Published in final edited form as:

Per Med. 2014 ; 11(5): 523–544. doi:10.2217/pme.14.34.

The road from next-generation sequencing to personalized medicine

Manuel L. Gonzalez-Garay

Center for Molecular Imaging, Division of Genomics & Bioinformatics, The Brown Foundation Institute of Molecular Medicine, University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Manuel L. Gonzalez-Garay: manuel.l.gonzalezgaray@uth.tmc.edu

Abstract

Moving from a traditional medical model of treating pathologies to an individualized predictive and preventive model of personalized medicine promises to reduce the healthcare cost on an overburdened and overwhelmed system. Next-generation sequencing (NGS) has the potential to accelerate the early detection of disorders and the identification of pharmacogenetics markers to customize treatments. This review explains the historical facts that led to the development of NGS along with the strengths and weakness of NGS, with a special emphasis on the analytical aspects used to process NGS data. There are solutions to all the steps necessary for performing NGS in the clinical context where the majority of them are very efficient, but there are some crucial steps in the process that need immediate attention.

Keywords

CADD; functional prediction program; genomics; GWAVA; NGS; personalized medicine; workflow management system

The current medical model focuses on the detection and treatment of pathologies. Treating disorders, especially on advanced states, is very expensive for patients and society in general. Screening for five of the most common disorders in the USA (cardiovascular disorders, stroke, cancer, chronic obstructive pulmonary disease and diabetes) could protect millions of lives and reduce the healthcare deficit [1]. Tailoring drug therapies by practicing personalized medicine (PM) has the potential to improve treatment of cancer and save lives by preventing drug-related fatalities. A new technology, next-generation sequencing (NGS), has the potential to accelerate the early detection of disorders and to detect pharmacogenetics markers to customize treatments [2].

For reprint orders, please contact: reprints@futuremedicine.com

Financial & competing interests disclosure

The research was supported by the Cullen Foundation for Higher Education. The funding organizations made the Awards to The University of Texas Health Science Center at Houston (UTHSCH). The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Initial work to generate the human genome template

In 1977, the Nobel laureate, Frederick Sanger developed the ‘dideoxy’ chain-termination method coupled with electrophoretic size separation for sequencing DNA molecules [3]. Sanger sequencing, as it is known today, started with low efficiency and high cost, but thanks to the work of a large number of scientists the cost of sequencing was reduced dramatically reaching a price of US\$0.0024/base by the mid-1990s [4]. The Human Genome Project started in 1990 after the scientific community recognized the urgent need for a complete map of the human genome. The project lasted 13 years with an astronomic cost of US\$3 billion and the involvement of thousands of international scientists [5]. The Human Genome Project transformed molecular biology by eliminating the need to individually clone and sequence genes of interest. During this period, there was a ferocious competition between the International Human Genome Sequencing Consortium (IHGSC), under the direction of Francis Collins (MD, USA), head of the National Human Genome Research Institute at the NIH and the private sector (Celera [CA, USA]) headed by Craig Venter (MD, USA). Both groups published the first draft of their human genome assemblies in 2001. IHGSC published the sequence in 15 February [6] while Venter published in 16 February [7]. Venter’s group used a shotgun clustering approach while the IHGSC used an independent bacterial artificial chromosome (BAC)-by-BAC approach. We now know that both groups produced mistakes in their first human genome drafts. There were hundreds of thousands of gaps and misassembled regions in both drafts [8].

It took 3 years for the IHGSC sequencing centers to finish filling the gaps in the draft. The finished version of the human assembly was published by the National Center for Biotechnology Information (NCBI) as NCBI build 35, also known as hg17 [9]. At the time of this writing, three subsequent versions have been released. The Genome Research Consortium (GRC) is the new organization in charge of working with genome assemblies, the latest version of the human assembly is known as GRCh38, and it was released on 24 December 2013. However, the majority of the sequencing groups still use GRCh37 (hg19) since it takes time and effort to migrate all the previously generated genomes to the new assembly.

Annotating the first human genome

Before and during the release of the first human genome assembly, thousands of scientists produced information about the structure and function of single genes. Projects like the expressed sequence tag generated millions of short subsequences of a cDNA sequence. Expressed sequence tag project identified the presence of thousands of genes and provided valuable information about alternative splice variants of genes [10,11]. During this period of time, bioinformaticians developed programs to scan the human genome assemblies for potential new genes. The IHGSC selected three gene prediction programs to scan the human assemblies: Genscan [12], a program developed by Burge *et al.* that identifies complete gene structures including exon–intron boundaries using a general probabilistic model of the gene structure and GC composition; Genie [13], a gene prediction program originally developed for the *Drosophila* genome, was selected to inspect the human assemblies. Genie was developed using generalized Hidden Markov models; and FGENES [14], a commercial

software developed by Softberry, Inc. (NY, USA) The predicted gene models are continually validated using biological data from well-annotated databases.

With the release of the first human genome, a group of human geneticists became interested in generating a map of human genetic variations or a haplotype map (HapMap). For the international HapMap project, four populations were selected with a total of 270 people. Two populations consisted of trios (a father, mother and an adult child), the Yoruba people of Ibadan, Nigeria, provided 30 trios and the USA provided 30 trios from US residents with northern and western European ancestry (Centre d'Étude du Polymorphisme Humain [CEPH]). The remaining two populations consisted of unrelated individuals. Japan provided 45 samples and China provided another 45 samples [15]. By 2005, approximately 1 million variants were genotyped and their linkage disequilibrium patterns characterized in Phase I of the project [16]. A second set of results was published in 2007 where more than 3 million variants were identified and characterized [17]. During the third phase of the HapMap project additional samples were genotyped, increasing total number of samples to 1301 from a variety of human populations [18]. For a more detailed review about the HapMap project and its impact on the discovery of SNP associated with common diseases, see Manolio *et al.* [19]. The information generated by HapMap project, including allele frequencies, have been incorporated into the public catalog of variant sites in the Database of SNPs (dbSNP) [20].

The birth of the NGS technology

The next logical objective to pursue, after the human genome was finished, was to sequence the diploid genome of a single person. However, the main problem was that the Sanger sequencing technology was expensive and slow. These arguments did not stop Venter from sequencing his own genome in September 2007. Venter published the first diploid human genome (called 'HuRef') [21]. The HuRef genome was the most expensive personal genome in history (US\$100 million).

On the other hand, visionaries like Jay Shendure (WA, USA) and George Church (MA, USA) concentrated their efforts into developing faster and more economical technologies. Church's group developed the first multiplex sequencing technology (Polony Sequencing). The Polony Sequencing combined the used of emulsion PCR, ligation and four-color imaging [22]. The sequencing machine was named Polonator. Polonator was a low cost sequencing machine (US\$170,000) [23].

Rothberg (CT, USA) developed an alternative sequencing technology based on miniaturized pyrosequencing reactions that run in parallel [24]. The technology captures the signals using charge-coupled device (CCD) camera-based imaging [25]. The final product was marked as 454 technologies, and it was quickly used to sequence multiple organisms including bacteria. In 2008, the entire genome of James Watson was sequenced using 454 technologies [26]. Watson's genome was sequenced in a record time of 4 months at a cost of US \$1,500,000 [27]. After 454 technologies was sold to Roche (Basel, Switzerland) and Rothberg departed, there was not a significant improvement in the technology and eventually in October 2013, Roche shut down 454.

Life technologies (CA, USA) developed a sequencing system borrowing the chemistry properties used by Polony Sequencing [28]. The machines were commercialized under the name SOLiD™ Instruments. SOLiD instruments allowed the sequencing of whole genomes at a lower price of US\$100,000. The first genome sequenced using SOLiD technology was the genome of Lupski, a geneticist from Baylor College of Medicine (TX, USA) [29]. Even though SOLiD technology was the most accurate sequencing technology, the major obstacles for the acceptance of SOLiD technology were the complexity of analyzing color space data and the large amount of computational resources required for its analysis. In addition, the read length was very short, 50 bp, in comparison with Illumina® (CA, USA) that normally generates reads over 100 bp for each side of every fragment (using the paired-end mode).

A fourth sequencing company emerged from the Cambridge Chemistry Department, Solexa with offices in Chesterford (UK) and Hayward (CA, USA). Solexa's technology was different from the existing NGS technologies. It was based on clonal arrays, and massively parallel sequencing of short reads using solid-phase sequencing by reversible terminators. The first machine was commercialized under the name Genome Analyzer and became commercially available in 2006. Solexa was acquired by Illumina in early 2007. Illumina eventually became the predominant sequencing technology, thanks to their aggressive marketing team, the simplicity of their technology and their constant efforts to improve their technology [30,31].

DNA nanoball sequencing is a technology developed by Complete Genomics Inc., (CGI; CA, USA) [32]. CGI's business strategy was different from other companies. Instead of selling machines, CGI exclusively sequenced human genomes and performed their downstream analysis delivering an annotated human genome as a final product. Their analysis included copy number variations, structural variations, variant calling, variant annotation, detection of mobile elements and multiple additional reports [33]. Their analysis reduced the computational challenges for customers. CGI was a very important player in the field; CGI's marketing forced competition to lower the price of whole human genomes. In addition, CGI changed the model of purchasing expensive equipment to a model of genome sequencing as a service. CGI is a very creative company but they were limited in that their only product was their genome services, in comparison with their competitors that had multiple sources of revenues (e.g., instruments, reagents, support and service, among others).

Other technologies like the Ion Torrent™ Systems entered the market at a later time (February 2010). Ion Torrent brought semiconductor based detection systems to the sequencing arena. Ion Torrent technology produced a significant improvement to the omnipresent and slow technology of image acquisition [34]. Ion Torrent keeps increasing its market share. Their system has the benefit of a very short turnaround time, an advantage when working with critical care patients that need an answer on the same day.

Single-molecule real-time (SMRT) sequencing is based on the sequencing by synthesis and real-time detection of the incorporation of fluorescent labels. The advantage of this technology is the continuous long reads generated by the instruments [35]. The technology

was developed by Pacific Biosciences® (PacBio; CA, USA) and recently, the latest machine PacBio RS II was released in April 2013. PacBio sequencing technology plays a very important role in filling the gaps in current assemblies [36].

There are many other new technologies on development that will make the sequencing even faster and more economical, such as Oxford Nanopore technologies (GridION™ System based on nanopore-based sensing), Fluidigm® (single-cell sequencing) and Nabsys (positional sequencing), among others. Figure 1 highlights the major events in next generation sequencing

Focus on the protein-coding genome

The best and more direct approach to study a person's genome would be to sequence the whole genome. However, since only roughly 2–3% of the human genome code for proteins, but harbor approximately 85% of the mutations with large effects on disease-related traits [37], it becomes a logical choice to focus efforts on a smaller subset of the genome that contains the exons (i.e., the exome). In addition, the interpretation of the functional effects of a mutation in a noncoding region of the genome is an extremely difficult task, as you will read in a further section of this review. This targeted approach reduced the cost and time to sequence samples but more importantly it reduced the computational processing time by at least 50 times.

The process of enrichment by hybridization has been commercialized mainly by three companies: Illumina, Nimble-Gen (Basel, Switzerland) and Agilent (CA, USA). Illumina offers three products: Nextera (target region 37 Mb); Nextera Expanded Exome Kit (target region of 62 Mb) and TruSight One (12 Mb including exons with known human disease genes) [38]. NimbleGen offers 'SeqCap EZ Exome v3' (target region 64 Mb) [39]. Agilent offers 'SureSelect Human All' (target region 75 Mb) [40]. All the enrichment kits, with the exception of TruSight One, are capable of capturing exons, 5' UTR, 3' UTR, miRNA and other noncoding RNA.

The challenge of working with billions of short reads

The development of new instruments capable of generating data in the gigabase-pair scale generated a new problem: the lack of software capable of aligning and assembling short reads. During the early days of NGS (2007–2008), there were direct requests from NIH to the scientific community especially the computational biologists to design short-read sequencing mapping tools (SRSMT) that work with NGS data. The bioinformatics community solved the problem very fast. By 2008, the first open source SRSMT was released 'Mapping and Assembly with Quality' (Maq) [41]. Maq is capable of mapping short reads to reference sequences and build an assembly. A recent survey estimates that the current number of SRSMT is over 70 [42]. Most of the current SRSMTs accelerate the mapping by creating indexes (hash tables) for the reads or the reference genome. Some bioinformaticians categorize the SRSMTs as genome-indexing or read-indexing. In general, the read-indexing SRSMTs like Maq or RMAP [43] perform better in short genomes and the genome indexing SRSMTs perform better with larger genomes like humans. The majority of the current SRSMTs are genome-indexing. Genome-indexing SRSMTs differ from each

other by the presence or absence of features or by the algorithm used to implement a feature of the software. The main differences between genome-indexing SRSMTs are in the following features: the technique used to create the index; the seeding algorithm; the usage of base-quality scores; the allowance of gaps during the alignment; and the quality threshold. The combination of each one of these features makes each SRSMT unique and a challenge for the user to select the right one. The most widely used SRSMTs are Bowtie2 [44], BWA [45], SOAP2 [46], GSNAP [47], Novoalign [48] and mrs-FAST/mrFAST [49,50]. Each one of them has its own strengths and weaknesses, and there is not a single best tool as each performs better under different conditions [51].

Variant callers

After the short reads have been aligned against the reference genome, variants need to be extracted from the alignments. Software packages that detect single nucleotide variations (SNV) and small insertion and deletions (Indels) are called SNV callers, while programs that determine the genotype for each site are called genotype callers. Before submitting information to the SNV callers, it is necessary to minimize the experimental errors in the alignment files or Binary files containing the Sequence Alignment/Map format (BAM files). Experimental errors and technology-specific artifacts could be introduced systematically or randomly.

SNV detection relies on the identification of statistical differences between the base found in a site of the template and the corresponding base found in the aligned reads. Any sequencing error can lead to an incorrect SNV identification. To avoid this problem, the Broad Institute (MA, USA) generated a programing suite PICARD [52] to identify and correct systematic errors on the initial BAM files. The PICARD suite complements and provides functionality to the Genome Analysis Toolkit (GATK) [53]. The GATK was developed at the Broad Institute to analyze NGS data and facilitate the identification of variant discovery. GATK was designed by geneticists and engineers with a very robust architecture. Some of the available high-quality variant callers are capable of identifying SNV and indels while others detect only SNVs. The most commonly used variant callers are listed in Table 1. High-quality BAM files with high levels of coverage are processed very well by all of them but BAM files with low levels of coverage and/or low quality are processed very poorly (for additional information and comparisons see [54–56]).

Distinguishing the forest from the trees: rare variants

As described in a previous section, population geneticists have been studying the distribution of variants in the population for many years, and they have found a correlation between the frequency of the variant and the expression of a phenotype (penetrance). Population geneticists postulated that a very low frequency allele is more likely to be responsible for a Mendelian phenotype with extreme and rare phenotype and that a common variant that it is fixed in the genome carries a low risk of being responsible for the phenotype [68,69]. This observation provides a perfect explanation for Mendelian disorders and has become the practical basis to identify potentially damaging mutations on NGS experiments. Common variants in a population are called SNP, the exact minor allele

frequency (MAF) used to distinguish a rare variant and a SNP is a subject of debate for the population geneticists. It has become common practice to filter out any variant that has a MAF bigger than 1.0%. The threshold of 1.0% for filtering is an arbitrary cutoff value, and the value depends on the source (population) and size of the samples used to generate the MAF information. Large sequencing centers, which have sequenced thousands or millions of local patients, will have better information about what frequency values to use as a cutoff value on such filters. A small laboratory has to use publicly available databases to estimate the MAF. Using publicly available data, as a sole source of frequency information, to filter NGS data increases the risk to over or under filter variants. Resources to obtain allele frequency information are listed in Table 2.

Information & material required to take NGS to the clinic

With the availability of many sequencing methods, short-read aligners and variant callers, there are significant differences between variant calls and interpretation of results. Efforts have been made to identify the most common practices between the top sequencing groups and suggest standards for best practices. A recent publication by the international CLARITY Challenge provides a comprehensive assessment of current practices for using genome sequencing to diagnose and report genetic diseases [83]. Their surveys and best practices provide important insights into clinical laboratories but do not provide the tools to evaluate their own implementation of the process. A universal, highly accurate set of genotypes across a genome that can be used as a benchmark is required to standardize clinical laboratories that offer clinical exomes and genomes.

The National Institute of Standards and Technology organized the ‘Genome in a Bottle Consortium’ (GBC) to develop such benchmarks. GBC developed and made publicly available the reference material, reference methods and reference data [84]. In a recent publication, GBC describes the sample selected for reference material, HapMap/Collection of European Samples (CEU) female NA12878, the 14 data sets generated by six different sequencing platforms, eight different mapping programs and various variant callers. GBC integrated all the information and provided a validated set of SNPs and indels, in addition they provided recommendations on how to deal with complex variants and genomic regions that are difficult to genotype [85]. Their work was essential for the recent authorization by the US FDA of the first next-generation sequencer Illumina’s MiSeqDx [86].

Distinguishing between benign & deleterious mutations

When a mutation occurs in the coding sequence of a protein, the result could be: a synonymous change (no amino acid change); a missense mutation (a single amino acid substitution in the protein); a premature chain termination; a frame-shift in the protein due to the addition or deletion of one or more nucleotides; and an altered exon–intron splice junction. The interpretation of the functional effect of all cases is readily done for all, except for the missense mutation(s). If a variant has not been studied before, it is considered a variant of unknown significance. Such variants are a source of diagnostic challenge and uncertainty for families.

The most straightforward approach to analyze a variant is to search databases that store information about known disease-causing mutations (DCM). Catalogs of DCMs are very useful, but the information has to be evaluated very carefully. DCM databases are very small and include errors that were carried over from the original scientific studies. The most widely used catalogs of DCMs are listed in Table 3. In most clinical laboratories pathogenic variants are detected using Human Genome Mutation Database (HGMD) Professional [87,88] and ClinVar databases [89]. HGMD is unquestionably the largest catalog of DCM mutations with approximately 116,000 DCM (release dated December 2013; variantType = DCM) while the latest release of ClinVar (March 2014) only has approximately 29,000 variants considered 'pathogenic'. Unfortunately, the number of pathogenic variants in both databases represents only a small fraction of the potential number of pathogenic mutations in a population of approximately 7 billion humans. Consequently, the majority of the missense mutations found in a NGS experiment will not be classified by DCM databases and alternative approaches are needed for the interpretation of such variants.

To perform the interpretation of the functional effect of variants that are not in a DCM catalog, functional prediction programs (FPPs) have to be used. FPP are capable of detecting pathogenic variations with some degree of certainty. Table 4 lists the majority of FPPs and few databases with precomputed scores. The method employed by each FPP is used to categorize them, and it is provided in the column label 'Category' of Table 4.

Under category 1 (protein stability), there are FPPs that evaluate how the stability of the protein is affected by an amino acid change. In an ideal situation, we would expect that the interpretation of the functional effect of a variant should be easily done by analyzing the 3D structure of a protein and query for the effect of the change on the 3D structure of the proteins. However, it is much more complicated process. The 3D structures of protein are stored in the protein data bank (PDB). PDB stores only 3D structures for a very small fraction of the entire set of human proteins (human proteome). In many cases, sections of a protein cannot be crystallized generating regions of a protein without a 3D structure. In addition, the majority of genes, during expression, will produce alternative splice variants. Alternative splice variants generate multiple protein isoforms from a single genetic locus. The vast majority of protein isoforms lack 3D structures. Furthermore, to be certain about the structural change of the amino acid substitution on the protein, we need the 3D structure of the wild-type protein and the 3D structure of the mutated protein. If we only have the 3D structure of the wild-type protein, it is possible to estimate the structural changes of the mutated protein by using molecular modeling [194] (for a recent review on molecular modeling, see [195]).

The FPPs under the category 2 (protein sequence and structure) evaluate the consequences of the amino acid changes by looking at individual amino acid properties and locations. For example, if an amino acid change is located in an important motif, of the protein or in a region associated with the activity of the protein, the probability that the change will affect the protein is high. The most widely use FPP in this category is PolyPhen-2. PolyPhen-2 is also a machine-learning FPP using a Bayesian classifier composed of eight sequence-based and three structure-based predictive features [147].

The FPPs grouped in category 3 are based on sequence and evolution conservation. The FPPs that use this method require multispecies sequence alignments, to calculate the divergence in a location. If the amino acid change occurred in a region that is highly conserved and the change is not observed in other species, the amino acid change is likely to affect the protein. Some of these FPPs use special matrices based on physicochemical properties to evaluate the changes. Others use Hidden Markov models to evaluate if the change is tolerated. The FPPs from this category that are more widely used are SIFT [137], MAPP [109] and PANTHER [103].

Category 8 (conservation and frequency) contains only one member Variant Annotation, Analysis and Search Tool 2 (VAAST2) [177]. VAAST2 employs a novel conservation-controlled AAS matrix (CASM), to incorporate information about phylogenetic conservation.

The new generation of FPPs has been developed using machine-learning algorithms (category 4). Learning algorithms include naïve Bayes classifiers, neural networks, support vector machines and random forests. Most often, the FPPs use a neural network or a support vector machine because these methods were designed to be trained with two data sets: for example, benign versus pathogenic variants. The FPPs learn to differentiate between both groups of variants. The most commonly used FPPs under this category are PMut [113], PhD-SNP [120], SNPs&GO [139] and MutationTaster [145].

Recently, several groups have begun developing methods to combine the scores of multiple FPPs into a single score (category 7). The Combined annotation scorIng toOL (CAROL) [169] combines the scores of two FPPs: PolyPhen-2 [147] and SIFT [137]. The Consensus deleteriousness score of missense mutations (Condel) [149] combines the scores of five FPPs: Logre [105], MAPP [109], Mutation assessor [157], PolyPhen- 2 [147] and SIFT [137]. The evaluation of tools that use a weighted average of the normalized scores from multiple FPPs indicates greater confidence levels in classifying missense mutations [196,197]. It is becoming a common practice to use this combinatorial approach.

In 2013, a group directed by Simpson evaluated seven predictive tools plus the two consensus tools, CAROL and Condel [182]. Their comparison showed that MutPred [135] had the highest sensitivity and the lowest number of false positives; PolyPhen-2 [147] was the second highest, and SNPs&GO [139] was the third best. The two combinatorial score programs CAROL [169] and Condel [149] performed very well but not as high as MutPred [135] by itself. Then Simpson's group developed their own Consensus Variant Effect Classification tools (CoVEC). CoVEC integrated the prediction results from four predictors SIFT [137], PolyPhen-2 [147], SNPs&GO [139] and Mutation assessor [157]. According to their evaluation of CoVEC, the tool performed almost as high as MutPred [135] and higher than CAROL [169] and Condel [149] and PolyPhen-2 [147].

The column labeled 'Access' in Table 4 pinpoints to several problems: many of the available FPPs are not released to users for running locally and the authors provide access through web servers. Unfortunately, many of the web servers are not consistent. Only one group provided web services application programming interfaces) to access their services.

Other groups provide simple batch processing, and some require that variants have been tested manually on their server, which is an impossible task when working with NGS where hundreds of missense mutations need to be evaluated. This problem is in part solved by databases with preprocessed variants like dbNSFP [180]. However, the major problem is the lack of standards between groups. Each group develops its own format and requires different input of the data. In addition, each group invents their own scoring system. In many cases, it is difficult to figure out what data sets were used to train their programs. An urgent call for standardization is required.

All the available FPPs are limited to evaluate the effect of single missense mutations. The effect of indels or multiple missense mutations in a single protein is beyond the scope of most, if not all, of the available programs. There is a lack of FPPs capable of evaluating the effect of variations in noncoding regulatory regions even when there is a plethora of annotations in the Encyclopedia of DNA elements (ENCODE) project. However, at the time of this writing, a new method was published, Genome Wide Annotation of Variants (GWAVA).

GWAVA uses a machine-learning algorithm (random forest) trained with annotations from ENCODE, GENCODE, and other sources to evaluate the effect of regulatory variants in noncoding portions of the genome. GWAVA uses a normalized score of 0–1 to report pathogenicity of variants. In addition, the group provides precomputed scores for all known noncoding variants that are available in Ensembl [190].

Very recently, the Combined Annotation-Dependent Depletion (CADD) framework was published [192]. CADD is based on the evolutionary principle that damaging mutations will be removed by natural selection from the gene pool. Shendure's group trained their support vector machines with two data sets. The first set was generated by the simulation of 14.7 million variants that reflect known mutational events. The second set of 14.7 million variants contains variants known to be fixed in the human genome. CADD framework incorporates the annotations from 63 different sources and generated a single metric score or C score. C score measures deleteriousness, a property that strongly correlates with both molecular functionality and pathogenicity. Shendure's group also precomputed and made available scores for all possible missense mutations that could occur at every position in the genome. In addition, CADD is capable to evaluate the effect of indels, but only a limited set of indels was precomputed at this time. The authors provided several examples between the correlation of C score with pathogenicity and tested CADD on several sets of known pathogenic variants. Their analysis shows that CADD outperform PolyPhen-2 [147] on distinguishing between pathogenic and benign variants. The precomputed data provide two types of scores: raw score, which goes from negative values to positive values (a negative value indicates that the variant is fixed in the population while a positive value indicates that the variant was simulated or rare), and a normalized Phred quality score scale. The advantage of using Phred scale, a ranking score, is that most of the people that work with sequence analysis are already familiar with Phred scale and the scores should be persistent between releases. For example, if a mutation ranks in the top 1% (CADD-20) of the whole set of mutations in the human genome and the program is updated the rank for the mutation

tested would be the same regardless of the absolute value of the raw score or the Phred value generated by the updated program [192].

Integrated software & commercial solutions to analyze your data

During the last few years, many institutions have been able to acquire NGS sequencers, but many of them lack the infrastructure and expertise to perform the bioinformatics analysis and the medical interpretation of the data. For a small laboratory that processes a small number of samples, annotating the variant call format (VCF) file and selecting a subset of variants to study is sufficient. There are several software packages, listed in Table 5, that annotate an entire VCF file (under type ‘VCF annotator’).

For a large laboratory that tries to analyze hundred or thousand of samples, the manual process is not a viable solution. A large laboratory wants to analyze every sample consistently and automatically. There are many bioinformatics steps between the raw data and the final report (Figure 2). For such laboratories the installation of a workflow management system is essential. In Table 5, there is a list of several workflow management systems, some of them free and others commercially available. Alternatively, there are many companies dedicated to providing a solution to analyze your data (Table 5). Several companies offer one-step solution like Genomatix and Knome. Others offer only the software and a third group offers to do the bioinformatics analysis and return the results.

Use of NGS to diagnose human disorders

One of the major concerns of medical diagnosis is to identify genes and mutations responsible for human disorders. Early identification of causative mutations enables the early detection of a myriad of disorders. We are living in an age of high healthcare cost. Early detection of genetic disorders, carrier status, genetic predispositions for cancer and cardiovascular disease could potentially reduce the healthcare cost.

The first proof of concept that the NGS technology could be used to detect genetic disorders was provided by Shendure’s group on September 2009 [225]. A few months later, the same group reported the detection of the first recessive disorder (Miller syndrome) detected by whole-exome sequencing (WES) [226]. These two papers marked a new era where NGS became the preferred tool for rare Mendelian disease gene identification. There are several excellent reviews that describe the exponential growth in disease gene identification that started in 2010 [227–229]. Up to 27 February 2014, the number of genes with phenotype-causing mutations has reached 3162 according to online Mendelian inheritance in man (OMIM) Mgene map statistics [230]. In a recent review, Rabbani *et al.* estimated that from January 2010 to May 2012, over 100 causative genes in various Mendelian disorders have been identified by means of exome sequencing [231].

WES is now a valid and standard diagnostic approach for the identification of molecular defects in patients with suspected genetic disorders. This fact was demonstrated last year by a publication in the *New England Journal of Medicine* by the Medical Genetics Laboratory group of Baylor College of Medicine. The group reported the WES sequencing of 250 probands referred by physician, 98% of the cases were billed to the insurance. They reported

a 25% molecular diagnostic rate (62 cases) [232]. In September 2013, the NIH funded four groups to explore the use of NGS for newborn screening [233]. With the cost per genome getting close to the US\$1000, it is becoming affordable to get sequenced at an early age, allowing for reanalysis of our genetic information at multiple intervals during the life of a person (Figure 3). A recent review outlines the approach, challenges, and benefits of such screening for adult genetic disease risks [2]. We also recently published a proof of concept project aimed to evaluate the benefits of screening healthy adults using WES. Our pilot project demonstrated that when WES is combined with medical and family history the findings are substantial. In a cohort of 81 unrelated individuals, we identified 271 recessive risk alleles (214 genes), 126 dominant risk alleles (101 genes) and three X-recessive risk alleles (three genes). In addition, we linked personal disease histories with causative disease genes in 18 volunteers [234].

Conclusion

The development of NGS was a monumental achievement that involved thousands of individuals from multiple professions and with a myriad of motivations, but with a common goal: to understand what make us unique. Definitively, the major milestone required for reaching our goal was to sequence the first human genome this was accomplished under the Human Genome Project (HGP). Reaching the first milestone took 13 years with a cost of US \$3 billion; however, we should not forget the overlapping project to annotate the human genome. Annotating the human genome was essential to understand and apply our newly acquired knowledge to improve human health. Before the end of the project, it became obvious that sequencing an individual genome was only the beginning of a long road to provide cures and prevention for genetic diseases. Two independent projects born after the completion of the HGP, one directed to understand the variability in the human population (HapMap Project) and a second project undertaken by commercial enterprises was able to develop the most economical massive parallel sequencing technology every seen. The success of both projects together with the growing catalog of human disorders merged to form what we now know now as clinical and medical genetics. Multiple commercial enterprises have been very successful in developing fast and affordable technology. We can now sequence the entire genome of an individual for approximately US\$1000 in less than two weeks (summarized in Figure 1). With such overwhelming success to generate large amounts of short reads several groups of developers were motivated to generate efficient tools to align and detect variants. Currently, we have excellent short-read sequencing mapping tools (SRSMT) and very accurate variant callers (Table 1). The process of interpreting an individual genome starts by separating the variations that are common in the population from the unique mutations, to complete this task resources developed by population geneticist are essential (Table 2). Only 5 years ago (from the publication of this review) the first proof of concept that NGS could be used to detect human disorders was provided by Shendure's group. Since that time an expansion in the number of pathogenic genes has surpassed the 3000 mark. Human catalogs of disease-causing mutations are also expanding very fast (Table 3) but since there are an extraordinary large number of potential damaging mutations in man, our repertoire of techniques to predict damaging mutations should become a priority. Currently, the number of functional prediction programs (FPPs)

capable of detecting pathogenic variants is over 40 (Table 4). However, there is a variable degree of accuracy and agreement between them, also the lack of standards; maintenance and form of distribution make it our biggest liability for the acceptance of personalized medicine. We have come a long way from 2007; we have now a large number of commercial and free workflows capable of analyzing the enormous amount of information from NGS sequencers (Table 5 & Figure 2). I feel confident that future generations will have a much more bright and healthy life with the incorporation of NGS into medicine. Figure 3 shows how the use of NGS in combination with additional information from the patient, at different stages of life, will improve early treatments and real on time personalized medicine.

Future perspective

Despite its early age, NGS has successfully extended our knowledge about disease phenotype–genotype relationships and disease gene discovery. The number of genetic disorders with a corresponding causative gene is growing very fast and will continue to grow exponentially during the next few years. The NGS technology has been adopted for clinical diagnosis of suspected genetic disorders with a 25% success rate [232]. The success rate will increase with the development of new sequencing technologies and better analytical tools. NGS is now moving to the area of carrier testing, newborn screening and prenatal screening. We expect that during the next few year NGS will become a part of the standard set of newborn screening tests.

Currently, many laboratories offer NGS panels for patients with different types of cardiomyopathies that could have a genetic cause and for patients with family histories of hereditary cancers. Some laboratories offer services for the detection of variants that could improve the treatment of cancer patients such as pharmacogenomics panels. Some groups like the Mayo Clinic (MN, USA) [235], Foundation Medicine (MA, USA) [236], Genekey (CA, USA) [237] and Molecular Health (TX, USA) [238] offer genetic tests and work with oncologists to improve the treatment of their patients and provide state-of-the-art technologies to personalize cancer treatments. Some of their analyses include molecular profiling, gene expression profiling, the identification of genetic rearrangements in tumor samples, the detection of circulating tumor cells and the detection of somatic mutations in tumor samples. During the next few years, we expect there to be an exponential increase in the number of organizations that not only offer NGS tests but also professional guidance to oncologists for the personalized treatment of cancer patients. The role of these professional counselors will extend from cancer to other genetic disorders, personalizing many medical treatments.

At the moment, screening healthy adults for genetic risks is a controversial issue. However, as patients become more aware of the benefits of using NGS for early detection of adult-onset disorders there will be an increase in the number of requests for NGS analyses, especially from healthy adults that are looking for new approaches to prevent disorders. Eventually, NGS will become part of the routine yearly physical examinations, or it may become a medical specialty on its own [234].

New technologies such as the GridION System (Oxford Nanopore technologies [Oxford, UK]), single-cell sequencing (Fluidigm), positional sequencing (Nabsys) and long fragment read (CGI) will provide cheaper, faster and more accurate sequencing data. The use of supercomputers, in conjunction with parallelization, will accelerate the analysis of genomic data. The increasing number of catalogs of causative and risk genes will provide a foundation for PM and pharmacogenomics. The use of NGS technology for patients in critical care units will become possible with the presence of three elements: high-quality whole-genome sequences delivered at a very fast rate; fast analysis time; and large catalogs of DCM and pharmacogenomics markers. Predicting the functional effects of a mutation is a complex area in need of standardization, but of crucial importance for the identification of variants with high impact. New developments in this area such as GWAVA and CADD are helping to provide light at the end of a dark tunnel.

References

Papers of special note have been highlighted as:

- of interest;
- of considerable interest

1. Bloom, DE.; Cafiero, ET.; Jané-Llopis, E., et al. The global economic burden of noncommunicable diseases. Geneva: World Economic Forum; www3.weforum.org/docs/WEF_Harvard_HE_GlobalEconomicBurdenNonCommunicableDiseases_2011.pdf
2. Caskey CT, Gonzalez-Garay ML, Pereira S, McGuire AL. Adult genetic risk screening. *Annu Rev Med.* 2014; 65:1–17. [PubMed: 24188662]
3. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA.* 1977; 74(12):5463–5467. [PubMed: 271968]
4. Wetterstrand, KA. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). www.genome.gov/sequencingcosts
5. NHGRI: all about the Human Genome Project (HGP). www.genome.gov/10001772
6. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001; 409(6822):860–921. [PubMed: 11237011]
7. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science.* 2001; 291(5507):1304–1351. [PubMed: 11181995]
8. Stein LD. Human genome: end of the beginning. *Nature.* 2004; 431(7011):915–916. [PubMed: 15496902]
- 9••. IHGSC. Finishing the euchromatic sequence of the human genome. *Nature.* 2004; 431(7011):931–945. Authored by the members of the International Human Genome Sequencing Consortium (IHGSC). It describes the finishing of the human genome, marking the last milestone in an historical project. This article reports how the gaps were filled up in both genome drafts, one generated by Celera and other by IHGSC. Both drafts were missing 10% of euchromatin and 30% of the genome. [PubMed: 15496913]
10. Nagaraj SH, Gasser RB, Ranganathan S. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinform.* 2007; 8(1):6–21. [PubMed: 16772268]
11. Adams MD, Kelley JM, Gocayne JD, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science.* 1991; 252(5013):1651–1656. [PubMed: 2047873]
12. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* 1997; 268(1):78–94. [PubMed: 9149143]
13. Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in Genie. *J Comput Biol.* 1997; 4(3):311–323. [PubMed: 9278062]

14. Softberry. Commercial developer of Gene Prediction Programs FGENES. www.softberry.com/berry.phtml?topic=products&no_menu=on
15. The International Hapmap Consortium. The International HapMap Project. *Nature*. 2003; 426(6968):789–796. [PubMed: 14685227]
16. The International Hapmap Consortium. A haplotype map of the human genome. *Nature*. 2005; 437(7063):1299–1320. [PubMed: 16255080]
17. The International Hapmap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007; 449(7164):851–861. [PubMed: 17943122]
18. The International Hapmap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010; 467(7311):52–58. [PubMed: 20811451]
19. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest*. 2008; 118(5):1590–1605. [PubMed: 18451988]
20. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2014; 42(Database issue):D7–D17. [PubMed: 24259429]
21. Levy S, Sutton G, Ng PC, et al. The diploid genome sequence of an individual human. *PLoS Biol*. 2007; 5(10):e254. [PubMed: 17803354]
22. Shendure J, Porreca GJ, Reppas NB, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*. 2005; 309(5741):1728–1732. [PubMed: 16081699]
23. The Polonator G007. www.polonator.org
24. Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. *Science*. 1998; 281(5375):363–365. [PubMed: 9705713]
25. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005; 437(7057):376–380. [PubMed: 16056220]
26. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*. 2008; 452(7189):872–876. [PubMed: 18421352]
27. Wadman M. James Watson's genome sequenced at high speed. *Nature*. 2008; 452(7189):788. [PubMed: 18431822]
28. Valouev A, Ichikawa J, Tonthat T, et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res*. 2008; 18(7):1051–1063. [PubMed: 18477713]
29. Lupski JR, Reid JG, Gonzaga-Jauregui C, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med*. 2010; 362(13):1181–1191. [PubMed: 20220177]
30. Birney E, Stamatoyannopoulos JA, Dutta A, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. 2007; 447(7146):799–816. [PubMed: 17571346]
31. Davies, K. The Solexa Story. www.bio-itworld.com/BioIT_Content.aspx?id=101666
32. Drmanac R, Sparks AB, Callow MJ, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*. 2010; 327(5961):78–81. [PubMed: 19892942]
33. CGI. CGI Documentation. www.completegenomics.com/customer-support/documentation
34. Rothberg JM, Hinz W, Rearick TM, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 2011; 475(7356):348–352. [PubMed: 21776081]
35. Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009; 323(5910):133–138. [PubMed: 19023044]
36. English AC, Richards S, Han Y, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE*. 2012; 7(11):e47768. [PubMed: 23185243]
37. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? *J Med Genet*. 2011; 48(9):580–589. [PubMed: 21730106]
38. Illumina exomes comparative table. http://res.illumina.com/documents/products/datasheets/datasheet_illumina_exomes_comparative_table.pdf
39. NimbleGen. SeqCap EZ Human Exome Library v3.0. www.nimblegen.com/products/seqcap/ez/v3/index.html
40. Agilent Technologies. SureSelect DNA Panels. www.genomics.agilent.com/en/SureSelect-DNA-RNA/SureSelect-Human-All-Exon-Kits/?cid=AG-PT-177&tabId=AG-PR-120

41. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 2008; 18(11):1851–1858. [PubMed: 18714091]
42. Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. *Bioinformatics.* 2012; 28(24):3169–3177. [PubMed: 23060614]
43. Smith AD, Chung WY, Hodges E, et al. Updates to the RMAP short-read mapping software. *Bioinformatics.* 2009; 25(21):2841–2842. [PubMed: 19736251]
44. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012; 9(4):357–359. [PubMed: 22388286]
45. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics.* 2010; 26(5):589–595. [PubMed: 20080505]
46. Li R, Yu C, Li Y, et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics.* 2009; 25(15):1966–1967. [PubMed: 19497933]
47. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics.* 2010; 26(7):873–881. [PubMed: 20147302]
48. Novocraft. Novoalign. www.novocraft.com/main/index.php
49. Xin H, Lee D, Hormozdiari F, Yedkar S, Mutlu O, Alkan C. Accelerating read mapping with FastHASH. *BMC Genomics.* 2013; 14(Suppl 1):S13. [PubMed: 23369189]
50. Hach F, Hormozdiari F, Alkan C, Birol I, Eichler EE, Sahinalp SC. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods.* 2010; 7(8):576–577. [PubMed: 20676076]
51. Hatem A, Bozdog D, Toland AE, Catalyurek UV. Benchmarking short sequence mapping tools. *BMC Bioinformatics.* 2013; 14:184. [PubMed: 23758764]
52. Picard. Picard Tools. <http://picard.sourceforge.net>
53. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20(9):1297–1303. Description of Broad's Genome Analysis Toolkit (GATK). Detailed explanation of the analysis performed by the toolkit and requirements and capabilities. In addition, the authors explain details about the software that are essential to understand how the software works, for example, the transversals and the walkers. [PubMed: 20644199]
54. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet.* 2011; 12(6):443–451. [PubMed: 21587300]
55. Yu X, Sun S. Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics.* 2013; 14:274. [PubMed: 24044377]
56. Li Y, Chen W, Liu EY, Zhou YH. Single nucleotide polymorphism (SNP) detection and genotype calling from massively parallel sequencing (MPS) data. *Stat Biosci.* 2013; 5(1):3–25. [PubMed: 24489615]
57. Broad Institute. The Genome Analysis Toolkit (GATK). www.broadinstitute.org/gatk
58. GitHub. Freebayes, a haplotype-based variant detector. <https://github.com/ekg/freebayes>
59. Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. <http://arxiv.org/abs/1207.3907>
60. Baylor College of Medicine. Human Genome Center. Atlas 2. www.hgsc.bcm.edu/software/atlas2
61. Challis D, Yu J, Evani US, et al. An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics.* 2012; 13:8. [PubMed: 22239737]
62. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. <https://cgwb.nci.nih.gov/goldenPath/bamview/documentation/index.html>
63. Edmonson MN, Zhang J, Yan C, Finney RP, Meerzaman DM, Buetow KH. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. *Bioinformatics.* 2011; 27(6):865–866. [PubMed: 21278191]
64. SAMtools. <http://samtools.sourceforge.net>
65. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009; 25(16):2078–2079. [PubMed: 19505943]
66. SNVer. Rare and common variants detection in next generation sequencing. <http://snver.sourceforge.net>

67. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.* 2011; 39(19):e132. [PubMed: 21813454]
68. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009; 461(7265):747–753. [PubMed: 19812666]
69. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet.* 2007; 80(4):727–739. [PubMed: 17357078]
70. HapMap Homepage. <http://hapmap.ncbi.nlm.nih.gov/downloads/index.html.en>
71. 1000 Genomes. A deep catalog of human genetic variation. www.1000genomes.org
- 72•. Abecasis GR, Altshuler D, Auton A, et al. A map of human genome variation from population-scale sequencing. *Nature.* 2010; 467(7319):1061–1073. Latest paper from the 1000 Genomes project describing the sequencing of 1092 human genomes and the number of variations found and the methods used to identify the mutations and combine variants from different sequencing sources. [PubMed: 20981092]
73. Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012; 491(7422):56–65. [PubMed: 23128226]
74. NHLBI. Exome Sequencing Project (ESP). Exome Variant Server. <http://evs.gs.washington.edu/EVS>
75. Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet.* 2012; 91(2):224–237. [PubMed: 22863193]
76. Nhlbi_Esp: Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP). <http://evs.gs.washington.edu/EVS>
77. Personal Genome Project. www.personalgenomes.org
78. Ball MP, Thakuria JV, Zaranek AW, et al. A public resource facilitating clinical use of genomes. *Proc Natl Acad Sci USA.* 2012; 109(30):11920–11927. [PubMed: 22797899]
79. NextCode Health. www.nextcode.com
80. Sheridan C. Amgen punts on deCODE's genetics know-how. *Nat Biotechnol.* 2013; 31(2):87–88. [PubMed: 23392483]
81. DNAnexus. CHARGE project use case. <https://dnanexus.com/usecases-charge>
82. Reid JG, Carroll A, Veeraraghavan N, et al. Launching genomics into the cloud: deployment of mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics.* 2014; 15:30. [PubMed: 24475911]
83. Brownstein CA, Beggs AH, Homer N, et al. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. *Genome Biol.* 2014; 15(3):R53. [PubMed: 24667040]
84. Genome in a Bottle Consortium. www.genomeinabottle.org
- 85•. Zook JM, Chapman B, Wang J, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol.* 2014; 32(3):246–251. Describes the sample selected as the standard NA12878, the sequence information generated for the sample using multiple sequencing platforms, the mapping programs and callers used and how to use the resources to test your own tools. [PubMed: 24531798]
86. Collins FS, Hamburg MA. First FDA authorization for next-generation sequencer. *N Engl J Med.* 2013; 369(25):2369–2371. [PubMed: 24251383]
87. HGMD Human gene mutation database (HGMD® Professional) from BIOBASE Corporation. www.biobase-international.com/hgmd
- 88•. 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; 133(1):1–9. Describes the Human Gene Mutation Database, HGMD. A database of germline mutations that have been previously reported in the scientific literature as associated and in many cases responsible for a genetic disorders. [PubMed: 24077912]

89. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014; 42(Database issue):D980–D985. [PubMed: 24234437]
90. Qiagen® BioBase Biological databases. HGMD®. Human Gene Mutation Database. www.biobase-international.com/product/hgmd
91. NCBI. ClinVar aggregates information about sequence variation and its relationship to human health. www.ncbi.nlm.nih.gov/clinvar
92. Human Genome Variation Society (HGVS). Locus specific mutation databases. www.hgvs.org/dblist/glsdb.htm
93. Hgv: Human Genome Variation Society (HGV). www.hgvs.org/dblist/dblist.html. URL
94. Locus Specific Mutation Databases. http://grenada.lumc.nl/LSDB_list/lsdbs
95. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, Den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat.* 2011; 32(5):557–563. [PubMed: 21520333]
96. Catalogue of somatic mutations in cancer (COSMIC). <http://cancer.sanger.ac.uk/cancergenome/projects/cosmic>
97. Forbes SA, Tang G, Bindal N, et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res.* 2010; 38(Database issue):D652–D657. [PubMed: 19906727]
98. Diagnostic Mutation Database (DMuDB). <https://secure.dmu-db.net/ngml-rep/Home.do>
99. MITOMAP. A human mitochondrial genome database. www.mitomap.org/MITOMAP
100. Ruiz-Pesini E, Lott MT, Procaccio V, et al. An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res.* 2007; 35(Database issue):D823–D828. [PubMed: 17178747]
101. PhenCode: paving the path between phenotype and genome. <http://globin.bx.psu.edu/phencode>
102. Giardine B, Riemer C, Hefferon T, et al. PhenCode: connecting ENCODE data with mutations and phenotype. *Hum Mutat.* 2007; 28(6):554–562. [PubMed: 17326095]
103. Thomas PD, Campbell MJ, Kejariwal A, et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 2003; 13(9):2129–2141. [PubMed: 12952881]
104. PANTHER. Classification System. www.pantherdb.org/tools/csnpscoreForm.jsp
105. Clifford RJ, Edmonson MN, Nguyen C, Buetow KH. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics.* 2004; 20(7):1006–1014. [PubMed: 14751981]
106. Logre. <http://lpgws.nci.nih.gov/cgi-bin/GeneViewer.cgi>
107. Stitzel NO, Binkowski TA, Tseng YY, Kasif S, Liang J. topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res.* 2004; 32(Database issue):D520–D522. [PubMed: 14681472]
108. topoSNP database. <http://gila.bioengr.uic.edu/snp/toposnp>
109. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* 2005; 15(7):978–986. [PubMed: 15965030]
110. Multivariate Analysis of Protein Polymorphism: MAPP. <http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html>
111. Bao L, Zhou M, Cui Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.* 2005; 33(Web Server issue):W480–W482. [PubMed: 15980516]
112. nsSNPAnalyzer: predicting disease-associated nonsynonymous single nucleotide polymorphisms. <http://snpanalyzer.uthsc.edu>
113. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, De La Cruz X, Orozco M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics.* 2005; 21(14):3176–3178. [PubMed: 15879453]
114. Karchin R, Diekhans M, Kelly L, et al. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics.* 2005; 21(12):2814–2820. [PubMed: 15827081]

115. Query LS-SNP for SNP annotations. <http://modbase.compbio.ucsf.edu/LS-SNP/Queries.html>
116. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. *Nucleic Acids Res.* 2005; 33(Web Server issue):W382–W388. [PubMed: 15980494]
117. A force field for energy calculations and protein design (FoldX). <http://foldx.crg.es>
118. Tavtigian SV, Deffenbaugh AM, Yin L, et al. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet.* 2006; 43(4):295–305. [PubMed: 16014699]
119. International Agency for Research on Cancer. Align-GVGD. http://agvgd.iarc.fr/agvgd_input.php
120. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics.* 2006; 22(22):2729–2734. [PubMed: 16895930]
121. PhD-SNP. Predictor of human deleterious single nucleotide polymorphisms. <http://snps.biofold.org/phd-snp/phd-snp.html>
122. Yuan HY, Chiou JJ, Tseng WH, et al. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res.* 2006; 34(Web Server issue):W635–W641. [PubMed: 16845089]
123. FASTSNP. http://fastsnp.ibms.sinica.edu.tw/pages/input_CandidateGeneSearch.jsp
124. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins.* 2006; 62(4):1125–1132. [PubMed: 16372356]
125. MUpro: prediction of protein stability changes for single-site mutations from sequences. www.ics.uci.edu/~baldig/mutation.html
126. Yue P, Melamud E, Moulton J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics.* 2006; 7:166. [PubMed: 16551372]
127. snps3D. www.snps3d.org
128. Kaminker JS, Zhang Y, Waugh A, et al. Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res.* 2007; 67(2):465–473. [PubMed: 17234753]
129. Tian J, Wu N, Guo X, Guo J, Zhang J, Fan Y. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. *BMC Bioinformatics.* 2007; 8:450. [PubMed: 18005451]
130. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* 2007; 35(11):3823–3835. [PubMed: 17526529]
131. SNAP SERVICE. www.rostlab.org/services/SNAP/submit
132. Cheng TM, Lu YE, Vendruscolo M, Lio P, Blundell TL. Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS Comput Biol.* 2008; 4(7):e1000135. [PubMed: 18654622]
133. Kristensen DM, Ward RM, Lisewski AM, et al. Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics.* 2008; 9:17. [PubMed: 18190718]
134. The Evolutionary Trace Server. <http://mammoth.bcm.tmc.edu/ETserver.html>
135. Li B, Krishnan VG, Mort ME, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics.* 2009; 25(21):2744–2750. [PubMed: 19734154]
136. MutPred. <http://mutpred.mutdb.org>
137. 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):1073–1081. [PubMed: 19561590]
138. J. Craig Venter Institute. SIFT. <http://sift.jcvi.org>
139. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat.* 2009; 30(8):1237–1244. [PubMed: 19514061]
140. SNPs&GO. <http://snps.biofold.org/snps-and-go/snps-and-go.html>

141. Wainreb G, Ashkenazy H, Bromberg Y, et al. MuD: an interactive web server for the prediction of non-neutral substitutions using protein structural data. *Nucleic Acids Res.* 2010; 38(Web Server issue):W523–W528. [PubMed: 20542913]
142. MuD. Mutation Detector. <http://mud.tau.ac.il>
143. Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics.* 2010; 11:548. [PubMed: 21059217]
144. NBIC. Project HOPE. www.cmbi.ru.nl/hope/input;jsessionid=8dd3352af2158fd6b4a526fae212?0
145. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods.* 2010; 7(8):575–576. [PubMed: 20676075]
146. Mutation taster. www.mutationtaster.org
147. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010; 7(4):248–249. [PubMed: 20354512]
148. PolyPhen-2 prediction of functional effects of human nsSNPs. <http://genetics.bwh.harvard.edu/pph2>
149. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet.* 2011; 88(4):440–449. [PubMed: 21457909]
150. Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med.* 2012; 4(11):89. [PubMed: 23181723]
151. CONsensus DELeteriousness score of missense SNVs (Condel). <http://bg.upf.edu/condel/home>
152. TRANSformed Functional Impact for Cancer (TransFIC). <http://bg.upf.edu/fannsdB>
153. Worth CL, Preissner R, Blundell TL. SDM – a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res.* 2011; 39(Web Server issue):W215–W222. [PubMed: 21593128]
154. SDM. <http://mordred.bioc.cam.ac.uk/~sdm/sdm.php>
155. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics.* 2009; 25(19):2537–2543. [PubMed: 19654118]
156. Prediction of Protein Mutant Stability Changes (PopMusic). <http://babylone.ulb.ac.be/popmusic>
157. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 2011; 39(17):e118. [PubMed: 21727090]
158. Functional impact of protein mutations. <http://mutationassessor.org/v1>
159. Olatubosun A, Valiaho J, Harkonen J, Thusberg J, Vihinen M. PON-P: integrated predictor for pathogenicity of missense variants. *Hum Mutat.* 2012; 33(8):1166–1174. [PubMed: 22505138]
160. PON-P2. <http://structure.bmc.lu.se/PON-P2>
161. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE.* 2012; 7(10):e46688. [PubMed: 23056405]
162. J. Craig Venter Institute. Protein Variation Effect Analyzer (PROVEAN). <http://provean.jcvi.org/index.php>
163. Luu TD, Rusu A, Walter V, et al. KD4v: comprehensible knowledge discovery system for missense variant. *Nucleic Acids Res.* 2012; 40(Web Server issue):W71–W75. [PubMed: 22641855]
164. KD4v: comprehensible knowledge discovery system for missense variants. <http://decryphon.igbmc.fr/kd4v/cgi-bin/prediction>
165. Schaefer C, Meier A, Rost B, Bromberg Y. SNPdbe: constructing an nsSNP functional impacts database. *Bioinformatics.* 2012; 28(4):601–602. [PubMed: 22210871]
166. nsSNP database of functional effects (SNPdbe). www.rostlab.org/services/snpdbe
167. Sasidharan Nair P, Vihinen M. VariBench: a benchmark database for variations. *Hum Mutat.* 2013; 34(1):42–49. [PubMed: 22903802]
168. A benchmark database for variations (VariBench). <http://structure.bmc.lu.se/VariBench>

169. Lopes MC, Joyce C, Ritchie GR, et al. A combined functional annotation score for non-synonymous variants. *Hum Hered.* 2012; 73(1):47–51. [PubMed: 22261837]
170. Wellcome Trust Sanger Institute. Combined Annotation scoRing toOL (CAROL). www.sanger.ac.uk/resources/software/carol
171. Acharya V, Nagarajaram HA. Hansa: an automated method for discriminating disease and neutral human nsSNPs. *Hum Mutat.* 2012; 33(2):332–337. [PubMed: 22045683]
172. HANSA. <http://hansa.cdfd.org.in:8080>
173. De Baets G, Van Durme J, Reumers J, et al. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res.* 2012; 40(Database issue):D935–D939. [PubMed: 22075996]
174. SNPeffect4. <http://snpeffect.switchlab.org>
175. Capriotti E, Altman RB, Bromberg Y. Collective judgment predicts disease-associated single nucleotide variants. *BMC Genomics.* 2013; 14(Suppl 3):S2. [PubMed: 23819846]
176. Meta-SNP. <http://snps.biofold.org/meta-snp/pages/methods.html>
177. Hu H, Huff CD, Moore B, Flygare S, Reese MG, Yandell M. VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet Epidemiol.* 2013; 37(6):622–634. [PubMed: 23836555]
178. Variant Annotation, Analysis and Search Tool – VAAST 2. www.yandell-lab.org/software/vaast.html
179. Li MX, Kwan JS, Bao SY, et al. Predicting Mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet.* 2013; 9(1):e1003143. [PubMed: 23341771]
180. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat.* 2013; 34(9):E2393–E2402. [PubMed: 23843252]
181. dbNSFP. <https://sites.google.com/site/jpopgen/dbNSFP>
182. Frousios K, Iliopoulos CS, Schlitt T, Simpson MA. Predicting the functional consequences of non-synonymous DNA sequence variants – evaluation of bioinformatics tools and development of a consensus strategy. *Genomics.* 2013; 102(4):223–228. [PubMed: 23831115]
183. Variant Effect Prediction. CoVEC. www.dcs.kcl.ac.uk/pg/frousiok/variants/index.html
184. Bendl J, Stourac J, Salanda O, et al. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput Biol.* 2014; 10(1):e1003440. [PubMed: 24453961]
185. PredictSNP. Consensus classifier for prediction of disease-related mutations. <http://loschmidt.chemi.muni.cz/predictsnp/>
186. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics.* 2014; 30(3):335–342. [PubMed: 24281696]
187. mCSM. Protein stability change upon mutation. <http://bleoberis.bioc.cam.ac.uk/mcsm/stability>
188. Liu M, Watson LT, Zhang L. Quantitative prediction of the effect of genetic variation using hidden Markov models. *BMC Bioinformatics.* 2014; 15:5. [PubMed: 24405700]
189. Quantitative prediction of the effect of genetic variation using hidden Markov models. <https://bioinformatics.cs.vt.edu/zhanglab/hmm>
190. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. *Nat Methods.* 2014; 11(3):294–296. [PubMed: 24487584]
191. Wellcome Trust Sanger Institute. Genome Wide Annotation of VARIants (GWAVA). www.sanger.ac.uk/sanger/StatGen_Gwava
192. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014; 46(3):310–315. Describes a new method – combined annotation-dependent depletion. This new method distinguishes between benign variants and variants that could affect the functionality of a protein. [PubMed: 24487276]
193. Combined Annotation Dependent Depletion (CADD). <http://cadd.gs.washington.edu/home>

194. Pirolli D, Carelli Alinovi C, Capoluongo E, et al. Insight into a novel p53 single point mutation (G389E) by molecular dynamics simulations. *Int J Mol Sci.* 2010; 12(1):128–140. [PubMed: 21339981]
195. Friedman R, Boye K, Flatmark K. Molecular modelling and simulations in cancer research. *Biochim Biophys Acta.* 2013; 1836(1):1–14. [PubMed: 23416097]
196. Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB. *In silico* analysis of missense substitutions using sequence-alignment based methods. *Hum Mutat.* 2008; 29(11):1327–1336. [PubMed: 18951440]
197. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat.* 2011; 32(4):358–368. [PubMed: 21412949]
198. Cassandra. www.hgsc.bcm.edu/software/cassandra
199. AnnTools. <http://anntools.sourceforge.net>
200. Variant Effect Predictor. www.ensembl.org/info/docs/tools/vep/index.html
201. SnpEff. Genetic variant annotation and effect prediction toolbox. <http://snpeff.sourceforge.net>
202. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. www.openbioinformatics.org/annovar
203. Home of variant tools. <http://varianttools.sourceforge.net/Main/HomePage>
204. Galaxy. Data intensive biology for everyone. <http://galaxyproject.org>
205. Mercury. www.hgsc.bcm.edu/software/mercury
206. GitHub. BauerLab/ngsane. <https://github.com/BauerLab/ngsane/wiki>
207. Seven Bridges. www.sbgenomics.com
208. Chipster. Open Source platform for data analysis. <http://chipster.csc.fi>
209. Anduril. www.anduril.org/anduril/site
210. Genomatix. www.genomatix.de
211. CLCbio. www.clcbio.com
212. Knome. The Human Genome Interpretation Company. www.knome.com
213. SoftGenetics. www.softgenetics.com
214. DNASTAR. www.dnastar.com
215. Partek. www.partek.com
216. Complete Genomics, a BGI company. www.completegenomics.com
217. Personalis. Pioneering genome guided medicine. www.personalis.com
218. Omicia. www.omicia.com
219. Invitae. www.invitae.com/en
220. Genformatic. www.genformatic.com/index.html
221. bina. www.binatechnologies.com
222. RealTime Genomics. <http://realtimegenomics.com>
223. DNAnexus. www.dnanexus.com
224. Ingenuity. www.ingenuity.com
- 225••. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature.* 2009; 461(7261):272–276. Describes the first proof of concept that exome sequencing could be able to detect variants associated or responsible for Mendelian disorders. [PubMed: 19684571]
- 226••. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet.* 2010; 42(1):30–35. Describes the detection of the first recessive disorder detected by whole-exome sequencing (Miller syndrome). [PubMed: 19915526]
227. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Unlocking Mendelian disease using exome sequencing. *Genome Biol.* 2011; 12(9):228. [PubMed: 21920049]
228. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet.* 2012; 20(5):490–497. [PubMed: 22258526]
229. Ku CS, Naidoo N, Pawitan Y. Revisiting Mendelian disorders through exome sequencing. *Hum Genet.* 2011; 129(4):351–370. [PubMed: 21331778]

230. OMIM Gene Map Statistics. www.omim.org/statistics/geneMap
231. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *J Hum Genet.* 2012; 57(10):621–632. [PubMed: 22832387]
232. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med.* 2013; 369(16):1502–1511. [PubMed: 24088041]
233. NIH program explores the use of genomic sequencing in newborn healthcare. www.nih.gov/news/health/sep2013/nhgri-04.htm
234. Gonzalez-Garay ML, Mcguire AL, Pereira S, Caskey CT. Personalized genomic disease risk of volunteers. *Proc Natl Acad Sci USA.* 2013; 110(42):16957–16962. [PubMed: 24082139]
235. Mayo Clinic. Center for individualized medicine. <http://mayoresearch.mayo.edu/center-for-individualized-medicine/medical-genome-facility.asp>
236. Foundation Medicine. Foundation One tests. <http://foundationone.com>
237. GeneKey. Unlocking new treatment approaches for your cancer. www.genekey.com/our-process
238. Molecular Health. Step-by-step process to better treatment decisions. www.molecularhealth.com/oncologists/order-treatment-decision-support

Executive summary

Moving from traditional medicine to personalized medicine

- With an overburdened and overwhelmed healthcare system new alternative strategies are required to reduce the cost and improve the well-being of the patients.
- Personalized medicine is a medical model that proposes the customization of healthcare by using biological markers and pharmacogenomics to direct the customized treatment of patients.
- A new technology, next-generation sequencing (NGS), has the potential to make personalized medicine a reality by accelerating the early detection of disorders and the identification of pharmacogenetics markers to customize treatments.

Brief history of NGS

- The Human Genome Project lasted 13 years with a cost of US\$3 billion and the involvement of thousands of international scientists.
- The Human Genome Project provided the first draft of the human genome assemblies in 2001.
- During the Human Genome Project the cost of sequencing was reduced dramatically with the development of better chemistry, the involvement of robotics and automation.
- Bioinformatics and functional genomics flourished during this period, resulting in a myriad of biological annotations for the human genome.
- The engagement of visionaries and entrepreneurs in the development of novel sequencing technologies bootstrapped the birth of NGS technology.

The goal of having an affordable diploid genome of a single person

- The first diploid human genome of Dr Craig Venter (MD, USA) was published in 2007 with a cost of US\$100 million.
- In 2008, 454 technologies enabled the sequencing of the second human genome at a cost of US\$1,500,000.
- In 2010, SOLiD™ technology reduced the cost of a genome to US\$100,000.
- The developments of targeted sequencing of all human exons lowered the price of sequencing to few thousand dollars.
- By 2012, a furious competition between Complete Genomics (CA, USA) and Illumina® (CA, USA) reduced the cost of a genome to US\$3000.

The use of NGS to diagnose human disorders

- The streamlining and the standardization of the sequencing analysis allowed detecting variations in a single individual.

- The comparison of variants from an individual against those found in populations allows the identification of rare variants.
- The evaluation of rare variants, using functional prediction programs, had identified a small subset of variants that could explain pathology.
- The demonstration that NGS analysis could be used to detect genetic disorders was provided by Shendure's laboratory (WA, USA) in September 2009.
- Since 2010, NGS has identified hundreds of causative genes in various Mendelian disorders.

Future perspective

- The identification of causative genes will continue to increase exponentially.
- The involvement of NGS on generating personalized pharmacogenomics profiles will increase and move to standard medical practice.
- NGS will become part of the standard set of newborn screening tests and ethicists; politicians and geneticists will debate for years to come about the value and risks of creating national databases for all newborn babies.
- The role of NGS in prenatal screening will increase along with the debates between pro-life and pro-choice groups on whether or not we should use NGS for prenatal screening.
- NGS will become part of the standard repertoire of techniques to guide the treatment of cancer patients.
- Patients' requests to primary care physicians for an NGS analysis will increase, especially from healthy adults looking for early detection or prevention of disorders.

Figure 1. Timeline: the major events in next-generation sequencing. On the left is the year of the event

EST: Expressed sequence tag; IHGSC: International human genome sequencing consortium; ENCODE: Encyclopedia of DNA elements; NCBI: National Center for Biotechnology Information; WGS: Whole-exome sequencing.

Figure 2. Generic pipeline for the analysis of next-generation sequencing

Multiple steps involved in the analysis of data from the next-generation sequencing. The paired-end short reads, from the sequencing machine, are submitted to a quality control process. The adaptors are removed from the reads, and then the reads are mapped to the human reference by using short-read sequencing mapping tools. The alignments in the sequence alignment/map format are cleaned with tools like Pickard and transformed into a binary version of the sequence alignment/map format BAM. The BAM file is processed with tools like the Genome Analysis Toolkit to clean up the alignments. Quality control reports are generated, and variants are extracted by the use of variant callers. The document containing the variants or variant call format is annotated and filtered. Low-frequency variants that are known or predicted to be damaging are validated and used to generate a final report to the physicians or genetic counselors.

BAM: Binary Sequence Alignment/Map format; dbNSFP: Lightweight database of human nonsynonymous SNPs and their functional predictions; GATK: Genome Analysis Toolkit; HGMD: Human gene mutation database; HPG: High performance genomics; MAF: Minor allele frequency; QC: Quality control; SAM: Sequence Alignment/Map format; SRSMT: Short read sequencing mapping tools; VEF: Variant effect predictor; VCF: Variant call format.

Figure 3. The road from next-generation sequencing to personalized medicine

An overall view of how next-generation sequencing will be incorporated into the medical healthcare system. At the time of birth, a small sample of blood is taken from the patient and submitted to whole genome sequencing. The physicians and genetic counselors will provide a detailed family and medical history to an entity that will store and analyze the next-generation sequencing data. This entity will receive additional information such as metabolomics, proteomics and transcriptomes, among others, as well as new bioinformatics interpretation will be performed in collaboration with molecular biologist, physicians and genetic counselors. The physicians will review the reports and formulate recommendations and treatments for the patient. The process will be interactive with constant communication between the doctor, patient and entity in charge of the data interpretation.

Table 1

The most frequently used variant callers.

Name	Institution	Comments	Ref.
GATK	Broad Institute	GATK is a suite of tools designed by geneticist and engineers with a very robust architecture. It provides two widely used tools to detect variants: UnifiedGenotyper – a Bayesian genotype likelihood program; HaplotypeCaller – it uses an affine gap penalty pair Hidden Markov models	[53,57]
FreeBayes	Boston College	FreeBayes is a Bayesian haplotype-based variant discovery program. It solves the problem of detecting haplotypes on regions where multiple alignments are possible	[58,59]
Atlas2	HGSC, Baylor College of Medicine	Atlas2 uses a logistic regression model that has been trained on a group of validated variants	[60,61]
Bambino	The National Cancer Institute's Center for Biomedical Informatics and Information Technology	Bambino takes advantages of pooling samples. It is specially designed for detection of somatic mutations. It takes a new approach of padding the reads to improve detection of insertions and deletions	[62,63]
SAMtools	The Wellcome Trust Sanger Institute	SAMtools provides an additional tool, bcftools, and an perl script to extract the variants from a multialignment format (mpileup) generated from bamfiles	[64,65]
SNVer	New Jersey Institute of Technology	It takes a statistical approach using a binomial–binomial model and test the significance of the of each allele generating a p-value	[66,67]

GATK: Genome Analysis Toolkit; HGSC: Human Genome Sequencing Center.

Table 2

Resources for allele frequency information.

Name	License	Comments	Ref.
HapMap project	Free access	HapMap project focus on the characterization of common SNPs with a minor allele frequency of $\geq 5\%$	[15,18,70]
1000 Genomes project	Free access	Based on the Extended HapMap Collection. 1000 Genome project captured up to 98% of the SNPs with a minor allele frequency frequency of $\geq 1\%$ in 1092 individuals from 14 populations	[71–73]
The NHLBI (MD, USA) Exome Sequencing Project	Free access	A project directed to discover genes responsible for heart, lung and blood disorder, decided to release the allele frequency of each variant detected in their exome sequencing project	[74–76]
The Personal Genome Project	Free access	Currently, the Personal Genome Project has the genomes of 174 individuals and the exomes of over 400 volunteers available for download	[77,78]
NextCode Health	Commercial	40 million validated variants collected from the genotype of 140,000 volunteers from Iceland	[79,80]
CHARGE consortia	Fee for access and require permission from CHARGE consortia	1000 whole exome data sets of well-phenotyped individuals from the CHARGE consortium	[81,82]

CHARGE: Cohorts for Heart and Aging Research in Genomic Epidemiology; HapMap: Haplotype map; NHLBI: National Heart, Lung, and Blood Institute.

Table 3

Human catalogs of disease-causing mutations.

Name	License	Ref.
Human Genome Mutation Database (HGMD)	Commercial	[87,88,90]
ClinVar database	Open	[89,91]
Human Genome Variation Society has a Locus Specific Mutation Database	Open	[92,93]
Leiden Open source Variation Database (LOVD)	Open	[94,95]
Catalogue of Somatic Mutations in Cancer	Open	[96,97]
The Diagnostic Mutation Database (DMuDB)	Commercial	[98]
A human mitochondrial genome database (MITOMAP)	Open	[99,100]
PhenCode	Open	[101,102]

Table 4

Functional prediction programs.

Tool	Date	Access[†]	Category[‡]	Ref.
PANTHER	2003	A and C	3	[103,104]
Logre	2004	H	3	[105,106]
topoSNP	2004	C	3	[107,108]
MAPP	2005	A and C	3	[109,110]
nsSNPAnalyzer	2005	C	4	[111,112]
PMut	2005	H	4	[113]
LS-SNP	2005	C	2	[114,115]
FoldX	2005	A and F	1	[116,117]
Align-GVGD	2006	C	3	[118,119]
PhD-SNP	2006	A and B and C	4	[120,121]
FASTSNP	2006	C and H	4	[122,123]
Mupro	2006	A and C	1	[124,125]
snps3D	2006	C	1	[126,127]
CanPredict	2007	H	4	[128]
Parepro	2007	H	4	[129]
SNAP	2007	A and B and C	4	[130,131]
BONGO	2008	H	2	[132]
ETA	2008	C	1 and 4	[133,134]
MutPred	2009	C	4	[135,136]
SIFT	2009	A and B and C and E	3	[137,138]
SNPs&GO	2009	C	4	[139,140]
MuD	2010	C and H	4	[141,142]
Hope	2010	C	2	[143,144]
MutationTaster	2010	C	4	[145,146]
PolyPhen-2	2010	A and B and C and E	2 and 4	[147,148]
Condel & FannsDb	2011	B and C	7	[149–152]
SDM	2011	C	1	[153,154]
PopMuSic	2011	C and F	1	[155,156]
Mutation-assessor	2011	C	3	[157,158]
PON-P	2012	C	2	[159,160]
PROVEAN	2012	A and B and C and E	3	[161,162]
KD4v	2012	C and D and I	1 and 4	[163,164]
SNPdbe	2012	C and G	6	[165,166]
VariBench	2012	C and G	5	[167,168]
CAROL	2012	B	7	[169,170]
Hansa	2012	C	4	[171,172]

Tool	Date	Access [†]	Category [‡]	Ref.
SNPeffect 4	2012	C and F	2	[173,174]
Meta-SNP	2013	C	7	[175,176]
VAAST 2.0	2013	A and F	8	[177,178]
logit	2013	H	7	[179]
dbNSFP v2.0	2013	G	6	[180,181]
CoVEC	2013	A and B and C	7	[182,183]
PredictSNP	2014	C	7	[184,185]
mCSM	2014	C	1	[186,187]
HMM	2014	A	3	[188,189]
GWAVA	2014	B and C and E	4	[190,191]
CADD	2014	C and E	4	[192,193]

[†] Access keys = A: Executables; B: Source; C: Web interface; D: Web services; E: Precomputed scores; F: Require registration; G: Download entire database; H: Site not available; I: Access to rules and training sets.

[‡] Category keys = 1: Protein stability; 2: Protein sequence and structure; 3: Sequence and evolution conservation; 4: Machine learning; 5: Data for benchmark; 6: Database; 7: Consensus classifier; 8: Conservation and frequency.

Table 5

Software to annotate variant call format files and manage workflow.

Name	Type of analysis or system provided	Access	Ref.
Cassandra	VCF annotator	Free	[198]
AnnTools	VCF annotator	Free	[199]
Ensembl SNP Effect Predictor	VCF annotator	Free	[200]
snpEff	VCF annotator/predictor	Free	[201]
ANNOVAR	VCF annotator	Commercial and free	[202]
Varianttools	VCF annotator	Free	[203]
Galaxy	Workflow management system	Free	[204]
Mercury	Workflow management system	Free	[205]
NGSANE	Workflow management system	Free	[206]
Seven Bridges Genomics, Inc.	Workflow management system	Commercial	[207]
Chipster	Workflow management system	Free	[208]
Anduril	Workflow management system	Free	[209]
Genomatix	Hardware and software	Commercial	[210]
CLC Bio	Hardware and software	Commercial	[211]
Knome, Inc.	Hardware and software	Commercial	[212]
SoftGenetics	Software	Commercial	[213]
DNASStar, Inc.	Software	Commercial	[214]
Partek, Inc.	Software	Commercial	[215]
Complete Genomics, Inc.	Whole genome and analysis	Commercial	[216]
Personalis	Exome sequencing and analysis	Commercial	[217]
Omicia	Analysis	Commercial	[218]
NextCODE Health	Analysis	Commercial	[79]
Invitae Corp.	Analysis	Commercial	[219]
Genformatic	Analysis	Commercial	[220]
Bina	Analysis	Commercial	[221]
Real Time Genomics	Analysis	Commercial	[222]
DNAnexus	Cloud service, storage and analysis	Commercial	[223]
Ingenuity	Analysis	Commercial	[224]

VCF: Variant call format.