NGS data analysis: a review of major tools and pipeline frameworks for variant discovery

Ángel Canal-Alonso[1], Noelia Egido[1], Pedro Jiménez[1], Juan Manuel Corchado

[1] Departamento de Bioinformática y Biología Computacional, AIR Institute, Carbajosa de la Sagrada, España

E-mail: acanal@air-institute.com

Abstract: The analysis of genetic data has always been a problem due to the large amount of information available and the difficulty in isolating that which is relevant. However, over the years progress in sequencing techniques has been accompanied by a development of computer techniques to the current application of artificial intelligence. We can summarize the phases of sequence analysis in the following: quality assessment, alignment, pre-variant processing, variant calling and variant annotation. In this article we will review and comment on the tools used in each phase of genetic sequencing, and analyze the drawbacks and advantages offered by each of them.

Keywords: Machine learning; Bioinformatics; Next-Generation Sequencing, Pipeline

1. Introduction

In the last few years we have experienced an unprecedented evolution in terms of DNA sequencing, to the point that it has become one of the main tools in multiple areas of biomedical science. It all began with the discovery of the structure of deoxyribonucleic acid - DNA - by Watson and Crick [1], the genetic material from which every living organism develops, consisting of a double helix with two polymer chains complementary to each other. Each of these strands is formed by the union of nucleotides, organic molecules composed of a carbohydrate, a phosphate group and a nitrogen base; the latter are cyclic compounds whose identity - adenine (A), thymine (T), cytosine (C) or guanine (G) - will determine the type of nucleotide, and therefore, each of the strands that make up DNA can be specified as a sequence of these four letters. Thus, the DNA sequencing process consists of determining the precise order of these four nucleotides along a DNA molecule. The complementarity we have previously mentioned is due to the fact that these two chains are naturally linked to each other by specific chemical bonds between their nucleotides, in the form A-T and C-G - mainly due to their chemical properties, so we might said that if we know the sequence of one of the chains we will be able to find out the sequential order of the other due to the complementarity of its nitrogen bases.

In parallel, Sanger [2] and Maxam and Gilbert [3] managed to develop different methods for sequencing DNA molecules in the late 1970s, although the Sanger sequencing technique or dideoxi method became the prevalent tool during the following years until the emergence of next-generation sequencing methods. This technique is based on the DNA polymerization reaction itself, a process known as replication that occurs in all types of biological organisms, in which a chain of DNA is synthesized using its complementary chain as a template; in this process, the nucleotides used carry with them a different fluorescent molecule according to its nitrogen base, in addition to a small chemical modification that causes, when it joins the chain in synthesis, that it no longer has the

capacity to bind to the next nucleotide, stopping the process at this point. Consequently, our reaction mixture will contain a multitude of DNA molecules that differ in length, since this type of modified nucleotides join the chain by chance. Subsequently, a molecular biology technique known as capillary electrophoresis will be used, in which the DNA fragments are separated by size and go through a detector that collects the luminous signal emitted by the last nucleotide of each fragment, thus knowing the sequential order of the bases according to their passage through the detector. In the end we will obtain as a result the set of signals collected by the detector ordered from the smallest fragment to the complete molecule.

During the coming years DNA sequencing brought with it an enormous amount of achievements and applications that culminated in the completion of the Human Genome Project in 2004 [4], where the human genome sequence was first obtained - a genome is defined as the complete DNA sequence of an organism -. The development of this enormous and long project - which lasted almost 15 years -, in addition to the enormous variety of applications that it might have in the future, revealed the urgent need for more advanced sequencing technologies that would allow us to obtain the genome of an organism in a fast and relatively accessible way in economic terms for most laboratories. Thus, after the completion of the project, the National Human Genome Research Institute (NHGRI) initiated a funding program known as the $1000 Genome project [5], with the goal of having high-precision sequencing methods - less than 1 error per 10000 bases -, long read lengths, high performance, and a reduction in the cost of sequencing a genome to $1,000 within 10 years. This accelerated the arrival of new sequencing technologies, coined with the term of next generation sequencing to differentiate them from the classical first generation sequencing methods of Sanger and Maxam and Gilbert.

2. Next Generation Sequencing technologies

The new sequencing methods that began to emerge from this moment share a series of characteristics that ostensibly improve the performance of classical techniques [6]. Firstly, they are based on the preparation of DNA libraries to adapt the genome according to the sequencing technique to be used; secondly, the sequencing process occurs in parallel on multiple fragments of the initial genome, allowing thousands to millions of sequencing reactions to occur at the same time; and finally, the detection of the sequenced bases is carried out directly without the need for electrophoresis, greatly accelerating the entire process. Within this wide range of new sequencing techniques we will differentiate two main groups: on the one hand the second generation sequencing methods, also called short-reads sequencing or based on PCR amplification, and on the other hand the latest techniques, known as third generation sequencing, long-read based or real time sequencing.

2.1. Second Generation Sequencing

This type of technology was the first to emerge after the completion of the Human Genome Project since 2005, and its most characteristic features are the amplification of the fragments to be sequenced by PCR and the high parallelization of the process. Polymerase Chain Reaction (PCR) is a fundamental technique in molecular biology in which DNA molecules are amplified by carrying out an in vitro replication process; the main goal is to imitate this cellular process by mixing all the necessary molecular components in a test tube, resulting in a large number of exact copies of the original fragment. This stage will allow the optical signal detected by the machine to be much greater in each sequencing reaction, greatly improving the precision of the process. The preparation of the DNA library is specific to the sequencing method, although they all have several characteristics in common. First, DNA must be fragmented into pieces ranging from 400 to 1200 base pairs in a process known as 'shotgun sequencing', because fragmentation occurs at random in multiple positions in the genome, resulting in multiple pieces of DNA that overlap each other. It should be pointed out that in a sequencing experiment it is traditionally necessary to extract the DNA from the biological sample being studied, normally made up of millions of cells, each one with its own DNA molecule that is intended to be sequenced; this is why each region of the genome will be represented by multiple

fragments, which may or may not be overlapping each other. Secondly, all these fragments are attached to adaptive sequences, small fragments of DNA of known sequence that have a double function; on the one hand, they are in charge of initiating the process of amplification by PCR - the cellular replication machinery always needs a small sequence to initiate the DNA synthesis - and on the other hand, they will allow all the fragments to be sequenced to be anchored to a solid support where the sequencing process itself takes place [7].

Second generation sequencing technologies fall into two broad categories [8]: sequencing by ligation methods - SBL - where the detection of bases is performed by binding oligonucleotides marked with a fluorescent molecule, and sequencing by synthesis - SBS - where the detection of the signal is produced by incorporating a nucleotide into an elongating chain. Specifically, the main SBL - SOLiD - and SBS technologies will be described below, the latter classified into cyclic reversible termination (CRT) methods, with the Illumina platform leading, and single-nucleotide addition (SNA), with 454 pyrosequencing and Ion Torrent as major flagships.

SOLiD is not one of the most commonly used techniques today, being widely displaced by other sequencing methods for various reasons that will be understood after describing its functioning. First of all, it is necessary to highlight an aspect of the preparation of the DNA library previously explained, since in these methods the amplification of fragments by PCR takes place in tiny resin spheres that act as micro-reactors, in a process known as emulsion PCR; its main difference is that a mixture formed by an aqueous phase is added, which contains all the chemical reagents necessary for the amplification and will include each one of the microspheres. In this way, the process is completely parallel and independent for every sphere, each carrying a different DNA fragment. Subsequently, the sequencing process itself is carried out using chemically modified oligonucleotides, in such a way that they contain at one end a pair of known nucleotides - 1 of 16 possible combinations of existing nucleotides - followed by a series of universal molecules that have no specificity for the DNA sequence. When the oligonucleotide hybridizes through this pair of bases the system detects the fluorescent colour, after which it is cleaved out and follows the oligonucleotide binding process until it reaches the end of the fragment. Once a cycle has ended, a new sequencing process begins with the same fragment, but this time the oligonucleotide will join at base n + 1 to detect the rest of the bases that have not been sequenced in the first cycle [8]. After finishing the series of cycles the result is that the same fragment of DNA has been sequenced several times but changing the order of the pair of nucleotides that is detected, so in the end each nucleotide will be read several times, and therefore greatly reducing the error rate of the sequencing process. It is undoubtedly the method with the highest precision after Sanger sequencing, with a value of 99.94%, although this advantage is notably overshadowed by the many drawbacks it has, such as the short length of its reads or the long duration per run [9].

On the other hand, within the second generation technologies we find SBS, whose best known and most currently implemented method is the cyclic reversible termination with the Solexa/Illumina platform at the front (Illumina acquired Solexa in 2007). This method has a common feature with respect to the classical Sanger sequencing, the use of chemically modified nucleotides that when added to the elongating chain they prevent the union of another nucleotide behind. The following steps are carried out in each sequencing cycle: union of the DNA fragments and their adapters to a solid surface - amplification process within the preparation of the library, generating clusters with every original fragment -, addition of the necessary components for the synthesis of new DNA strands, including the fluorescently marked modified nucleotides, hybridization of the complementary nucleotide to the template sequence, washing of the unincorporated bases and detection of the fluorescent molecule, and finally separation of the terminal part of the nucleotide so that a new cycle can begin and the whole fragment can be sequenced [10]. The Illumina sequencing platform, with its wide range of sequencers with very disparate features and applications, is currently a leader in the high-throughput sequencing industry and most library preparation protocols are

compatible with Illumina technology. With the release of its new HiSeq X Ten sequencer, it was able to achieve to a large extent with the premises of NHGRI; specifically, it is undoubtedly the technology with the highest throughput to date - number of fragments sequenced per run -, the lowest cost per sequenced base - reaching the $1000 barrier -, and its average length of 300 nucleotides per fragment sequenced in its best performance make it valid for most applications [9].

Second generation technologies also brought another approach, the single-nucleotide addition sequencing method - SNA -, whose most important flagships are the pyrosequencing system 454 and the IonTorrent technology; both methods rely on a single signal to mark the incorporation of a base to an elongating DNA chain, so this time each of the nucleotides will be added one by one sequentially. The pyrosequencing technology has the honour of being the first to be released in 2005 after the completion of the HGP, becoming the first next-generation sequencing instrument. In this method the bases are always incorporated in the same order, and their detection occurs when a pyrophosphate is released in the formation of the chemical bond between nucleotides - pyrophosphates are molecules composed of two phosphate groups that are released when the original nucleotide becomes part of DNA -, after which they undergo a chemical reaction and are transformed into another compound, luciferase, which is capable of emitting a bioluminescent signal. Depending on where this luminous signal has been detected it will be possible to know in which fragment such a nucleotide has been added, that we know beforehand, as well as the intensity will tell us if multiple bases of the same type have been added. Ion Torrent's technology, on the other hand, was the first NGS instrument without using optical signals for base detection, substituted instead by a semiconductor system that will detect the union of a nucleotide by a change of pH in the medium. This variation is due to the release of protons in the DNA synthesis process, which will allow a chemical signal in this case to be transformed into digital information [8]. This type of systems arose from the pyrosequencing methodology, having as main advantages the high speed of the runs, a lower cost and a more compact instrumentation. However, this type of SNA sequencing, especially the 454 pyrosequencing technology, has lagged behind Illumina and other newer technologies, mainly because of the big difference in execution performance and other technical issues that we will see below [6].

2.2 Third Generation Sequencing

One of the main drawbacks of short-read sequencing techniques is related to one of the later stages in bioinformatic data analysis, namely the alignment or mapping of DNA fragments to a reference genome. As will be described below, most applications of high-throughput sequencing require these sequenced reads to be aligned to a reference genome, a process in which searching algorithms are used to map or know the specific position within a genome of the sequenced fragments. In this sense, there are two types of problems to be taken into account with this sequencing paradigm, either by the sequenced reads themselves or by the reference genome used. In the first place, the high complexity of the genomes makes them having regions that complicate to a great extent the correct mapping of fragments, such as highly repeated zones or structural variations; the first ones can measure several hundred base pairs, so our much shorter reads have the option of aligning in several different positions without reaching a unique alignment, something that also occurs with structural variations. This type of genomic variants, unlike single-nucleotide polymorphism, consist of fragments of a certain length that are duplicated or deleted - a deletion is the elimination of a DNA fragment - in different positions in the genome, often even distributed among several chromosomes. It has been shown that these types of genomic variants are involved in a large number of diseases [11], so their incorrect mapping and/or interpretation can have serious clinical implications. In addition to this issue regarding the length of sequenced fragments, another potential improvement that could be made in second generation methods is the use of PCR for clonal amplification of fragments. This process will be avoided in third generation technology, as it is a relatively error-sensitive technique in DNA zones with high GC-base content, and would result in considerable time savings in the preparation of libraries [12].

The third generation sequencing technologies or long-read based methods will be characterized, as its name suggests, in obtaining DNA fragments sequenced with a much longer length than the techniques described above, of the order of kilobases - 1 kb corresponds to 1000 base pairs -. Furthermore, as mentioned above, the fragments to be sequenced are not amplified by PCR, but the bases are detected from the single original molecule obtained in the preparation of the library; this also makes the sequencing process to be at real time, i.e. without washing and scanning cycles as was done in the second generation methods. The first approach to this type of technology was born in early 2011 with the release of the first PacBio RS sequencer by Pacific Biosciences, which uses the technology known as SMRT sequencing - single molecule real-time sequencing - a very similar method to the one used by the Illumina platform. Unlike the Illumina system, the fragments to be sequenced form a structure known as SMRTbell - the result of joining open adaptive sequences at both ends that make up a linear and circular molecule - which will be individually loaded into wells that will act as detectors. Within these structures, a DNA polymerase will be placed, which will be prepared to replicate the linear DNA sequence that has entered, making use of the corresponding fluorescently marked nucleotides, whose optical signals will be detected by a camera system in real time. A particularity of this platform is that the polymerase, once it has finished replicating the original chain, can start the process again with the resulting molecules, thus sequencing the same sequence several times; this will result in an unprecedented degree of precision, reaching a level of 99.999% with approximately 25 sequencing cycles. This great advantage is limited by the high cost involved during the first few years and the low performance obtained; although recent advances have improved them considerably these are still the main pitfalls to implement this technology in various mass sequencing projects [12].

The second major third-generation sequencing approach is the nanopore-based technology known as Oxford Nanopore Technology - ONT, name given by the company that developed it - which emerged in 2014 with the appearance of its first MinION sequencer. This novel sequencing method allows the detection of each nucleotide in the DNA chain as it passes through a nanopore, thanks to the voltage changes experienced by an electric current caused by the passage of these molecules. The DNA fragments to be sequenced carry with them adaptive sequences that will allow the union of motor proteins, whose function is to transport the DNA chain through the nanopore; this tiny opening is part of a large protein complex, and has inside an area more sensitive to the passage of nucleotides with a detector that measures changes in the current voltage, to different degrees depending on the nature of the nucleotide itself. This technology has features that make it the most promising at present, being perfectly valid in multiple sequencing projects. With it the longest readings to date have been achieved - reaching even 1 Mb [13] -, and there is no upper limit if the quality and quantity of the DNA of origin is good; the cost of sequencing is relatively low, making it possible for small laboratories to acquire and use it; and finally, its great scalability - it has from portable sequencers that fit in the palm of the hand like the MinION to high-throughput devices such as the PromethION - makes it a truly flexible technology to the needs required by the project. On the other hand, it still has a relatively high average error rate compared to other platforms, although lately attempts are being made to implement a system similar to PacBio in which both DNA strands are sequenced through the nanopore, thus reducing the error level to approximately 3% [12].

In addition to these two main technologies, in recent years new methods of long-read sequencing have emerged that is interesting to note. These new systems are known as SLR, or synthetic long-read sequencing, since the sequenced fragments are not really of this length, but short reads assembled in silico to generate a larger fragment. They are based on the Illumina sequencing platform of short fragments generation, but they present a series of changes regarding the preparation of the library. Illumina's own company acquired the Moleculo sequencing system, in which the initial DNA is fragmented into long molecules, up to 10 kb, and then introduced into micro-wells where they will be specially marked with adapters as a barcode system. Later they are fragmented again to be able to

be sequenced by Illumina and at the end, as each fragment is labelled according to its molecule of origin, they can be reconstructed synthetically to generate the long reads. The 10X Genomics technology, on the other hand, does not classify the original long fragments in micro-wells, but in a system of micelles in emulsion, similar to the system of sequencing by ligation with microspheres. As for the advantages of this type of platform, its main characteristic is that it is based on Illumina sequencing, taking advantage of its low level of error and its enormous performance per run; however, it requires the acquisition of new equipment to prepare the libraries, relatively increasing their cost, in addition to the fact that they depend on an amplification process by PCR, so that sometimes these technologies are not considered as really long-reads based methods [12].

3. NGS data analysis

The rapid advances in high-throughput sequencing following the completion of the Human Genome Project have allowed this technology to settle as a routine tool in multiple research laboratories and genetic centers, regardless of their area of work or their ability to address large or small projects. It has been discovered that in our genome are the biological answers to many of the questions that humanity continually asks itself regarding all types of medical problems, from the basis of any disease to the reason for our intelligence. However, the new paradigm has brought with it an enormous amount of data that classical computational approaches have not managed to handle, which has led to the emergence of multiple tools and algorithms to try to analyze and manage all these data from different sequencing platforms. This set of tools will be classified according to the stage in which it intervenes in the processing of data generated by next-generation sequencing methods, from the analysis of the reads from the sequencer to the obtaining of relevant biological information according to the project in question. The applications where the full potential of high performance sequencing can be exploited are very diverse: genomic DNA sequencing, which also includes the obtaining of new unknown genomes, the study of genetic variants at a population level or the clinical diagnosis of both Mendelian diseases and more complex syndromes such as cancer; RNA sequencing, also known as RNA-seq, where we may be able to analyze a complete transcriptome as a complement to the use of classical microarrays - RNA is a polymer of nucleotides, as is DNA, whose function is to serve as an intermediary between the genome and proteins, cellular components responsible for all the functions of an organism - or, more specifically, we can analyze by means of sequencing projects the interactions that occur between proteins and DNA [6].

In this paper we will focus on analyzing the existing pipeline for the analysis of mass sequencing data from genomic variants identification studies based on clinical applications, i.e. DNA variations related to various human diseases - when we compare the DNA of different individuals within the same species we see that it is exactly the same except in certain positions of the genome, whose variation is responsible for certain cellular proteins not working, generating a disease, or for the differences among individuals -. The stages that make up this type of analysis, and which will be described below, are the following: evaluation of the quality of the reads, alignment against a reference genome, identification of the variants and, finally, their annotation to give biological significance to the data. In addition to this main workflow, it is normally necessary to follow intermediate steps of filtering and various pre-processing of the data, following the standards of some of the main tools.

3.1. Quality assessment

The files resulting from any of the sequencing technologies described above contain all the reads detected by the sequencer in a standardized format known as FASTQ. This format presents an input for each sequenced read, in which each nucleotide brings with it an associated value of its quality generated by the sequencer itself. The reason for introducing a measurement of the quality of the bases is that, as we commented previously, each sequencing technology has a certain precision value when it comes to detecting true positives, so sequencing errors are inherent to all techniques, even more so taking into account both the human factor and failures related to the instrumentation or

chemical reagents used - it is common for example in certain platforms, such as Illumina, that as sequencing advances along the fragment the probability of error increases due to wear and tear on the molecular components used -. Therefore, our FASTQ file will have an associated value of the error probability for each nucleotide of the reads, thus being necessary to carry out this first step to ensure the quality of the fragments.

There are now numerous tools that allow us both to evaluate the overall quality of the reads and to perform a trimming of them based on certain parameters to filter, for example areas that do not reach a certain quality threshold. Tools have been developed that allow us to perform both processes together, such as NGSQC Toolkit [14], PRINSEQ [15] or the Galaxy environment [16], which produce general reports of the reads and are capable of filtering them. However, the most commonly used in sequencing data analysis pipelines today is the evaluation of reads using FASTQC [17] and their subsequent filtering with Trimmomatic [18], two totally independent tools but with a wide range of very interesting modules. FASTQC is a software that provides a large number of graphs and statistics showing the average quality of the reads, the average quality per base, the distribution of indeterminate nucleotides (Ns) or GC content, etc; on the other hand, Trimmomatic is a very powerful tool for filtering sequences based on multiple parameters, quite optimized for Illumina sequencing platform data, in addition to allowing the systematic elimination of adapters, sequences with no real biological value that come from the preparation of libraries in any of the sequencing platforms, whose elimination avoids the entry of a large signal of noise in subsequent stages [19]. Albeit it is not the most critical stage of the whole pipeline, it is necessary to clean the reads and facilitate the work of the tools that will follow, but there are currently no exhaustive reviews that compare the performance of different quality control software. Even so, different tools have been developed that might improve the performance of Trimmomatic, such as PathoQC [20], AfterQC [21], or the newest ones, fastp [22] and FastProNGS [23], which show several improvements in computational cost per run, a relevant aspect when analyzing several samples at the same time and hardware requirements start to increase considerably.

3.2. Aligment

Once the reads have been properly processed it is necessary to perform a mapping or alignment against an existing reference genome, for which there are two main sources, the University of Santa Cruz (UCSC) and the Genome Reference Consortium. Both institutions provide the scientific community with an assembly of the reference of human genome, on which they continuously apply improvements and various optimizations in order to know the specific genomic position of millions of reads from a sequencer. As for the process of aligning itself, it is necessary to highlight the enormous computational complexity involved in having to accurately place the reads in their correct position within the genome, something that is not as simple as it might seem. The human genome possesses enormous complexity, with regions so odd that have not yet been possible to characterize, such as repetitions of one or several nucleotides in the intergenic regions or duplications of the same gene in different chromosomes, so the process carried out by an aligning software is incredibly costly. To facilitate the work of this type of tools there are several concepts that are interesting to clarify. On the one hand, the longer the fragments, the easier it is to map them, just as it is easier to find in a book a unique coincidence of a specific phrase than of a single word, since in the latter case it is more likely that several options will be found where to place the word. On the other hand, a sequencing mode widely used today is the paired-end sequencing, in which each DNA fragment sequenced has a pair known and labelled beforehand, so we know exactly the extent that separates the two reads. Therefore, they are fragments that go hand in hand, so their mapping is more accurate because the genomic coordinates of one can help locate the other, if it falls in a compromised area.

The different alignment software will be classified according to the type of algorithm they implement for the mapping of the reads [24]. First of all there are hashing-based algorithms - the result of the hash function that generates keys to unequivocally represent a set of data - that elaborate

an index to quickly find the position of each read, but in exchange for mapping them very promptly they are very sensitive to errors; within this category we would find RMAP [24], SOAP [26], Novoalign [27] or SHRiMP [28]. In second place we find those based on the Smith-Waterman algorithm, which applies dynamic programming methods to ensure that local alignment is optimal with respect to a given scoring system, so they will be more precise and less sensitive to errors but more time-consuming; an example of this type of software is BFAST [29], whose peculiarity is that it exclusively implements this algorithm. Finally, the algorithms based on the Burrows-Wheeler transform optimize the use of memory, being currently the preferred for short reads to offer a balance between efficiency, sensitivity and specificity; currently there are many tools that implement this algorithm, such as BWA [30], Bowtie [31], or SOAP2 [32] and SOAP3 [33]. Numerous studies have evaluated the performance of various aligners for the identification of variants, although in the end the most commonly used today are those that offer the best performance - BWA, Bowtie, Novoalign and SOAP. Generally, most NGS data analysis projects use BWA or Bowtie2, the improved version of their predecessor, although several reviews seem to indicate that BWA delivers slightly better results [34] at a significantly faster speed [35]; however, it seems that this software is not as accurate even with low error rates - a characteristic of a good aligner is that they are able to map reads correctly, even when they contain errors from sequencing or genetic polymorphisms that diminish coincidences with respect to the reference genome - so it is a tool that does not let any potential alignment to escape, with the trade-off of generating many incorrectly mapped reads [36]. Novoalign, on the other hand, has also shown a very good performance with respect to the rest of the tools when GATK is subsequently used for the identification of variants [37], in addition to presenting a greater sensitivity or proportion of true positives when the reads are very short [35]. Finally, SOAP and its improved versions have a high accuracy even with high error rates in the mapping, so it seems the best choice for the identification of SNPs or single-nucleotide polymorphisms in later stages [36].

3.3. Post-alignment and pre-variant calling processing

The results obtained from the mapping algorithms contain the reads aligned against the reference genome in a quasi-standard format known as SAM, or Sequence Alignment/Map format, in which diverse information is presented about each aligned read, such as the specific position in the reference, its orientation - remember that the DNA molecule is bicatenary, so it is said to have a positive chain and a negative chain, also called forward and reverse, respectively - or the quality of such alignment. This information is stored in labels known as flags, whose resulting value is going to be the result of the sum of all the individual labels, each one of them representing a type of information that will serve us later to manage and filter them. Once these SAM files have been obtained, it is almost always necessary to carry out a series of pre-processes before identifying the variants themselves, either because they are essential requirements for subsequent tools or because they greatly facilitate their work. Most NGS data analysis projects will focus the attention of this pre-processing on three fundamental tools, such as SAMtools [38], GATK [39] - a suite of analysis tools that will later be fundamental for variant identification - and Picard [40]. Below we will describe the workflow followed in most variant identification studies from these tools, as they tend to be more standardized and revised pipelines within the bioinformatics community [41].

First, most variant calling algorithms require the mapped reads be ordered by genomic position and indexed, i.e. an index file is created to facilitate the search for information on the aligned reads. In addition, it is also common that the SAM file is transformed to its binary version BAM, which contains exactly the same information but in a compressed manner to make the management of the data easier; all this can be done through the SAMtools package, which even provides functions to offer summaries of the main alignment statistics, such as the percentage of correctly mapped reads or the proportion of correctly aligned pairs. These statistics will give us the possibility of eliminating certain biases from the alignment software itself, for example keeping all those readings mapped correctly or uniquely in the reference genome, all from SAMtools [41]. Subsequently, and due to the fact that the use of GATK for the identification of variants is very standardized, it is common to follow

the protocol or best practices pipeline designed by the creators of this software [42], in which several stages of preprocessing of the aligned reads before the identification of variants are detailed, using tools from GATK itself and from Picard. Therefore, using the files from SAMtools as input, the following processes will be carried out: creation of a reference sequence dictionary and preparation of the appropriate information from the reads, which will update the information in our files; marking or labelling of duplicate sequences using Picard as these are DNA fragments that have been sequenced several times during the sequencing process, giving rise to reads that do not provide any type of information and may falsify the coverage values of certain regions of the genome; local realignment around the indels - insertions and deletions -, since this type of structural variations causes the adjacent zones to be mapped incorrectly, a typical problem in the majority of aligners existing at present; and finally, a specific GATK process known as BQSR, or Base Quality Score Recalibration, is carried out, which as its own name indicates will determine the real value of error probability associated with each sequenced base, which sometimes are not completely precise. This will be essential because the variant identification algorithms will later use these quality values, together with another set of parameters, to obtain the degree of reliability of each identified variant, so it is logical that the process of detecting variants is required to be as accurate as possible and obtain a large proportion of true positives [41].

3.4. Variant calling

Thus far, the identification of variants and SNPs was typically done in microarrays, but their density limited the detection of genetic polymorphisms to a certain amount; however, the emergence of mass sequencing techniques has made possible a new approach of exhaustive variant identification, covering all possible points of a genome where there is a variation with respect to the reference, and also being able to obtain variants that we call rare - due to their low proportion in the population - whose role in complex diseases has recently been demonstrated [24] [43]. Therefore, and thanks to numerous tools developed in recent years, we may be able to obtain a complete map of the genomic variants of any individual in a much more precise and reliable way, although depending on complex algorithms to mitigate the enormous computational cost involved in the new paradigm of mass sequencing.

Genomic variants can be classified into several groups, depending on both their genetic nature and the type of algorithm needed to identify them. In the first place there is a first large group constituted by variants of small length, from a single nucleotide - what we know as SNP, single-nucleotide polymorphism, or SNV, single-nucleotide variation - to several pairs of bases - called indels, by the conjunction of insertions and deletions -. The single-nucleotide polymorphisms are the most common and therefore best known genomic variants, based simply on the substitution of one nucleotide base for another; the cellular machinery, as is known, translates this nucleotide sequence into a sequence of another type of molecule, the amino acids, constituting what we know as proteins. In this way, the change from one nucleotide to another will in turn cause a variation in the sequence of amino acids, which can have both negative effects - the protein is truncated and ceases to perform its function, giving way to a disease - and neutral - the change of amino acid does not affect the protein as a whole and can continue to perform its function -, or even positive - the new amino acid enhances the existing protein, either by adding a new function or optimizing the one it already had, which ultimately causes the new sequence to be maintained in evolution by the basic principle of natural selection -. On the other hand, so-called indels are small insertions or deletions of several nucleotides in a particular position, which will commonly cause a negative effect by altering the sequential reading of the DNA chain. Both types of variants in turn will be divided into two groups for technical reasons, since the algorithms that detect them will be different: on the one hand, germline variants are those that occur in the germ cells of an organism - ova and sperm - and are therefore those that are inherited from the offspring and will be present in all the cells of your body; on the other hand, somatic variants are those that arise as their name indicates in the somatic cells - the rest of cells of an organism - during the adult life of any living being, but will not pass on to the

offspring. These latter variants, however, are key to understanding the emergence and development of complex diseases such as cancer. In second place are the CNVs or copy-number variants, based on repeated fragments of relative size that are distributed along the genome, whose difference between individuals lies in the number of repetitions that each one presents. It has been shown that these types of variants represent up to 9.5% of our entire genome [44], and like the rest of variants can be the cause of certain diseases or have no visible effect on the body, simply representing genetic variation between individuals. Finally, structural variants or SV are based on genetic rearrangements of large areas of our genome, which may move from one chromosome to another or even be completely eliminated, clearly causing serious problems in the individual.

Different variant calling tools will typically be grouped according to their ability to detect a particular type of variant, although some have specific modules that allow the identification of different types of variants from the same sample. As for the first large group of SNPs and indels, because they appear more frequently and are better known than structural variants, we can say that numerous tools have been developed based fundamentally on two approaches: on the one hand, heuristic methods assign variants based on multiple sources of information related to data quality, such as VarScan2 [45], which also implements statistical methods such as Fisher's test to compare variants with theoretical distributions [24]; on the other hand, probabilistic methods are based on Bayesian approaches to optimize the probability of identified genotypes, where we find more tools currently and widely used as SAMtools or GATK. Speaking specifically of germline callers, whose detection is the most standardized of all, we find various software such as the aforementioned GATK, SAMtools or VarScan2, in addition to others such as SNVer [46] or FreeBayes [47]; of all of them, the GATK algorithm is usually the one that always offers the most reliable and accurate results [37], in addition to having modules to detect other variants and various functions for filtering and recalibrating the results, so it seems to be the best option in most studies. However, other reviews have highlighted the good role of FreeBayes in detecting a good number of truly high-quality variants, so it may be a good option in cases where greater precision is needed to the detriment of the number of variants obtained [34]. On the other hand, at the time of detecting somatic variants only acceptable results were seen in the previously mentioned tools, such as GATK, SAMtools and VarScan2; even so, an attempt was made to test the efficacy of another software, SomaticSniper [48], which offered acceptable results by identifying SNPs between tumor samples and controls. For the identification of CNVs some specific tools have also been developed, such as CNVnator [49], CONTRA [50], ExomeCNV [51], or RDXplorer [52], while for structural variants we have several softwares such as Breakpointer [53], CLEVER [54] or SVMerge [55]. In conclusion, according to numerous studies and revisions, it is highly recommended to approach the problem of identifying variants with a multiple approach, that is, to apply a set of algorithms on our data set to maximize the pool of potential variants and then carry out a series of filters to retain the highest possible proportion of true positives; these filtering processes can be carried out using specific tool modules such as GATK or SAMtools, or a more manual filtering can also be done in which we keep those variants that are present in a certain number of tools [19].

3.5. Variant annotation

The next-generation sequencing data analysis pipeline culminates with the process of variant annotation to bring some biological significance to the results obtained. Thanks to certain applications and tools it is possible to perform what is known as biological or functional annotation of variants, in which a large amount of information is searched for on these variants based on multiple parameters, such as the genomic region where it is found, the gene and the protein it affects, its effect according to the nature of the variant, etc. All this is possible thanks to all the information available in different databases and online resources, such as dbSNP [56] or the 1000 genomes project [57], which in turn will provide us with metrics to evaluate the possible clinical impact of the variant in question, something essential if we are talking about sequencing projects for clinical research, where it is necessary to know the potential relationship or causality between the disease of a patient and its

genomic variants. These metrics, such as Condel [58], PolyPhen [59] or SIFT [60], provide a prediction score based on the variant annotation that classifies it according to its potential clinical impact, from variants with great certainty of being pathogenic, to neutral or possibly benign variants, and even variants with unknown function or VUS - Variant of Uncertain Significance -. This classification is currently standardized and there are consensus guidelines for its evaluation and application in different NGS data analysis pipelines [61].

For this process we also have numerous tools, whose main difference with the rest of NGS analysis software is that many offer a graphical interface or a web platform that allow the functional annotation being more intuitive and not requiring so much computational knowledge; however, in most cases the sequencing projects offer such a large amount of data and variants that this type of platform cannot support it, so command line tools are going to be widely used when a high parallelization or computation is required in the process. There are many tools to carry out this step, such as ANNOVAR [62], NGS-SNP [63], snpEff [64] or VEP [65]; of all of them, the most revised and currently used are ANNOVAR and VEP - Variant Effect Predictor -, for its degree of comprehensiveness and the possibility of making annotation both in command line and through a graphical interface [19].

4. Next-generation sequencing data analysis frameworks

The analysis of next-generation sequencing data involves, as we have seen, numerous stages in which the output generally becomes the input of the next step, giving rise to what is known as a pipeline, a flow composed of a series of stages of analysis until finally arriving at the result we need, the biological and clinical information of the genomic variants detected. This is why the process of bioinformatic analysis of data from NGS is a task that requires a minimal IT expertise to know how to handle all the files generated, implement all the third-party software that have been mentioned previously for each stage and, in most cases, build a script that can be executed in command line to make the process more automated. Therefore, during the last years the bioinformatics community has been developing analytical pipelines to face this problem, generating tools in which the only task is to import the raw reads coming from the sequencer and let it work to finally obtain a set of identified variants with relevant biological information, allowing their use and application by researchers without any computational background. Often this type of software brings with it a graphical interface so that the user can modify parameters and the interpretation of the results is much more intuitive, avoiding what is known in computing as black box, a system in which only the inputs and outputs are studied without being able to know or take into account their internal functioning.

The need for the development of these pipelines and workflows also arises from the large number of challenges posed by the new paradigm of genomic data analysis. The numerous applications that are being discovered in this era of mass sequencing is causing a constant emergence of new tools, the evolution and optimization of existing platforms or the development of increasingly innovative algorithms to address new problems that are emerging. All of this results in an increase in the complexity of the analysis and an increasing difficulty in selecting the appropriate tools for each subprocess of the pipeline, since for each step new algorithms arise, each time more sophisticated and optimized, reaching the point that in 2017 there were already more than 11,000 tools for the analysis of omic data catalogued in the OMICtools platform [66]. This is aggravated by the realization that this high complexity is often left in the hands of researchers, who in addition to their own line of research must be able, with their scarce IT knowledge, to assemble these pipelines and choose the right tool at each step, thus making clear the urgency of standardizing analyses and increasing reproducibility in computational biology [67]. Finally, this complexity of use that we have mentioned makes it even more necessary to develop technologies that do not require a high technical level, without having to apply intricate command line instructions, so that the group of users who can apply this type of analysis is greatly expanded [68].

Some existing analytical pipeline bioinformatics analyses often offer a predefined order of steps and processes to be carried out, not allowing great flexibility to modify or replace certain modules; it is the case of pipelines such as HugeSeq [69], SIMPLEX [70], TREAT [71], bcbio-nextgen [72] or Sam2bam [73], which implement an automatic analysis of NGS data from the reception of reads to the identification of different types of variants, having the ability to receive different formats, be used on cloud platforms, carry out specific sections of the entire pipeline and offer researchers comprehensive results in the form of summary reports. However, they are usually not very flexible tools when inserting new modules or modifying certain stages to adapt it to the needs of the project in question, so they may lag behind especially for the bioinformatics community due to their great rigidity. To solve this, new platforms known as workflow management systems or pipeline frameworks emerge, tools that offer greater openness and flexibility to accommodate different pipelines, both in series and parallel, complex dependencies, varied software or parameters modified by the user, in addition to more advanced features such as the visualization of the process in real time, the possibility of working in the cloud and with graphical user interface or the ability to containerize various tools [74]. There is a large amount of workflow management systems at present, some of them more standardized and others newer and more innovative. Galaxy [75] is a web platform widely used in bioinformatics analysis with more than 100 tools available for the different stages of NGS analysis, with the possibility of creating custom pipelines, reproducing them and sharing them later with the community. Being a web platform, the graphical interface allows its use to be simple and intuitive even in the creation and customization of scripts, so it has become a benchmarking system for the rest of workflow frameworks for its wide use in the scientific community. SEQprocess [76] is a framework for carrying out NGS data analysis that already offers several pre-installed pipelines, as well as the possibility of generating them in a personalized way. It is an R package whose main characteristic is that it implements specific analyses for new oncological applications based on the TGCA, The Cancer Genome Atlas, although in the case of making your own pipeline requires a certain computational background to install the specific software and modify the parameters in configuration files. Closha [77], another recently developed workflow framework, is a system optimized for use in the cloud through high-performance computing clusters, also with a graphical interface and the possibility of running both existing pipeline and customized by the user. It presents certain technical advantages, such as the implementation of a new system known as KoDS for fast file transfer or the scalability of resources - it increases its performance as computational requirements increase -, which makes its execution speed slightly higher than in Galaxy. NGS-pipe [78] is another analysis framework that allows custom pipelines to be designed automatically and user-friendly, ensures reproducibility in clinical applications and allows parallelization in clusters; however, it also requires the installation of software manually and the modification of a config file to adjust the parameters needed. Finally, another more innovative framework in this sense is Bio-Docklet [79], a tool that allows managing pipelines from other systems such as Galaxy in Docker containers, encapsulating all the necessary pre-configured software and being a very interesting approach in the current scenario so that the researcher does not have to worry about manually installing all the required software.

As we have seen, there are currently multiple tools and platforms to face the arduous task of analyzing data from a mass sequencing project, each of them more sophisticated than the previous one; this makes it a field of computational research that evolves very quickly, so it lacks much of the standardization and reproducibility presented by other scientific fields, as may be the case of clinical and biomedical research, in which there are usually well-structured protocols and agreed guidelines to follow before a particular experiment. Therefore, in view of this situation our recommendation would be to carry out an exhaustive study of the applications in which next-generation sequencing is present, evaluating each specific case and optimizing the different parameters required. Therefore, it can be said that the elaboration of guidelines and pipelines for each of the applications would be a great step towards improving transparency and reproducibility between different sequencing

projects. Secondly, in relation to the different existing framework systems, it has been seen that they have numerous tools to become common instruments in any biomedical laboratory, such as their ease of use, the possibility of running parallel work in the cloud or the implementation of graphical user interfaces; however, these interfaces are only contemplated during the analysis process and the creation of the personalized pipeline, as it has been seen that in most cases the final information is not offered in such a graphical and intuitive way. In cases such as clinical research where the conclusions of an experiment or even the health of a patient depend on the final information obtained, the interpretation of the results is essential, so it would be necessary to focus efforts on developing a robust and optimized pipeline for clinicians in which one of the basic pillars is the biological annotation of the variants and an optimal interpretation of them, generating graphical and intuitive reports where the most clinically relevant information of a patient appears. Finally, a very interesting approach to apply in these NGS data analysis pipelines would be the implementation of machine learning algorithms, something quite scarce in the current standard frameworks. Artificial intelligence is making its way in recent years in the field of genomics, and due to the enormous amount of data generated by these platforms, which currently continues to grow at an exponential rate, it would be interesting to generate models and training algorithms that further improve the results obtained from classic pipelines.

Author Contributions: Conceptualization, A.C. and M.S.; methodology, A.C.; writing—original draft preparation, V.M.D.; writing—review and editing, A.C., J.M.C and M.S.; supervision, M.S.; project administration, J.M.C.; funding acquisition, J.M.C.

Funding: This research received no external funding

Acknowledgments: This work was carried out under the frame of the "Towards Sustainable Intelligent Mobility: Blockchain-based framework for IoT Security" Ref. RTI2018-095390-B-C32" project. The project was supported and funded by the Spanish Ministerio de Economía, Industria y Competitividad. Retos de investigación, Proyectos I+D+i.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Watson, J. D. & Crick, F. H. The structure of DNA. Cold Spring Harb. Symp. Quant. Biol. 18, 123–131 (1953).
2. F. Sanger, et al. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A., 74 (1977), pp. 5463-5467.
3. A.M. Maxam, W. Gilbert. A new method for sequencing DNA. Proc. Natl. Acad. Sci. U.S.A., 74 (1977), pp. 560-564.
4. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature, 431 (2004), pp. 931-945.
5. J.A. Schloss. How to get genomes at one ten-thousandth the cost. Nat. Biotechnol., 26 (2008), pp. 1113-1115.
6. Erwin L. van Dijk, et al. Ten years for next-generation sequencing. Trends in Genetics, Volume 30, Issue 9 (2014), pp. 418-426.
7. Steven R. Head, et al. Library construction for next-generation sequencing: Overviews and challenges. Biotechniques. 2014; 56(2): 61–passim.
8. Sara Goodwin, et al. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 17, 333-351 (2016).
9. Lin Liu, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012.
10. Michael L. Metzker. Sequencing technologies – the next generation. Nature Reviews Genetics 11, 31-46 (2010).
11. J. Weischenfeldt, et al. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet., 14 (2013), pp. 125-138.
12. Erwin L. van Dijk, et al. The third revolution in sequencing technology. Volume 34, Issue 9, September 2018, pp. 666-681.

13. M. Jain, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol., 36 (2018), pp. 338-345.
14. Dai M, Thompson RC, Maher C, et al. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics. 2010;11(Suppl 4):S7.
15. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27:863–4.
16. Blankenberg D, Gordon A, Von Kuster G, et al. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010;26:1783–5.
17. Babraham Bioinformatics, FastQC:A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
18. Anthony M. Bolger, et al. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1; 30(15): 2114–2120.
19. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2014 Mar; 15(2): 256-278.
20. Hong C, et al. PathoQC: Computationally Efficient Read Preprocessing and Quality Control for High-Throughput Sequencing Data Sets. Cancer Inform. 2015 May 12;13(Suppl 1):167-76.
21. Shifu Chen, et al. AfterQC: automatic filtering, trimming, error removing and quality control for fatsq data. BMC Bioinformatics 18, Article number: 80 (2017).
22. Shifu Chen, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, Volume 34, Issue 17, September 2018, pp i884–i890.
23. Xiaoshuang Liu, et al. FastProNGS: fast processing of next-generation sequencing reads. BMC Bioinformatics 2019; 20: 345.
24. M. Mielczarek, J. Szyda. Review of alignment and SNP calling algorithms for next-generation sequencing data. Journal of Applied Genetics, Volume 57, Issue 1 (2016), pp 71-79.
25. Smith AD, et al. Updates to the RMAP short-read mapping software. Bioinformatics 25 (2009):2841–2842.
26. Li R, Li Y, Kristiansen K, Wang J (2008b) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714.
27. Novocraft (2010), http://www.novocraft.com/.
28. Rumble S.M., et al. Shrimp: accurate mapping of short color-space reads, PLoS Comput. Biol., 2009, vol. 5 pg. e1000386.
29. Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS One 4(11):e7767.
30. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15; 25(14):1754-60.
31. 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(3):R25.
32. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009 Aug 1; 25(15):1966-7.
33. Liu CM, et al. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics. 2012 Mar 15;28(6):878-9.
34. Sohyun Hwang, et al. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015; 5: 17875.
35. Subazini Thankaswamy-Kosalai, et al. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. Genomics, Volume 109, Issues 3–4, July 2017, pp 186-191.
36. Matthew ruffalo, et al. Comparative analysis of algorithms for next-generation read alignment. Bioinformatics, Volume 27, Issue 20 (2011), pp. 2790-2796.
37. Adam Cornish and Chittibabu Guda. A comparison of variant calling pipelines using Genome in A Bottle as a reference. Biomed Res Int. 2015; 2015: 456479.
38. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics 25 (2009):2078–2079.
39. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20 (2010):1297–1303.
40. http://picard.sourceforge.net/

41. André Altman, et al. A beginners guide to SNP calling from high-throughput DNA-sequencing data. Human Genetics, Volume 131, Issue 10 (2012), pp. 1541-1554.
42. Geraldine A. Van der Auwera, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013 Oct 15; 11(1110): 11.10.1–11.10.33.
43. Handel AE, Disanto G, Ramagopalan SV. Next-generation sequencing in understanding complex neurological disease. Expert Rev Neurother 13(2):215–227 (2013).
44. Zrrei M, et al. A copy number variation map of human genome. Nat Rev Genet. 2015 Mar;16(3):172-83.
45. 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 22(3):568–576 (2012).
46. Wei Z, Wang W, Hu P, et al. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res. 2011;39:e132.
47. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN] 2012.
48. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012;28:311–7.
49. Abyzov A, Urban AE, Snyder M, et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21:974–84.
50. Li J, Lupat R, Amarasinghe KC, et al. CONTRA: copy number analysis for targeted resequencing. Bioinformatics. 2012;28:1307–13.
51. Sathirapongsasuti JF, Lee H, Horst BAJ, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011;27:2648–54.
52. Yoon S, Xuan Z, Makarov V, et al. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009;19:1586–92.
53. Sun R, Love MI, Zemojtel T, et al. Breakpointer: using local mapping artifacts to support sequence breakpoint discovery from single-end reads. Bioinformatics. 2012;28:1024–5.
54. Marschall T, Costa I, Canzar S, et al. CLEVER: clique-enumerating variant finder. Bioinformatics. 2012;28(22):2875–288.
55. Wong K, Keane TM, Stalker J, et al. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol. 2010;11:R128.
56. https://www.ncbi.nlm.nih.gov/snp/
57. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature, volume 526, pages 68–74 (01 October 2015).
58. González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011;88:440–9.
59. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.
60. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–81.
61. Sue Richards, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine, volume 17, pages 405–423 (2015).
62. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
63. Grant JR, Arantes AS, Liao X, et al. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. Bioinformatics. 2011;27:2300–1.
64. Cingolani P, Patel VM, Coon M, et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front Genet. 2012;3:35.
65. McLaren W, Pritchard B, Rios D, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. Bioinformatics. 2010;26:2069–70.
66. https://omictools.com/

67. Geir Kjetil Sandve, et al. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013 Oct;9(10).
68. Jeremy Davis-Turak, et al. Genomics pipelines and data integration: challenges and opportunities in the research setting. Expert Rev Mol Diagn. 2017 Mar; 17(3): 225–237.
69. Lam HYK, Pan C, Clark MJ, et al. Detecting and annotating genetic variations using the HugeSeq pipeline. Nat Biotechnol. 2012;30:226–9.
70. Fischer M, Snajder R, Pabinger S, et al. SIMPLEX: cloud-enabled pipeline for the comprehensive analysis of exome sequencing data. PLoS ONE. 2012;7:e41948.
71. Asmann YW, Middha S, Hossain A, et al. TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data. Bioinformatics. 2012;28:277–8.
72. https://github.com/bcbio/bcbio-nextgen
73. Ogasawara T, Cheng Y, Tzeng TK. Sam2bam: High-Performance Framework for NGS Data Preprocessing Tools. PLoS One. 2016 Nov 18;11(11).
74. Leipzig J. A review of bioinformatic pipeline frameworks. Brief Bioinform. 2017 May 1;18(3):530-536.
75. Goecks J, Nekrutenko A, Taylor J; Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86.
76. Joo T, Choi JH, Lee JH, Park SE, Jeon Y, Jung SH, Woo HG. SEQprocess: a modularized and customizable pipeline framework for NGS processing in R package. BMC Bioinformatics. 2019 Feb 20;20(1):90.
77. Ko G, Kim PG, Yoon J, Han G, Park SJ, Song W, Lee B. Closha: bioinformatics workflow system for the analysis of massive sequencing data. BMC Bioinformatics. 2018 Feb 19;19(Suppl 1):43.
78. Singer J, Ruscheweyh HJ, Hofmann AL, Thurnherr T, Singer F, Toussaint NC, Ng CKY, Piscuoglio S, Beisel C, Christofori G, Dummer R, Hall MN, Krek W, Levesque MP, Manz MG, Moch H, Papassotiropoulos A, Stekhoven DJ, Wild P, Wüst T, Rinn B, Beerenwinkel N. NGS-pipe: a flexible, easily extendable and highly configurable framework for NGS analysis.
79. Kim B, Ali T, Lijeron C, Afgan E, Krampis K. Bio-Docklets: virtualization containers for single-step execution of NGS pipelines. Gigascience. 2017 Aug 1;6(8):1-7.