Personalized Medicine. Medical opportunities and challenges in the massive sequencing era

Authors

Rosario Perona[1], Elena G. Árias-Salgado[2] and Leandro Sastre[1]

Affiliations

[1] Instituto de Investigaciones Biomédicas CSIC/UAM. Arturo Duperier, 4. Madrid. Spain. Centro de Investigación en Red de Enfermedades Raras (CIBERER), Spain. Biomarkers and Experimental Therapeutics in Cancer, IdiPaz, Madrid, Spain.

[2] Advanced Medical Projects, Madrid, Spain.

Corresponding author:

Leandro Sastre
Instituto de Investigaciones Biomedicas CSIC/UAM
Arturo Duperier, 4
28029 – Madrid. Spain
Email: lsastre@iib.uam.es

Abstract

DNA sequencing capacity has increased tremendously in the last ten years due to the development of massive sequencing techniques, also known as Next Generation Sequencing. A large amount of information on the sequence of the human genome, its organization, transcription and regulation of gene expression has been generated from healthy individuals and also from patients suffering a broad group of diseases. The technical capacity to determine the sequence of patients' DNA at decreasing cost and the knowledge already generated is making possible a precise molecular diagnosis for many diseases in clinical settings. Determining the molecular basis of the disease for each particular patient has important implications in diagnosis, prognosis, genetic counselling and for the determination of an optimal treatment, moving towards a personalized medicine. In this review the available sequencing platforms will be briefly analyzed. DNA sequencing for clinical practice can be extended to different levels and the advantages and disadvantages of Whole Genome Sequencing (WGS), Whole Exome Sequencing (WES), RNA sequencing, Targeted Panel Sequencing and Mehylated DNA Sequencing will be discussed. The application of these technologies to Monogenic, Multigenic diseases and Cancer will be reviewed. Finally, the clinical implementation of massive parallel sequencing technologies and the present challenges will be discussed.

Introduction

The development of DNA Massive Parallel Sequencing techniques, also known as Next Generation Sequencing (NGS), has represented an enormous breakthrough in the knowledge of the human genome. Previous sequencing methods, often referred as Sanger's sequencing (1), determined the DNA sequence of a DNA molecule at a time and required previous bacterial DNA cloning. In contrast, massive parallel sequencing simultaneously determines the sequence of millions of DNA molecules and no cloning step is required. Therefore, these new techniques can determine the nucleotide sequence of large regions of DNA in a short time and at much reduced cost. As a prominent example, sequencing the first humane genome, reported in 2001 (2), was a collaborative project that involved 20 Institutes working during 13 years with a cost of nearly 3000 millions US dollars. With the new techniques sequencing a human genome takes a few days with a price that is approaching the 1000 US dollars boundary.

The application of NGS techniques has generated of a huge amount of information of the humane genome in the last few years through Whole Genome Sequencing (WGS) studies. For example, the genomes of 2504 people from 26 different populations had been reported by 2015 as result of the 1000 Genomes Project (3, 4) and the UK10K Consortium (5). The effort will continue in the next years through new projects such as the 100,000 genomes project aimed to sequence 100,000 whole genomes from patients and families of the England National Health Service (available from https://www.genomic-sengalnd.co.uk/the-100000-genomes-project/). There is also a similar GenomeAsia 100K (GA100K) project aimed to sequence and analyze 100,000 Asian individuals' genomes (available from http://www.genomasia100k.com) (6). Whole genome sequencing studies are complemented by sequencing projects involving specific regions of the genome. Particularly important is Whole Exome Sequencing (WES) where the nucleotide sequence of the exons contained in protein-coding genes is determined. In this respect, the NHBLBI project described the nucleotide sequence of 6,515 exomes (7) and the Exome Aggregation Consortium (ExAC) the corresponding to 60,706 humans (8). These studies are basic to describe the heterogeneity and variability of the human genome (4). The number of genomic sequence variants (Single Nucleotide-Variants, SNVs) present

between any two individuals has been estimated in 4-5x10^6 (9) and 20,000-25,000 are present in the exons (10). Among them, it has been estimated that a typical human genome contains 100 loss-of-function variants that completely inactivate around 20 genes (11). In addition, over 500,000 variants in predicted promoter, insulator and enhancer regions could alter gene regulation (4). It is, therefore, important to determine the nucleotide variants present in the general population and their frequencies as a reference to compare the variants found in a particular individual and their possible relevance. The human reference genomes are maintained updated by the Genome Reference Consortium (GRC) (available from: http://www.ncbi.nlm.nhi.gov/projects/genome/assembly/grc/humna/).

Massive sequencing projects have also provided a large amount of valuable information on genome structure and expression. Initial efforts were centred to the identification of protein-coding genes and their intron/exon structure (2). Subsequently, the ENCODE (Encyclopedia of DNA Elements) project was aimed to identify all the functional regions of the genome (12) (see (13) for a short summary). 32 research groups and over 440 scientists were involved in this project for 5 years. Genomic regions that were transcribed in 147 cell types were identified and shown to represent about 60% of the genome including protein-coding regions and many other regions coding for non-coding RNAs possibly involved in the regulation of gene expression, including small-interfering RNAs (siRNAs) and long non-coding RNAs (lncRNAs). Transcription initiation sites were also determined to a genome-wide scale as well as possible transcription-regulatory regions. The study analyzed chromatin structure including the distribution of 13 chromatin-associated histone variations, DNaseI hypersensitivity and DNA methylation. The binding site of 14 RNA polymerase and basal-transcription factors and 87 sequence-specific transcription factors was determined. Therefore, this study described the landscape of transcription regulation at a genome-wide perspective. Gene expression programs are specific for each cell type but also depend on the physiological situation and can be greatly altered by pathological stages. The functional annotation of the mammalian genome 5 (FANTOM5) project was aimed to provide comprehensive RNA expression profiles of mammalian cells types. The activity of transcriptional regulatory regions was analyzed in 573

human primary cell types, 250 cancer cell lines and 152 tissues, both at the promoter (14) and at the enhancers level (15). These groundbreaking studies set the reference for the comparative analyses of gene expression profiles in pathological situations. The data obtained by NGS are also useful to determine the length of the telomeres (16), the terminal region of the chromosomes, that get shorter with age what is related to premature ageing diseases, known as telomeropathies (17).

The development of NGS techniques open new perspectives to the molecular analyses of the different patients to determine the genetic basis of their disease that is at the basis of personalized medicine, defined by the US National Institute of Health as an "approach to treatment based on individual differences in a patients genome" (https://www.nih.gov/precission-medicine-initiative-cohort-program/).

Genetic analyses of the nucleotide sequence of the patient's genome, at different levels of complexity, as well as changes in gene expression regulation can be performed. The results can be compared to the enormous amount of information available for the general population, as described above. Possible pathological genetic alteration can be identified that can be the basis for the diagnosis, prognosis and treatment of the patient. The different NGS techniques available, their application to the study of genetic diseases and several recent examples of clinical utility will be described in this review. Several recent reviews have also discussed different clinical applications of next generation sequencing (6, 18-21)-

Next Generation Sequencing Platforms

Several NGS platforms are commercially available and each of them has specific characteristics. The first platforms that were developed are based on the incorporation of fluorescent nucleotides in successive rounds of DNA polymerization. The reaction is carried out on solid surfaces were millions of DNA molecules are loaded and the platforms simultaneously determine the incorporation of the nucleotides in each of the spots (massive parallel sequencing). These platforms require the previous "in situ" amplification of the substrate DNA molecules by PCR to generate sequencing libraries. Among these platforms, the most frequently used are commercialized by Roche (Roche Diagnostics Co., Branford, CT, USA) and Illumina (San Diego, CA. USA). The most recent Roche sequencer, GS FLX+, based on the previous 454 model,

generates reads of 700 base pairs (bp) and generates about 1,000,000 reads per run (available from http://454.com/products/gs-flx-system/).

Illumina commercialize several platforms such as HiSeq2500, HiSeq4000 and HiSecX (available from http://www.illumina.com/systems/). The last available system consists of 10 HiSeqX and generates 6,000 millions of reads of 2x150 bp per run in 3 days (18 Tb of sequence per run). A third available platform detects a hydrogen ion that is released at each polymerization step and is detected by semiconductor technology (Ion Torrent technology). It is commercialized by Life Technologies (Thermo Fisher, Waltham, MA, USA) and the present models include the Personal Genome Machine (PGM) and Proton (available from https://www.thermofisher.com/us/en/home/life-science/sequencing/). This sequencer generates 60-80 millions of reads up to 200 bp (10Gb of sequence) per run. The strategy of these platforms highlights the options of generating fewer runs of higher length of a larger number of shorter runs. The election of one or the other option depends on the specific aims of the sequencing project.

A new generation of platforms, named Third Generation sequencing technologies are being developed that do not require substrate DNA amplification and are also named single molecule sequencers. The main advantage is that any possible bias introduced by amplification of the substrate DNA by PCR is avoided. In addition, these platforms allow much longer reads than the previous ones. The only platform presently available of this generation is the PacBio RS II from Pacific Biosciences technology (Menlo Park, CA, USA) commercialized by Roche. This sequencer generates reads of up to 20.000 bp. Other single-molecule sequencers are being developed based on nanopore (22) and optical technologies but presently are not commercially available.

DNA sequencing extension options

Sequencing of a human genome is still a compelling project in a clinical laboratory. The amount of sequence generated is very large and the complex analysis requires expertise and time. Because of this reason, many clinical projects are centred in sequencing specific regions of the genome that can provide enough information to solve the clinical problem. The decision on the sequencing approach to be used depends, then, on the region to be tested, the type of alterations expected and turnaround

time. In general terms, sequencing longer regions provides more information but requires more complex analyses and takes more time to obtain the results, as schematically shown in Figure 1. On the contrary, sequencing smaller regions is easier and faster but the information obtained is more limited. Another general consideration is that sequencing shorter DNA regions provides more sensitivity because of increasing sequencing depth, the average number of times that each residue is sequenced in the run. The advantages and limitations of the different sequencing approaches will be discussed in next paragraphs, are summarized in Table I and have been discussed in recent reviews (6, 19, 20, 23).

Whole Genome Sequencing (WGS)

Sequencing the whole genome provides complete information on the possible genetic alterations present in an individual genome. Variations in the nucleotide sequence can be found in protein-coding exons but also in introns and regulatory regions. Structural alterations of the genome such as small and large deletions, insertions and translocations can be precisely identified. In addition, alteration in the copy number of specific regions (Copy number variations, CNVs) can be determined. The main disadvantage is that the amount of sequencing required is larger that with other options which makes their analyses more difficult and time consuming. This is exemplified by the number of SNVs that can be found in a typical genome, about $4x10^{6,}$ that have to be analyzed to find a few potential pathological variants. Besides, sequencing depth is usually 30-60x (24) and is limited by the amount of reads that can be obtained which makes this methods less sensitive and accurate than others. Lower sequencing depth can be limiting when analyzing heterogeneous samples such as those obtained for cancer studies. The development of improved sequencing platforms that generate a larger number of sequences provide increased sequencing depth which, together with more efficient programs for sequence analysis and decreased prices is making of WGS an option more affordable for clinical application.

Whole Exome Sequencing (WES)

Mutations in protein-coding regions (exons) are expected to be most frequently involved in the development of pathologies. Therefore, many project are centred in sequencing the exons of the protein-coding genes or exome. Actually, it has been estimated than WES can detect up to 85% of the pathological

mutations (20). In this approach, exons are isolated from the sample DNA by hybridization to a set of oligonucleotides that represent all the exome after DNA fragmentation. Hybridized DNA is then isolated and sequenced (19). Depending on the protocols used for exome isolation, the exome contains between 40 and 50 millions of bp, between 1 and 2 % of the 3,300 millions of the whole genome (13). Therefore, WES requires the generation of many less sequences that WGS and the sequencing depth usually obtained is larger, 100-150x. The complexity of data analysis and the time required to complete the sequencing process is also considerably lower. The analysis of WES sequencing provides information on the existence of nucleotide variation and small deletions and insertions (indels) in protein-coding regions that might result in missense, frameshift and non-sense mutations (18). Exome sequencing also provide information on the sequence of proximal intronic regions whose mutation can alter splicing sites resulting in alteration of mRNA processing and the generation of mutated proteins. Many exome capture kits also target regulatory regions such as those coding for miRNAs. In addition, probes specific for selected non-coding regions can be included. Apart for these specific regions, variations in intronic or other regulatory regions cannot be detected by WES. Insertion/deletions larger than 20 bp are also difficult to detect because WES is usually based in short sequences (less than 150 bp) that are difficult to align to the reference genome if a significant insertion/deletion is present. One technical limitation of WES is the possible variability of the exome capture process so that a fraction of the exons might not be represented in the exon library. However, with present methods more than 95% of exons are sequenced with high coverage.

Targeted panel sequencing

Panel sequencing project are centred in the analysis of a reduced number of genes, usually between 20 and 500 genes. This approach can be used in diseases where most patients present mutations in a limited number of genes. In this case, the exons of the selected genes are isolated and used to construct the sequencing libraries. Isolation can be done by hybridization to specific oligonucleotide probes, as explained in WES. Alternatively, exons can be amplified in multiplexed PCR reactions (21, 25). Because of the limited amount of regions to be sequenced, panel sequencing provides a large sequencing depth (in the order of 500x) that is a great advantage for heterogeneous samples.

For example in a tumor sample where only a fraction of the cells are cancerous, tumor-specific mutations will be present only in a fraction of the DNA molecules. Therefore, a high sequencing depth will be necessary to confidently detect these mutations over a background of wild-type non-tumoral sequences. Another clinical advantage is that panel sequencing can be performed in DNA obtained from formalin-fixed paraffin-embedded tissue (26). The information provided by Panel sequencing is similar to that previous described of WES except that non-coding regulatory regions can be included in the panel if recurrent mutations are found in these regions, including gene fusions (27). The main limitation of this approach is that the design of the panel is based on previous knowledge of the genetic of the disease and mutations in genes not represented in the panel are not analyzed. However, the smaller amount of data generated makes the sequencing project and the analyses of the data faster than in WES and WGS. Because of these reasons, targeted panel sequencing is the NGS protocol that can be implemented most easily in a clinical setting (Fig 1) (18, 21).

Transcriptomics. RNA sequencing

The nucleotide sequence of the RNA molecules expressed in a cell or tissue can be determined previous generation of DNA copies of the RNA by reverse transcription. This technique is commonly named RNA sequencing and the sequences generated are known as the transcriptome of the cell type or tissue (28). The human genome is predicted to contain about 20,100 protein-coding genes, 12,900 regions coding for long-non coding RNAs (lncRNAs) over 500 regions coding for Micro RNAs (miRNAs), apart from the genes coding for small nuclear, nucleolar, antisense, ribosomal and transfer RNAs (29). RNA sequencing provides information about all of these transcribed regions (20) although mRNAs, coding for proteins, are expressed at higher levels than lncRNAs, siRNAs and other non-ribosomal RNAs. It is important to have into account that all these genes are not transcribed in every cell. For example, the average number of protein-coding genes transcribed in a specific cell has been estimated between 1,000 and 10,000, depending on the tissue (30). RNA sequencing provides information exclusively about the RNA molecules being expressed, which are different for each cell type or tissue. This is a significant difference with respect to WES where all protein-coding genes are analyzed (Table I). However, RNA sequencing provides additional

information with respect to other DNA sequencing techniques. The first one is that RNA sequencing data can be analyzed quantitatively to determine the expression levels of every gene. This parameter is determined by quantifying the number of sequences generates for each mRNA that is proportional to the number of mRNA molecules present in the RNA preparation. This number is normalized to the length of the mRNA and the total number of reads obtained in the experiment and is usually expressed as the relative number of reads by kb of each mRNA. The clinical relevance of these data is that RNA expression is frequently affected by pathological situations (29, 31) and the expression level of specific genes can be used as molecular biomarker for diagnosis, prognosis, treatment and to study the evolution of the disease (32-34).

The second specific property is that RNA sequencing provides information on the processed RNA molecules. It is well established that most transcripts can be processed in different ways by alternative splicing (35). Also, many genes are transcribed from different initiation sites. All these alternative transcripts can be detected and quantified by RNA sequencing but not by DNA sequencing techniques that are centred on the coding DNA (36). Similarly, many DNA translocations result in the generation of fusion RNAs, coding for fusion proteins, that can be detected by RNA sequencing (21). Most of these translocations occurs in intronic regions and cannot be detected by techniques that sequence coding exons such as WES. Chromosome translocations have an important pathological role in many diseases, such as some cancer types, and their detection can be, therefore, of clinical relevance.

Another difference is that RNA sequencing techniques require the isolation of high-quality RNA from the samples. Fresh of quickly frozen clinical samples are required. The RNA isolated from formalin-fixed paraffin-embedded tissue is usually of low quality for RNA sequencing which can represent a limitation for the clinical use of this technique (20).

The analysis of siRNA (also named miRNAs) requires some specific techniques for RNA isolation and for the construction of the sequencing libraries. The reason is that miRNAs have a size of 20-22 bp and are not highly represented in the total RNA population. siRNAs regulate mRNA stability and translation by binding to homologous regions of specific mRNAs. Each siRNA can regulate a number of mRNAs so that their expression plays an important role in the regulation of gene expression. Because of

this reasons, the expression levels of specific siRNAs is also used as molecular biomarker in clinical studies.

RNA sequencing data can be also analyzed to detect variation in the nucleotide sequence in comparison to the reference human genome. This information would be additional to the data on alternative splicing, the use of different transcription initiation sites and the presence of gene fusions that are specific to this NGS technology.

Global analysis of DNA methylation

Cytidine methylation at CpG dinucleotides is an important mechanism that mediates gene expression regulation. This modification takes mainly place in CpG-rich regions known as CpG islands. About 50% of these regions are associated to protein-coding genes and regulate their expression so that DNA methylation generally results in transcription inhibition. Alterations in DNA methylation, resulting in changes in gene expression, can be at the bases of some pathologies (37). A classical example is the methylation of the regulatory region of tumor-suppressor genes that inhibits the expression of these genes contributing to the development of cancer (38). DNA methylation cannot be detected by general DNA sequencing techniques since methyl-cytidine is detected as cytidine. However, treatment of DNA with bisulfite allows the differential detection of methyl-cytidine and the identification of the specific residues of cytidine that are methylated. This technique has been applied to WGS protocols so that all methylated residues can be identified in the whole genome (39). DNA methylation pattern is specific for each cell type and physiological condition and is known as the methylome of that cell type. The NIH Roadmap Epigenomic Program is presently generating the epigenomic landscapes of primary human tissues and cells (39). As mentioned above, the pattern of DNA methylation can be altered in pathological stages so that the methylation or de-methylation of specific DNA regions is being used as molecular markers of clinical relevance. Genome-wide DNA methylation studies are presently one of the more specialized NGS techniques and, therefore, more difficult to implement in the clinic but it provides additional regulatory information that cannot be obtained by other NGS approaches (Table I).

Clinical applications of NGS technology

The technical capacity offered by NGS for high throughput generation of nucleotide sequence data at decreasing

prices open tremendous opportunities for clinical diagnosis, prognosis and treatment (see (18-20, 40) for recent reviews). There are, however, different techniques with their own advantages and challenges (Table I) that have to be chosen for each specific clinical situation. In this section we will discuss the application of NGS technology to the clinic centred in three different types of diseases, monogenic rare hereditary diseases, multigenic common diseases and cancer, as summarized in Table II.

Monogenic diseases.

Monogenic diseases that follow a Mendelian inheritance pattern are the more obvious candidates for the use of NGS technology. These diseases are caused by single mutations and have been genetically studied by more than 40 years. The classical strategy required to first narrow down the mutated region from microsatellite- based linkage studies that required samples from large pedigrees. This analysis would focus the interest in one or a few candidate genes that had to be sequenced by the Sanger's method in the affected patient and control healthy donors. The study of mutations in the identified genes required the amplification and sequencing of the gene exons (or other mutated regions). In addition, many diseases can be caused by mutations in several genes and all of them have to be analyzed in every patient. NGS offers now the possibility to analyze the nucleotide sequence of all these genes in a single sequencing reaction. Sequencing of a panel of genes previously shown to be involved in the specific disease would be the most direct and simple approach.

In most hereditary diseases, not all the causative genes have been identified at the present time and a proportion of the patients do not carry mutations in the known genes. Therefore, the analysis of the above mentioned panel of genes would not allow the genetic diagnosis of these patients. These patients would require a broader genetic analysis (for example (41)). WES would allow to determine the nucleotide sequence of all the protein-coding regions of the genome that are the cause of most monogenic diseases (42). Mutations could be also present in gene regulatory regions and their identification would require WGS. As mentioned above, each of these magnifications in the sequence scope imply increased complexity in the analysis of the data generated.

Independently of the sequencing extension, the next step is to identify the genetic variations present in the patient. The general steps followed in this analysis are schematically shown in Fig 2

for monogenic diseases but can be also applied to Multigenic diseases and Cancer. The sequence data generated are analyzed for quality requirements in first place. Clinical projects are re-sequencing studies and the next step is to align each sequence to a reference human sequence such as the GRCh38.p8 from the Genome Reference Consortium, released on June 30, 2016 (available from: http://www.ncbi.nlm.nhi.gov/projects/genome/assembly/grc/humna/). This comparison can detect single nucleotide variations (SNVs), small insertions or deletions (Indels), genomic reorganizations and variations in the number of copies of some genes (CNVs, copy number variations). A limitation is that most NGS platforms generate short nucleotide reads that make difficult the alignment of the sequences to the reference genome when there are Indels longer than 20-30 nucleotides, specially in Gene panels and WES. The detection of CNVs is also difficult in NGS experiments because of the intrinsic variation in the number of reads obtained for each DNA region.

Probably the most challenging step in the analysis of NGS data is the identification of pathogenic mutations. The reason is the large amount of sequence variants found between any two individuals, as previously discussed (11, 43). This large number of variants has to be filtered to select the ones that might have clinical significance (reviewed in (44)). The first criterion is the functional significance of the nucleotide variant so that missense, nonsense coding variants, start-loss, stop-loss, frameshift insertions and deletions and splice-site variants are selected (Fig 2). A second criterion is the abundance of the nucleotide variant in the general population. Monogenic diseases are rare, with a frequency of less than 1 case each 100.000 inhabitants. Causative mutations are expected to be similarly infrequent in the general population. To determine the frequency, the presence of each specific variant is searched in general databases such as the Exome Aggregation Consortium (ExAC) (8) or that generated by the 1000 genomes project (45). Also, more specific databases centred on geographic regions or on ethnicity are being developed (46). Variants present in more than 5% of the population are considered Single Nucleotide Polymorphisms (SNPs) and are not considered of functional significance in the study of monogenic diseases. Some studies even consider SNPs when the frequency is higher than 1%.

The number of rare variants present in each individual is still high and additional criteria need to be applied for the identification of pathogenic variants. One

of them is the genetic mode of inheritance. In diseases inherited with a recessive pattern, the patient has to be homozygous for the pathogenic variation and the carrier fathers heterozygous. If the disease is inherited with a dominant pattern, the patient can be heterozygous for the variant as well as one of the parents. In the case of X-linkage, the patient should be hemizygous, which is detected as homozygous in the sequence data, and the carrier mother heterozygous. There can also be novel mutations not present in the parents but the frequency is very low and has been estimated in less than one novel mutation per individual (43). The application of inheritance criteria decreases the number of possible pathogenic variants but usually is still not sufficient to determine the causative one.

In the case of families with several affected members, an important criterion is the correlation between the presence of specific sequence variants and the presentation of the disease. This consideration is usually very informative but can be problematic in some diseases of low penetrance where individuals carrying the pathogenic variation may not develop the disease or, at least, not all the symptoms. Usually the next step is to search if any of the sequence variants identified has been previously associated with the disease (Fig 2). Several databases are being generated that compile all the pathogenic variants found in the numerous studies already carried out (representative examples are shown in Table III). Among these databases are the Online Mendelian Inheritance in Man (OMIM, https://www.omim.org/), Human Gene Mutation Database (HGMD) (http://www.hgmd.cf.ac.uk/ac/index.php) (47), Decipher (48), ClinVar (49) or DisGeNET (http://www.disgenet.org) (50). Other databases have a more restricted geographic distribution like the German "VarWatch" project or the Belgian database "SymBioSys" (http://www.kuleuven.be/symbiosys/).

The interpretation of the possible relevance of variants that have not been characterized required additional information. The functional consequences of each specific variant can only be ascertained by experimental approaches. However, there are several in silico prediction tools that can be used as an approximation to predict the functional consequences of amino acid changes on protein structure. Among them are SIFT (51), Polyphen-2 (52), DANN (53), CADD (54) and FATHMM (55) (Fig 2). Other tools predict the possible significance of splice-altering variants (56, 57). These analyses classify

the variants as benign, possibly damaging or damaging which can help to make predictions about their possible pathogenic significance. After these analysis usually several functional variants remain whose significance has not been determined. They are named VUS (variants of unknown significance) and might have pathogenic effect. Therefore, unless the patient carries a well-characterized pathogenic variant, the molecular diagnosis based on the data obtained by NGS is still challenging.

Multigenic diseases.

A number of relatively common diseases also have a genetic basis and their frequency is higher in families with patients of the disease than in the general population. In these cases the disease is caused by variations in more than one gene each of which contributes partially to the pathology. The transmission of these diseases does not follow a simple Mendelian inheritance. Examples of these diseases are inflammatory diseases such as the inflammatory bowel diseases (IBD), including ulcerative colitis (UC), Crohn's disease (CD) (reviewed by (18, 58)) and systemic lupus erythematousus (reviewed by (59)). These diseases are considered caused by the interaction of environmental triggers and the activated immune system in genetically susceptible patients (60, 61). Another example is psychiatric disorders such as the autism spectrum disorder, bipolar disorder and schizophrenia (reviewed by (62)) or Intellectual disability (ID) (63).

The genetic heterogeneity of these diseases with a large number of genes possibly involved makes necessary the study of large populations of patients and healthy control and/or large families with several affected relatives. The only exception are some rare severe variations with high penetrance that show early onset in the patients and that can be identified by studying small families as in the case of monogenic diseases described above. Variations of lower penetrance in these genes can also be present in other patients, which has facilitated their study (64). In the last years the studies have been carried out mainly using microarray hybridization technology for genome-wide association studies (GWAS) focused on common SNPs that identified more than 25,000 significantly disease-associated genetic loci (MacArthur, et al, (65) available at http://www.ebi.ac.uk/gwas/). These studies gave an idea of the large number of genes involved in these diseases. For example, more than 200 loci have been identified for IBD (66) and 40 for autosomal ID although the estimated number may be over 2500 (67).

Because of this complexity, the diagnosis of these diseases, finding the causative variations, requires techniques such as whole exome (WES) or genome (WGS) sequencing (Table II). Among them, WES is the one more extensively utilized because of the easier analysis of the data generated. The criteria applied are the same explained for the study of monogenic diseases except that the variants are expected to be present at higher frequency in the healthy population and, therefore, the number of possible variants to be analyzed is much larger. Because of this reason, a larger number of patient and control samples have to be compared to find significant associations between specific variants and the development of the disease, as mentioned above. Large databases are being generated to facilitate the molecular diagnosis of these hereditary common diseases (Table III). For example, the Deciphering Developmental Disorders group from the United Kingdom has integrated the data generated by WES and CNV studies of 1,133 family trios (proband and both parents) (68). Presently the database contains more than 1,000 genes related to developmental disorders. Even with these limitations, recent studies using WES allowed the diagnosis of about 25% of patients with ID (69).

As mentioned in the Introduction, GWAS studies have associated many possible pathogenic regions to non protein-coding genomic areas. Therefore, WGS will be required in those cases. Presently several WGS experiments have been carried out in common diseases. For example Tan et al (70) studied patients with severe ID and observed a 42% of diagnosis yield. As mentioned above the study of common diseases requires the analysis of a large number of samples what greatly complicates the use of WGS. However, large genomic studies are beginning to be carried out to characterize the complete complexity of the genome and its alteration in the patients. One of them is the PsychENCODE project aimed to generate a public multidimensional genomic database from approximately 1,000 phenotypically well-characterized healthy and disease-affected human post-mortem brains. The project proposed to functionally characterized disease-associated regulatory elements. Initial focus will be on autism spectrum disorder, bipolar disorder and schizophrenia (62).

Cancer

Cancer is one of the most prevalent diseases of genetic origin and one of the most studied by NGS techniques. It is well established that cancer cells carry a

variable number of genetic mutations that provide them with unlimited proliferative capacity. In addition, other mutations can induce tumor progression and metastatic capacity. These mutations can be determined by comparison of tumor cells with proximal non-tumor cells obtained from patient's biopsies (for example, (71)). Numerous extensive studies of different tumors have been carried out using this strategy. The results have shown that tumors usually present a significant number of novel (non-germinal) genetic variations including SNVs and genomic reorganizations such as deletions, duplication, inversions or translocations within and between chromosomes. These reorganizations frequently results in gene copy number variations (CNV) and gene fusions. The frequency of SNVs is especially high in some tumors such as lung cancer and melanoma that are associated to the exposure to external mutagenic agents (tobacco smoke and sun radiation, respectively (72, 73)). Large genome structural changes, known as chromothripsis (74) and chromoplexy (75) are also characteristic of some tumors. The analyses of the results obtained in numerous studies has shown that some of the mutations found are important for cancer development, are frequently found in specific tumors and are considered as driver mutations. On the contrary, other mutations are sporadic and not related to tumor development and are considered passenger mutations (76). Systematic NGS studies aimed to discover the genetics causes of cancers have characterized over 200 cancer driver genes and their relation to specific cancer types (77). One of them is The Cancer Genome Atlas project (TCGA, http://cancergenome.nih.gov) (78). Other projects include the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) and International Cancer Genomic Consortium (ICGC)(htpps://dcc.icgc.org) for paediatric and adult cancers, respectively. The TCGA and ICGC projects will generate WGS data from 25,000 tumours (reviewed in (20)). These genes and the described mutations are available in cancer genome databases (Table III) such as the Catalogue of Somatic Mutations in Cancer (COSMIC) that presently offers curated information for the mutations found in more than one million tumour samples (79, 80). The currently available databases and Web tools for cancer study have been recently reviewed (81). These data have allowed the definition of new tumor subtypes, of tumor-specific biomarkers and the establishment of novel therapeutic targets (see for example (82) and (83)).

The generation of all these molecular data makes possible the use of NGS technology in the clinic. Sequencing of tumor DNA can allow the identification of driver mutations that can be used for the diagnosis of the patient, the precise identification of the tumor type and would help to determine the prognosis and the treatment of the patient, advancing towards the personalized medicine. Guidelines for the interpretation and reporting of sequence variants in cancer have been recently published (84).

All the NGS techniques previously described can be used for the study of cancer (Table II). As mentioned above, cancer is caused by mutation of a reduced number of driver genes some of which are more specifically associated to some tumors (see, for example recent reviews on Non-small Cell Lung Cancer (85) and Gastric Cancer (86)) . Therefore, sequencing of gene panels containing driver genes and even non-coding mutations associated to tumor development are being extensively used in the clinic. This approach presents the limitation that mutations in genes absent from the panel are not detected and might have clinical relevance in some patients. Therefore, WES that provides more comprehensive information is also frequently used. The use of WGS, that generates the most complete information, is presently limited by its greater price and complexity of the analysis of the data generated. Many tumors present gene rearrangements that results in the generation of fusion proteins. These alterations cannot be detected by panel or exome sequencing but mRNA sequencing (RNAseq) can demonstrate their existence. RNAseq can provide also information on transcription regulation that is also altered in many tumors. Because of these reasons, RNAseq (including siRNA) is also frequently used for the clinical study of cancer. The main disadvantage is that RNAseq requires fresh or frozen tissue samples while panel sequencing and WES can be made from DNA isolated from formalin-fixed paraffin-embedded tissue, as discussed above. Presently, a combination of WES and RNAseq is often the preferred option for the clinical diagnosis of cancer (20).

The study of DNA methylation and other epigenetic mechanisms, such as histone modification, provides important information about gene expression associated to tumor development and evolution (38). NGS technology is also being applied to the study of epigenetic alterations in cancer patients (reviewed in (87)). Epigenetic changes provide relevant information for cancer diagnosis and treatment and identify new

biomarkers of clinical use (see for example (88, 89)).

Tumors are very dynamic entities and progressively accumulate mutations so that their genome changes over the time generating intra-tumor heterogeneity as already shown in early NGS studies (90). Tumor diversification is involved in the development of metastasis and on the generation of resistance to therapeutic drugs and can be detected by NGS techniques (reviewed by (91)). An important contribution in this field has been the development of liquid biopsies. The blood contains circulating tumor cells (CTC) and also tumoral DNA and RNA, either as free molecules (fcDNA, free circulating DNA, and fcRNA) (92) or as part of tumoral circulating exosomes (93). The sensitivity of present NGS technology allows sequencing of circulating DNA and RNA (92, 94) and of DNA isolated from single CTCs (95, 96). Panel sequencing, WES and RNAseq (mRNA or siRNA), as well as multiplex PCR studies (97), can be performed from these samples. The data obtained can be used for the molecular diagnosis of the tumor but also to follow tumor development. For example, tumor recurrence after anti-oncogenic treatment can be precociously detected (98). Also, samples can be obtained at different times after tumor diagnosis to determine tumor evolution and the possible appearance of new driver mutations (99). DNA methylation can also be determined from plasma of cancer patients (100). Therefore, the combination of liquid biopsy and NGS represents a notorious advance in the treatment of cancer patients (recently reviewed by (91)).

NGS has been also applied to the study of familial cancer. Some genetic variations result in increased susceptibility to specific types of cancer. Because of this reason some familial pedigrees present a larger frequency of cancer patients than the general population. In some cases a single gene is involved and the study of these families would be similar to the previously described for monogenic diseases of Mendelian inheritance. Representative example would be the BRCA1 and BRCA2 genes, associated to increased frequency of breast and ovarian cancer (101). In many other families cancer susceptibility cannot be associated to variations in a single gene but to the accumulation of SNPs in several genes, more similarly to common multigenic disorders. The study of these families can be approached by the use of WES or WGS, as described for these common diseases (see for example (102)). The data generated in the study of familial cancer are listed in the database http://ghr.nlm.nih.gov.

Clinical implementation of NGS technologies

NGS techniques can be used for a large number of clinical applications including precise molecular diagnosis of the diseases, prognosis, selection of the most adequate treatment, study of disease progression and genetic counselling to the families. The implementation of this technology to the clinic is at the first stages, however, and a number of challenges are still ahead (for recent reviews (18) and (103)).

One important need is the establishment of international standards for the generation and analyses of NGS clinical data. As already mentioned, the interpretation of NGS data is a global enterprise and the data from each patient have to be compared to those of other patients at general databases. Correct data comparison requires that all the studies have been performed with similar criteria and compatible methodologies. Therefore, all the different steps required have to be normalized, from DNA/RNA isolation to the generation of libraries, the length of the reads, sequencing depth, data storage or analysis pipelines. Internal standards are also needed to certify the quality of the analysis. Some of these Standards are already available like the External RNA Control Consortium RNA mixtures (104) and the Genome in a Bottle (GIAB) Consortium standard human genome (105). Several standardization efforts have been recently published by the Sequencing Quality Control Consortium from the USA Food and Drug Administration (106), the Centre for Disease Control and Prevention's Next Generation Sequencing Standardization of Clinical Testing group (107) or the Association of Biomolecular Resources Facilities NGS group (108). The Next-Generation Sequencing Standardization of Clinical Testing II (Nex-StoCT II) informatics workgroup has recently presented a report on the principles and pipelines for NGS data analyses (109). The different aspects of International standardization of NGS studies have been recently reviewed (40, 103).

A second important challenge is the interpretation of the NGS data. These analyses generate a large amount of data and each patient present a large number of sequence variations with respect to the general population. The analysis is based on the comparison of the sequence data from the patient to those of a cohort of healthy and diseased people. The discriminatory capacity of the analysis frequently depends on the size of the reference population. Because of this

reason, large sequence databases are being built such as Decipher (48), HGMD (47), ClinVar (49), as already mentioned (Table III), or the Cinical Genome Resource (ClinGen; http://www.nih.gov/news/health/sep2013/nhgri-25.htm) and the Human Variome Project (http://www.humanvariomeproject.org/) aimed to collect and curate humane genetic variants and alleles that affect health. Efforts are also being made to provide standards and software for sharing and analyzing clinical data. One of these international consortiums is the "global alliance for Genomics and Health (GA4GH). Another example is the "Beacon" network that has developed the "MatchMaker Exchange" (MME) (110). The aim of these comparatives studies would be to correlate the clinical data of the patient with the genetic information using the data contained in these large databases. An important challenge in the process is the homogenization of the terms used for the clinical description of the patient. One of the attempts to create a unified vocabulary is the Unified Medical Language System (UMLS) (111). A more comprehensive approach is the Human Phenotype Ontology (HPO) project aimed to provide bioinformatics resources for the analysis of human diseases and phenotypes. The project is developing a phenotype vocabulary, disease-phenotype annotations and algorithms to operate on this and to offer a bridge between genome biology and clinical medicine (112). Despite this huge progress, much larger databases are still needed, for example in the study of rare diseases where the genetic information available for patients and close relatives is still very limited. Another example is the study of common diseases associated to a large number of genetic variations, as already mentioned.

The development of infrastructures in clinical settings is another challenge. One first aspect would be the acquisition of NGS platforms and required reagents. A second requirement is the availability of a technical stuff, including bioinformaticians and computational biologists, for the interpretation of the data and the elaboration of evidence-based diagnostics reports. In addition, the clinical relevance of the genetic findings in terms of their utility for patient treatment or clinical trial enrolment should be informed by expert clinicians as discussed by Dienstmann et al (113) and Hynes et al (114). As mentioned before, data analysis and interpretation are dependent on the comparison of the data obtained worldwide and contained in large databases. Therefore, the inter-connection of NGS services in both

national and international networks will be necessary for the successful clinical implementation of NGS technologies.

The cost of NGS technology is an important consideration in clinical practice. The most cost-effective NGS technology has to be chosen for each specific case. For example, gene panels that contain between 3 and 100s of genes are commonly used for cancer diagnosis since they can be designed for specific tumor categories. Some monogenic diseases of know etiology can also be analyzed using a panel of genes. These approaches are cheaper, easier to analyze and the turnaround time to diagnoses is faster than in other NGS technologies. However, broader approaches such as RNA sequencing, WES or WGS are needed for many other clinical situations. These approaches are more expensive but it is important to have into account that NGS technologies can help to get a more accurate diagnosis and to select an effective treatment that would result in a lower cost for the health system so that NGS technology can be in a longer term a good investment.

The application of NGS technology also raises some ethical concerns (115). The main reason is that NGS projects generate many more data than those required for diagnosis of the disease. These incidental findings could indicate predisposition to additional pathologies or information otherwise relevant for the future of the patient and its family. Should all this information be offered to the patient, even if the evidence is not completely solid? Alternatively, should the information be filtered to select only clinically relevant information? Usually this problem, that is not always easy, is discussed with the patient. The ownership, access and storage of the data are also controversial. Guidelines for interpretation and reporting of this information are being developed (116, 117).

Concluding remarks

NGS technologies offer an unprecedented capacity to determine the molecular basis of patient's diseases. This information, combined with histopathological and clinical findings, can greatly improve diagnosis, prognosis and the election of a personalized treatment. This promise is becoming a reality in many hospitals for some groups of diseases such as cancer and many monogenic diseases. There are, however some challenges and limitations that have to be improved to get a more general application of these technologies. From a clinical point of view, the more important one is to improve the analytical capacity in order to better determine

pathogenic variants among the many genetic variations found in each patient. This goal requires extended functional analysis of the variants that significantly associate with each disease, as determined by the analysis of a larger number of sequence data from patients, relatives and healthy controls. The development of more comprehensive databases, incorporating sequence and phenotype data, as well as more powerful and accessible computer tools is necessary in the near future. Another important consideration is the cost and the time required for NGS studies. In this respect, the development of new methodologies, based on sequencing of single DNA molecules and on the generation of long reads, are aimed to decrease sequencing cost and to make NGS cheaper and the analysis of the data generated faster than in present technologies. The interpretation of the large amount of data generated by NGS requires the combined expertise of bioinformatics, biologists and clinical experts that probably will require the formation of molecular diagnosis units at the Hospitals. The implementation of these new technological advances, together with the application of standardization and ethical criteria and the organization of specialized units will probably results in the general application of NGS in health systems in a future not too far away.

Acknowledgments

Work at the laboratory of the authors is funded by grant P14-01495 (Fondo de Investigaciones Sanitarias, Instituto de Salud Carlos III, Spain supported by FEDER funds) and grant CIBER 576/805_ER16PE06P2016.

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463-7.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921.
3. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061-73.
4. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015 Oct 01;526(7571):68-74.
5. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015 Oct 01;526(7571):82-90.
6. Park ST, Kim J. Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. Int Neurourol J. 2016 Nov;20(Suppl 2):S76-83.
7. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2012 Jan 10;493(7431):216-20.
8. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285-91.
9. Marian AJ. Challenges in medical applications of whole exome/genome sequencing discoveries. Trends Cardiovasc Med. 2012 Nov;22(8):219-23.
10. Singleton AB. Exome sequencing: a transformative technology. Lancet Neurol. 2011 Oct;10(10):942-6.
11. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012 Feb 17;335(6070):823-8.
12. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74.
13. Sastre L. Clinical implications of the ENCODE project. Clin Transl Oncol. 2012 Nov;14(11):801-2.
14. Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, et al. A promoter-level mammalian expression atlas. Nature. 2014 Mar 27;507(7493):462-70.
15. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014 Mar 27;507(7493):455-61.
16. Ding Z, Mangino M, Aviv A, Spector T, Durbin R. Estimating telomere length from whole genome sequence data. Nucleic Acids Res. 2014 May;42(9):e75.
17. Perona R, Iarriccio L, Pintado-Berninches L, Rodriguez-Centeno J, Manguan-Garcia C, Garcia E, et al. Molecular diagnosis and precission therapeutic approaches for telomere biology disorders. In: Larramendy M, Soloneski S, editors. Telomeres: INTECH; 2016. p. 77-117.
18. Petersen BS, Fredrich B, Hoeppner MP, Ellinghaus D, Franke A. Opportunities and challenges of whole-genome and -exome sequencing. BMC Genet. 2017 Feb 14;18(1):14.
19. Sheikine Y, Kuo FC, Lindeman NI. Clinical and Technical Aspects of Genomic Diagnostics for Precision Oncology. J Clin Oncol. 2017 Mar 20;35(9):929-33.
20. Horak P, Frohling S, Glimm H. Integrating next-generation sequencing into clinical oncology: strategies, promises and pitfalls. ESMO Open. 2016;1(5):e000094.

21. Surrey LF, Luo M, Chang F, Li MM. The Genomic Era of Clinical Oncology: Integrated Genomic Analysis for Precision Cancer Care. Cytogenet Genome Res. 2016;150(3-4):162-75.
22. Fuller CW, Kumar S, Porel M, Chien M, Bibillo A, Stranges PB, et al. Real-time single-molecule electronic DNA sequencing by synthesis using polymer-tagged nucleotides on a nanopore array. Proc Natl Acad Sci U S A. 2016 May 10;113(19):5233-8.
23. Sastre L. New DNA sequencing technologies open a promising era for cancer research and treatment. Clin Transl Oncol. 2011 May;13(5):301-6.
24. Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun. 2015 Dec 09;6:10001.
25. Kluk MJ, Lindsley RC, Aster JC, Lindeman NI, Szeto D, Hall D, et al. Validation and Implementation of a Custom Next-Generation Sequencing Clinical Assay for Hematologic Malignancies. J Mol Diagn. 2016 Jul;18(4):507-15.
26. Wagle N, Berger MF, Davis MJ, Blumenstiel B, Defelice M, Pochanard P, et al. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer Discov. 2012 Jan;2(1):82-93.
27. Beadling C, Wald AI, Warrick A, Neff TL, Zhong S, Nikiforov YE, et al. A Multiplexed Amplicon Approach for Detecting Gene Fusions by Next-Generation Sequencing. J Mol Diagn. 2016 Mar;18(2):165-75.
28. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57-63.
29. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013 Mar 14;152(6):1237-51.
30. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015 May 08;348(6235):660-5.
31. Consoritium TG. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015 May 08;348(6235):648-60.
32. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000 Aug 17;406(6797):747-52.
33. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. N Engl J Med. 2015 Nov 19;373(21):2005-14.
34. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N Engl J Med. 2016 Aug 25;375(8):717-29.
35. Gallego-Paez LM, Bordone MC, Leote AC, Saraiva-Agostinho N, Ascensao-Ferreira M, Barbosa-Morais NL. Alternative splicing: the pledge, the turn, and the prestige : The key role of alternative splicing in human biological systems. Hum Genet. 2017 Apr 03.
36. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet. 2016 May;17(5):257-71.
37. Robertson KD. DNA methylation and human disease. Nat Rev Genet. 2005 Aug;6(8):597-610.
38. Herman JG, Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. N Engl J Med. 2003 Nov 20;349(21):2042-54.
39. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A,

et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb 19;518(7539):317-30.
40. Smith M. DNA Sequence Analysis in Clinical Medicine, Proceeding Cautiously. Front Mol Biosci. 2017;4:24.
41. Stuart BD, Choi J, Zaidi S, Xing C, Holohan B, Chen R, et al. Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. Nat Genet. 2015 May;47(5):512-7.
42. Kuhlenbaumer G, Hullmann J, Appenzeller S. Novel genomic techniques open new avenues in the analysis of monogenic disorders. Hum Mutat. 2011 Feb;32(2):144-51.
43. Sastre L. Exome sequencing: what clinicians need to know. Advances in Genomics and Genetics. 2014;4:15-27.
44. Chakravorty S, Hegde M. Gene and Variant Annotation for Mendelian Disorders in the Era of Advanced Sequencing Technologies. Annu Rev Genomics Hum Genet. 2017 Apr 17.
45. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65.
46. Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. 2011Kaviar: an accessible system for testing SNV novelty. Bioinformatics. Nov 15;27(22):3216-7.
47. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet. 2014 Jan;133(1):1-9.
48. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet. 2009 Apr;84(4):524-33.
49. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016 Jan 04;44(D1):D862-8.
50. Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2016 Jan 04;45(D1):D833-D9.
51. 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(7):1073-81.
52. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248-9.
53. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015 Mar 01;31(5):761-3.
54. 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 Mar;46(3):310-5.
55. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015 May 15;31(10):1536-43.
56. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014 Dec 16;42(22):13534-44.
57. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, et al. RNA splicing. The human splicing

code reveals new insights into the genetic determinants of disease. Science. 2014 Jan 09;347(6218):1254806.
58. Katsanos KH, Papadakis KA. Pharmacogenetics of inflammatory bowel disease. Pharmacogenomics. 2014 Dec;15(16):2049-62.
59. Morris DL, Sheng Y, Zhang Y, Wang YF, Zhu Z, Tombleson P, et al. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. Nat Genet. 2016 Aug;48(8):940-6.
60. Baumgart DC, Sandborn WJ. Crohn's disease. Lancet. 2012 Nov 03;380(9853):1590-605.
61. Ellinghaus D, Bethune J, Petersen BS, Franke A. The genetics of Crohn's disease and ulcerative colitis--status quo and beyond. Scand J Gastroenterol. 2015 Jan;50(1):13-23.
62. Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, et al. The PsychENCODE project. Nat Neurosci. 2015 Dec;18(12):1707-12.
63. Harripaul R, Noor A, Ayub M, Vincent JB. The Use of Next-Generation Sequencing for Research and Diagnostics for Intellectual Disability. Cold Spring Harb Perspect Med. 2017 Mar 01;7(3).
64. Essers JB, Lee JJ, Kugathasan S, Stevens CR, Grand RJ, Daly MJ. Established genetic risk factors do not distinguish early and later onset Crohn's disease. Inflamm Bowel Dis. 2009 Oct;15(10):1508-14.
65. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2016 Jan 04;45(D1):D896-D901.
66. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015 Sep;47(9):979-86.
67. Musante L, Ropers HH. Genetics of recessive cognitive disorders. Trends Genet. 2014 Jan;30(1):32-9.
68. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet. 2015 Apr 04;385(9975):1305-14.
69. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med. 2013 Oct 17;369(16):1502-11.
70. Tan CA, Topper S, Del Gaudio D, Nelakuditi V, Shchelochkov O, Nowaczyk MJ, et al. Characterization of patients referred for non-specific intellectual disability testing: the importance of autosomal genes for diagnosis. Clin Genet. 2015 Feb 19;89(4):478-83.
71. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012 Sep 14;150(6):1107-20.
72. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. 2010 Jan 14;463(7278):184-90.
73. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010 Jan 14;463(7278):191-6.
74. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell. 2011 Jan 7;144(1):27-40.
75. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y,

et al. Punctuated evolution of prostate cancer genomes. Cell. 2013 Apr 25;153(3):666-77.
76. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013 Oct 17;502(7471):333-9.
77. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014 Jan 23;505(7484):495-501.
78. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113-20.
79. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004 Mar;4(3):177-83.
80. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015 Jan;43(Database issue):D805-11.
81. Yang Y, Dong X, Xie B, Ding N, Chen J, Li Y, et al. Databases and web tools for cancer genomics study. Genomics Proteomics Bioinformatics. 2015 Feb;13(1):46-50.
82. Genome CLC. A genomics-based classification of human lung tumors. Sci Transl Med. 2013 Oct 30;5(209):209ra153.
83. Ellis MJ, Perou CM. The genomic landscape of breast cancer as a therapeutic roadmap. Cancer Discov. 2013 Jan;3(1):27-34.
84. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn. 2017 Jan;19(1):4-23.
85. Sabari JK, Santini F, Bergagnini I, Lai WV, Arbour KC, Drilon A. Changing the Therapeutic Landscape in Non-small Cell Lung Cancers: the Evolution of Comprehensive Molecular Profiling Improves Access to Therapy. Curr Oncol Rep. 2017 Apr;19(4):24.
86. Katona BW, Rustgi AK. Gastric Cancer Genomics: Advances and Future Directions. Cell Mol Gastroenterol Hepatol. 2017 Mar;3(2):211-7.
87. Soto J, Rodriguez-Antolin C, Vallespin E, de Castro Carpeno J, Ibanez de Caceres I. The impact of next-generation sequencing on the DNA methylation-based translational cancer research. Transl Res. 2016 Mar;169:1-18 e1.
88. Ibanez de Caceres I, Battagli C, Esteller M, Herman JG, Dulaimi E, Edelson MI, et al. Tumor cell-specific BRCA1 and RASSF1A hypermethylation in serum, plasma, and peritoneal fluid from ovarian cancer patients. Cancer Res. 2004 Sep 15;64(18):6476-81.
89. Ibanez de Caceres I, Cortes-Sempere M, Moratilla C, Machado-Pinilla R, Rodriguez-Fanjul V, Manguan-Garcia C, et al. IGFBP-3 hypermethylation-derived deficiency mediates cisplatin resistance in non-small-cell lung cancer. Oncogene. 2010 Mar 18;29(11):1681-90.
90. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature. 2009 Oct 8;461(7265):809-13.
91. Roukos DH. Spatiotemporal diversification of intrapatient genomic clones and early drug development concepts realize the roadmap of precision cancer medicine. Drug Discov Today. 2017 Apr 08.
92. Lebofsky R, Decraene C, Bernard V, Kamal M, Blin A, Leroy Q, et al. Circulating tumor DNA as a non-invasive

substitute to metastasis biopsy for tumor genotyping and personalized medicine in a prospective trial across all tumor types. Mol Oncol. 2015 Apr;9(4):783-90.
93. Kalluri R. The biology and function of exosomes in cancer. J Clin Invest. 2016 Apr 01;126(4):1208-15.
94. Rothe F, Laes JF, Lambrechts D, Smeets D, Vincent D, Maetens M, et al. Plasma circulating tumor DNA as an alternative to metastatic biopsies for mutational analysis in breast cancer. Ann Oncol. 2014 Oct;25(10):1959-65.
95. Lohr JG, Adalsteinsson VA, Cibulskis K, Choudhury AD, Rosenberg M, Cruz-Gordillo P, et al. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. Nat Biotechnol. 2014 May;32(5):479-84.
96. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016 Mar;17(3):175-88.
97. Sequist LV, Heist RS, Shaw AT, Fidias P, Rosovsky R, Temel JS, et al. Implementing multiplexed genotyping of non-small-cell lung cancers into routine clinical practice. Ann Oncol. 2011 Dec;22(12):2616-24.
98. Chia PL, Do H, Morey A, Mitchell P, Dobrovic A, John T. Temporal changes of EGFR mutations and T790M levels in tumour and plasma DNA following AZD9291 treatment. Lung Cancer. 2016 Aug;98:29-32.
99. Chan KC, Jiang P, Zheng YW, Liao GJ, Sun H, Wong J, et al. Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. Clin Chem. 2013 Jan;59(1):211-24.
100. Legendre C, Gooden GC, Johnson K, Martinez RA, Liang WS, Salhia B. Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer. Clin Epigenetics. 2015;7:100.
101. Melchor L, Benitez J. The complex genetic landscape of familial breast cancer. Hum Genet. 2013 Aug;132(8):845-63.
102. Phelan CM, Kuchenbaecker KB, Tyrer JP, Kar SP, Lawrenson K, Winham SJ, et al. Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. Nat Genet. 2017 May;49(5):680-91.
103. Mason CE, Afshinnekoo E, Tighe S, Wu S, Levy S. International Standards for Genomes, Transcriptomes, and Metagenomes. J Biomol Tech. 2017 Apr;28(1):8-18.
104. Munro SA, Lund SP, Pine PS, Binder H, Clevert DA, Conesa A, et al. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nat Commun. 2014 Sep 25;5:5125.
105. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016 Jun 07;3:160025.
106. Consortium SM-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol. 2014 Sep;32(9):903-14.
107. t Hoen PA, Friedlander MR, Almlof J, Sammeth M, Pulyakhina I, Anvar SY, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol. 2013 Nov;31(11):1015-22.
108. Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat Biotechnol. 2014 Sep;32(9):915-25.
109. Gargis AS, Kalman L, Bick DP, da Silva C, Dimmock DP, Funke BH, et al. Good laboratory practice for clinical

next-generation sequencing informatics pipelines. Nat Biotechnol. 2015 Jul;33(7):689-93.
110. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. Hum Mutat. 2015 Oct;36(10):915-21.
111. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 01;32(Database issue):D267-70.
112. Kohler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Ayme S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017 Jan 04;45(D1):D865-D76.
113. Dienstmann R, Dong F, Borger D, Dias-Santagata D, Ellisen LW, Le LP, et al. Standardized decision support in next generation sequencing reports of somatic cancer variants. Mol Oncol. 2014 Jul;8(5):859-73.
114. Hynes SO, Pang B, James JA, Maxwell P, Salto-Tellez M. Tissue-based next generation sequencing: application in a universal healthcare system. Br J Cancer. 2017 Feb 28;116(5):553-60.
115. Shahmirzadi L, Chao EC, Palmaer E, Parra MC, Tang S, Gonzalez KD. Patient decisions for disclosure of secondary findings among the first 200 individuals undergoing clinical diagnostic exome sequencing. Genet Med. 2014 Oct 10;16(5):395-9.
116. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013 Jul;15(7):565-74.
117. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, 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. Genet Med. 2015 May;17(5):405-24.

Tables

Table I. Advantages and limitation of the different NGS technologies

Technology	Advantages	Limitations
Whole Genome Sequencing (WGS)	- Complete information on coding and regulatory regions - Detects structural alterations and Copy Number variations	- Large amount of sequence generated - Complex analyses of data - Lower sequencing depth
Whole Exome Sequencing (WES	- Information of all protein-coding genes - Smaller amount of sequencing and data analysis required than in WGS - Intermediate sequencing depth	- No information in non - coding regions - Large insertions and deletions are not detected
RNA Sequencing (RNAseq)	- Information on coding and non-coding RNAs expressed in the analyzed sample - Information on alternative splicing and on the use of alternative promoters. - Detects translocations - Information on gene expression levels - Variable sequencing depth	- No information in non-protein-coding regions and in genes not expressed in the sample - Requires the isolation of high-quality RNA
Targeted Panel Sequencing	- Analysis of a reduced number of genes simplify interpretation of the data - The highest sequencing depth - Low cost and easy implementation in clinical settings	- Analysis restricted to pre-determined genome regions. - Possible variation in other regions is not detected
Global Analysis of DNA methylation	- Provides information on gene expression regulation that is specific to this technology	- Higher technical complexity requiring DNA of higher quality - Information on nucleotide variability is not reported

Table II. NGS technologies most frequently used for different groups of diseases

Group of diseases	NGS Technologies
Monogenic diseases	Targeted Panel Sequencing Whole Exome sequencing (WES) Whole Genome sequencing (WGS)
Complex diseases	Whole Exome sequencing (WES) Whole Genome sequencing (WGS)
Cancer	Targeted Panel Sequencing RNA sequencing (RNAseq) Whole Exome sequencing (WES) Whole Genome sequencing (WGS) Genome-wide DNA methylation

Table III. Examples of presently available databases for clinical diagnosis

Group of diseases	Databases
Monogenic diseases	OMIM; https://www.omim.org Decipher; https://decipher.sanger.ac.uk/ HGMD; http://www.hgmd.cf.ac.uk/ac/index.php ClinVar; https://www.ncbi.nlm.nih.gov/clinvar/ DisGeNET; http://www.disgenet.org
Complex diseases	PsychENCODE; https://www.nimhgenetics.org/available_data/psychencode/ Deciphering Developmental Disorders; https://www.ddduk.org/
Cancer	TCGA; http://cancergenome.nih.gov TARGET; http://archive.broadinstitute.org/cancer/cga/target ICGC; htpps://dcc.icgc.org COSMIC; http://cancer.sanger.ac.uk/cosmic MethyCancer; http://methycancer.psych.ac.cn

Figures

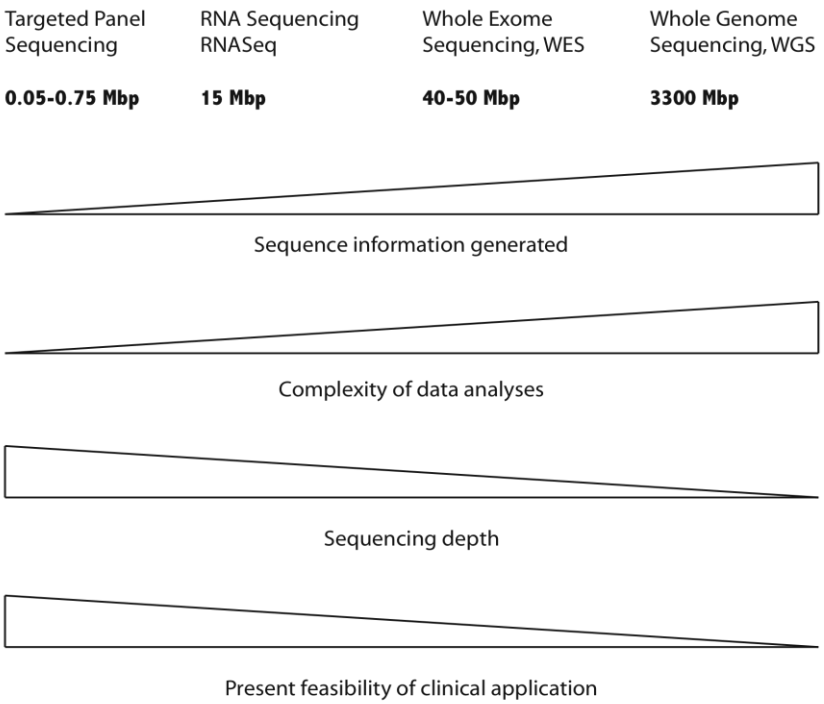

Figure 1

Figure 1. Schematic comparison of the different NGS methods presently available for clinical diagnosis. The four more generally used methods of NGS are indicated at the upper part of the figure. Underneath each methods name the extension of DNA being sequenced in indicated in Mega base pairs (Mbp, 10^6 bp). The triangles indicate increasing amount of information or feasibility of the parameter indicated under each of them from Targeted Panel Sequencing to Whole Genome Sequencing. These parameters include the amount of sequence information generated; the complexity involved in the analysis of the data generated; sequencing depth, that makes reference to the number of times that each specific nucleotide in independently sequenced; and the possibility for clinical application at the present time.

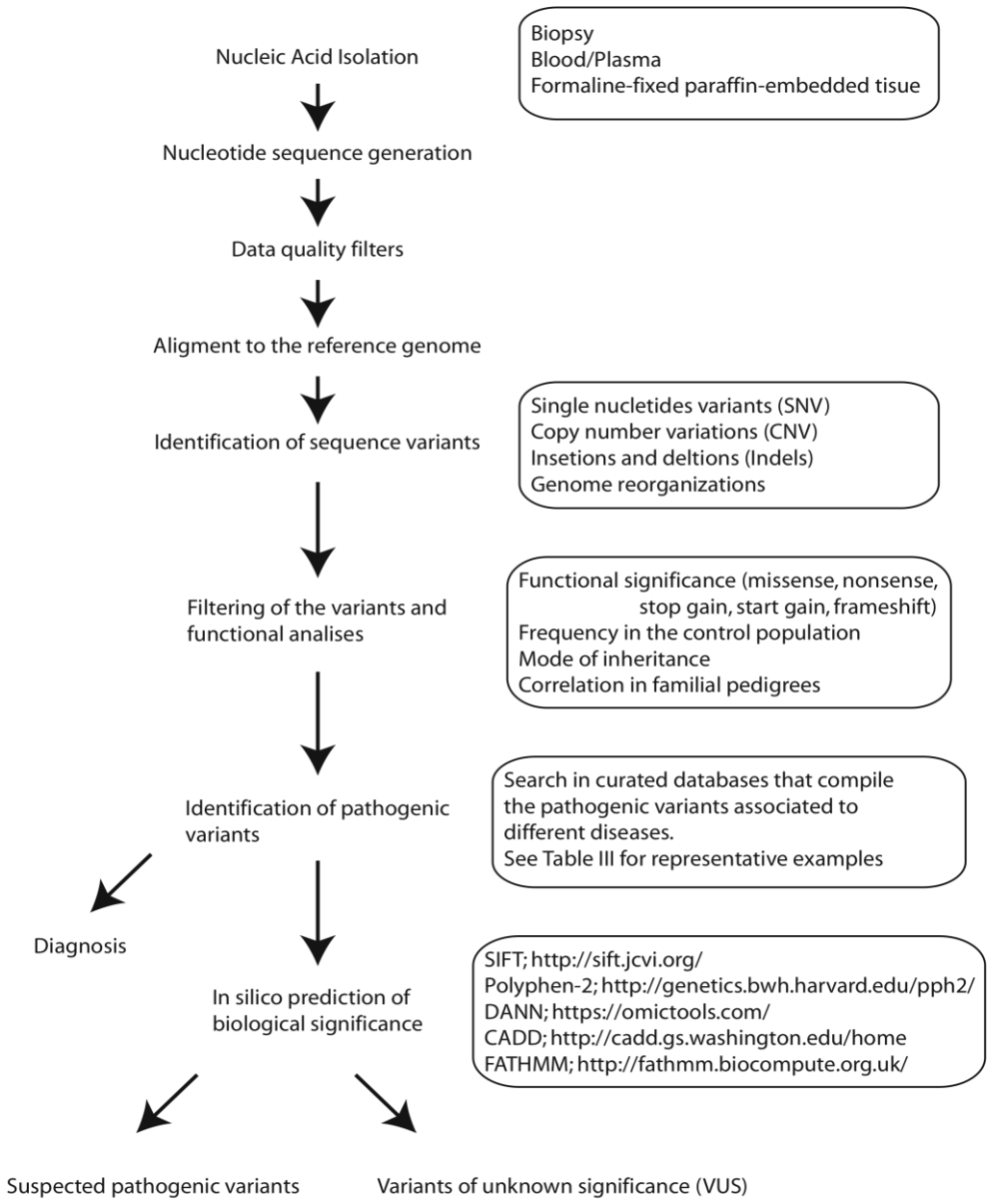

Figure 2

Figure 2. Workflow of a typical NGS experiment
The different steps involved in a typical NGS experiment, from the isolation of the samples (DNA or RNA), sequencing and data analysis are schematically shown. Some of the variables involved in these experiments are shown in the boxes at the right side of the figure. The upper box indicates the biological samples used for DNA or RNA isolation. The second one the type of sequence variants found in the samples in comparison to the reference genome. The third one the criteria more often used for the selection of variants of possible clinical relevance. The fourth box indicates that pathological variants can be identified by searching a number of databases that are shown in more detail in Table III. The lower box indicates several informatic programs that are used to determine the possible functional significance of the nucleotide variants found. The technical steps required for the generation of the nucleotide sequence, including the isolation of exons and other DNA regions (exome and panel sequencing), reverse transcription of RNA (RNAseq) and the preparation of libraries are not shown for simplicity.