PERSPECTIVE

DNA sequencing technologies: 2006–2016

Elaine R Mardis

Recent advances in the field of genomics have largely been due to the ability to sequence DNA at increasing throughput and decreasing cost. DNA sequencing was first introduced in 1977, and next-generation sequencing technologies have been available only during the past decade, but the diverse experiments and corresponding analyses facilitated by these techniques have transformed biological and biomedical research. Here, I review developments in DNA sequencing technologies over the past 10 years and look to the future for further applications.

Introduction

The development of Sanger sequencing almost 40 years ago has revolutionized biological research. The implications of this technique became even more far reaching with the introduction of next-generation sequencing (NGS) methods in 2005 (**Fig. 1**); these techniques have markedly increased the amount of sequencing data produced per instrument and have dramatically decreased the costs of generating sequencing data. Coupled with innovative methodological and computational developments, NGS platforms have facilitated an explosion in biological knowledge over the past decade. This unique combination of methods, analyses and biological interpretations is a logical evolution of the collaborative 'team science' approaches that characterized large-scale genome projects in the 1980s and 1990s, including the Human Genome Project. At that time, teams combined their expertise in large-insert clone mapping to scaffold the genome, used Sanger sequencing to produce data for each clone in the scaffold and performed 'finishing' to assemble the data for each clone and to close the gaps. The result was a high-quality reference-genome sequence that was then annotated to identify genes and other content. In the current NGS era, teams working in genomics include new disciplines, owing to the types of biological questions being addressed and the larger sizes of the NGS data sets and their associated analytical complexity. Now, data integration is critical to producing higher-level data interpretations, and expertise in biological systems is needed to further interpret biological meaning in the context of the integrated results. These higher-order findings in turn generate new biological hypotheses that require an ever-evolving suite of functional assays to correlate genotype with phenotype. Hence, although the teams differ in their makeup, sequencing technology continues to play a central role in producing biological findings that are rapidly advancing the state of knowledge and setting the stage for further discoveries.

McDonnell Genome Institute, Washington University, St. Louis, Missouri, USA. Correspondence should be addressed to E.R.M. (emardis@wustl.edu).

Received 17 August 2016; accepted 20 October 2016; published online 5 January 2017; doi:10.1038/nprot.2016.182

NGS: library construction and input DNA

A key procedural difference in NGS methods in comparison to Sanger sequencing is in the construction of the sequencing library. Sanger sequencing libraries require multiple steps that combine molecular biology with microbiological culture to represent the DNA sample of interest as a series of subclones in a bacterial plasmid or phage vector. These subclones then require growth in culture and DNA isolation before sequencing. This multistep process can be completed in approximately one week, at which point the purified DNAs are ready for sequencing. By contrast, the simplicity and speed of NGS library construction is striking. Starting from a variety of input DNA sources ranging from high-molecular-weight genomic DNA to a pool of PCR products, to short stretches of histone-bound DNA released after chromatin immunoprecipitation (ChIP) or reverse transcriptase–converted RNA (for example), only a few preparatory molecular biology steps are performed on the input DNA, and synthetic adapters are then ligated to the library fragment ends. These common adapters facilitate subsequent PCR amplification cycles that produce a library ready for quantification, dilution and sequencing[1] within approximately 2 days. More recently, the use of transposons to simultaneously insert the synthetic adaptor sequences during fragmentation of chromosomal DNA has further simplified and accelerated NGS library construction[2]. The simplicity of NGS library construction enables numerous preparatory assays such as ChIP, reverse transcription and bisulfite treatment, among many others, to provide the input DNA for an NGS library. As a result, NGS has enhanced the ability to transition from site-specific assays to genome-wide assays that can characterize, for example, all of a specific protein's binding sites on DNA, the flux of methylated loci under specific cell growth conditions or stresses, and the mapping of open versus closed chromatin.

Sequencing by synthesis and sequencing by ligation: fragment amplification, data production and alignment

The first NGS platforms were instruments performing sequencing by synthesis (SBS). Unlike Sanger sequencing, SBS libraries are amplified *in situ* on a solid support that is integrated within

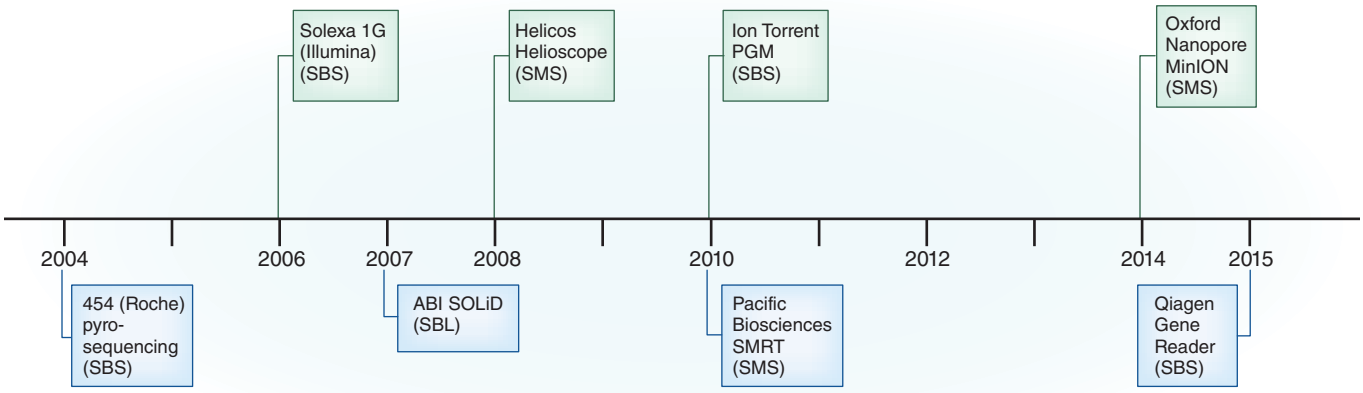

Figure 1 | NGS instruments introduced over the past decade. This timeline describes the year of introduction of each of the NGS platforms that successfully achieved commercial introduction during the past decade. SBS, sequencing by synthesis; SMS, single-molecule sequencing; SBL, sequencing by ligation.

the fluidics needed for the synthesis and detection steps of the sequencing process[1,3]. A fragment-amplification step is needed to produce sufficient copies of each library fragment, which in turn produce sufficient signal strength to enable the detection of incorporated nucleotides. Fragment amplification occurs either directly on the surface of the sequencing fluidics device or indirectly on the surface of a spherical support (bead). In either case, the solid support is typically coated with synthetic DNA adapters that are complementary to those used in the library construction process. For on-device amplification, the enzyme cocktail is supplied directly to the hybridized library fragments and incubated. For on-bead amplification, each bead is emulsified in an aqueous mixture of library fragments and enzyme cocktail before amplification. Here, the input fragment concentration is adjusted to result in, on average, one bead and one library fragment in each micelle. Amplification occurs through temperature cycling the micelles *en masse*. After amplification, the emulsion is separated, and the beads are enriched to eliminate beads that have no amplified library fragments. Because each sequencing read originates from a single library fragment, SBS data are digital, and the resulting ability to quantify read data from different preparatory inputs is one important advantage of SBS over Sanger sequencing.

Data production by Sanger sequencing involves two decoupled steps: (i) a discrete molecular biology step that produces fluorescently labeled fragments from each isolated subclone DNA template and (ii) a sequencing-platform-based data-collection step that combines simultaneous electrophoretic separation of the fragments with their fluorescence-based detection. By contrast, each commercially available SBS platform (for example, 454, Illumina, Ion Torrent and GeneReader; **Fig. 1** and **Table 1**) uses a unique combination of sequencing-reaction chemistry and detection modality applied in a stepwise manner to produce sequencing data. The SOLiD platform (**Fig. 1** and **Table 1**), is an exception that uses a similar stepwise approach but determines sequence on the basis of ligation rather than nucleotide addition (referred to as sequencing by ligation (SBL)). The need for electrophoresis in SBS and SBL platforms is obviated because each on-device-amplified library fragment either is defined by positional (*x*,*y*) coordinates on the surface of the fluidic device or is held at a fixed bead-in-well position. The instrument collects data from all *x*,*y* coordinates corresponding to amplified on-surface or on-bead fragments throughout the duration of the sequencing run, then converts these position-unique signal data into sequencing-read data for downstream analyses. This unique aspect of coupled sequencing and detection on SBS and SBL platforms permits simultaneous data production from very large numbers of library fragments, thereby explaining the commonly used descriptor 'massively parallel sequencing'.

Increased numbers of library fragments sequenced, and increased read lengths obtained from each template, owing to improvements in labeling chemistries and/or detection sensitivities have resulted in ever-increasing data output per sequencing run in NGS platforms. Details of platform-specific library construction, sequencing chemistries and detection schemes have previously been described[1,4,5]. Whereas Sanger sequencing read lengths typically range between 600 and 800 bp, most SBS instruments deliver read lengths in the range of 100–400 bp. This key difference in read length means that for all but the least complex and smallest genomes (for example, viral genomes), one must align SBS and SBL reads to a reference genome before performing downstream data analysis and interpretation[6], rather than assembling the genome by using shared sequence overlaps, as is done for Sanger sequencing reads. Because of the requirement for alignment to a reference genome, and because of the increasing amount of data generated per experiment, the progress of NGS has also rejuvenated the field of computational biology.

Computational analysis of SBS data

Along with the advances in SBS technology, the computational and algorithmic approaches developed to align and analyze the sequence-read data resulting from a diverse number of upstream preparatory experiments have been remarkable. However, genomes of increasing complexity in terms of size and repetitive content pose problems for the correct alignment of short read

TABLE 1 | Comparison of available NGS platforms

Company	Read length	Applications	Website
454/Roche	400 bp (single end)	Bacterial and viral genomes, multiplex-PCR products, validation of point mutations, targeted somatic-mutation detection	http://www.454.com/
Illumina	150–300 bp (paired end)	Complex genomes (human, mouse and plants) and genome-wide NGS applications, RNA-seq, hybrid capture or multiplex-PCR products, somatic-mutation detection, forensics, noninvasive prenatal testing	http://www.illumina.com/
ABI SOLiD	75 bp (single end) or 50 bp (paired end)	Complex genomes (human, mouse, plants) and genome-wide NGS applications, RNA-seq, hybrid capture or multiplex-PCR products, somatic-mutation detection	http://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/solid-next-generation-sequencing.html/
Pacific Biosciences	Up to 40 kb (single end or circular consensus)	Complex genomes (human, mouse and plants), microbiology and infectious-disease genomes, transcript-fusion detection, methylation detection	http://www.pacb.com/
Ion Torrent	200–400 bp (single end)	Multiplex-PCR products, microbiology and infectious diseases, somatic-mutation detection, validation of point mutations	http://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing.html/
Oxford Nanopore	Variable: depends on library preparation (1D or 2D reads)	Pathogen surveillance, targeted mutation detection, metagenomics, bacterial and viral genomes	http://nanoporetech.com/
Qiagen GeneReader	107 bp (single end)	Targeted mutation detection, liquid biopsy in cancer	http://www.genereaderngs.com/

sequences. Correct alignment is required to appropriately identify the genomic origin of fragments sequenced in an NGS experiment so that any variation can be identified and characterized. Although genome partitioning methods such as hybrid capture sequencing[7–9], reduced-representation methods[10,11] or multiplex PCR amplification of specific loci[12,13] can decrease not only the cost per experiment but also the challenges posed by large data sets and the accuracy of read placement, there remain problems with incorrect read alignments to members of gene families that share high homology and to pseudogenes.

Although computational prowess with NGS data has been critical to the success of SBS and its widespread use in research, and more recently in the clinic, the increasing data volumes resulting from widespread adoption of NGS have come at a steep price. Early NGS adopters developed substantial computational infrastructure that scaled with the increasing output of the sequencing instruments, albeit at high cost for both data storage and data management. At present, with the decreases in the costs of generating data, we are at a crossroads at which data are cheaper to reproduce than to store.

Limitations of NGS

Although costs of NGS data production have fallen, and the scale of data production has dramatically increased over the past 10 years, some technological limitations remain. The human genome in particular poses a number of major barriers to correct read alignment, owing to its size (~3 GB) and complexity (~48% repetitive sequences), as do other vertebrate and plant genomes. In the absence of an available high-quality reference human genome[14], the impact of SBS on biomedical research would have been minimal, because making sense of human-genome-scale data sets would have been largely impossible. The aforementioned improvements in read length and the ability to produce data from both ends of each library fragment ('paired-end reads') on some platforms have positively affected the certainty of read placement on complex reference genomes and have increased the numbers and types of variants that can be detected downstream of this step through the appropriate computational analysis approach (**Table 1**). However, even given the now relatively advanced state of data analysis after short-read alignment, the complexity of many genomes results in high false-positive rates in detecting certain types of variation, such as structural alterations[15–17]. As discussed below, comparison of short-read alignment to long-read assembly of human genomes has indicated that certain types and sizes of alterations are being missed because of the limitations of SBS read lengths and library fragment sizes. Finally, having only a single human reference genome (albeit one that is an amalgamation of DNA sourced from several individuals) limits the discovery of novel genomic content because sequences that are truly unique to the individual genomes studied by SBS will not yield an alignment

match on the reference[18,19]. As a result of these limitations stemming from the fixed reference genome, efforts to represent alternative human sequence content have populated more recent builds of the human genome that also contain improved sequence contiguity at complex (and therefore challenging) loci that originally were not complete[20]. More recently, efforts to sequence and assemble additional human-genome references of high quality and contiguity from individuals of different ancestries are underway (http://genome.wustl.edu/projects/detail/reference-genomes-improvement/), as outlined below.

Single-molecule sequencing platforms

In addition to stepwise sequencing platforms that produce data from amplified library fragments, there also are platforms that sequence individual molecules of DNA in parallel, by using various schemes. Single-molecule sequencing (SMS) in principle has substantial difficulties in producing sufficient signal for detection of the sequencing reaction, and as a result, the error rates of the individual sequencing reads are higher than those for SBS sequencing data. The original SMS platform, introduced by Helicos (**Fig. 1**), is no longer commercially available. Currently available SMS platforms (Pacific Biosciences and Oxford Nanopore; **Fig. 1** and **Table 1**) produce substantially longer reads than those produced by SBS platforms, thus enabling the assembly of sequencing-read data to produce long contiguous sequences, even for large complex genomes[16,21,22]. Computational approaches have been developed to assemble these long-read sequences, thus demonstrating that although individual reads may have relatively high error rates, the consensus sequence generated from the assembly can have an error rate comparable to those of both Sanger assemblies and SBS or SBL alignments[23–26].

One such SMS platform, the Pacific Biosciences RSII (**Fig. 1** and **Table 1**), can produce contiguous sequencing reads in excess of 40,000 bp by using a highly specialized library preparation procedure that begins with high-molecular-weight DNA. This approach is being used to sequence and assemble complete human genomes, including samples from the 1,000 Genomes Project cell lines, which were originally sequenced and characterized with Illumina SBS technology[27]. One early study of this type first mapped ~40-fold genome coverage of Pacific Biosciences long reads from a haploid human (hydatidiform mole) cell line, CHM1 to the human reference genome, then performed localized assemblies of the mapped reads. Subsequent comparisons of this reference-guided assembly to the human reference genome (GrCh37) and to high-coverage Illumina sequencing of the same cell line have been informative[28]. These comparisons have demonstrated the potential advantage of the long-read-assembly approach over Illumina paired-end-read alignment to a fixed reference by identifying over 26,000 insertion and deletion events of between 50 and 5,000 bp in euchromatin that are novel to the CHM1 genome assembly. In particular, the study has revealed a substantial increase in novel insertion events identified genome wide (92% novel insertions compared with 69% novel deletions). Recently, through new assembly-based approaches, the diploid genome of 1000 Genomes sample NA12878 has been sequenced and assembled with Pacific Biosciences data[25] and compared with the high-quality Illumina-based sequence of this individual. Although the long-read-assembly-based approach to human whole-genome sequencing has merit in addressing the need for additional complete high-quality human-genome reference assemblies, the current costs in terms of DNA quantity, time and money to generate Pacific Biosciences sequencing data are substantially higher than the costs with Illumina.

More recently, progress has been made in another SMS approach called nanopore sequencing. The concept, first described 16 years ago[29], of translocating a DNA strand through a limiting-diameter, membrane-bound biological pore while detecting the change in electrical conductance across the membrane as a function of the changing sequence content is beautiful in its simplicity but has proven to be extremely difficult in practice. The Oxford Nanopore technology now appears to be approaching feasibility, with decreased error rates and the development of specialized assembly algorithms more tolerant of the range of error types and read lengths produced by this device (http://arxiv.org/abs/1512.01801/). Additional challenges in nanopore sequencing include scalability in throughput and improved robustness. To date, this platform has not demonstrated sufficient data quality or read lengths to produce human assemblies, but such a potential exists.

Finally, a new preparatory approach to achieve long stretches of contiguous haplotype-resolved sequencing data from complex genomes has recently been commercialized by 10X Genomics (http://www.10xgenomics.com/). This library preparation concept provides a methodology and device to generate libraries from individual long DNA fragments by inserting specific barcodes throughout the length of each fragment, and then Illumina SBS reads are produced from the resulting library. These data are subsequently analyzed through specific proprietary algorithms to provide contiguous read assemblies for further analysis. In particular, the device begins with longer fragmented DNA molecules (as is done for the other single-molecule libraries) but isolates each individual DNA molecule into a micelle that is subsequently fused with another micelle containing DNA-barcoded random oligomers, nucleotides and polymerase. The resulting mixture is incubated to produce a set of barcoded fragments in each micelle that are short complementary copies of the original long input fragment. The collection of micelle-derived barcoded amplimers from each fragment are pooled, and an Illumina sequencing library is generated. After sequencing-data generation, the barcoded fragments are computationally isolated into read pools that are then assembled to create contiguous sequences representing the original long fragment. The resulting long fragments can be used to produce haplotype-resolved assemblies of chromosomal segments. One potentially notable advantage of this approach is the ability to generate whole-human-genome coverage from very low input amounts of fragmented genomic DNA (1 ng), owing to the library-construction efficiencies in micelles.

Future prospects for sequencing technology

These recent developments coalesce around a trend in NGS-based technology development, namely clever variations on

existing locus-specific techniques that now permit genome-wide readout. Examples include ChIP–seq (antibody-specific pulldown of proteins bound to chromatin)[30], RNA immunoprecipitation (RIP)–seq (protein–RNA interactions)[31], assay for transposase-accessible chromatin (ATAC)–seq (transposon insertion of barcoded sequencing adapters into open chromatin)[32], methyl-seq (bisulfite conversion of 5-methylcytosine residues on DNA)[33] and many more. These techniques reinforce the critical interplay among preparatory methods, sequencing-data production and the corresponding computational analysis. By integrating results from different NGS data sets, higher-level interpretations of cell- or tissue-specific changes in disease or health can be obtained; for example, shifts in chromatin packaging identified by ChIP–seq data analysis can be compared with corresponding RNA-seq data to interpret how changing chromatin packaging results in either silencing or expression of specific genes. Indeed, short-term computational challenges include developing tools for genome-wide integration of such data sets, thus leading to higher-level understanding and prediction of genome-scale biology.

The future may bring new and innovative platforms for sequence-data generation. In particular, this research and development area invites great innovation as well as interest from investors. I would argue, however, that regardless of the emergence of any new platform, the excitement around sequencing is not simply focused on instrumentation but instead is focused on the concept of generating integrated data sets that combine the results of multiple preparatory, readout, analysis and interpretation experiments for a given model organism or disease state, and are suitable for addressing different biological questions. Beyond the application of such solutions to basic research, the translation of these new sequencing technologies into the clinical realm highlights their transformative potential. Integrated sequencing solutions may provide a new type of clinical diagnostic assay that can add valuable evidence to that derived from conventional clinical assays or indeed may replace existing less comprehensive, less precise or less cost- and time-effective approaches. There are already different integrated systems being reported that are poised for clinical translation, to complement and enrich conventional clinical diagnostic assays.

One recently published example involves deep clinical phenotyping of 41 people diagnosed with neurometabolic disorders (i.e., combined intellectual developmental delay and undiagnosed metabolic abnormalities) coupled with exome sequencing and analysis using an integrated computational pipeline[34]. The resulting evaluation has identified new candidate genes, new variants in known metabolic genes, known genes linked to new metabolic anomalies and rare coexisting monogenic conditions. Overall, 68% of the affected individuals in the study obtained a diagnosis, and 44% were found to be potentially treatable. Had the diagnosis been available at their birth, many of these patients might have been spared their resulting symptoms or might have experienced substantially milder symptoms. Clinical translation of such results may allow for newborn screening that combines conventional blood-spot cards with indicated metabolite screening in blood or urine samples, coupled with exome sequencing and integrated analysis (http://www.tidebc.org/Ph/physicians.html). The resulting decrease or elimination of phenotypic sequelae for such patients would probably be substantial, as would the resulting cost savings in their long-term clinical care.

Similarly, in cancer care, analysis of NGS-based mutational profiling of tumor-derived DNA is gaining clinical acceptance for several endpoints. First, NGS-based mutational profiling can be used to identify targeted therapies that inhibit known cancer driver mutations[35,36]. Second, NGS panels have increasingly been used to identify underlying genetic susceptibility to developing cancer, both inherited and *de novo*[37]. And third, recently, the mutational load of cancer patients has been explored as a predictor of the response to checkpoint-blockade immune therapies[38,39] or even for personalized vaccine design[40,41]. A recent study by Stadler *et al.*[42] has illustrated how a single cancer-gene-panel NGS assay result can be broadly applicable as an assay for all three types of clinical decision points. This cost-effective approach may provide an efficient means of maximizing the diagnostic yield from precious biopsy materials, which often are in limited supply.

These results provide a glimpse into the future of genomics that has been enabled by the past ten years of sequencing-technology developments, thereby encouraging future applied research combining genome-wide preparatory techniques with combinatorial analytical approaches of the corresponding sequence data to reveal the underlying biology. To maintain this ever-accelerating trajectory of discovery, it will be essential to ensure that data sets from NGS experiments and their corresponding phenotypic data are made accessible; that relevant open-source analytical software is correctly packaged for convenient, easy local installation and user feedback; and that databases of variants identified from genomic studies are compiled and made available, to provide a source of relevant genomic alterations in the disease or model organism of interest. This powerful combination of resources should greatly facilitate the further acceleration of genomic studies, thereby enhancing understanding of biological and biomedical systems.

ACKNOWLEDGMENTS The author wishes to acknowledge her PhD mentor, B.A. Roe, whose encouragement and enthusiasm for technology and its applications to biology have inspired her career.

AUTHOR CONTRIBUTIONS E.R.M. conceptualized, wrote and edited the manuscript in its entirety.

COMPETING FINANCIAL INTERESTS The author declares competing financial interests: details are available in the online version of the paper.

1. Mardis, E.R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem. (Palo Alto Calif)* **6**, 287–303 (2013).
2. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
3. Head, S.R. *et al.* Library construction for next-generation sequencing: overviews and challenges. *Biotechniques* **56**, 61–64, 66, 68 passim (2014).
4. Metzker, M.L. Sequencing technologies: the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
5. Goodwin, S., McPherson, J.D. & McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

6. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
7. Bainbridge, M.N. *et al.* Whole exome capture in solution with 3 Gbp of data. *Genome Biol.* **11**, R62 (2010).
8. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
9. Hodges, E. *et al.* Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat. Protoc.* **4**, 960–974 (2009).
10. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
11. Springer, N.M., Xu, X. & Barbazuk, W.B. Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. *Plant Physiol.* **136**, 3023–3033 (2004).
12. Baetens, M. *et al.* Applying massive parallel sequencing to molecular diagnosis of Marfan and Loeys-Dietz syndromes. *Hum. Mutat.* **32**, 1053–1062 (2011).
13. Hollants, S., Redeker, E.J. & Matthijs, G. Microfluidic amplification as a tool for massive parallel sequencing of the familial hypercholesterolemia genes. *Clin. Chem.* **58**, 717–724 (2012).
14. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
15. Mardis, E.R. Sequencing the AML genome, transcriptome, and epigenome. *Semin. Hematol.* **51**, 250–258 (2014).
16. Wong, K., Keane, T.M., Stalker, J. & Adams, D.J. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* **11**, R128 (2010).
17. Alkan, C., Sajjadian, S. & Eichler, E.E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).
18. Huddleston, J. & Eichler, E.E. An incomplete understanding of human genetic variation. *Genetics* **202**, 1251–1254 (2016).
19. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
20. Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* **24**, 688–696 (2014).
21. Chaisson, M.J., Wilson, R.K. & Eichler, E.E. Genetic variation and the *de novo* assembly of human genomes. *Nat. Rev. Genet.* **16**, 627–640 (2015).
22. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
23. Goodwin, S. *et al.* Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res.* **25**, 1750–1756 (2015).
24. Pirola, Y. *et al.* HapCol: accurate and memory-efficient haplotype assembly from long reads. *Bioinformatics* **32**, 1610–1617 (2016).
25. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
26. Madoui, M.A. *et al.* Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* **16**, 327 (2015).
27. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
28. Chaisson, M.J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
29. Akeson, M., Branton, D., Kasianowicz, J.J., Brandin, E. & Deamer, D.W. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. *Biophys. J.* **77**, 3227–3233 (1999).
30. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
31. Zhao, J. *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell* **40**, 939–953 (2010).
32. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
33. Harris, R.A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.* **28**, 1097–1105 (2010).
34. Tarailo-Graovac, M. *et al.* Exome sequencing and the management of neurometabolic disorders. *N. Engl. J. Med.* **374**, 2246–2255 (2016).
35. Tsimberidou, A.M. *et al.* Personalized medicine in a phase I clinical trials program: the MD Anderson Cancer Center initiative. *Clin. Cancer Res.* **18**, 6373–6383 (2012).
36. Wagle, N. *et al.* High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer Discov.* **2**, 82–93 (2012).
37. Susswein, L.R. *et al.* Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing. *Genet. Med.* **18**, 823–832 (2016).
38. Le, D.T. *et al.* PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
39. Rizvi, N.A. *et al.* Cancer immunology: mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
40. Carreno, B.M. *et al.* Cancer immunotherapy: a dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* **348**, 803–808 (2015).
41. Fritsch, E.F., Hacohen, N. & Wu, C.J. Personal neoantigen cancer vaccines: the momentum builds. *OncoImmunology* **3**, e29311 (2014).
42. Stadler, Z.K. *et al.* Reliable detection of mismatch repair deficiency in colorectal cancers using mutational load in next-generation sequencing panels. *J. Clin. Oncol.* **34**, 2141–2147 (2016).