprisingly, a consistent fraction of such variants cannot be explained by a high frequency in the population. Thus, although the subsequent analyses were performed on the total cohort of the 264303 variants, we mainly focused on these recurrent variants in order to find a feasible explanation about their presence.

Comparison with Illumina and SOLiD datasets

To understand if the 9313 variants shared by at least the 90% of exomes could derived from Ion Proton platform specific biases, they were searched for confirmation in two independent control datasets. These latter were produced by Illumina and SOLiD sequencing and analyzed with their corresponding pipelines, leading to collect 124935 and 189512 variants respectively. Both datasets presented a lower number of variants in respect to the whole Ion Proton dataset. This event could be due to different causes: I) the Illumina samples were considerably fewer than the samples in our study dataset; II) despite the high number of samples in the SOLiD dataset, it should be pointed out that they belong to trios - that will implicitly share most of their variants - collected for the study of a particular mental disorder.

We considered variants as confirmed if they were present in at least the 50% of Illumina and SOLiD samples, separately. The number of confirmed variants is 6008 for Illumina, 5733 for SOLiD and 4607 considering both (Figure 21). The majority of not confirmed variants were localized in regions peculiar of Ion AmpliSeq Exome panel that were not enriched with the Illumina and SOLiD target. Only 41 variants localized in common target regions were not confirmed by both Illumina and SOLiD, so they could be platform specific systematic errors.

A

Ion Proton dataset	Illumina control dataset			SOLiD control dataset			Both control dataset	
	confirmed	not confirmed		confirmed	not confirmed		confirmed	not confirmed
9313	6008 (64.51%)	3305 (35.49%)		5733 (61.56%)	3580 (38.44%)		4607 (49.47%)	4706 (50.53%)
		OOT	IT		OOT	IT		
		3228 (97.67%)	77 (2.33%)		2267 (63.32%)	1313 (36.68%)		

B

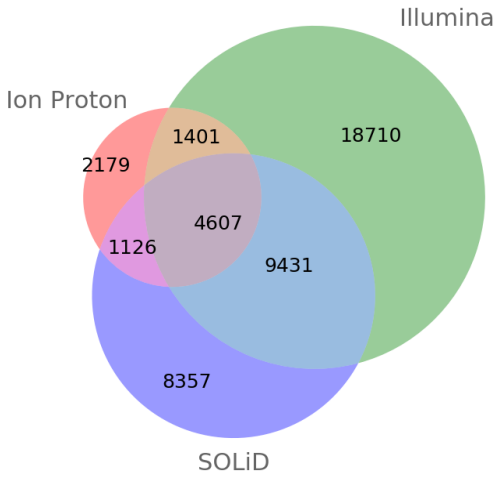


Figure 21|**Recurrent variants shared among Ion Proton, Illumina and SOLiD datasets.** **A**|Paired comparisons are shown in the first two blocks where a distinction between not confirmed variants in target (IT) and out of target (OOT) is also appreciable. The last box considers all the three technologies together. **B**|Graphical representation of the comparison among the datasets.

GRCh38 variant comparison

Alignment and variant calling of the study dataset were performed also using GRCh38 as reference genome. We identified 255124 variants, a smaller number compared to the previous genome release. This was somehow expected as data published by Guo and colleagues [281] claimed a lower number of SNPs due to the improvements introduced in the latest release of the human reference genome, thus reducing the number of false positive variants. The number of variants shared between GRCh37 and GRCh38 was 242259 (91.66% of the GRCh37 variants dataset). Similarly, the number of recurrent variants fell

from 9313 to 8132 (87.32%), indicating that the improvement of the latest human genome assembly allowed to sensibly diminish the number of false positives, as previously reported [134].

European and total population allele frequencies

All the exomes of the study belong to European people. We wondered if variants in the dataset could present a higher alternative allele frequency in the European population compared to the general population (only ExAC database frequencies were considered). In fact, the high number of shared variants in our samples could be explained as European-specific polymorphisms. The plot in Figure 22 shows the almost perfect correlation between the frequencies in the two populations, indicating that there is no evidence of a possible bias due to ethnic origin of the samples.

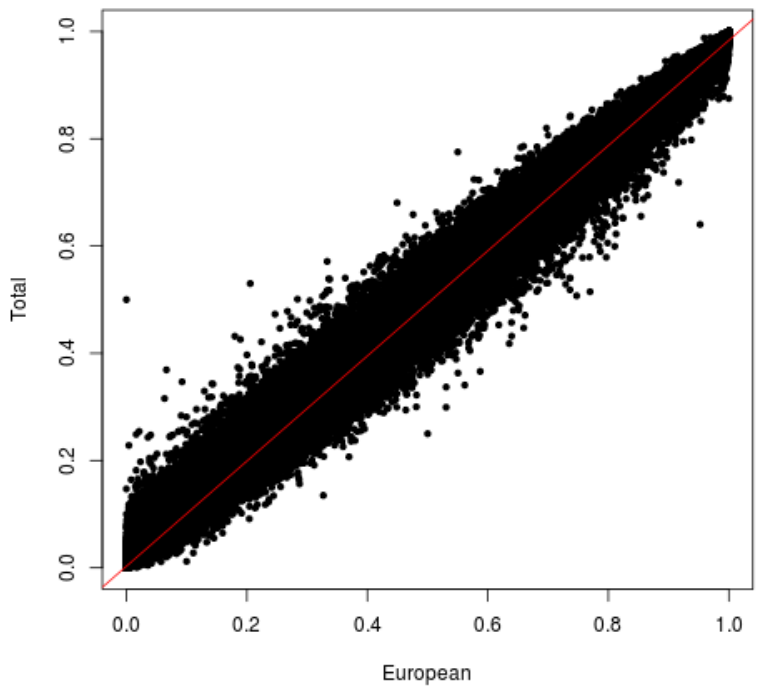


Figure 22|**Correlation of allele frequencies between European and Total populations.** Red line shows the high correlation between the two datasets.

Identification of Minor Allele in Reference positions

It was estimated that erroneous reference bases, responsible also for incorrect variant calls, occur at a rate of 10^{-5} [15]. Consequently, several thousand positions of the reference genome do not carry the major allele of the population, as we previously reported [178]. In these positions, variant callers will identify an alternative allele that indeed should represent

the reference, thus increasing the number of false positives. We wondered if some variants in the dataset could be due to the presence of these erroneous reference positions. As explained in the Materials and Methods section, we partially revised our previous definition of Minor Allele in Reference (MAiR) positions. Thus, we marked 18839 (7.13%) variants as MAiRs. Interestingly, MAiR variants represent the majority (94.09%) of the variants shared by at least the 90% of the samples. Moreover, we checked if these erroneous reference positions have been corrected in the latest release of the reference genome. Only 1808 (9.60%) MAiR positions have been solved in GRCh38, while the remaining 17031 (90.40%) MAiRs are still present, including 7876 highly recurrent variants.

Furthermore, in order to understand if these MAiR variants were frequently found not only at genomic level but also at protein level, we investigated the UniProt protein variation database in search of correlation between predicted missense variants and known amino acid substitutions. In particular, the MAiR variants retained in GRCh38 were annotated using two different annotations (UCSC and RefSeq transcripts) thus avoiding transcript-dependent biases. After the selection of missense variants, we obtained a comparable number of mutations in the two databases: 3814 with RefSeq and 3761 with UCSC. When we compared these missense variants with the protein variation database, we found that ~74% of these variants were already known also at protein level as natural variants. More interestingly the 2.6% of these mutated residues were included in the protein primary sequence, indicating that the alternative allele in MAiR position actually corresponds to the most frequent amino acid in the protein sequence. The remaining part of variants (~23%) were not confirmed at protein level: the main reason could derive from the fact that not all transcripts used during the annotation have a corresponding curated protein sequence in the Swiss-Prot database.

Analysis of heterozygous genotype frequencies

We proposed the presence of gene and region duplications not yet annotated in the reference genome as one of the possible causes for a misleading variant calling in the target regions: since these duplicated regions can be enriched and sequenced together with the original target gene, the corresponding reads will align to an improper position causing the identification of variants not really present in the gene. Consequently, we expected a heterozygous genotype for these variants, with the reference allele deriving from the original target gene and the alternative allele from the duplicated region. For each variant in

the study dataset we performed a statistical test to compare the observed and the expected heterozygous genotype frequency and we found that 767 variants had the heterozygous genotype frequency significantly higher than the expected one. These interesting variants were used in the following variants selection steps.

Analysis of interesting variants and read realignment

In order to identify possible duplicated regions accountable for the misleading variant calling, we analyzed those variants with the heterozygous genotype frequency significantly higher than expected. Among these we excluded from the subsequent analysis the 14 variants not confirmed neither by Illumina and SOLiD control datasets, but localized in Illumina and SOLiD target regions, as they could be Ion Proton specific systematic errors. We obtained 753 variants that we believed to be very reliable. We then focused on enriched target regions containing more than one of the selected variant, collecting 145 different regions spanning over 45 genes. In the process of investigating these regions, we observed that two different groups of aligned reads were distinguishable: reads having all the selected variants or reads having none of them, thus suggesting a possible different genomic origin of these reads even if they aligned on the same region. This observation agreed with our hypothesis of duplicated regions not present in the human reference genome used for the analysis (Figure 23).

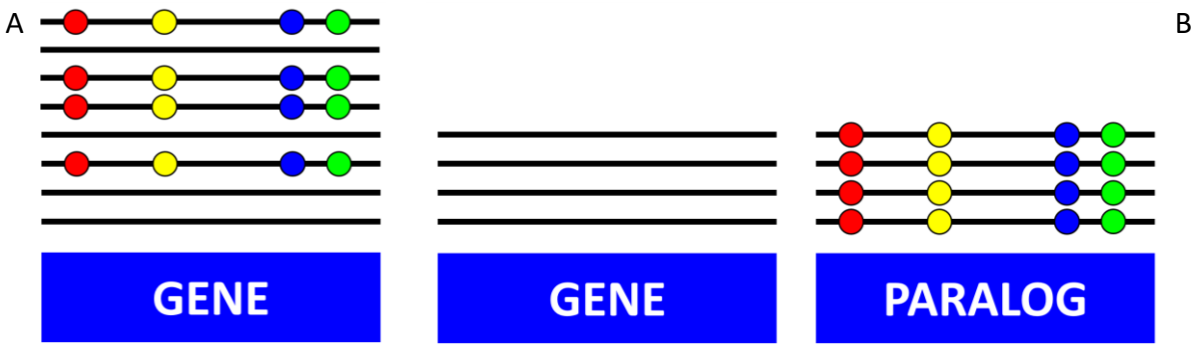


Figure 23|**Gene duplication hypothesis**. **A**|*Wrong assembly*. Both kinds of reads align on the target gene. **B**|*Correct assembly*. Reads with no variants map on the original target gene, while reads affected by variations properly align on a paralog gene.

We thus wondered if these duplicated regions had been identified and inserted in the most recent reference genome releases. In fact, the introduction of assembly patches in the latest

GRCh37 release and their resolution in the GRCh38 reference made possible to include segmental duplications and 'missing' sequences, such as paralogous sequences [134]. For this reason, we decided to realign both the two groups of reads against the GRCh37 and GRCh38 toplevel human reference genomes using BLAST. For each of the 45 analyzed genes, we compared the alignments of the two pools of reads in both the references. Assuming that the highest identity percentage indicates the real genomic origin of that read, BLAST results showed two possible scenarios: both groups of reads derived from the same region corresponding to the target gene or they derived from different regions. We realized that these different regions could be not only patches and other regions in the same or in a different chromosome, but also alternate loci scaffolds that provide more representation for population variation in the reference [134].

Among the 45 genes (Table 8), we classified 34 genes as *solved* in GRCh38 since reads with variants aligned to a position different from the original gene, indicating the presence of duplicated regions or haplotypes, whereas reads with none variant aligned to the original target gene. These new sequences added to the reference have been able to capture several reads that otherwise aligned to the original gene, thus preventing variants to be called. In fact, almost all variants localized in these solved genes were not identified using the GRCh38 toplevel reference, while the remaining variations could be real private variants. Most of these genes were solved because the toplevel references present different haplotypes that likely account for variants that we saw in the original gene. For example, the KIR2DL3 gene, coding the killer cell immunoglobulin like receptor, is known to be highly polymorphic [290, 291] and many alternate loci for this gene were introduced in the toplevel releases. Other solved genes are known to have paralogous genes that were not reported in the GRCh37 genome release. For instance, the PRIM2 paralog, missing in GRCh37 [292] is present in the modeled centromere for chromosome 3 in GRCh38 [134]. This paralog contains only exons 6-14 of the original transcripts [292], that actually are the exons covered by the enriched target regions we selected.

We classified 6 genes as *unsolved* in GRCh38 since both the two pools of reads aligned only to the target gene, indicating that neither duplicated regions nor haplotypes are known. In fact, all variants localized in these unsolved genes were still present analyzing the dataset with the GRCh38 toplevel reference. For this reason, we suppose that also such genes could

present duplicated or haplotype sequences not yet reported even in the latest version of the human genome.

We classified the remaining 5 genes as *partially solved* in GRCh38 since only a portion of the enriched target regions for these genes was solved by the introduction of new sequences in the references, whereas the reads aligning on the remaining target regions behave as those aligning on genes we classified as unsolved. Also in this group we found interesting genes worthy of being further investigated, as for example the MAP2K3 gene which is known to play an important role in tumor invasion and progression [293, 294].

Solved genes			Partially solved genes	Unsolved genes
BCLAF1	KIR2DS4	OR4C45	FAM104B	ALG1L2
CCDC144NL	KRT6B	OR9G1	FRG2B	ANKRD36
CES1	KRTAP4-11	PDE4DIP	FRG2C	FAM131C
CTBP2	KRTAP9-2	PPYR1	KCNJ12	PDPR
FRG1	MUC20	PRIM2	MAP2K3	PCDH11X
GPRIN2	NBPF10	PRSS3		PER3
HLA-DQA2	NBPF1	SEC22B		
HNRNPCL1	NOTCH2NL	TPTE		
HYDIN	OR1D5	ZDHHC11		
KIR2DL3	OR4C3	LOC653486		
FAM194B	MLL3	OR4M2		
TNXB				

Table 8|**Genes classification into solved, partially solved and unsolved cases.**

Variant database

The results of the analyses, performed on all the variants included into the study dataset, have been gathered within an exhaustive database, which should help geneticists and clinicians to discriminate interesting variants from false positives. The process of database creation is shown in Figure 24, which also allows to summarize all the steps carried out during the study.

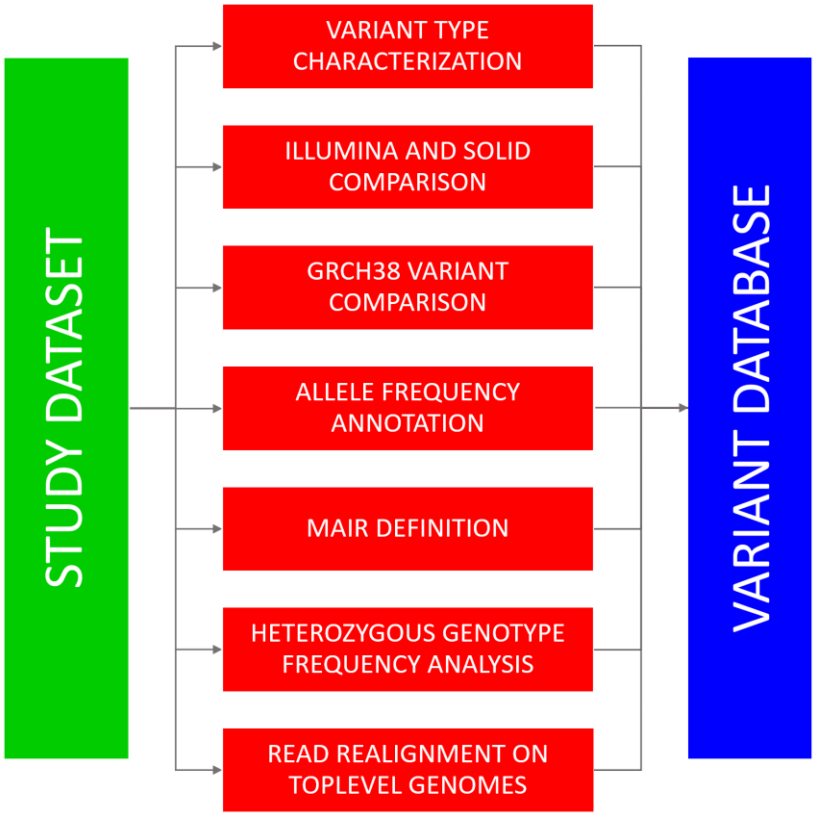


Figure 24|**Variant database creation process**. The building of the variant database passes through several analyses (red boxes) performed on the 264303 variants of the study dataset. Each investigation can return one or more results which constitute the columns of the final table.

4.3.4. Discussion

Whole exome sequencing is a powerful tool for analyzing first and foremost human genetic variations and rare hereditary diseases. Nevertheless, finding the appropriate answer is a complex task while handling such a big amount of data as those obtained in exome sequencing projects. It is thus mandatory to perform the most reliable analysis as possible in order to reduce errors. In particular, during alignment and variant calling, a central role is played by the reference used, as it should be representative of the total possible variations in order to highlight known and unknown mutations.

In this work, we presented a comprehensive study on a dataset composed by variants found in 222 exomes. Our investigations were targeted to find possible explanations about the presence of anomalous variants and to develop strategies able to individuate, characterize and filter them. For clarifying such strangeness, the existence of inconsistencies in the version GRCh37 of the human reference genome, which is suggested by the companies for alignment and variant calling steps, was put at the basis of our hypothesis. Following this assumption, we tried to achieve a feasible explanation for the 9313 variants that we found unexpectedly shared by at least the 90% of the samples. In particular, through a wide set of

analyses (Figure 24), ranging from allele and genotype frequencies comparison to read realignment on different reference genomes, we demonstrated that: I) 8680 variants are MAiR, meaning that the reference does not carry the most frequent allele in the population, II) 316 are possible indicators of gene or region duplications, III) 82 are both MAiR and with an unbalanced heterozygous genotype frequency, thus involving the issues of points I and II, IV) 16, among which 1 is also MAiR, could be Ion Proton specific errors as they are absent in Illumina and SOLiD samples, V) only 219 stand without a clear explanation. Among the latter, 41 variants have never been previously reported, while for the remaining we can hypothesize they could be population specific polymorphisms. In fact, although we did not detect any differences between the two populations for the whole set of variants (Figure 22), the frequencies of these 178 variants are significantly higher for the Europeans compared to the total population as reported in Figure 25.

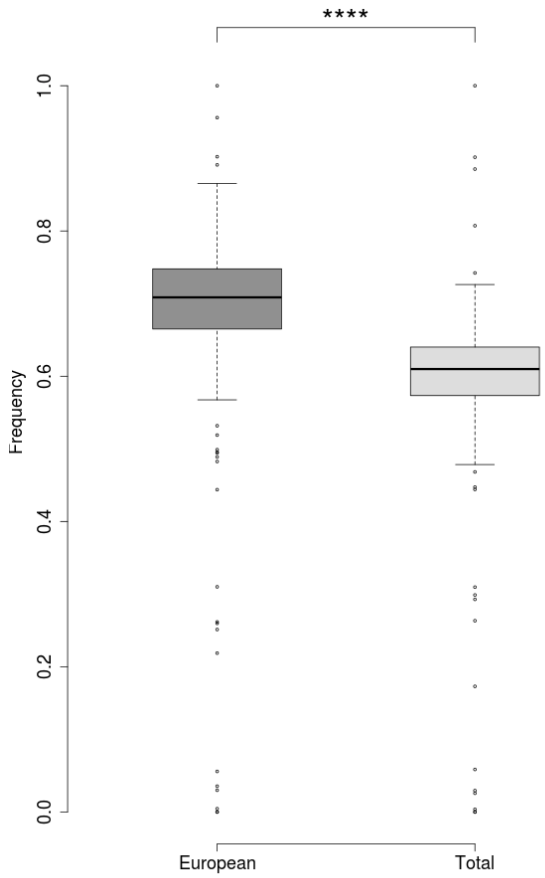


Figure 25|**European and Total frequencies of the 178 possible population specific polymorphisms.** Contrary to what shown in Figure 22, the difference between allele frequencies in the two analyzed datasets is highly significant (p-value<0.0001).

These findings strongly supported our theory regarding the presence of possible errors in the GRCh37 genome: in fact, the identification of MAiR variants can be essentially associated with uncorrected bases, while the detection of variants with an unbalanced

heterozygous genotype frequency showed that at least 45 genes were affected by assembly problems, mainly due to the presence of collapsed repeats within the primary sequence of the reference.

Furthermore, since in 2013 a new version of human genome has been released, we also decided to investigate if the issues affecting the GRCh37 have been solved in the GRCh38. Indeed, important ameliorations in the latest version of the human reference have been declared in the works published by Guo [281] and Schneider [134].

Nevertheless, our results indicated that although more than 8000 bases have been corrected in the most recent release of the human genome [134], other efforts are necessary to further reduce the base-pair-level errors, because more than 90% of MAiR found in GRCh37 are still kept in GRCh38. However, it is important to remember that the new definition of MAiR triggered the inclusion in such group of a higher number of variants than what we previously published [178]. So, to understand if these positions should be corrected in the reference, we planned to carry out in the near future a deeper analysis on the frequencies of variants associated to these positions, as only alternative alleles with a really high frequency in the population are worthy to be included within the primary sequence of the human genome.

Regarding the problem of the collapsed repeats, many suspicious genes (34) marked as duplicate in GRCh37 have been solved in the most recent release of human reference, indicating a significant improvement of the assembly. Anyway, as happened for the previous task, even this question has not been completed answered, because 11 genes could present a partial or an entire duplication since they behave similarly to the “solved” genes when the GRCh37 is used. Thus, it is important to find these possible new regions and to correctly report them in the assembly as their absence leads to improper reads alignment and variant calling, complicating the discovery of the real disease-causing variations.

4.3.5. Conclusion

We believe that the human reference genome should be the best possible representation of the known global variation. GRCh38 clearly goes in this way and we encourage the usage of this latest release in whole exome and whole genome studies. However, our results indicate that some inconsistencies are still there. Duplications could be deeply investigated when

whole genome sequencing and analysis will be the regular practice in human genetic studies. Only with this kind of data it will be possible to have a thorough insight on gene duplication and gene copy number variation. However, we are aware that some genetic differences could be ascribed to natural polymorphisms. A possible solution could come from the development of regional variant databases more comprehensive of the allelic frequencies of specific populations than the global frequencies. An essential step in this path is the free access to human exome data: we have no doubt that these data contain sensitive information but they are also extremely useful to improve the knowledge on human genetic variability and hereditary diseases. In this view, we hope that the variant catalog produced with this study could represent a little step forward to the solution of this issue, helping researchers in discriminating between really interesting variants and those that could mislead their work.

5. Concluding remarks and future perspectives

Since the middle of 2000s the development of high-throughput sequencing technologies led to a very fast decreasing of the sequencing cost, allowing nowadays to obtain a whole genome with less than $1000. This event shifted the main bottleneck of the workflow from the sequencing process to the data analysis [88], raising at the same time new issues regarding in particular the manipulation, the understanding and the storage of the produced data. Although various applications can be reached applying the NGS, one of the most widespread is surely the study of DNA variations for discovering the molecular basis of genetic diseases [45].

As the interest in this sector has been growing steadily over the years, in 2014 the BioInfoGen Strategic Project was funded by the University of Padua, in order to establish new expertise in the areas of bioinformatics and molecular biology for approaching the personal genomics. My PhD project was framed within the BioInfoGen, with the goal to implement a series of tools for variant analysis and prioritization, easily applicable not only to the medical studies which I was involved on, but also to a wider spectrum of clinical cases. Obviously, the implementation of the various programs has required diversified developments, in order to adapt them to the state-of-the-art and to the specific needs and for this reason they have been separately described in the several chapters of the manuscript. Nevertheless, it is important to remember that such tools are not stand-alone software uniquely implemented to carry out a peculiar function, but they can be considered single step of a more complex pipeline specifically designed to approach to the personal genomics, which is finally the highest goal of the BioInfoGen project.

The first problem I faced up derived from the purchase of a new sequencer, the Illumina NextSeq 500, which has been affixed at the CRIBI center to the already working Ion Proton machine. Since few Illumina exome/genome data had been processed by the bioinformatics unit before the advent of NextSeq, no pipeline for variant detection was available, raising the need to implement a new one. Initially a deep bibliographic research was performed to understand which were the mandatory steps for an accurate data analysis, giving particular importance to the software parameters. Then, thanks to the possibility of using a HPC cluster, various strategies to parallelize the pipeline were evaluated, finding in the scatter-gather procedure, a performing and secure way to speed-up the process. The obtained

pipeline is able to process both paired-end and single-end reads and it can be used to detect germline and somatic mutations, exploiting two different variant calling suites, GATK HaplotypeCaller/MuTect2 and VarDict [143]. Another strength of the pipeline is the usability, as in its simple form it requires only few parameters, allowing also to the non-bioinformatician personnel to execute it. The reliability of the final results is guaranteed by several internal controls which permit to check out if each step has been correctly concluded. Finally, such workflow has been applied for studying patients with poor prognostic factors from the Italian Trial of Adjuvant Chemotherapy Adenocarcinoma (ITACA-S) trial, leading to the publication of a congress abstract [186].

Considering the issues raised during the pipeline development, one of the most challenging task was the selection of the best parameters for performing the variant calling step. Moreover, even if variant callers currently integrate many statistical and machine learning approaches useful to define quite affordable cohorts of variations, for obtaining the most reliable results, the starting material, including the reference genome, must be as precise as possible. Indeed, during our analyses of exome data produced by both Ion Proton and Illumina, it was surprising to notice that a group of variants was detected in more than 90% of the samples, even if not all of them were reported as common variants in the allele frequency databases. Excluding that such false-positives were due to platform specific errors, an extensive investigation of these variants was performed in order to create a clear repertoire of the recurrent miscalls for helping geneticists in analyzing exome data, but also to understand their origin, highlighting possible inconsistencies and errors of the reference genome. To validate our assumption, a lot of analyses were performed on the collected variants ranging from statistical tests on genotype frequencies, to read realignment on top-level genomes, passing through the comparison between two different versions of the reference genome and an extensive evaluation of allele frequencies reported in widely used databases, such as dbSNP [70], ExAC [68] and ESP6500 [69]. We demonstrated that collapsed repeats could be responsible for the call of many false-positive variants, several of which still remain using the latest release of the human assembly as reference. However, the majority of them are due to the inclusion of MAIR positions in the reference genome, suggesting the needing of a revision in the future updates. Furthermore, we gathered all the results of the various investigations within an exhaustive variant database, which should help geneticists and clinicians to discriminate interesting variants from false positives.

Although the analysis is essentially completed and almost ready to be published, it was planned to extend the study to variants included into bigger cohorts such as the 2504 individuals of the 1000 genomes project (http://www.internationalgenome.org/data) [86] and the 15496 genomes collected by gnomAD (http://gnomad.broadinstitute.org/downloads) [197]. This should allow to obtain a lot of information also on non-coding regions and not only on coding sequences analyzed by WES. In this way, it should be possible to have a really wide and complete overview of assembly problems along the whole genome sequence, thus identifying feasible assembly modifications which will ameliorate the final results of all sequencing projects. In fact, an improved reference will help the researchers to discriminate true variants, diminishing at the same time the number of false positives.

The strategies proposed in the previous paragraph, together with the application of suitably defined parameters in the variant calling process, should allow the selection of the most affordable variants. With the following step of annotation, variations are enriched with a lot of information, such as the position at transcript and protein level, the impact on protein sequence, the pathogenicity and many more. Several programs are currently available for performing this analysis, but at the end of 2014, when my PhD project was at the beginning, these tools were afflicted by many troubles mainly regarding the indels realignment, the variant normalization and the annotation of complex rearrangements. To solve these issues and to improve the overall quality of the annotation step, a new variant annotator, referred as VarPred, was developed. From the comparison with state-of-the-art software, including VEP [162], ANNOVAR [163] and SnpEff [188], VarPred showed the best consistency in variant annotation (>99%), but also good performance in running times, in particular at exome level. Also a web interface for an easy data filtration has been implemented, even if it is not completed yet. Further developments of both the stand-alone software and the web platform will mainly regard the introduction of new sources of annotation, including dbNSFP [161] and ENCODE project [295] data, in order to extend the information provided by VarPred also to pathogenic prediction scores and non-coding features.

The variant annotation process is fundamental for achieving a valuable prioritization, allowing to discover mutations which can be associated to the genetic disease under investigation. For this purpose, QueryOR [178] has been developed. QueryOR is the web-platform with the highest number and the widest spectrum of selectable criteria. In

particular, differently from the other tools which apply arbitrary thresholds to filter out variants, QueryOR works on the whole cohort of variants, sorting them for the number of the satisfied criteria. In this way, variants with the highest score will be shown on the top of the list, even if they do not satisfy all the imposed criteria. In addition, the comprehensiveness of the implemented criteria and the aptness to add new features together with a user-friendly interface make QueryOR very suitable to support researchers, clinicians and geneticists engaged in variant analyses.

So QueryOR has been successfully used in the case studies included within the BioInfoGen project. Although the objectives of the two investigations were different, the flexibility of the platform allowed to reach in both cases interesting results. Indeed, on one side it helped the detection of causative variants in LSD patients, pointing out also the good performance of the designed panel, while on the other hand QueryOR facilitated the identification of the LRP2 gene as possible candidate for explaining the phenotype of DD subjects with no mutation in the previously identified disease-associated genes, CLCN5 and OCRL.

Concluding, the proposed thesis covers the various steps of variant analysis from the raw data manipulation to the final prioritization process, passing through the read alignment, the variant calling and the annotation. Many issues identified only a few years ago have been solved through the development of appropriate tools, such as VarPred and QueryOR, while others, raised during the PhD, have laid the groundwork for future projects. In fact, most of the work done during these three years was centered on the study of the coding regions, but thanks to the continuous decreasing of the sequencing cost, new scenarios over the non-coding side of genomics will be opened, leading to a continuous updating of all the bioinformatic programs including those I described in this manuscript.

The road to personal genomics is still hard and long, but I will work for it!

6. References

1. Jay E, Bambara R, Padmanabhan R, Wu R: **DNA sequence analysis: a general, simple and rapid method for sequencing large oligodeoxyribonucleotide fragments by mapping**. *Nucleic Acids Res.* 1974, **1**:331–353.

2. Salser W, Fry K, Brunk C, Poon R: **Nucleotide Sequencing of DNA: Preliminary Characterization of the Products of Specific Cleavages at Guanine, Cytosine, or Adenine Residues**. *Proceedings of the National Academy of Sciences* 1972, **69**:238–242.

3. Salser WA: **DNA Sequencing Techniques**. *Annu. Rev. Biochem.* 1974, **43**:923–965.

4. Maxam AM, Gilbert W: **A new method for sequencing DNA**. *Proc. Natl. Acad. Sci. U. S. A.* 1977, **74**:560–564.

5. Sanger F, Coulson AR: **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase**. *J. Mol. Biol.* 1975, **94**:441–448.

6. Maxam AM, Gilbert W: **Sequencing end-labeled DNA with base-specific chemical cleavages**. *Methods Enzymol.* 1980, **65**:499–560.

7. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc. Natl. Acad. Sci. U. S. A.* 1977, **74**:5463–5467.

8. Sanger F, Coulson AR: **The use of thin acrylamide gels for DNA sequencing**. *FEBS Lett.* 1978, **87**:107–110.

9. Martin WJ, Warmington JR, Galinski BR, Gallagher M, Wayne Davies R, Beck MS, Oliver SG: **Automation of DNA Sequencing: A System to Perform the Sanger Dideoxysequencing Reactions**. *Biotechnology* 1985, **3**:911–915.

10. Cohen AS, Najarian D, Smith JA, Karger BL: **Rapid separation of DNA restriction fragments using capillary electrophoresis**. *J. Chromatogr.* 1988, **458**:323–333.

11. Swerdlow H, Gesteland R: **Capillary gel electrophoresis for rapid, high resolution DNA sequencing**. *Nucleic Acids Res.* 1990, **18**:1415–1419.

12. Smith LM: **Automated DNA sequencing and the analysis of the human genome**. *Genome* 1989, **31**:929–937.

13. Lander ES, Linton LM, Birren B, et al.: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860–921.

14. Venter JC: **The Sequence of the Human Genome**. *Science* 2001, **291**:1304–1351.

15. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**:931–945.

16. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C: **Ten years of next-generation sequencing**

technology. *Trends Genet.* 2014, **30**:418–426.

17. Service RF: **Gene sequencing. The race for the $1000 genome**. *Science* 2006, **311**:1544–1546.

18. Grada A, Weinbrecht K: **Next-generation sequencing: methodology and application**. *J. Invest. Dermatol.* 2013, **133**:e11.

19. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M: **Comparison of next-generation sequencing systems**. *J. Biomed. Biotechnol.* 2012, **2012**:251364.

20. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nat. Biotechnol.* 2008, **26**:1135–1145.

21. Metzker ML: **Sequencing technologies - the next generation**. *Nat. Rev. Genet.* 2010, **11**:31–46.

22. Khan Z, Poetter K, Park DJ: **Enhanced solid phase PCR: mechanisms to increase priming by solid support primers**. *Anal. Biochem.* 2008, **375**:391–393.

23. Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, Griffiths AD: **Amplification of complex gene libraries by emulsion PCR**. *Nat. Methods* 2006, **3**:545–550.

24. Mitra RD, Church GM: **In situ localized amplification and contact replication of many individual DNA molecules**. *Nucleic Acids Res.* 1999, **27**:e34.

25. Lizardi PM, Huang X, Zhu Z, Bray-Ward P, Thomas DC, Ward DC: **Mutation detection and single-molecule counting using isothermal rolling-circle amplification**. *Nat. Genet.* 1998, **19**:225–232.

26. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, Dahl F, Fernandez A, Staker B, Pant KP, Baccash J, Borcherding AP, Brownley A, Cedeno R, Chen L, Chernikoff D, Cheung A, Chirita R, Curson B, Ebert JC, Hacker CR, Hartlage R, Hauser B, Huang S, Jiang Y, Karpinchyk V, Koenig M, Kong C, Landers T, Le C, Liu J, McBride CE, Morenzoni M, Morey RE, Mutch K, Perazich H, Perry K, Peters BA, Peterson J, Pethiyagoda CL, Pothuraju K, Richter C, Rosenbaum AM, Roy S, Shafto J, Sharanhovich U, Shannon KW, Sheppy CG, Sun M, Thakuria JV, Tran A, Vu D, Zaranek AW, Wu X, Drmanac S, Oliphant AR, Banyai WC, Martin B, Ballinger DG, Church GM, Reid CA: **Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays**. *Science* 2010, **327**:78–81.

27. Leamon JH, Lee WL, Tartaro KR, Lanza JR, Sarkis GJ, deWinter AD, Berka J, Weiner M, Rothberg JM, Lohman KL: **A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions**. *Electrophoresis* 2003, **24**:3769–3777.

28. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome**. *Science* 2005, **309**:1728–1732.

29. Bentley DR, Balasubramanian S, Swerdlow HP, et al.: **Accurate whole human genome sequencing using reversible terminator chemistry**. *Nature* 2008, **456**:53–59.

30. Kircher M, Kelso J: **High-throughput DNA sequencing--concepts and limitations**. *Bioessays* 2010, **32**:524–536.

31. Morozova O, Marra MA: **Applications of next-generation sequencing technologies in functional genomics**. *Genomics* 2008, **92**:255–264.

32. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**:376–380.

33. Shao K, Ding W, Wang F, Li H, Ma D, Wang H: **Emulsion PCR: a high efficient way of PCR amplification of random DNA libraries in aptamer selection**. *PLoS One* 2011, **6**:e24910.

34. Merriman B, Ion Torrent R&D Team, Rothberg JM: **Progress in ion torrent semiconductor chip based sequencing**. *Electrophoresis* 2012, **33**:3397–3417.

35. Jia H, Guo Y, Zhao W, Wang K: **Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer**. *Sci. Rep.* 2014, **4**:5737.

36. Caruccio N: **Preparation of next-generation sequencing libraries using Nextera™ technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition**. *Methods Mol. Biol.* 2011, **733**:241–255.

37. Adey A, Shendure J: **Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing**. *Genome Res.* 2012, **22**:1139–1143.

38. Thompson JF, Steinmann KE: **Single molecule sequencing with a HeliScope genetic analysis system**. *Curr. Protoc. Mol. Biol.* 2010, **Chapter 7**:Unit7.10.

39. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J: **An integrated semiconductor device enabling non-optical genome sequencing**. *Nature* 2011, **475**:348–352.

40. Ronaghi M: **Pyrosequencing sheds light on DNA sequencing**. *Genome Res.* 2001, **11**:3–11.

41. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM: **Accuracy and quality of**

massively parallel DNA pyrosequencing. *Genome Biol.* 2007, **8**:R143.

42. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: **A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers**. *BMC Genomics* 2012, **13**:341.

43. McKernan K, Blanchard A, Kotler L, Costa G: **Reagents, Methods, and Libraries for Bead-Based Sequencing**. *Patent* 2008.

44. Housby J: **Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides**. *Nucleic Acids Res.* 1998, **26**:4259–4266.

45. Mardis ER: **The impact of next-generation sequencing technology on genetics**. *Trends Genet.* 2008, **24**:133–141.

46. Chen F, Dong M, Ge M, Zhu L, Ren L, Liu G, Mu R: **The history and advances of reversible terminators used in new generations of sequencing technology**. *Genomics Proteomics Bioinformatics* 2013, **11**:34–40.

47. Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, Li X, Marma MS, Shi S, Wu J, Edwards JR, Romu A, Turro NJ: **Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators**. *Proc. Natl. Acad. Sci. U. S. A.* 2006, **103**:19635–19640.

48. Guo J, Xu N, Li Z, Zhang S, Wu J, Kim DH, Sano Marma M, Meng Q, Cao H, Li X, Shi S, Yu L, Kalachikov S, Russo JJ, Turro NJ, Ju J: **Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides**. *Proc. Natl. Acad. Sci. U. S. A.* 2008, **105**:9145–9150.

49. Wang Y, Yang Q, Wang Z: **The evolution of nanopore sequencing**. *Front. Genet.* 2014, **5**:449.

50. Schadt EE, Turner S, Kasarskis A: **A window into third generation sequencing**. *Hum. Mol. Genet.* 2010, **20**:853–853.

51. Krivanek OL, Chisholm MF, Nicolosi V, Pennycook TJ, Corbin GJ, Dellby N, Murfitt MF, Own CS, Szilagyi ZS, Oxley MP, Pantelides ST, Pennycook SJ: **Atom-by-atom structural and chemical analysis by annular dark-field electron microscopy**. *Nature* 2010, **464**:571–574.

52. Tanaka H, Kawai T: **Partial sequencing of a single DNA molecule with a scanning tunnelling microscope**. *Nat. Nanotechnol.* 2009, **4**:518–522.

53. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S: **Real-time DNA sequencing from single polymerase molecules**. *Science* 2009, **323**:133–138.

54. Thompson JF, Milos PM: **The properties and applications of single-molecule DNA sequencing**. *Genome Biol.* 2011, **12**:217.

55. Travers KJ, Chin C-S, Rank DR, Eid JS, Turner SW: **A flexible and efficient template format for circular consensus sequencing and SNP detection**. *Nucleic Acids Res.* 2010, **38**:e159.

56. Lu H, Giordano F, Ning Z: **Oxford Nanopore MinION Sequencing and Genome Assembly**. *Genomics Proteomics Bioinformatics* 2016, **14**:265–279.

57. Rhoads A, Au KF: **PacBio Sequencing and Its Applications**. *Genomics Proteomics Bioinformatics* 2015, **13**:278–289.

58. Korlach J, Bibillo A, Wegener J, Peluso P, Pham TT, Park I, Clark S, Otto GA, Turner SW: **Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides**. *Nucleosides Nucleotides Nucleic Acids* 2008, **27**:1072–1083.

59. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR: **Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome**. *Genome Res.* 2015, **25**:1750–1756.

60. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H: **Continuous base identification for single-molecule nanopore DNA sequencing**. *Nat. Nanotechnol.* 2009, **4**:265–270.

61. Derrington IM, Craig JM, Stava E, Laszlo AH, Ross BC, Brinkerhoff H, Nova IC, Doering K, Tickman BI, Ronaghi M, Mandell JG, Gunderson KL, Gundlach JH: **Subangstrom single-molecule measurements of motor proteins using a nanopore**. *Nat. Biotechnol.* 2015, **33**:1073–1075.

62. Pennisi E: **Genome sequencing. Search for pore-fection**. *Science* 2012, **336**:534–537.

63. Bayley H: **Nanopore sequencing: from imagination to reality**. *Clin. Chem.* 2015, **61**:25–31.

64. Derrington IM, Butler TZ, Collins MD, Manrao E, Pavlenok M, Niederweis M, Gundlach JH: **Nanopore DNA sequencing with MspA**. *Proc. Natl. Acad. Sci. U. S. A.* 2010, **107**:16060–16065.

65. Wendell D, Jing P, Geng J, Subramaniam V, Lee TJ, Montemagno C, Guo P: **Translocation of double-stranded DNA through membrane-adapted phi29 motor protein nanopores**. *Nat. Nanotechnol.* 2009, **4**:765–772.

66. Feng Y, Zhang Y, Ying C, Wang D, Du C: **Nanopore-based fourth-generation DNA sequencing technology**. *Genomics Proteomics Bioinformatics* 2015, **13**:4–16.

67. Di Ventra M, Taniguchi M: **Decoding DNA, RNA and peptides with quantum tunnelling**. *Nat. Nanotechnol.* 2016, **11**:117–126.

68. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, Hamamsy T, Lek M, Samocha KE, Cummings BB, Birnbaum D, The Exome Aggregation

Consortium, Daly MJ, MacArthur DG: **The ExAC browser: displaying reference data information from over 60 000 exomes**. *Nucleic Acids Res.* 2017, **45**:D840–D845.

69. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, NHLBI Exome Sequencing Project, Akey JM: **Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants**. *Nature* 2013, **493**:216–220.

70. Sherry ST: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Res.* 2001, **29**:308–311.

71. Meienberg J, Bruggmann R, Oexle K, Matyas G: **Clinical sequencing: is WGS the better WES?** *Hum. Genet.* 2016, **135**:359–362.

72. Rabbani B, Tekin M, Mahdieh N: **The promise of whole-exome sequencing in medical genetics**. *J. Hum. Genet.* 2014, **59**:5–15.

73. Bodi K, Perera AG, Adams PS, Bintzler D, Dewar K, Grove DS, Kieleczawa J, Lyons RH, Neubert TA, Noll AC, Singh S, Steen R, Zianni M: **Comparison of commercially available target enrichment methods for next-generation sequencing**. *J. Biomol. Tech.* 2013, **24**:73–86.

74. Mertes F, ElSharawy A, Sauer S, van Helvoort JMLM, van der Zaag PJ, Franke A, Nilsson M, Lehrach H, Brookes AJ: **Targeted enrichment of genomic DNA regions for next-generation sequencing**. *Brief. Funct. Genomics* 2011, **10**:374–386.

75. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ: **Target-enrichment strategies for next-generation sequencing**. *Nat. Methods* 2010, **7**:111–118.

76. Niedzicka M, Fijarczyk A, Dudek K, Stuglik M, Babik W: **Molecular Inversion Probes for targeted resequencing in non-model organisms**. *Sci. Rep.* 2016, **6**:24051.

77. Xue Y, Ankala A, Wilcox WR, Hegde MR: **Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing**. *Genet. Med.* 2014, **17**:444–451.

78. Gómez J, Reguero JR, Morís C, Martín M, Alvarez V, Alonso B, Iglesias S, Coto E: **Mutation analysis of the main hypertrophic cardiomyopathy genes using multiplex amplification and semiconductor next-generation sequencing**. *Circ. J.* 2014, **78**:2963–2971.

79. Zhao Y, Feng Y, Zhang Y-M, Ding X-X, Song Y-Z, Zhang A-M, Liu L, Zhang H, Ding J-H, Xia X-S: **Targeted next-generation sequencing of candidate genes reveals novel mutations in patients with dilated cardiomyopathy**. *Int. J. Mol. Med.* 2015, **36**:1479–1486.

80. Fernández-Marmiesse A, Morey M, Pineda M, Eiris J, Couce M, Castro-Gago M, Fraga J, Lacerda L, Gouveia S, Pérez-Poyato M, Armstrong J, Castiñeiras D, Cocho JA: **Assessment of a targeted resequencing assay as a support tool in the diagnosis of lysosomal storage disorders**. *Orphanet J. Rare Dis.* 2014, **9**:59.

81. Hildebrand MS, Myers CT, Carvill GL, Regan BM, Damiano JA, Mullen SA, Newton MR, Nair U, Gazina EV, Milligan CJ, Reid CA, Petrou S, Scheffer IE, Berkovic SF, Mefford HC: **A targeted resequencing gene panel for focal epilepsy**. *Neurology* 2016, **86**:1605–1612.

82. Møller RS, Larsen LHG, Johannesen KM, Talvik I, Talvik T, Vaher U, Miranda MJ, Farooq M, Nielsen JEK, Svendsen LL, Kjelgaard DB, Linnet KM, Hao Q, Uldall P, Frangu M, Tommerup N, Baig SM, Abdullah U, Born AP, Gellert P, Nikanorova M, Olofsson K, Jepsen B, Marjanovic D, Al-Zehhawi LIK, Peñalva SJ, Krag-Olsen B, Brusgaard K, Hjalgrim H, Rubboli G, Pal DK, Dahl HA: **Gene Panel Testing in Epileptic Encephalopathies and Familial Epilepsies**. *Mol. Syndromol.* 2016, **7**:210–219.

83. Yohe S, Hauge A, Bunjer K, Kemmer T, Bower M, Schomaker M, Onsongo G, Wilson J, Erdmann J, Zhou Y, Deshpande A, Spears MD, Beckman K, Silverstein KAT, Thyagarajan B: **Clinical validation of targeted next-generation sequencing for inherited disorders**. *Arch. Pathol. Lab. Med.* 2015, **139**:204–210.

84. Pallen MJ: **Reply to Updating benchtop sequencing performance comparison**. *Nat. Biotechnol.* 2013, **31**:296–296.

85. Feuk L, Carson AR, Scherer SW: **Structural variation in the human genome**. *Nat. Rev. Genet.* 2006, **7**:85–97.

86. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, Konkel MK, Malhotra A, Stütz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer E-W, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, 1000 Genomes Project Consortium, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO: **An integrated map of structural variation in 2,504 human genomes**. *Nature* 2015, **526**:75–81.

87. Marx V: **Biology: The big challenges of big data**. *Nature* 2013, **498**:255–260.

88. Mardis ER: **The 1,000 genome, the 100,000 analysis?** *Genome Med.* 2010, **2**:84.

89. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**. *Nucleic Acids Res.* 2010, **38**:1767–1771.

90. **FASTX-Toolkit** [http://hannonlab.cshl.edu/fastx_toolkit/].

91. **dcjones/fastq-tools** [https://github.com/dcjones/fastq-tools].

92. **lh3/seqtk** [https://github.com/lh3/seqtk].

93. Droop AP: **fqtools: an efficient software suite for modern FASTQ file manipulation**.

Bioinformatics 2016, **32**:1883–1884.

94. Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y: **Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data**. *BMC Genomics* 2012, **13 Suppl 8**:S8.

95. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads**. *EMBnet.journal* 2011, **17**:10.

96. Criscuolo A, Brisse S: **AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads**. *Genomics* 2013, **102**:500–506.

97. Chen C, Khaleel SS, Huang H, Wu CH: **Software for pre-processing Illumina next-generation sequencing short read sequences**. *Source Code Biol. Med.* 2014, **9**:8.

98. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**:2114–2120.

99. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Galaxy Team: **Manipulation of FASTQ data with Galaxy**. *Bioinformatics* 2010, **26**:1783–1785.

100. Koparde VN, Parikh HI, Bradley SP, Sheth NU: **MEEPTOOLS: a maximum expected error based FASTQ read filtering and trimming toolkit**. *Int. J. Comput. Biol. Drug Des.* 2017, **10**:237.

101. Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J: **AfterQC: automatic filtering, trimming, error removing and quality control for fastq data**. *BMC Bioinformatics* 2017, **18**.

102. **Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data** [https://www.bioinformatics.babraham.ac.uk/projects/fastqc/].

103. **GATK | Best Practices** [https://software.broadinstitute.org/gatk/best-practices/].

104. Reinert K, Langmead B, Weese D, Evers DJ: **Alignment of Next-Generation Sequencing Reads**. *Annu. Rev. Genomics Hum. Genet.* 2015, **16**:133–151.

105. Li H, Homer N: **A survey of sequence alignment algorithms for next-generation sequencing**. *Brief. Bioinform.* 2010, **11**:473–483.

106. Ajtai M: **The complexity of the pigeonhole principle**. In *[Proceedings 1988] 29th Annual Symposium on Foundations of Computer Science*. IEEE; 1988:346–355.

107. Ukkonen E: **Approximate string-matching with q-grams and maximal matches**. *Theor. Comput. Sci.* 1992, **92**:191–211.

108. Baeza-Yates and G. Navarro R, R Baeza-Yates And: **Faster Approximate String Matching**. *Algorithmica* 1999, **23**:127–158.

109. Jiang H, Wong WH: **SeqMap: mapping massive amount of oligonucleotides to the genome**. *Bioinformatics* 2008, **24**:2395–2396.

110. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program**. *Bioinformatics* 2008, **24**:713–714.

111. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Res.* 2008, **18**:1851–1858.

112. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat. Methods* 2012, **9**:357–359.

113. Siragusa E, Weese D, Reinert K: **Fast and accurate read mapping with approximate seeds and multiple backtracking**. *Nucleic Acids Res.* 2013, **41**:e78.

114. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M: **SHRiMP: accurate mapping of short color-space reads**. *PLoS Comput. Biol.* 2009, **5**:e1000386.

115. Weese D, Emde A-K, Rausch T, Döring A, Reinert K: **RazerS--fast read mapping with sensitivity control**. *Genome Res.* 2009, **19**:1646–1654.

116. Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC: **mrsFAST: a cache-oblivious algorithm for short-read mapping**. *Nat. Methods* 2010, **7**:576–577.

117. Hach F, Sarrafi I, Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC: **mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications**. *Nucleic Acids Res.* 2014, **42**:W494–500.

118. Weiner P: **Linear pattern matching algorithms**. In *14th Annual Symposium on Switching and Automata Theory (swat 1973)*. 1973.

119. Manber U, Myers G: **Suffix Arrays: A New Method for On-Line String Searches**. *SIAM J. Comput.* 1993, **22**:935–948.

120. Abouelhoda MI, Kurtz S, Ohlebusch E: **The Enhanced Suffix Array and Its Applications to Genome Analysis**. In *Lecture Notes in Computer Science*. 2002:449–463.

121. Ferragina P, Manzini G: **An experimental study of a compressed index**. *Inf. Sci.* 2001, **135**:13–28.

122. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes**. *Genome Biol.* 2004, **5**:R12.

123. Meek C, Patel JM, Kasetty S: **OASIS**. In *Proceedings 2003 VLDB Conference*. 2003:910–921.

124. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J: **Fast mapping of short sequences with mismatches, insertions and deletions using index structures**. *PLoS Comput. Biol.* 2009, **5**:e1000502.

125. Abouelhoda MI, Kurtz S, Ohlebusch E: **Replacing suffix trees with enhanced suffix arrays**. *J. Discrete Algorithms* 2004, **2**:53–86.

126. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**:1754–1760.

127. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment**. *Bioinformatics* 2009, **25**:1966–1967.

128. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol.* 2009, **10**:R25.

129. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**:2078–2079.

130. **[No title]** [https://samtools.github.io/hts-specs/SAMv1.pdf].

131. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler a. D: **The Human Genome Browser at UCSC**. *Genome Res.* 2002, **12**:996–1006.

132. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. *Nat. Genet.* 2011, **43**:491–498.

133. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data**. *Nat. Rev. Genet.* 2011, **12**:443–451.

134. Schneider VA, Lindsay TG, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, Fulton RS, Kremitzki M, Magrini V, Markovic C, McGrath S, Steinberg KM, Auger K, Chow W, Collins J, Harden G, Hubbard T, Pelan S, Simpson JT, Threadgold G, Torrance J, Wood J, Clarke L, Koren S, Boitano M, Li H, Chin C-S, Phillippy AM, Durbin R, Wilson RK, Flicek P, Church DM: **Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly**. 2016.

135. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Keira Cheetham R: **Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs**. *Bioinformatics* 2012, **28**:1811–1817.

136. Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, Dugas M: **Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data**. *Sci. Rep.* 2017, **7**:43169.

137. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK-S, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J: **The diploid genome sequence of an Asian individual**. *Nature* 2008, **456**:60–65.

138. Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing**. *arXiv preprint* 2012.

139. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing**. *Genome Res.* 2012, **22**:568–576.

140. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N: **LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets**. *Nucleic Acids Res.* 2012, **40**:11189–11201.

141. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H: **SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data**. *Nucleic Acids Res.* 2011, **39**:e132–e132.

142. **GATK | GATK | Tool Documentation Index** [https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php].

143. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, Dry JR: **VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research**. *Nucleic Acids Res.* 2016, **44**:e108.

144. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, WGS500 Consortium, Wilkie AOM, McVean G, Lunter G: **Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications**. *Nat. Genet.* 2014, **46**:912–918.

145. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, Wei Z, Wang K, Lyon GJ: **Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing**. *Genome Med.* 2013, **5**:28.

146. Hwang S, Kim E, Lee I, Marcotte EM: **Systematic comparison of variant calling pipelines using gold standard personal exome variants**. *Sci. Rep.* 2015, **5**.

147. Goldfeder RL, Priest JR, Zook JM, Grove ME, Waggott D, Wheeler MT, Salit M, Ashley EA: **Medical implications of technical accuracy in genome sequencing**. *Genome Med.* 2016, **8**:24.

148. Callari M, Sammut S-J, De Mattos-Arruda L, Bruna A, Rueda OM, Chin S-F, Caldas C: **Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers**. *Genome Med.* 2017, **9**:35.

149. Popitsch N, Schuh A, Taylor JC, WGS500 Consortium: **ReliableGenome: annotation of genomic regions with high/low variant calling concordance**. *Bioinformatics* 2016, **33**:155–160.

150. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter

G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group: **The variant call format and VCFtools**. *Bioinformatics* 2011, **27**:2156–2158.

151. **Website**.

152. Salgado D, Bellgard MI, Desvignes J-P, Béroud C: **How to Identify Pathogenic Mutations among All Those Variations: Variant Annotation and Filtration in the Genome Sequencing Era**. *Hum. Mutat.* 2016, **37**:1272–1282.

153. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation**. *Nature* 2015, **526**:68–74.

154. **Exome Variant Server** [http://evs.gs.washington.edu/EVS/].

155. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC: **SIFT web server: predicting effects of amino acid substitutions on proteins**. *Nucleic Acids Res.* 2012, **40**:W452–7.

156. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants**. *Nat. Genet.* 2014, **46**:310–315.

157. Quang D, Chen Y, Xie X: **DANN: a deep learning approach for annotating the pathogenicity of genetic variants**. *Bioinformatics* 2015, **31**:761–763.

158. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies**. *Genome Res.* 2010, **20**:110–121.

159. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res.* 2005, **15**:1034–1050.

160. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S: **Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP**. *PLoS Comput. Biol.* 2010, **6**:e1001025.

161. Liu X, Wu C, Li C, Boerwinkle E: **dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs**. *Hum. Mutat.* 2016, **37**:235–241.

162. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F: **The Ensembl Variant Effect Predictor**. *Genome Biol.* 2016, **17**:122.

163. Yang H, Wang K: **Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR**. *Nat. Protoc.* 2015, **10**:1556–1566.

164. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Juettemann T, Keenan S, Laird MR, Lavidas I, Maurel T, McLaren W, Moore B, Murphy DN,

Nag R, Newman V, Nuhn M, Ong CK, Parker A, Patricio M, Riat HS, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Wilder SP, Zadissa A, Kostadima M, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Cunningham F, Yates A, Zerbino DR, Flicek P: **Ensembl 2017**. *Nucleic Acids Res.* 2017, **45**:D635–D642.

165. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: **The UCSC Known Genes**. *Bioinformatics* 2006, **22**:1036–1046.

166. NCBI Resource Coordinators: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res.* 2016, **44**:D7–19.

167. Wright JC, Mudge J, Weisser H, Barzine MP, Gonzalez JM, Brazma A, Choudhary JS, Harrow J: **Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow**. *Nat. Commun.* 2016, **7**:11778.

168. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner M-M, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D: **The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes**. *Genome Res.* 2009, **19**:1316–1323.

169. den Dunnen JT: **Sequence Variant Descriptions: HGVS Nomenclature and Mutalyzer**. In *Current Protocols in Human Genetics*. 2016:7.13.1–7.13.19.

170. Münz M, Ruark E, Renwick A, Ramsay E, Clarke M, Mahamdallie S, Cloke V, Seal S, Strydom A, Lunter G, Rahman N: **CSN and CAVA: variant annotation tools for rapid, robust next-generation sequencing analysis in the clinical setting**. *Genome Med.* 2015, **7**:76.

171. Gene Ontology Consortium: **Gene Ontology Consortium: going forward**. *Nucleic Acids Res.* 2015, **43**:D1049–56.

172. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI: **DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants**. *Nucleic Acids Res.* 2017, **45**:D833–D839.

173. Robinson PN, Mundlos S: **The Human Phenotype Ontology**. *Clin. Genet.* 2010, **77**:525–534.

174. GTEx Consortium: **Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans**. *Science* 2015, **348**:648–660.

175. Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, Füllgrabe A, Fuentes AM-P, Jupp S, Koskinen S, Mannion O, Huerta L, Megy K, Snow C, Williams E, Barzine M, Hastings E, Weisser H, Wright J, Jaiswal P, Huber W, Choudhary J, Parkinson HE, Brazma A: **Expression Atlas update--an integrated database of gene and protein expression in**

humans, animals and plants. *Nucleic Acids Res.* 2016, **44**:D746–52.

176. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M: **KEGG as a reference resource for gene and protein annotation**. *Nucleic Acids Res.* 2016, **44**:D457–62.

177. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P: **The Reactome pathway Knowledgebase**. *Nucleic Acids Res.* 2016, **44**:D481–7.

178. Bertoldi L, Forcato C, Vitulo N, Birolo G, De Pascale F, Feltrin E, Schiavon R, Anglani F, Negrisolo S, Zanetti A, D'Avanzo F, Tomanin R, Faulkner G, Vezzi A, Valle G: **QueryOR: a comprehensive web platform for genetic variant analysis and prioritization**. *BMC Bioinformatics* 2017, **18**:225.

179. Fokkema IFAC, Ivo F A, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT: **LOVD v.2.0: the next generation in gene variant databases**. *Hum. Mutat.* 2011, **32**:557–563.

180. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR: **ClinVar: public archive of interpretations of clinically relevant variants**. *Nucleic Acids Res.* 2016, **44**:D862–8.

181. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res.* 2010, **20**:1297–1303.

182. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA: **From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline**. *Curr. Protoc. Bioinformatics* 2013, **43**:11.10.1–33.

183. **Picard Tools - By Broad Institute** [http://broadinstitute.github.io/picard/].

184. Frelinger JA: **Big Data, Big Opportunities, and Big Challenges**. *J. Investig. Dermatol. Symp. Proc.* 2015, **17**:33–35.

185. **GATK | Queue + GATK** [https://software.broadinstitute.org/gatk/download/queue].

186. Di Bartolomeo M, Devecchi A, Canevari S, Pellegrinelli A, Pietrantonio F, Martinetti A, Paoli A, Penso D, Valle G, Disciglio V, Bertoldi L, Feltrin E, Rosati G, De Vita F, Dominoni F, De Braud FG, Berenato R, Niger M, Miceli R, De Cecco L: **Whole-exome sequencing in radically resected gastric cancer (GC): Analysis of patients (pts) with poor prognostic factors from the Italian Trial of Adjuvant Chemotherapy Adenocarcinoma (ITACA-S) trial**. *J. Clin. Oncol.* 2017, **35**:64.

187. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J,

Roux A-F, Smith T, Antonarakis SE, Taschner PEM: **HGVS Recommendations for the Description of Sequence Variants: 2016 Update**. *Hum. Mutat.* 2016, **37**:564–569.

188. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3**. *Fly* 2012, **6**:80–92.

189. Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, Marth GT, Stein L, Flicek P, Yandell M, Eilbeck K: **A standard variation file format for human genome sequences**. *Genome Biol.* 2010, **11**:R88.

190. Schaafsma GCP, Vihinen M: **VariOtator, a Software Tool for Variation Annotation with the Variation Ontology**. *Hum. Mutat.* 2016, **37**:344–349.

191. Vihinen M: **Variation Ontology for annotation of variation effects and mechanisms**. *Genome Res.* 2014, **24**:356–364.

192. **[No title]** [http://snpeff.sourceforge.net/VCFannotationformat_v1.0.pdf].

193. Tan A, Abecasis GR, Kang HM: **Unified representation of genetic variants**. *Bioinformatics* 2015, **31**:2202–2204.

194. Wei L, Liu LT, Conroy JR, Hu Q, Conroy JM, Morrison CD, Johnson CS, Wang J, Liu S: **MAC: identifying and correcting annotation for multi-nucleotide variations**. *BMC Genomics* 2015, **16**:569.

195. Frankish A, Uszczynska B, Ritchie GRS, Gonzalez JM, Pervouchine D, Petryszak R, Mudge JM, Fonseca N, Brazma A, Guigo R, Harrow J: **Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction**. *BMC Genomics* 2015, **16 Suppl 8**:S2.

196. Li H: **Tabix: fast retrieval of sequence features from generic TAB-delimited files**. *Bioinformatics* 2011, **27**:718–719.

197. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium: **Analysis of protein-coding genetic variation in 60,706 humans**. *Nature* 2016, **536**:285–291.

198. Forbes SA, Beare D, Bindal N, Bamford S, Ward S, Cole CG, Jia M, Kok C, Boutselakis H,

De T, Sondka Z, Ponting L, Stefancsik R, Harsha B, Tate J, Dawson E, Thompson S, Jubb H, Campbell PJ: **COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer**. *Curr. Protoc. Hum. Genet.* 2016, **91**:10.11.1–10.11.37.

199. **Help - Glossary - Homo sapiens - Ensembl genome browser 90** [http://www.ensembl.org/Help/Glossary?id=346].

200. Zhou W, Chen T, Chong Z, Rohrdanz MA, Melott JM, Wakefield C, Zeng J, Weinstein JN, Meric-Bernstam F, Mills GB, Chen K: **TransVar: a multilevel variant annotator for precision genomics**. *Nat. Methods* 2015, **12**:1002–1003.

201. Fonseca NA, Rung J, Brazma A, Marioni JC: **Tools for mapping high-throughput sequencing data**. *Bioinformatics* 2012, **28**:3169–3177.

202. Leung RKK, Tsui SKW: **Alns: a new searchable and filterable sequence alignment format**. *Int. J. Data Min. Bioinform.* 2013, **7**:135–145.

203. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC: **SIFT missense predictions for genomes**. *Nat. Protoc.* 2016, **11**:1–9.

204. Adzhubei I, Jordan DM, Sunyaev SR: **Predicting functional effect of human missense mutations using PolyPhen-2**. *Curr. Protoc. Hum. Genet.* 2013, **Chapter 7**:Unit7.20.

205. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE: **Rare-disease genetics in the era of next-generation sequencing: discovery to translation**. *Nat. Rev. Genet.* 2013, **14**:681–691.

206. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent L-C, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y: **Gene prioritization through genomic data fusion**. *Nat. Biotechnol.* 2006, **24**:537–544.

207. Sifrim A, Popovic D, Tranchevent L-C, Ardeshirdavani A, Sakai R, Konings P, Vermeesch JR, Aerts J, De Moor B, Moreau Y: **eXtasy: variant prioritization by genomic data fusion**. *Nat. Methods* 2013, **10**:1083–1084.

208. Zemojtel T, Köhler S, Mackenroth L, Jäger M, Hecht J, Krawitz P, Graul-Neumann L, Doelken S, Ehmke N, Spielmann M, Oien NC, Schweiger MR, Krüger U, Frommer G, Fischer B, Kornak U, Flöttmann R, Ardeshirdavani A, Moreau Y, Lewis SE, Haendel M, Smedley D, Horn D, Mundlos S, Robinson PN: **Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome**. *Sci. Transl. Med.* 2014, **6**:252ra123.

209. Yang H, Robinson PN, Wang K: **Phenolyzer: phenotype-based prioritization of candidate genes for human diseases**. *Nat. Methods* 2015, **12**:841–843.

210. Robinson PN, Köhler S, Oellrich A, Sanger Mouse Genetics Project, Wang K, Mungall CJ, Lewis SE, Washington N, Bauer S, Seelow D, Krawitz P, Gilissen C, Haendel M, Smedley D: **Improved exome prioritization of disease genes through cross-species phenotype comparison**. *Genome Res.* 2014, **24**:340–348.

211. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, Siragusa E, Zemojtel T, Buske OJ, Washington NL, Bone WP, Haendel MA, Robinson PN: **Next-generation diagnostics and disease-gene discovery with the Exomiser**. *Nat. Protoc.* 2015, **10**:2004–2015.

212. Singleton MV, Guthery SL, Voelkerding KV, Chen K, Kennedy B, Margraf RL, Durtschi J, Eilbeck K, Reese MG, Jorde LB, Huff CD, Yandell M: **Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families**. *Am. J. Hum. Genet.* 2014, **94**:599–610.

213. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat. Genet.* 2000, **25**:25–29.

214. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GCM, Brown DL, Brudno M, Campbell J, FitzPatrick DR, Eppig JT, Jackson AP, Freson K, Girdea M, Helbig I, Hurst JA, Jähn J, Jackson LG, Kelly AM, Ledbetter DH, Mansour S, Martin CL, Moss C, Mumford A, Ouwehand WH, Park S-M, Riggs ER, Scott RH, Sisodiya S, Van Vooren S, Wapner RJ, Wilkie AOM, Wright CF, Vulto-van Silfhout AT, de Leeuw N, de Vries BBA, Washingthon NL, Smith CL, Westerfield M, Schofield P, Ruef BJ, Gkoutos GV, Haendel M, Smedley D, Lewis SE, Robinson PN: **The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data**. *Nucleic Acids Res.* 2014, **42**:D966–74.

215. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, Parkinson H, Schriml LM: **Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data**. *Nucleic Acids Res.* 2014, **43**:D1071–D1078.

216. Kornblihtt AR, Schor IE, Alló M, Dujardin G, Petrillo E, Muñoz MJ: **Alternative splicing: a pivotal step between eukaryotic transcription and translation**. *Nat. Rev. Mol. Cell Biol.* 2013, **14**:153–165.

217. Alemán A, Garcia-Garcia F, Salavert F, Medina I, Dopazo J: **A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies**. *Nucleic Acids Res.* 2014, **42**:W88–93.

218. Santoni FA, Makrythanasis P, Nikolaev S, Guipponi M, Robyr D, Bottani A, Antonarakis SE: **Simultaneous identification and prioritization of variants in familial, de novo, and somatic genetic disorders with VariantMaster**. *Genome Res.* 2014, **24**:349–355.

219. Liu X, Jian X, Boerwinkle E: **dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations**. *Hum. Mutat.* 2013, **34**:E2393–E2402.

220. **FTP Download** [http://grch37.ensembl.org/info/data/ftp/].

221. **dbSNP Home Page** [http://www.ncbi.nlm.nih.gov/SNP/].

222. **OMIM - Online Mendelian Inheritance in Man** [http://www.omim.org/].

223. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, Teague JW, Stratton MR, McDermott U, Campbell PJ: **COSMIC: exploring the world's knowledge of somatic mutations in human cancer**. *Nucleic Acids Res.* 2014, **43**:D805–D811.

224. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJA, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C: **InterPro: the integrative protein signature database**. *Nucleic Acids Res.* 2009, **37**:D211–D215.

225. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic Acids Res.* 2011, **40**:D109–D114.

226. Sales G, Calura E, Cavalieri D, Romualdi C: **graphite - a Bioconductor package to convert pathway topology to gene network**. *BMC Bioinformatics* 2012, **13**:20.

227. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol.* 2004, **5**:R80.

228. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat. Methods* 2008, **5**:621–628.

229. Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI: **DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes**. *Database* 2015, **2015**:bav028.

230. de Ligt J, Willemsen MH, van Bon BWM, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, del Rosario M, Hoischen A, Scheffer H, de Vries BBA, Brunner HG, Veltman JA, Vissers LELM: **Diagnostic exome sequencing in persons with severe intellectual disability**. *N. Engl. J. Med.* 2012, **367**:1921–1929.

231. **EGAS00001000287 < Studies < EMBL-EBI** [https://www.ebi.ac.uk/ega/studies/EGAS00001000287].

232. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics**. *Nucleic Acids Res.* 2011, **39**:e118.

233. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**:1061–1073.

234. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *Nat. Biotechnol.* 2011, **29**:24–26.

235. Schwarz JM, Cooper DN, Schuelke M, Seelow D: **MutationTaster2: mutation prediction for the deep-sequencing age**. *Nat. Methods* 2014, **11**:361–362.

236. Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zöllner S, Rosenberg NA, Li JZ: **Long runs of homozygosity are enriched for deleterious variation**. *Am. J. Hum. Genet.* 2013, **93**:90–102.

237. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat. Protoc.* 2009, **4**:44–57.

238. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ: **Exome sequencing identifies the cause of a mendelian disorder**. *Nat. Genet.* 2010, **42**:30–35.

239. Antanaviciute A, Watson CM, Harrison SM, Lascelles C, Crinnion L, Markham AF, Bonthron DT, Carr IM: **OVA: integrating molecular and physical phenotype data from multiple biomedical domain ontologies with variant filtering for enhanced variant prioritization**. *Bioinformatics* 2015, **31**:3822–3829.

240. Meikle PJ, Hopwood JJ, Clague AE, Carey WF: **Prevalence of lysosomal storage disorders**. *JAMA* 1999, **281**:249–254.

241. **Ion AmpliSeq Designer** [https://www.ampliseq.com/browse.action].

242. **Website** [http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/placentalM ammals/].

243. **Website** [ftp://ftp.ensembl.org/pub/release-75/gtf/homo_sapiens/Homo_sapiens.GRCh37.75.gtf.gz].

244. Filocamo M, Mazzotti R, Corsolini F, Stroppiano M, Stroppiana G, Grossi S, Lualdi S, Tappino B, Lanza F, Galotto S, Biancheri R: **Cell Line and DNA Biobank From Patients Affected by Genetic Diseases**. *Open Journal of Bioresources* 2014, **1**:e2.

245. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q, Barash Y, Krainer AR, Jojic N, Scherer SW, Blencowe BJ, Frey BJ: **RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease**. *Science* 2015, **347**:1254806.

246. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes**. *Nature* 2012, **491**:56–65.

247. Ramachandran A, Micsinai M, Pe'er I: **CONDEX: Copy number detection in exome sequences**. In *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*. 2011.

248. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll SA, O'Donovan MC, Owen MJ, Kirov G, Sullivan PF, Hultman CM, Sklar P, Purcell SM: **Discovery and statistical genotyping of copy-number variation from whole-exome**

sequencing depth. *Am. J. Hum. Genet.* 2012, **91**:597–607.

249. Ivakhno S, Royce T, Cox AJ, Evers DJ, Cheetham RK, Tavaré S: **CNAseg--a novel framework for identification of copy number changes in cancer from second-generation sequencing data**. *Bioinformatics* 2010, **26**:3051–3058.

250. Gambin T, Akdemir ZC, Yuan B, Gu S, Chiang T, Carvalho CMB, Shaw C, Jhangiani S, Boone PM, Eldomery MK, Karaca E, Bayram Y, Stray-Pedersen A, Muzny D, Charng W-L, Bahrambeigi V, Belmont JW, Boerwinkle E, Beaudet AL, Gibbs RA, Lupski JR: **Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort**. *Nucleic Acids Res.* 2016:gkw1237.

251. Ceyhan-Birsoy O, Pugh TJ, Bowser MJ, Hynes E, Frisella AL, Mahanta LM, Lebo MS, Amr SS, Funke BH: **Next generation sequencing-based copy number analysis reveals low prevalence of deletions and duplications in 46 genes associated with genetic cardiomyopathies**. *Mol Genet Genomic Med* 2016, **4**:143–151.

252. Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C: **Human Splicing Finder: an online bioinformatics tool to predict splicing signals**. *Nucleic Acids Res.* 2009, **37**:e67–e67.

253. Devuyst O, Thakker RV: **Dent's disease**. *Orphanet J. Rare Dis.* 2010, **5**:28.

254. Picollo A, Pusch M: **Chloride/proton antiporter activity of mammalian CLC proteins ClC-4 and ClC-5**. *Nature* 2005, **436**:420–423.

255. Hichri H, Rendu J, Monnier N, Coutton C, Dorseuil O, Poussou RV, Baujat G, Blanchard A, Nobili F, Ranchin B, Remesy M, Salomon R, Satre V, Lunardi J: **From Lowe syndrome to Dent disease: correlations between mutations of the OCRL1 gene and clinical and biochemical phenotypes**. *Hum. Mutat.* 2011, **32**:379–388.

256. Anglani F, D'Angelo A, Bertizzolo LM, Tosetto E, Ceol M, Cremasco D, Bonfante L, Addis MA, Del Prete D, Dent Disease Italian Network: **Nephrolithiasis, kidney failure and bone disorders in Dent disease patients with and without CLCN5 mutations**. *Springerplus* 2015, **4**:492.

257. Kantarci S, Al-Gazali L, Hill RS, Donnai D, Black GCM, Bieth E, Chassaing N, Lacombe D, Devriendt K, Teebi A, Loscertales M, Robson C, Liu T, MacLaughlin DT, Noonan KM, Russell MK, Walsh CA, Donahoe PK, Pober BR: **Mutations in LRP2, which encodes the multiligand receptor megalin, cause Donnai-Barrow and facio-oculo-acoustico-renal syndromes**. *Nat. Genet.* 2007, **39**:957–959.

258. Densupsoontorn N, Sanpakit K, Vijarnsorn C, Pattaragarn A, Kangwanpornsiri C, Jatutipsompol C, Tirapongporn H, Jirapinyo P, Shah NP, Sturm AC, Tanner SM: **Imerslund-Gräsbeck syndrome: new mutation in amnionless**. *Pediatr. Int.* 2012, **54**:e19–21.

259. Christensen EI, Birn H: **Megalin and cubilin: multifunctional endocytic receptors**. *Nat. Rev. Mol. Cell Biol.* 2002, **3**:256–266.

260. Lloyd SE, Pearce SH, Fisher SE, Steinmeyer K, Schwappach B, Scheinman SJ, Harding B, Bolino A, Devoto M, Goodyer P, Rigden SP, Wrong O, Jentsch TJ, Craig IW, Thakker RV: **A common molecular basis for three inherited kidney stone diseases**. *Nature* 1996, **379**:445–449.

261. Thakker RV: **Pathogenesis of Dent's disease and related syndromes of X-linked nephrolithiasis**. *Kidney Int.* 2000, **57**:787–793.

262. Hoopes RR, Shrimpton AE, Knohl SJ, Hueber P, Hoppe B, Matyus J, Simckes A, Tasic V, Toenshoff B, Suchy SF, Nussbaum RL, Scheinman SJ: **Dent Disease with Mutations in OCRL1**. *Am. J. Hum. Genet.* 2005, **76**:260–267.

263. Longoni M, High FA, Russell MK, Kashani A, Tracy AA, Coletti CM, Hila R, Shamia A, Wells J, Ackerman KG, Wilson JM, Bult CJ, Lee C, Lage K, Pober BR, Donahoe PK: **Molecular pathogenesis of congenital diaphragmatic hernia revealed by exome sequencing, developmental data, and bioinformatics**. *Proc. Natl. Acad. Sci. U. S. A.* 2014, **111**:12450–12455.

264. **GATK | Home** [https://software.broadinstitute.org/gatk/].

265. **seqr** [https://seqr.broadinstitute.org/].

266. **Individual #00131892 - Shared database** [https://databases.lovd.nl/shared/individuals/00131892].

267. **Individual #00131950 - Shared database** [https://databases.lovd.nl/shared/individuals/00131950].

268. Dachy A, Paquot F, Debray G, Bovy C, Christensen EI, Collard L, Jouret F: **In-depth phenotyping of a Donnai-Barrow patient helps clarify proximal tubule dysfunction**. *Pediatr. Nephrol.* 2015, **30**:1027–1031.

269. Kantarci S, Donnai D, Noonan KM, Pober BR: **Donnai-Barrow Syndrome**. In *GeneReviews(®)*. edited by Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mefford HC, Stephens K, Amemiya A, Ledbetter N Seattle (WA): University of Washington, Seattle; 2008.

270. Pober BR, Longoni M, Noonan KM: **A review of Donnai-Barrow and facio-oculo-acoustico-renal (DB/FOAR) syndrome: clinical features and differential diagnosis**. *Birth Defects Res. A Clin. Mol. Teratol.* 2009, **85**:76–81.

271. Gorvin CM, Wilmer MJ, Piret SE, Harding B, van den Heuvel LP, Wrong O, Jat PS, Lippiat JD, Levtchenko EN, Thakker RV: **Receptor-mediated endocytosis and endosomal acidification is impaired in proximal tubule epithelial cells of Dent disease patients**. *Proc. Natl. Acad. Sci. U. S. A.* 2013, **110**:7014–7019.

272. Saito A, Pietromonaco S, Loo AK, Farquhar MG: **Complete cloning and sequencing of rat gp330/"megalin," a distinctive member of the low density lipoprotein receptor gene family**. *Proc. Natl. Acad. Sci. U. S. A.* 1994, **91**:9725–9729.

273. Schrauwen I, Sommen M, Claes C, Pinner J, Flaherty M, Collins F, Van Camp G: **Broadening the phenotype of LRP2 mutations: a new mutation in LRP2 causes a predominantly ocular phenotype suggestive of Stickler syndrome**. *Clin. Genet.* 2014, **86**:282–286.

274. Khalifa O, Al-Sahlawi Z, Imtiaz F, Ramzan K, Allam R, Al-Mostafa A, Abdel-Fattah M, Abuharb G, Nester M, Verloes A, Al-Zaidan H: **Variable expression pattern in Donnai-Barrow syndrome: Report of two novel LRP2 mutations and review of the literature**. *Eur. J. Med. Genet.* 2015, **58**:293–299.

275. Vasli N, Ahmed I, Mittal K, Ohadi M, Mikhailov A, Rafiq MA, Bhatti A, Carter MT, Andrade DM, Ayub M, Vincent JB, John P: **Identification of a homozygous missense mutation in LRP2 and a hemizygous missense mutation in TSPYL2 in a family with mild intellectual disability**. *Psychiatr. Genet.* 2016, **26**:66–73.

276. Leheste JR, Melsen F, Wellner M, Jansen P, Schlichting U, Renner-Müller I, Andreassen TT, Wolf E, Bachmann S, Nykjaer A, Willnow TE: **Hypocalcemia and osteopathy in mice with kidney-specific megalin gene defect**. *FASEB J.* 2003, **17**:247–249.

277. Storm T, Tranebjærg L, Frykholm C, Birn H, Verroust PJ, Nevéus T, Sundelin B, Hertz JM, Holmström G, Ericson K, Christensen EI, Nielsen R: **Renal phenotypic investigations of megalin-deficient patients: novel insights into tubular proteinuria and albumin filtration**. *Nephrol. Dial. Transplant* 2013, **28**:585–591.

278. **Human Genome Overview - Genome Reference Consortium** [http://www.ncbi.nlm.nih.gov/grc/human].

279. Li H: **Toward better understanding of artifacts in variant calling from high-coverage samples**. *Bioinformatics* 2014, **30**:2843–2851.

280. Miga KH, Eisenhart C, Kent WJ: **Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments**. *Nucleic Acids Res.* 2015, **43**:e133.

281. Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y: **Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis**. *Genomics* 2017, **109**:83–90.

282. Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P, 1000 Genomes Project Consortium: **Alignment of 1000 Genomes Project reads to reference assembly GRCh38**. *Gigascience* 2017, **6**:1–8.

283. **Website** [ftp://ftp.ensembl.org/pub/release-88/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.toplevel.fa.gz].

284. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L: **CrossMap: a versatile tool for coordinate conversion between genome assemblies**. *Bioinformatics* 2014, **30**:1006–1007.

285. **Website**

[ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/variants/humsavar.txt].

286. **taxonomy:"Homo sapiens (Human) [9606]" AND reviewed:yes in UniProtKB** [http://www.uniprot.org/uniprot/?query=taxonomy:%22Homo%20sapiens%20(Human)%20[9606]%22&fil=reviewed%3Ayes].

287. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *J. R. Stat. Soc. Series B Stat. Methodol.* 1995, **57**:289–300.

288. **FTP Download** [https://grch37.ensembl.org/info/data/ftp/index.html].

289. **FTP Download** [https://www.ensembl.org/info/data/ftp/index.html].

290. Middleton D, Gonzelez F: **The extensive polymorphism of KIR genes**. *Immunology* 2010, **129**:8–19.

291. Keaney L, Williams F, Meenagh A, Sleator C, Middleton D: **Investigation of killer cell immunoglobulin-like receptor gene diversity III. KIR2DL3**. *Tissue Antigens* 2004, **64**:188–194.

292. Genovese G, Handsaker RE, Li H, Kenny EE, McCarroll SA: **Mapping the Human Reference Genome's Missing Sequence by Three-Way Admixture in Latino Genomes**. *Am. J. Hum. Genet.* 2013, **93**:411–421.

293. Bossi G: **MKK3 as oncotarget**. *Aging* 2016, **8**:1–2.

294. Wysk M, Yang DD, Lu HT, Flavell RA, Davis RJ: **Requirement of mitogen-activated protein kinase kinase 3 (MKK3) for tumor necrosis factor-induced cytokine expression**. *Proc. Natl. Acad. Sci. U. S. A.* 1999, **96**:3763–3768.

295. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski LL, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard T, Kent J, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos JA, Weng Z, White KP, Hardison RC: **Defining functional DNA elements in the human genome**. *Proc. Natl. Acad. Sci. U. S. A.* 2014, **111**:6131–6138.