Next Generation Sequencing for Next Generation Diagnostics and Therapy

Marianna Garonzi, Cesare Centomo and Massimo Delledonne

Abstract DNA sequencing technologies are evolving at a prodigious rate. First-generation approaches have now been largely replaced by second-generation technologies (still known as "next generation sequencing" (NGS) even though they are now current and commonplace), and third-generation technologies (sometimes called "next-next generation sequencing") are starting to arrive. This has led to global boom in whole genome or exome sequencing, boosting the discovery of sequence variants associated with disease that will eventually be translated into new diagnostic, prognostic, and therapeutic targets for individual patients in "precision medicine." Acknowledgement of disease predisposition and specific therapeutic behavior for each individual addresses a more preventive approach. Adoption of such novel means represents an anticipation-relevant outcome as it can affect our healthcare on many different levels, ranging from a simple lifestyle adjustment to a well-defined clinical guideline. In this chapter we summarize current and emerging sequencing technologies for clinical applications, and some of the challenges that lie ahead.

Keywords Precision medicine · Genomics · Sequencing technologies

M. Garonzi (✉) · C. Centomo · M. Delledonne
Department of Biotechnology, University of Verona, Strada Le Grazie 15, 37134 Verona, Italy
e-mail: marianna.garonzi@univr.it

C. Centomo
e-mail: cesare.centomo@univr.it

M. Delledonne
e-mail: massimo.delledonne@univr.it

M. Nadin (ed.), *Anticipation and Medicine*, DOI 10.1007/978-3-319-45142-8_6

1 Introduction

Next generation diagnostics and therapy—the object of the conference on Anticipation and Medicine that is the origin of this volume—can be defined as any set of clinical approaches informed by consulting the sequence of a patient's genome. The origins of this approach can therefore be traced to the first medical applications of individual gene sequences. Comprehensive applications required two major developments: the sequencing of the human genome and the arrival of technologies allowing access to individual genomes quickly and inexpensively. We shall focus on the development of technology, connecting it to the anticipatory aspects of genetic research. As we shall see, issues such as individualized treatment, new screening methods, individual genetic disease, among other aspects pertain to the anticipatory perspective (see Nadin[1] and [1]). How technology evolved is not irrelevant to how all the questions related to the role of the "genetic map" are articulated.

The first project to sequence the entire human genome arose from several debates, some doubting whether such a project would be worthwhile in terms of downstream applications, and others arguing that it would expedite cancer research and help to identify medically relevant mutations. The advent of recombinant DNA technology in the 1970s, followed by development of methods to clone larger fragments of DNA in 1980s, provided the tools necessary to isolate and assemble genomic DNA sequences hundreds to several thousands of base pairs in length. Larger DNA inserts, in the megabase range, became accessible following the development of artificial chromosome vectors, allowing the construction of physical maps of the human genome upon which individual clones could be assembled.

The human genome project (HGP) was the first step towards the application of genome technology in medicine, but the initial aim was to assemble a highly accurate reference sequence (one error per 10,000 bases) that spanned the majority of each human chromosome. This sequence was predicted to offer valuable information concerning human biology, thus facilitating applications in other fields, including medicine, drug development and forensics. The HGP officially commenced in 1990 and lasted 13 years, with total funding of US$3.8 billion. A "working draft" of the human genome DNA sequence was completed in June 2000 and published in February 2001 [2].

Alongside the map-based sequencing approach adopted by the publicly funded International Human Genome Sequencing Consortium (IHGSC), Celera Genomics (founded in 1998 by Dr. Craig Venter) declared its intent to sequence the human genome using the comparatively new method of whole-genome shotgun sequencing. This does not require the prior development of a physical map for assembly, but relies instead on the generation of large numbers of overlapping reads that can be assembled de novo. This approach was faster than the IHGSC strategy but much

[1]Nadin, M.: Medicine: The Decisive Test of Anticipation. In: Nadin, M.: (ed.) Anticipation and Medicine, pp. 1–25. Springer, Cham (2016).

more computationally demanding, and therefore only became possible towards the end of the publicly funded project when sufficient computing power became available. The Celera Genomics project was able to sequence the human genome in three years at a cost of approximately US$300 million. However, this progress would not have been possible without access to the sequencing data already produced by the IHGSC [3].

Before the analysis of the draft sequence, the human genome was expected to contain ~120,000 genes [4], but sequence annotation only revealed ~20,500 [5]. Only 1.1 % of the genome was represented by exons, whereas 24 % was represented by introns and 75 % by intergenic DNA. The alignment of the human genome with other sequenced genomes (a relatively new field at the time, known as comparative genomics) revealed vertebrate-specific evolutionary expansions in gene families associated with neuronal functions, tissue-specific developmental regulation, hemostasis, and the immune system [3].

The draft sequences also provided locations for 2.1 million single-nucleotide polymorphisms (SNPs), showing that single-base differences between any randomly selected human genomes occur at an average frequency of 1 every 1250 bp. A SNP map has therefore been integrated with the human genome sequence to highlight how nucleotide diversity varies across the genome, in a manner broadly consistent with the standard population genetics model of human history. This high-density SNP map provided a public resource that defined variation across the genome and identified markers for disease diagnosis and therapy [3].

Both human genome projects relied on several cumulative improvements in the Sanger chain-termination sequencing method to improve accuracy and throughput. This increased the output of a single sequencing machine to about 1.6 million bp per day. But even at that rate it would take 15,000 days of continuous operation for one Sanger sequencer running 96 reactions in parallel to cover the three billion base pairs of the human genome with the minimum eight-fold redundancy required to ensure accuracy. Entirely new sequencing methods were therefore required to gain access to the genomic information of individuals at a cost suitable for standard healthcare practices and rapidly enough to facilitate diagnosis in time for effective therapy, heralding the development of next-generation sequencing (NGS). To reach these goals, a new initiative was founded by the National Human Genome Research Institute (NHGRI) in 2004 aiming to reduce the cost of sequencing a human genome to US$1,000 within 10 years [6].

2 The Advent of NGS Technologies

Next-generation sequencing provided the basis for new diagnostic and therapeutic strategies by accelerating the rate of sequence generation and reducing the cost per base to the extent that individual genomes became accessible for the first time. This step change meant that sequencing technology could be used for the discovery of

medically relevant sequences at the level of individual patients, rather that the erstwhile approach of testing short stretches of DNA for previously discovered sequence variants. Several different NGS platforms have been developed but they all share one property that differs from the Sanger method, i.e., the ability to produce millions of short reads in parallel.

The first NGS platform was commercialized in 2005 by 454 Life Sciences [7]. The 454 technology combined emulsion PCR (allowing the amplification of DNA fragments in massively parallel arrays without cloning) with pyrosequencing, a real-time sequencing by synthesis method developed almost 10 years before [8, 9]. Emulsion PCR is based on the in vitro amplification of up to 10 million copies of single DNA fragments attached to the surface of beads encapsulated in water droplets in an oil–water emulsion, thus avoiding time-consuming standard cloning methods. Millions of beads, each decorated with millions of copies of a different genomic fragment, are then placed in picoliter-sized wells where the sequencing reaction takes place, achieving massive parallelization and throughput. The sequencing process is based on the real-time detection of pyrophosphate release following the addition of a nucleotide by DNA polymerase to the growing DNA strand. A sulfurylase converts the pyrophosphate to ATP which acts as a substrate for the luciferase-mediated conversion of luciferin to oxyluciferin thus generating a measurable flash of light.

The 454 sequencing technology overcomes two of the bottlenecks in Sanger sequencing: the need to prepare individual templates, which is avoided by the multiplex emulsion PCR format, and the need to complete a chain-termination reaction and separate the products by capillary electrophoresis, which was addressed by the real-time optical detection of nucleotide insertion in a high density multiwell plate. This early example of NGS increased the throughput by 100-fold compared to Sanger sequencing. In more recent 454 instruments, up to one million reads can be generated per run, each 700–800 bp in length. This remains one of the longest read lengths among all the current NGS technologies, not far short of the $\sim$1,000 bp maximum achieved by Sanger capillary sequencing, and there is a low rate of substitution errors. However, the intrinsic limitations of pyrosequencing mean that 454 sequencing is sensitive to indel errors due to the misinterpretation of homopolymer sequence runs.

Other NGS technologies followed hot on the heels of the 454 method including the Genome Analyzer launched in 2006 by Illumina [10] and *Sequencing by Oligo Ligation Detection* (SOLiD) marketed by Applied Biosystems in 2007 [11]. Illumina offered a novel strategy for preparing the sequencing template, using a "bridge PCR" to amplify the signal directly on the solid surface of a flow cell where the sequencing reaction takes place. The sequencing method is analogous to the Sanger approach because it is based on chain termination. However, it uses reversible terminators and achieves sequencing in real time through cycles of fluorescent nucleotide incorporation, imaging, and cleavage of the terminator group containing the dye. Both technologies have been streamlined and improved to simplify template preparation, remove awkward bead-handling steps and increase the number of reads generated per cycle, thus reducing the per-base cost of

sequencing even further. In the original Illumina method, the read length was limited to ~35 bases, although millions of reads were generated per run, but more recent machines can generate billion of reads of up to 250 bases. Compared to 454 sequencing, the Illumina method is less prone to indel errors in homopolymer runs but there is a higher substitution error rate.

Both Illumina and SOLiD use expensive fluorophore-based labeling technologies and optical imaging. In 2010, Ion Torrent sequencing introduced advanced semiconductor technology to detect the hydrogen ions released during nucleotide incorporation, thus further simplifying the overall detection process and providing sequences more rapidly at a lower cost [12]. However, Ion Torrent shares with 454 sequencing the tendency to suffer a high indel error rate in homopolymer runs.

As stated briefly above, the advent of new sequencing technologies that are simpler, faster, and more scalable than the Sanger method has caused the per-base cost of sequencing to fall rapidly. In 2001, when the first draft Human Genome Sequence was published, the cost to sequence 1 Mb was US$5,292.39. During the transition from Sanger-based sequencing to NGS technologies, the cost per Mb fell to ~US$100. During 2008, which is arguably when NGS became "mainstream", the cost per Mb fell from US$100 in January to less than US$4.00 in December. This trend has continued, and as of June 2015, the cost per Mb had declined to a remarkable US$0.015. The cost to sequence a human genome has therefore fallen from US$95,263,072.00 in 2001 to US$1,363.00 in 2015, very close to the US $1,000 target set by the NHGRI in 2004 [13].

3 Limits of NGS Applications

Although NGS has precipitated astonishing advances in the last ten years, all the techniques are limited by the relatively short length of the reads. This is not an issue when sequencing unique or well-characterized regions of the genome, but short reads cannot resolve repetitive regions longer than the read length, such as trinucleotide repeat expansions associated with diseases known as trinucleotide repeat disorders (e.g., Huntington's disease, fragile X syndrome, and neurodegenerative progressive disorders known as ataxias) [14]. For the same reason, only short indels can be detected and the technology struggles with larger structural variants, such as translocations, because of the small spatial resolution achieved by short reads. Structural variants are less common at the population level than SNPs and indels, but recent studies indicate they are associated with a number of human diseases ranging from sporadic syndromes and Mendelian diseases to complex traits, including neurodevelopmental disorders. Chromosomal aneuploidies, such as trisomy 21 (Down syndrome), and monosomy X (Turner syndrome) are well characterized; but de novo copy number variations are now known to be enriched in autism spectrum disorders [15] and structural variations may contribute to other complex traits including cancer, schizophrenia, epilepsy, Parkinson's disease, and immune disorders such as psoriasis [16, 17].

Gene fusions caused by somatic translocations are associated with tumorigenesis, e.g., chronic myeloid leukemia (CML) and acute myeloid leukemia (AML) [18, 19]. Although several strategies based on whole genome sequencing and/or transcriptome sequencing have been used to discover gene fusion events, they remain limited by the high frequency of false positives and low sensitivity of computational approaches based on short sequence reads.

Whole-genome sequencing using NGS platforms also provides little if any haplotype information at the level of an individual genome. Haplotype data facilitates linkage analysis and association studies, and is a key component of population genetics and clinical genetics [20]. Haplotype data can be used to predict the severity and prognosis of certain genetic disorders. For example, intragenic cis-interactions between common polymorphisms and pathogenic mutations in the prion protein (*PRNP*) and cystic fibrosis transmembrane conductance regulator (*CFTR*) genes greatly influence the penetrance and expressivity of hereditary Creutzfeldt-Jakob disease and cystic fibrosis, respectively [21]. Similarly, the gene encoding the protease inhibitor α-2-macroglobulin is located within the Alzheimer's disease (AD) susceptibility locus on chromosome 12p, and a series of studies using SNP markers show that certain haplotypes (especially those containing a 5-bp deletion in intron 18 and a non-synonymous SNP in exon 24) have a high-risk association with AD [22–24]. Although haplotypes can be inferred by population-based methods or by genotyping multiple individuals from the same family, data interpretation can be hindered by low-frequency variants, private variants and de novo variants that are poorly resolved.

4 Third-Generation Sequencing

The second-generation sequencing technologies described above rely on PCR to amplify signals from individual templates. This means that the accuracy of sequencing is dependent on the accuracy of the PCR step and the read length is limited by the need for template amplification. Third-generation technologies overcome this limitation by using ultrasensitive imaging technologies or electrochemical sensors that allow the sequencing of individual molecules without prior amplification. The main advantages of third-generation technologies include minimal sample preparation, the ability to use smaller amounts of biological materials, faster sequence acquisition, increased throughput, longer read lengths and the potential to reduce the cost of sequencing the human genome to US$100.00 within a few years.

The first commercially available third-generation sequencing technology was the Helicos Genetic Analysis Platform [25–28], which achieved single-molecule sequencing by using a high-resolution camera to detect the incorporation of a single fluorophore during DNA synthesis. Although the imaging of fluorescent dyes was reminiscent of NGS, the innovation of the Helicos platform was the use of a single-molecule template, removing the need for an initial PCR amplification

step. There was no improvement over NGS in terms of read lengths, throughput, and accuracy, but one unique advantage was the ability to directly sequence RNA, as well as DNA [29]. The Helicos platform is no longer available because the company has ceased trading.

Another example of third-generation sequencing is the Single-Molecule Real-Time (SMRT) technology developed by Pacific Biosciences, which enables the direct observation of a single molecule of DNA polymerase synthesizing a strand of DNA. DNA polymerases are attached to the bottom of ~50-nm wells, which act as zero-mode wave guides (ZMWs). The DNA polymerase utilizes γ-phosphate fluorescently labeled nucleotides to synthesize the nascent DNA strand. The narrow width of the ZMW prevents light propagation through the waveguide, but energy penetration over a short distance excites the fluorophores attached to the nucleotides in the area near the DNA polymerase at the bottom of the well. The fluorescence pulse that follows nucleotide incorporation can thus be detected in real time [30].

The SMRT method has been improved by simplifying the sample preparation, scanning, and washing steps to produce results more quickly and with less effort [31]. The absence of template amplification allows the processivity of DNA polymerase to be fully exploited, resulting in reads with an average length of ~10 kb and often exceeding 20 kb. This facilitates de novo assembly, the direct detection of haplotypes and the phasing of entire chromosomes. However, one drawback of this technology is that indel errors can exceed 13 % [32].

In 2015, Pacific Biosciences launched a new SMRT-based sequencer claiming higher throughput and lower costs. Although the chemistry has not changed, the SMRT cells have been redesigned to contain one million ZMWs compared to 150,000 in the previous system, increasing throughput 7-fold. Each SMRT cell therefore has a throughput of 5–10 Gb and initial average read lengths of 8–12 kb; both throughput and average read length should increase over time.

Nanopore sequencing is another third-generation technology based on the direct detection of DNA nucleotides passing through a nanoscale pore. The sequence can be recorded directly as a current fluctuation or converted into an optical signal [33]. The Oxford Nanopore Technologies MinION platform is the first available commercial example of nanopore sequencing, following beta-testing in 2014. MinION is a hand-held device which produces much longer reads than other technologies (tens of kilobases) in a short time. Once a sample is charged in the MinION flow cell, initial results are provided in minutes and a run can be completed in a few hours. The main disadvantage of current nanopore sequencing is the high error rate due to the low spatial resolution of the biological pore. New nanopore technologies that aim to overcome such problems replace the protein channels with artificial non-organic pores small enough to report the intervals between consecutive nucleotides. The Oxford Nanopore Technologies MinION platform is simple, portable, much less expensive and capable of producing much longer reads than the other NGS technologies currently available [34, 35].

5 Clinical Analysis of the Human Genome Sequence

There are many technological differences among first-, second-, and third-generation sequencing platforms, but in practical terms there are two main advances that can be brought to bear in clinical diagnosis and therapy—the anticipation-relevant outcomes of such advances. Next-generation sequencing technologies are (i) much faster and (ii) much less expensive than Sanger sequencing, which means a patient could undergo genome sequencing at approximately the same cost and in approximately the same timescale as standard laboratory assays, a consideration that would have been inconceivable 10 years ago. The new sequencing technologies therefore bring more diseases than ever before into the domain of sequence-based diagnosis and therapy.

More than 5,000 human single-gene disorders have been resolved to causative mutations, and others have been associated with structural aberrations or aneuploidies [36]. Although the availability of the human genome sequence has greatly improved our understanding of the genetic basis of disease—including the anticipatory aspects—second- and third-generation sequencing technologies have made the identification of genetic variants feasible on a genomic scale because the sequencing of individual genomes is now possible, making it easier to identify rare SNPs, indels and structural variations with a high degree of confidence. However, detecting variants is only the first part of a complex interpretation process. The number of polymorphisms and rare sequence variants per individual ranges from few hundred thousand in the exome to millions in the entire genome, so comprehensive biochemical characterization and the assessment of a causal link between a gene variant and a disease is not always possible. The comprehensive prioritization of candidate genes prior to experimental testing is therefore necessary, but this requires the screening of genome sequence data to select the most likely clinically relevant variants. Candidates are prioritized using correlative evidence that associates each variant and gene with the given disease based on the integration of molecular, genetic, biochemical, functional and epidemiological data.

Several resources have been developed to facilitate the identification and reporting of variants by collecting human variations and associated outcomes. These resources include ClinVar [37], an archive of relationships between medically important variants and phenotypes along with supporting evidence, and the Human Genes Mutation Database (HGMD) [38], which collates known mutations associated with human inherited diseases. Specific databases have also been developed for variants associated with drug responses, and such resources can be tailored to individual genomic profiles, e.g., the Pharmacogenomics Mutation Database (PGMD) and PharmGKB [39, 40], and the COSMIC database, which collects somatic mutations identified in cancer research [41]. These databases can be screened to determine the relevance of the variants in a given genome sequence; but they have no predictive capability, i.e., they provide no information about novel and unassociated variants.

One way to predict the potential impact of genome sequence variants is to determine their frequencies in population data, as demonstrated by the 1000 Genomes Project discussed in more detail below [42]. At the population level, natural selection removes deleterious alleles, so filtering for low-frequency variants in populations can help to enrich the dataset for potentially dangerous variants. However, a rare allele can also be present for other reasons. It may be a harmless variant that is rare because it has arisen recently or is close to elimination by genetic drift. Indeed, genetic drift may result in a particular variant becoming rare in one population but part of a common polymorphism in others. Therefore, rarity per se is not necessarily evidence for disease association and additional epidemiological and functional evidence should also be sought.

Functional prediction algorithms for non-synonymous variants use different forms of evidence such as sequence conservation (SIFT [43]) or the predicted impact of amino acid substitutions on protein structure and function (PolyPhen2 [44]). Others use a classifier trained with known disease mutations as well as harmless SNPs and indels to predict the likelihood of disease association (MutationTaster [45]). More confidence can be assigned to predictions that are generated using more than one of these tools. Therefore a database of pre-calculated scores obtained from many different predictors for all possible nonsynonymous substitutions in the current human genome sequence has been developed to process queries more rapidly (dbNSFP [46, 47]).

6 Population-Scale Sequencing Projects

The discovery of relevant variations in individual genomes requires data from the analysis of large groups of people because it is necessary to correlate variations with phenotypes in a statistically significant manner. Population-scale sequencing projects thus increase the power of research on diseases and provide the foundations of personalized medicine. This is one of the claims of those who advance the anticipatory perspective (see Nadin [48]).

The first international project aiming to collect and analyze genome data from a large cohort of human subjects was the 1000 Genomes Project mentioned above. This was launched in 2008, and its primary objective was to produce a comprehensive human genetic variation database by identifying all polymorphisms, i.e., genetic variants that have frequencies of at least 1 % in the populations included in the project. The analysis of 2,504 samples from 13 different populations allowed the creation of the first complete catalog of genetic variations and their frequencies, which can be used to filter genomic data from patients afflicted by rare diseases to remove common variants and enrich for rarer variations more likely to be associated with the disease [49]. Many similar projects have been initiated more recently, such as the NHLBI GO Exome Sequencing Project (ESP) which focuses on variants contributing to heart, lung and blood disorders. In this project, the exomes of 6,503 unrelated individuals in diverse well-characterized populations were sequenced and

the resulting datasets and frequency tables have been shared with the scientific community [50].

These first massive sequencing projects facilitated the discovery of rare variants by sampling individuals from diverse populations. However, because there is high genetic diversity among populations due to genetic drift and natural selection, population-specific sequencing projects may help with the interpretation of variants in individual genomes. One of the largest studies based on a single population was carried out by the company deCODE Genetics Inc., which sequenced more than 2,600 genomes from the Icelandic population to a median coverage of 20-fold, and used comprehensive national genealogies to accurately impute even rare variants throughout the population [51]. The project discovered novel diseases-associated rare variants such as mutations in *ABCA7* that increase the risk of Alzheimer's disease [52]. A similar example is Genomics England, a UK company owned by the UK Department of Health, which aims to sequence 100,000 whole genomes by 2017 in collaboration with the UK National Health Service and 11 Genomic Medicine Centers across the country. This is the first genomics initiative which is tightly integrated with a national health system to accelerate the translation of research into clinical practice [53].

The main limitation of population studies is that rare variants contributing to quantitative traits can be difficult to identify even in large cohorts. This can be overcome by studying founder populations, in which variants that are rare or absent elsewhere may be more common due to the founder effect. One example of this approach is the analysis of the Sardinian population in Italy, in which 2,120 individual genomes were sequenced with low coverage. The project identified $\sim$3.8 million variants that were not detected in previous sequencing-based compilations such as dbSNP 142 and ExAC [54], which are also enriched for predicted functional mutations. The Sardinian project also revealed the presence of 76,286 variants with a frequency exceeding 5 % which are rare (frequency lower than 0.5 %) or absent in other populations.

7 NGS and Precision Medicine

Precision medicine is a new healthcare approach that matches the genomic data and clinical records of individual patients. In this way, treatments are tailored to the patients based on their genetic profile or other molecular and cellular information. The concept of precision medicine is based on the fact that diseases affect individuals in different ways and that different patients show distinct responses to the same treatments. Clinicians have known for many years that individual patients respond differently to the same treatments, but genome analysis now provides data that may allow the development of individualized therapies. Precision medicine encompasses screening for inherited conditions, carrier screening and prenatal testing, through to the identification of targets for cancer treatment, and the

diagnosis, and treatment of rare diseases. The possible future that affects a current state [55] is the goal of the PMI initiative [56].

Many of the ~5,000 known genetic disorders manifest during the first 28 days of life, but the full clinical symptoms may not be evident in newborns. Screening at this stage can therefore identify babies with genetic disorders that have silent, heterogeneous, or ambiguous phenotypes at birth, but which benefit from early intervention to avoid an irreversible impact on health. One example is sickle cell disease, which causes blood clotting and a shortage of red blood cells. Early identification can prevent the onset of complications caused by increased susceptibility to infections through proper and timely treatment. This is where the anticipatory perspective plays an important role. Newborn screening has been integrated into postnatal healthcare for many years, but currently it only targets about 50 of the most severe genetic disorders that require urgent clinical decisions [57].

The use of NGS-based newborn screening (NBS) companion to the current biochemical testing regime would dramatically increase the quantity and diversity of information parents and clinicians could derive from screening. A targeted NGS assay based on panels of hundreds of relevant sequence variants could improve diagnostic testing in newborns in a cost-effective manner by selectively sequencing the corresponding genomic regions, mainly exons, following enrichment in a physical DNA capture step [58].

Traditional molecular biology assays such as PCR can be used to identify a limited range of known cancer-related mutations and rearrangements but NGS could reveal comprehensive, individualized mutational landscapes, including both known and novel variations. Integrated high-throughput sequencing of tumor biopsy genomes could also facilitate biomarker-driven clinical trials in oncology [59].

Although individual genetic diseases are rare, they are collectively common, affecting millions of people worldwide. Many rare genetic diseases have escaped traditional gene discovery approaches due to heterogeneity, a limited number of patients or families for analysis, and the loss of reproductive fitness as a result of such diseases. NGS-based gene discovery partially overcomes such limitations and has enabled the discovery of hundreds of novel, rare disease mutations [60].

Finally, NGS is revolutionizing pharmacogenomics as a disease management concept. Pharmacogenomics correlates human genome sequence data with drug responses and aims to improve therapeutic efficacy and reduce side effects by developing qualitatively and quantitatively tailored treatment regimens. Pharmacogenomics has the potential to transform medical practice by replacing broad methods of screening and treatment with a more personalized approach that takes into account both clinical factors and genome data. For example, in cancer treatment, small-molecule inhibitors and antibodies that bind to “druggable” targets are revolutionizing medicine. Personalized anti-cancer therapy requires the identification of cancer-specific driver mutations in each patient. For example, among patients diagnosed with non-small-cell lung cancer, only those harboring the ALK gene fusion (present in less than 5 % of the affected population) respond to treatment with targeted inhibitors such as crizotinib [61]. The identification of patients

suitable for this treatment before therapeutic selection would avoid administering ineffective drugs to the >95 % of nonresponsive patients on a trial and error basis, not only hastening the deployment of more suitable treatment regimens, but also reducing the costs associated with wasting the drug.

Another example of the application of pharmacogenomics in healthcare is the prediction of individual responses to warfarin, a commonly prescribed oral anticoagulant which is used to prevent thromboembolic diseases in patients with deep vein thrombosis, atrial fibrillation, or recurrent stroke [62]. Although warfarin is effective, the optimal dose differs widely among individuals and is often determined on a trial and error basis. However, several studies have shown that the *VKORC1* locus is the single most significant predictor of warfarin tolerance, accounting for ~25 % of the variance in a stabilized warfarin dose [63, 64]. As sequencing technologies become less expensive and more widely available, pharmacogenomics will transition from a niche research area to a main player in drug development and clinical decision making. Whole genome sequencing can reveal rare and even unique markers that would not be detected by conventional genetic screening methods.

8 Future Perspectives

Most current medical treatments are generalized, but one size does not fit all: therapies that are highly successful in some patients may have no effect or even a deleterious effect in others. The outlook for precision medicine has been dramatically improved by the recent development of high-throughput methods to characterize patients individually, including NGS, proteomics and metabolomics, as well as the availability of comprehensive databases containing information derived from large screening projects.

Even so, there remains a lack of broad research programs translating the outcomes from these large-scale projects into clinical practice. In the future, there should be more effort to integrate whole-genome sequencing initiatives with clinical applications, as shown by the Genomics England initiative in the UK and the recently announced Precision Medicine Initiative in the USA [56, 65].

In this future scenario, the tighter integration of research programs, precision medicine initiatives, and health services will allow physicians to consult patient genome data as well as conventional medical records. Whole genome sequencing will be part of routine medical screening, and health insurance companies will cover the (declining) costs of genomic analysis because early investment in preventive medicine would save the greater costs of therapy later down the line (see [66, pp. 75, 101, 111]). New diagnostic and prognostic markers will become available, so physicians will know which diseases present the most risk to their patients and which drugs, at which doses, are likely to be most effective. The details of anticipation-driven medicine were not entered into here, rather, aspects of such an approach were suggested.

References

1. Nadin, M.: The anticipatory profile. An attempt to describe anticipation as process. In: Nadin, M. (ed.) Anticipation (special issue of the International Journal of General Systems), vol. 41, no. 1, pp. 43–75. Taylor and Francis, London (2012). http://www.tandfonline.com/doi/abs/10.1080/03081079.2011.622093, doi:10.1080/03081079.2011.622093
2. Lander, E.S., Heaford, A., Sheridan, A., Linton, L.M., Birren, B., Subramanian, A., Coulson, A., Nusbaum, C., Zody, M.C., Dunham, A., Baldwin, J., et al.: Initial sequencing and analysis of the human genome. Nature **409**, 860–921 (2001)
3. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., et al.: The sequence of the human genome. Science **291**, 1304–1351 (2001)
4. Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., Quackenbush, J.: Gene index analysis of the human genome estimates approximately 120,000 genes. Nat. Genet. **25**, 239–240 (2000)
5. Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K., Lander, E.S.: Distinguishing protein-coding and noncoding genes in the human genome. Proc. Natl. Acad. Sci. U.S.A. **104**, 19428–19433 (2007)
6. Schloss, J.A.: How to get genomes at one ten-thousandth the cost. Nat. Biotechnol. **26**, 1113–1115 (2008)
7. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., et al.: Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**, 376–381 (2005)
8. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., Nyrén, P.: Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem. **242**, 84–89 (1996)
9. Ronaghi, M., Uhlén, M., Nyrén, P.: A sequencing method based on real-time pyrophosphate. Science **281**, 363, 365 (1998)
10. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., et al.: Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456**, 53–59 (2008)
11. Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., Sidow, A., et al.: A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. **18**, 1051–1063 (2008)
12. Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., Hoon, J., et al.: An integrated semiconductor device enabling non-optical genome sequencing. Nature **475**, 348–352 (2011)
13. Wetterstrand, K.A.: DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). www.genome.gov/sequencingcosts
14. Budworth, H., McMurray, C.T.: A brief history of triplet repeat diseases. Methods Mol. Biol. **1010**, 3–17 (2013)
15. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., et al.: Strong association of de novo copy number mutations with autism. Science **316**, 445–449 (2007)
16. Stankiewicz, P., Lupski, J.R.: Structural variation in the human genome and its role in disease. Annu. Rev. Med. **61**, 437–455 (2010)
17. Girirajan, S., Campbell, C.D., Eichler, E.E.: Human copy number variation and complex genetic disease. Annu. Rev. Genet. **45**, 203–226 (2011)
18. Rowley, J.D.: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. Nature **243**, 290–293 (1973)
19. Rowley, J.D.: Identification of a translocation with quinacrine fluorescence in a patient with acute leukemia. Ann. génétique **16**, 109–112 (1973)

20. Browning, S.R., Browning, B.L.: Haplotype phasing: existing methods and new developments. Nat. Rev. Genet. **12**, 703–714 (2011)
21. Lee, J.-E., Choi, J.H., Lee, J.H., Lee, M.G.: Gene SNPs and mutations in clinical genetic testing: haplotype-based testing and analysis. Mutat. Res. **573**, 195–204 (2005)
22. Pericak-Vance, M.A.: Complete genomic screen in late-onset familial Alzheimer disease. Evidence for a new locus on chromosome 12. JAMA **278**, 1237 (1997)
23. Blacker, D., Wilcox, M.A., Laird, N.M., Rodes, L., Horvath, S.M., Go, R.C., Perry, R., Watson, B., Bassett, S.S., McInnis, M.G., Albert, M.S., et al.: Alpha-2 macroglobulin is genetically associated with Alzheimer disease. Nat. Genet. **19**, 357–360 (1998)
24. Saunders, A.J., Bertram, L., Mullin, K., Sampson, A.J., Latifzai, K., Basu, S., Jones, J., Kinney, D., MacKenzie-Ingano, L., Yu, S., Albert, M.S., et al.: Genetic association of Alzheimer's disease with multiple polymorphisms in alpha-2-macroglobulin. Hum. Mol. Genet. **12**, 2765–2776 (2003)
25. Bowers, J., Mitchell, J., Beer, E., Buzby, P.R., Causey, M., Efcavitch, J.W., Jarosz, M., Krzymanska-Olejnik, E., Kung, L., Lipson, D., Lowman, G.M., et al.: Virtual terminator nucleotides for next-generation DNA sequencing. Nat. Methods **6**, 593–595 (2009)
26. Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W., Giladi, E., et al.: Single-molecule DNA sequencing of a viral genome. Science **320**, 106–109 (2008)
27. Lipson, D., Raz, T., Kieu, A., Jones, D.R., Giladi, E., Thayer, E., Thompson, J.F., Letovsky, S., Milos, P., Causey, M.: Quantification of the yeast transcriptome by single-molecule sequencing. Nat. Biotechnol. **27**, 652–658 (2009)
28. Tessler, L.A., Reifenberger, J.G., Mitra, R.D.: Protein quantification in complex mixtures by solid phase single-molecule counting. Anal. Chem. **81**, 7141–7148 (2009)
29. Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M., Milos, P.M.: Direct RNA sequencing. Nature **461**, 814–818 (2009)
30. Korlach, J., Marks, P.J., Cicero, R.L., Gray, J.J., Murphy, D.L., Roitman, D.B., Pham, T.T., Otto, G.A., Foquet, M., Turner, S.W.: Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc. Natl. Acad. Sci. U.S.A. **105**, 1176–1181 (2008)
31. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., et al.: Real-time DNA sequencing from single polymerase molecules. Science **323**, 133–138 (2009)
32. Quail, M., Smith, M.E., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y.: A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genom. **13**, 1 (2012)
33. Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S.B., et al.: The potential and challenges of nanopore sequencing. Nat. Biotechnol. **26**, 1146–1153 (2008)
34. Mikheyev, A.S., Tin, M.M.Y.: A first look at the Oxford Nanopore MinION sequencer. Mol. Ecol. Resour. **14**, 1097–1102 (2014)
35. Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B., Akeson, M.: Improved data analysis for the MinION nanopore sequencer. Nat. Methods **12**, 351–356 (2015)
36. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., Hamosh, A.: OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res. **43**, D789–D798 (2015)
37. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., Maglott, D.R.: ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. **42**, D980–D985 (2014)
38. Cooper, D.: The human gene mutation database. Nucleic Acids Res. **26**, 285–287 (1998)
39. Kaplun, A., Hogan, J.D., Schacherer, F., Peter, A.P., Krishna, S., Braun, B.R., Nambudiry, R., Nitu, M.G., Mallelwar, R., Albayrak, A.: PGMD: a comprehensive manually curated pharmacogenomic database. Pharmacogenomics J. 1–5 (2015)

40. Hewett, M., Oliver, D.E., Rubin, D.L., Easton, K.L., Stuart, J.M., Altman, R.B., Klein, T.E.: PharmGKB: the pharmacogenetics knowledge Base. Nucleic Acids Res. **30**, 163–165 (2002)
41. Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P.A., Stratton, M.R., Wooster, R.: The COSMIC (catalogue of somatic mutations in cancer) database and website. Br. J. Cancer **2**, 355–358 (2004)
42. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.: An integrated map of genetic variation from 1,092 human genomes. Nature **491**, 56–65 (2012)
43. Ng, P.C.: SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. **31**, 3812–3814 (2003)
44. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R.: A method and server for predicting damaging missense mutations. Nat. Methods **7**, 248–249 (2010)
45. Schwarz, J.M., Rödelsperger, C., Schuelke, M., Seelow, D.: MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods **7**, 575–576 (2010)
46. Liu, X., Jian, X., Boerwinkle, E.: dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Hum. Mutat. **32**, 894–899 (2011)
47. Liu, X., Jian, X., Boerwinkle, E.: dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum. Mutat. **34**, 2393–2402 (2013)
48. Nadin. M.: Anticipation and the brain. In: Nadin, M. (ed.): Anticipation and Medicine, pp. 135–162. Springer, Cham (2016)
49. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.-Y., Konkel, M.K., et al.: An integrated map of structural variation in 2,504 human genomes. Nature **526**, 75–81 (2015)
50. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D.A., Bamshad, M.J., et al.: Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature **493**, 216–220 (2012)
51. Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B.V., Hjartarson, E., Sigurdsson, G.T., et al.: Large-scale whole-genome sequencing of the Icelandic population. Nat. Genet. **47**, 435–444 (2015)
52. Steinberg, S., Stefansson, H., Jonsson, T., Johannsdottir, H., Ingason, A., Helgason, H., Sulem, P., Magnusson, O.T., Gudjonsson, S.A., Unnsteinsdottir, U., Kong, A., et al.: Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. Nat. Genet. **47**, 445–447 (2015)
53. Siva, N.: UK gears up to decode 100 000 genomes from NHS patients. Lancet **385**, 103–104 (2015)
54. Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziewska, M., Mulas, A., Pistis, G., Steri, M., Danjou, F., Kwong, A., et al.: Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat. Genet. **47**, 1272–1281 (2015)
55. Nadin, M.: Anticipation and dynamics: Rosen's anticipation in the perspective of time. In: Klir, G. (ed.) Special issue of International Journal of General Systems, vol. 39, no. 1, pp. 3–33. Taylor and Blackwell, London (2010)
56. Precision Medicine Cohort Initiative. https://www.nih.gov/precision-medicine-initiative-cohort-program
57. Bhattacharjee, A., Sokolsky, T., Wyman, S.K., Reese, M.G., Puffenberger, E., Strauss, K., Morton, H., Parad, R.B., Naylor, E.W.: Development of DNA confirmatory and high-risk diagnostic testing for newborns using targeted next-generation DNA sequencing. Genet. Med. **17**, 337–347 (2014)
58. Saunders, C.J., Miller, N.A., Soden, S.E., Dinwiddie, D.L., Noll, A., Alnadi, N.A., Andraws, N., Patterson, M.L., Krivohlavek, L.A., Fellis, J., Humphray, S., et al.: Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Sci. Transl. Med. **4**, 154ra135 (2012)

59. Roychowdhury, S., Iyer, M.K., Robinson, D.R., Lonigro, R.J., Wu, Y.-M., Cao, X., Kalyana-Sundaram, S., Sam, L., Balbin, O.A., Quist, M.J., Barrette, T., et al.: Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci. Transl. Med. **3**, 111ra121 (2011)
60. Boycott, K.M., Vanstone, M.R., Bulman, D.E., MacKenzie, A.E.: Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat. Rev. Genet. **14**, 681–691 (2013)
61. Méndez, M., Custodio, A., Provencio, M.: New molecular targeted therapies for advanced non-small-cell lung cancer. J. Thorac. Dis. **3**, 30–56 (2011)
62. Kamali, F.: Genetic influences on the response to warfarin. Curr. Opin. Hematol. **13**, 357–361 (2006)
63. Tatarunas, V., Lesauskaite, V., Veikutiene, A., Grybauskas, P., Jakuska, P., Jankauskiene, L., Bartuseviciute, R., Benetis, R.: The effect of CYP2C9, VKORC1 and CYP4F2 polymorphism and of clinical factors on warfarin dosage during initiation and long-term treatment after heart valve surgery. J. Thromb. Thrombolysis **37**, 177–185 (2014)
64. Zhang, J., Tian, L., Zhang, Y., Shen, J.: The influence of VKORC1 gene polymorphism on warfarin maintenance dosage in pediatric patients: a systematic review and meta-analysis. Thromb. Res. (2015)
65. Collins, F.S., Varmus, H.: A New Initiative on Precision Medicine. N. Engl. J. Med. **372**, 793–795 (2015)
66. Nadin, M.: Anticipation—The End Is Where We Start From. Müller Verlag, Basel (2003)