Review in Advance first posted online on June 9, 2016. (Changes may still occur before final publication online and in print.)

Advancements in Next-Generation Sequencing

Shawn E. Levy and Richard M. Myers

HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806;
email: slevy@hudsonalpha.org, rmyers@hudsonalpha.org

Annu. Rev. Genom. Hum. Genet. 2016. 17:16.1–16.21

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi: 10.1146/annurev-genom-083115-022413

Keywords

sequencing, whole genome, exome

Abstract

The term next-generation sequencing is almost a decade old, but it remains the colloquial way to describe highly parallel or high-output sequencing methods that produce data at or beyond the genome scale. Since the introduction of these technologies, the number of applications and methods that leverage the power of genome-scale sequencing has increased at an exponential pace. This review highlights recent concepts, technologies, and methods from next-generation sequencing to illustrate the breadth and depth of the applications and research areas that are driving progress in genomics.

INTRODUCTION

Since the fundamental discovery of the structure of DNA (128) and the pioneering development of methods to detect the sequence of DNA bases by foundational approaches such as Maxam & Gilbert's technique (76) and Sanger sequencing (106), the field of DNA sequencing has rapidly evolved in capacity, capability, and applications. As with many technologies, advances across multiple fields were brought together to achieve routine sequencing at the genome scale. The development of the polymerase chain reaction (103, 104), the widespread availability of high-quality nucleic acid–modifying enzymes, and the development of fluorescent automated DNA sequencing enabled the Human Genome Project to deliver the first draft of the human genome sequence in 2001 (64, 123) and the first completed draft three years later (54). Since then, genomics has evolved at an amazing pace. Dozens of next-generation sequencing companies and technologies have been created, and the corresponding field of bioinformatics has exploded as a major scientific and training discipline. DNA sequence has even been proposed as a highly efficient storage mechanism for large-scale data (22).

The progression from the discovery of the structure of DNA to the ability to sequence it as a routine assay has had several inflection points. In the mid-to-late 1990s, microarrays were developed as highly parallel assays to measure RNA and DNA (91, 107). Between 2001 and 2006, microarrays offered the first genome-scale parallel analysis of DNA and RNA. In 2006, second- and third-generation sequencing techniques began to emerge that permitted an unbiased means to examine billions of templates of DNA and RNA. Although now almost a decade old, the term next-generation sequencing remains the popular way to describe very-high-throughput sequencing methods that allow millions to trillions of observations to be made in parallel during a single instrument run.

Since 2006, there has been an explosion of new methods, techniques, and protocols for the examination of virtually any question in basic genetics or clinical research involving nucleic acid. The rapid evolution of instruments, chemistries, and techniques led to next-generation sequencing instruments changing within months and chemistries and analysis algorithms changing within weeks, creating substantial challenges for both researchers and clinicians. The challenges arising from such rapid changes were amplified by a lack of widely available biological and biochemical standards and public data sets to assess these nascent technologies and methods. Over the last few years, technology platforms have been used and tested across a broad user market in a wide variety of research projects, helping the methods and instruments to mature and enabling a diversity of publications, methods, and applications of sequencing technology. Thousands of application, technical, informatic, and translational articles have been published that describe the use of sequencing technologies, with many hundreds more added each year.

Several excellent reviews over the last several years have described the technological landscape of sequencing (78, 80, 96). When examined in chronological order, these and other examples provide a superb history of the changing sequencing space and the amazing pace that has brought us from the first draft of the human genome sequence to the ability to routinely sequence human genomes with widely available technology at a cost decreasing from billions of dollars to thousands of dollars in less than 25 years. **Table 1** summarizes the past, present, and future of commercially launched sequencing platforms and their original references.

This review focuses on advancements in several areas, with the caveat that the breadth and depth of the field make it impossible to be comprehensive. The exclusion of any particular area or advance reflects not a lack of impact or importance but rather the sheer volume of information available and the desire to highlight specific areas. **Table 2** lists several excellent Internet resources, including blogs and electronic journals, that expand on the information in this review and are frequently updated with the latest developments and information.

Table 1 Summary of second-generation sequencing manufacturers

Manufacturer	Amplification	Detection	Chemistry	URL	Reference(s)
Commercial					
Illumina	Clonal	Optical	Sequencing by synthesis	**http://www.illumina.com**	12, 26, 47
Oxford Nanopore	Single molecule	Nanopore	Nanopore	**http://www.nanoporetech.com**	10, 55, 95, 125,
Pacific Biosciences	Single molecule	Optical	Sequencing by synthesis	**http://www.pacb.com**	16, 17, 29, 30, 33, 35, 60, 67, 71, 108, 117, 120
ThermoFisher Ion Torrent	Clonal	Solid state	Sequencing by synthesis	**http://www.thermofisher.com/us/en/home/brands/ion-torrent.html**	70, 77, 101
Precommercial					
Quantum Biosystems	Single molecule	Nanogate	Nanogate	**http://www.quantumbiosystems.com**	—
Base4	Single molecule	Optical	Pyrophosphorolysis	**http://base4.co.uk**	—
GenapSys (GENIUS)	Clonal	Solid state	Sequencing by synthesis	**http://www.genapsys.com**	—
QIAGEN (GeneReader)	Clonal	Optical	Sequencing by synthesis	**http://www.qiagen.com**	—
Roche Genia	Single molecule	Solid state	Nanopore	**http://geniachip.com**	—
Postcommercial					
Roche 454 (GS FLX)	Clonal	Optical	Sequencing by synthesis	**http://www.454.com**	75
Helicos BioSciences (Heliscope)	Single molecule	Optical	Sequencing by synthesis	—	48
Dover (Polonator)	Clonal	Optical	Sequencing by ligation	—	109
ThermoFisher Applied Biosystems (SOLiD)	Clonal	Optical	Sequencing by ligation	**http://www.thermofisher.com/us/en/home/brands/applied-biosystems.html**	121
Complete Genomics	Clonal	Optical	Sequencing by ligation	**http://www.completegenomics.com**	27

Dashes indicate that no URL or reference is available. Platforms listed as precommercial have been announced but at the time of writing have not been formally launched; platforms listed as postcommercial are no longer commercially available as new instrument sales.

SEQUENCING PLATFORMS AND CAPABILITIES

Current as well as some past commercially available sequencing platforms can be generally divided via three axes. Axis one is single-molecule detection per reaction, well, or sensor, as performed by Pacific Biosciences and Oxford Nanopore platforms, or the detection of clonally amplified DNA, as performed by Illumina, Ion Torrent, and Roche 454 platforms, among others. Axis two is the use of optical detection to make sequencing base calls, as performed by Illumina, Pacific

Table 2 Examples of well-known online resources for genomic technologies

Name	URL
Next Gen Seek	**http://nextgenseek.com**
Bits of DNA	**http://liorpachter.wordpress.com/seq**
RNA-Seq Blog	**http://www.rna-seqblog.com**
Journal of Next Generation Sequencing & Applications	**http://www.omicsonline.org/next-generation-sequencing-applications.php**
CoreGenomics	**http://core-genomics.blogspot.com**
Next-Gen Sequencing	**http://nextgenseq.blogspot.com**
Omics! Omics!	**http://omicsomics.blogspot.com**
In Between Lines of Code	**http://flxlexblog.wordpress.com**
Kevin's GATTACA World	**http://kevin-gattaca.blogspot.com**
Blog @ Illumina	**http://blog.illumina.com**
Next Generation Technologist	**http://www.yuzuki.org**

Biosciences (detection of fluorescently modified nucleotides), and Roche 454 (detection of light via pyrosequencing) platforms, or nonoptical detection, as performed by Ion Torrent (detection of the release of H^+ during a polymerization reaction via a solid-state sensor) and Oxford Nanopore (measurement of the translocation of DNA through a nanopore sensor) platforms. The third axis is the use of a polymerase or ligation process to drive a sequencing-by-synthesis reaction in which the products of the reaction are measured to produce sequencing data, or direct measurement of DNA molecules. Sequencing-by-synthesis reactions performed by Illumina, Ion Torrent, Pacific Biosciences, and Roche 454 platforms utilize a polymerase reaction, whereas the former Applied Biosystems SOLiD platform and the Polonator platform use a ligation-mediated synthesis. Direct measurement of DNA sequences is performed by the Oxford Nanopore platform. Each commercially available platform has similarities and differences relative to the others depending on the chemistries and detection methods used. These similarities and differences result in a spectrum of capabilities and specifications that lead to different strengths and weaknesses among the platforms. The differences between the platforms, particularly in their limitations, have resulted in frequent comparisons to evaluate their performance under similar conditions (94). It has also become efficient to use multiple platforms in a single experiment, with the goal of capitalizing on the strengths of each platform (60).

The two most common specifications used to compare platforms are the number of reads produced in a given instrument run and the length of those reads. Other metrics—such as cost per run, cost per base, instrument run time, presequencing sample preparation time, sample preparation cost, and platform bias or error modes—are far more difficult to compare across multiple platforms owing to the number of variables involved and great debate about how to consider those factors. Although illustrating the available platforms based on the number of reads and read length has limitations, it does provide a useful picture of comparative output and a convenient way to compare changes over time. **Figure 1** shows a graph of the commercially available instruments in terms of read length and depth. It includes the now discontinued Roche 454 platform for a point of comparison and because it was such a foundational platform in the development of the next-generation sequencing field. The graph is taken from an actively updated blog by Lex Nederbragt at the University of Oslo (**http://flxlexblog.wordpress.com**); the comprehensive and active updates provide confidence that the graph will continue to be updated in the future to illustrate the dynamic and changing nature of sequencing.

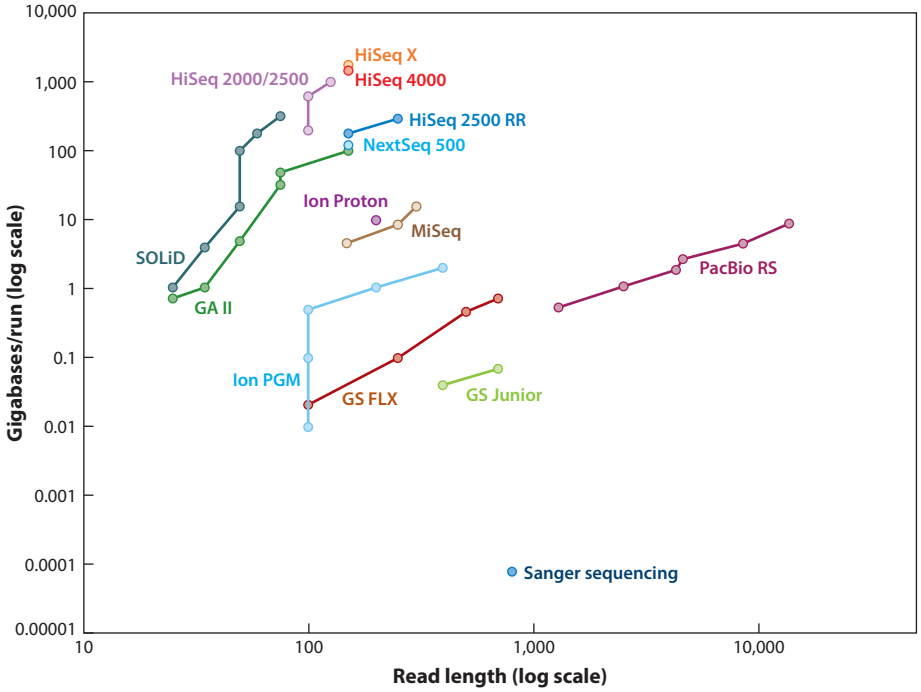

Figure 1

Developments in high-throughput sequencing. SOLiD is an Applied Biosystems platform; Ion PGM and Ion Proton are Ion Torrent platforms; GA II, HiSeq, NextSeq, and MiSeq are Illumina platforms; GS FLX and GS Junior are Roche 454 platforms; and PacBio RS is a Pacific Biosciences platform. Adapted from a figure created by Lex Nederbragt (**http://dx.doi.org/10.6084/m9.figshare.100940**) under the Creative Commons Attribution license (**http://creativecommons.org/licenses/by/4.0**).

The sequencing world is a dynamic and unforgiving space. Only a few short years ago, the winter of 2011–2012 brought tremendous hype and overpromise. At the annual Advances in Genome Biology and Technology conference, Oxford Nanopore announced that their GridION sequencing platform would sequence a human genome in 15 minutes and be commercial available by mid-2012. Swiss drug giant F. Hoffmann–La Roche (Roche) launched a takeover bid for Illumina. In the same month, Ion Torrent stated that by the end of the year they would begin selling a machine that could sequence an entire human genome in a day for less than $1,000, and Pacific Biosciences was approaching the one-year anniversary of their commercial launch. Many of those announcements proved to be aggressive and unrealistic, and the development efforts proved to be far more challenging than the manufacturers expected. However, 2015 saw a resurgence of platforms and innovations. Illumina has proven the viability and efficiency of their ultra-high-output HiSeq X and introduced patterned flow cells on the HiSeq X and HiSeq 3000/4000 platforms, likely paving the way for increased output from those instruments in the future. Pacific Biosciences announced the details of their second commercial platform, the Sequel, which has six to seven times the data output of their existing platform while decreasing the cost of the instrument by half. Oxford Nanopore has had a successful early access program, with their MinION sequencer being used by more than 1,000 researchers, and will launch the second iteration of the MinION in 2016 along with a higher-output instrument tentatively named the PromethION. Ion Torrent is launching the third iteration of its technology in the Ion S5 and Ion

S5 XL. These systems include the same core instrument, but the Ion S5 XL adds local computing to enhance analysis speed. Taking a step back from the whole-genome market, these Ion Torrent instruments are aimed at targeted sequencing workflows, with the goal of greatly simplifying the hands-on time. The first released applications will be amplicon based via Ion AmpliSeq technology, a well-tested and robust multiplex amplification method for creating sequencing templates (14, 119), and will require a reported 45 minutes of hands-on time when coupled with the supporting Ion Chef system. The output of both instruments will use one of three chips, with outputs per chip ranging from 600 Mb to 15 Gb and read lengths of 200 or 400 nucleotides (nt), although current specifications list the highest-output mode (10–15 Gb) as being limited to 200-nt reads.

Now more than two years old, the Illumina HiSeq X system remains the highest-output platform and the only sequencing technology available that can generate highly accurate data that allow sequencing at the human genome scale at reagent costs under $1,000. In 2015, the chemistry and flow cells on the platform were improved to increase the consistency in the amount and quality of the data produced by the instrument. Initially launched as a human-only platform, the HiSeq X can now be used with any species, provided that the targeted coverage per sample is at least 30×. The limitation for whole-genome use and the minimum coverage requirements per sample are contractual stipulations by Illumina and not based on any specific technological limitations. The capabilities of the HiSeq X have resulted in more than 200 instruments installed or sold during the last year, providing unprecedented sequencing scale to the worldwide markets. Several large-scale sequencing projects are ongoing or in the planning stages. Significant among them for both the volume of data produced and their transformative impact on genomics are the 1000 Genomes Project (1) and the Exome Sequencing Project (36, 115). In addition to the large-scale efforts, the available scale and platform capabilities have resulted in the development of several population-scale sequencing efforts. The two largest announced to date are the 100,000 Genomes Project, led by Genomics England (**http://www.genomicsengland.co.uk**), and the GenomeAsia 100K Initiative, led by the nonprofit consortium GenomeAsia 100K (**http://genomeasia100k.com**). Genomics England is a company founded and owned by the UK Department of Health. Using England's National Health Service, the project plans to perform whole-genome sequencing of 100,000 participant samples over a four-year period. The GenomeAsia 100K Initiative plans to also sequence 100,000 individuals. The project will initially include populations from 12 South Asian countries and at least 7 North and East Asian countries, in order to increase understanding of the population history and population substructure of the region and contribute to health and disease research.

Iceland has long been investing in the genetic and genomic analysis of its population, primarily through efforts supported by deCODE Genetics (now owned by Amgen) (44, 45). These efforts have been expanded to whole-genome sequencing of the Icelandic population, and the first publication from the group illustrated the potential of population sequencing as well as the power of combining population-scale data sets. Gudbjartsson et al. (43) described the insights gained from sequencing the whole genomes of 2,636 Icelanders to a median depth of 20×. The authors reported the discovery of 20 million single-nucleotide polymorphisms (SNPs) and 1.5 million insertions or deletions. The discovered variants were annotated with respect to functional impact (gene position, protein sequence impact, pathway, and conservation score), frequency, and density. The relatively small and homogeneous population of Iceland allowed an elegant and effective combination of existing microarray-based genotyping data and the newly annotated and phased variants to be imputed using the sequencing data, resulting in 104,220 individuals imputed down to a minor allele frequency of 0.1%. The combined results revealed in the Icelandic population a recessive frameshift mutation in *MYL4* that causes early-onset atrial fibrillation, several mutations in *ABCB4* that increase risk of liver diseases, and an intronic variant in *GNAS* associated with

increased thyroid-stimulating hormone levels when maternally inherited (43). These clinically relevant results illustrate the immediate value of this work, and the imputed genomes support powerful tests of association with an extensive range of traits and phenotypes, including parent-of-origin models as well as the recessive model. In combination, these data provide the foundation for more studies at the population scale that will enable a deep and robust understanding of how variation in the sequence of the human genome gives rise to human diversity.

The available sequencing power is not only being applied to humans at a grand scale; massive projects are under way to examine many other species. One exemplary project is the 100K Pathogen Genome Project, launched in mid-2012 by Bart Weimer at the University of California, Davis. This project aims to sequence the genomes of 100,000 infectious microorganisms to create a database of bacterial genome sequences for use in public health, outbreak detection, and bacterial pathogen detection. This information will be used to accelerate the diagnosis of food-borne illnesses and identify and trace the origins of pathogens more quickly, with the goal of better understanding and shortening infectious disease outbreaks. The project is a public-private collaborative project started by the University of California, Davis; Agilent Technologies; and the US Food and Drug Administration. The US Centers for Disease Control and Prevention and the US Department of Agriculture are notable collaborators. The project will enable worldwide collaboration to identify sets of genetic biomarkers associated with important pathogen traits and result in a public database as part of the National Center for Biotechnology Information resources.

These population-scale sequencing efforts are extensions of foundational projects such as the 1000 Genomes Project, the Cancer Genome Atlas (**http://cancergenome.nih.gov**), and the related Encyclopedia of DNA Elements (ENCODE) project. The 1000 Genomes Project recently published data from phase 3 of the project (113), discussing the completion of the project and its description of more than 88 million variants from 2,504 individuals from six populations, with all variants phased into high-quality haplotypes. The latest phase of the 1000 Genomes Project provided a broad diversity of data to observe what a "typical" genome looks like in different populations, a vital advancement toward effective personalized genomics or personalized medicine. There are an ever-increasing number of consortium-based projects, some of which are mentioned here and others of which have been reviewed elsewhere (see table 2 of Reference 96). These projects reflect the changing challenges and opportunities for team-based science with respect to sample availability, sequencing and data-sharing technologies, and funding resources. The efficiencies and impacts these and many other consortium projects bring to basic and translational research are profound. It is now inconceivable to perform any type of genomic analysis without using data from one or more consortium projects, be it a reference genome, a variant frequency, a sequence search, or any number of other data types available in the public domain.

SHORT- AND LONG-READ SEQUENCING

The amazing increase in data output and decrease in cost per base sequenced has been driven primarily by increases in parallelization in short-read sequencing technologies such as the Illumina and Ion Torrent platforms. Although there have been increases in read length, the highest-output platforms continue to have relatively short read lengths, on the order of 35–300 bases per read. As in many areas of genomics, the revisions to read lengths occur rapidly and will likely continue to do so as chemistries are optimized and improved. Illumina compensates for its short read lengths by supporting paired-end sequencing, in which each end of the same DNA molecule is sequenced to the full read length. Because the approximate size of the insert is known, the paired-end information greatly improves unique alignment rates compared with single reads alone. The dominance of the Illumina platform both in the literature and in the amount of data from the

platform submitted to the Sequence Read Archive (65) demonstrates its efficiency and power. The Illumina platform has proved to be powerful for both resequencing approaches (such as whole-genome and whole-exome sequencing) and read counting applications [such as RNA sequencing (RNA-seq) and chromatin immunoprecipitation sequencing (ChIP-seq)]. The genomics field has contributed dozens of methods for template preparation on the Illumina platform, leading to more than 50 different preparatory methods for template generation on this platform (53).

Although the first draft of the human genome sequence was completed in 2001, and a much more complete draft was reported in 2003, many areas of the genome remain poorly characterized or missing from the current assembly. This is due to challenges and biases in preparing, characterizing, and sequencing DNA, spanning each point of sample manipulation, extraction, sequencing, and analysis. Together, these regions represent the darker areas of the genome, where sequences are substantially more difficult to resolve with short-read technologies or have never been well resolved in the references. Several developments have expanded the diversity of technologies and applications available to bring light to these dark regions or enable the efficient de novo sequencing or characterization of nontraditional species.

Two commercially available, highly parallel sequencing technologies produce long sequencing reads: Instruments from both Oxford Nanopore and Pacific Biosciences produce read lengths in the thousands of bases per read. Both utilize single-molecule sequencing, albeit with very different detection methods. As its name implies, Oxford Nanopore uses nanopores for detection, whereas Pacific Biosciences uses optical detection of a sequencing-by-synthesis reaction that occurs inside a zero-mode waveguide (67). The details of the chemistries used on both platforms have been well reviewed elsewhere (96).

Although Illumina has a dominant position in terms of both current market share and the amount of sequence their platforms can output, there are limitations to the resolution that short-read technologies bring to many applications in genomics. De novo sequencing is probably the most widely appreciated limitation of short-read sequencing, followed by the resolution of structural variations in the genome. Short-read sequencing data can be effectively used to investigate structural variation, especially when applied in combination with genotyping data, as in a study recently published by the Structure Variation Analysis Group of the 1000 Genomes Project (113). However, appreciating the full resolution of genomic variation is assured only when a complete, reference-free, de novo assembly of a genome is possible (reviewed in detail in 18). Using significantly longer sequencing reads provides a way to increase resolution for assembly or structural variation. Short- and long-read technologies have been at opposing ends of the spectra for read length and read density. Illumina's two highest-output platforms, the HiSeq X for genome sequencing and the HiSeq 4000 for more general applications, are limited to 150-nt reads but can output more than 6 billion paired-end reads or more than 12 billion total reads per instrument run. The Pacific Biosciences platform, the most widely proven long-read technology, produces approximately 880,000 reads per 16 SMRT (Single-Molecule Real-Time) cell instrument run but at a read length that averages more than 10,000 nt and can exceed 40,000 nt in length. The current P6-C4 chemistry has a per-base error rate of approximately 15%, but the stochastic nature of the errors allows for highly accurate consensus sequencing, both when using the circular consensus sequencing (69, 72) and when generating consensus reads by sequencing samples to multiple times depth on the Pacific Biosciences platform.

The Pacific Biosciences RS platform was released in 2010 and has since undergone several iterations and chemistry revisions. Template preparation involves ligating hairpin adapters on either end of a DNA molecule, with the length of the DNA molecule defining the maximal read length of the sequencing run. These capped templates are called SMRTbells; a sequencing-by-synthesis reaction occurs on the SMRTbell template that is detected with an optical system via

a zero-mode waveguide (29). Because the SMRTbell was created with a hairpin at either end, a strand-displacing polymerase can sequence the template several times, providing multiple reads of each base of the template and increasing the accuracy of the read (72, 117). This approach has been applied to several areas, including assembly of chloroplast genomes (69). The circular consensus method is not limited to Pacific Biosciences sequencing and has been described as a preparatory method valuable in sequencing RNA viruses to very high accuracy, allowing the detection of ultrarare variants and accurate measurement of low-frequency variants (2). The long sequence reads of the Pacific Biosciences platform have allowed the analysis of challenging areas of the genome, such as the major histocompatibility complex (MHC) class I region transcripts (19, 129) and regions of segmental duplication (52). Studies performed to generate de novo assemblies have also illustrated the impact of the platform and its potential role in developing routine analysis of human genomes driven by de novo assembly rather than comparisons to a reference sequence (18, 31).

In late 2015, Pacific Biosciences announced a new platform to augment their RS II instrument. This new platform, named Sequel, is a significant change from the RS in both form and capabilities. It is a fraction of the size of the original RS platform and has a capital cost that is half that of the original RS. The platform will reportedly launch with a substantial increase in read density compared with the available RS, with each SMRT cell having 1 million zero-mode waveguides (compared with 150,000 on the RS II), increasing the read output by approximately seven times. Taken together, these improvements will reduce the cost of sequencing a 20× human genome by at least 50% and provide a sevenfold improvement in the speed of data production compared with the RS II. The Sequel is being developed in collaboration with Roche, and the first instruments will be provided to Roche for development of human in vitro diagnostics.

The use of nanopores to sequence DNA has been discussed or demonstrated in various forms since at least 1996 (59), and the potential and challenges of nanopore sequencing have been well reviewed elsewhere (13). The promise in nanopore sequencing is its sensitivity to native molecules and the potential to use inexpensive materials and reagents in the process. Nanopore sequencing is based on measuring changes in electrical properties as biomolecules such as DNA traverse the pore and then using those electrical changes to identify the exact DNA base going through the pore. Oxford Nanopore was the first company to commercialize nanopore sequencing technology, which they did in their handheld sequencer, the MinION. The MinION broke many barriers at its launch. It was the first DNA sequencer that could be handheld and not require anything more than an active USB port to operate. It is also the lowest-cost DNA sequencer to be released, with an instrument price of $1,000. These features remain exceptional in the genomics world.

Oxford Nanopore has been a somewhat nontraditional biotechnology company. It has alternated between grandiose and stealthy, and it introduced its instrument to the world via an early access program that allowed those selected for the program to acquire the sequencer for a deposit of $1,000. This program had robust online community support and by nearly all measures was a reasonable success. It provided an ever-expanding number of users with access to the platform while also providing the company with an ever-increasing amount of data and feedback to work with and base improvements on. As stated above, the detailed chemistry of the Oxford Nanopore platform has been reviewed in detail elsewhere (96).

As with the other commercially launched platforms, several features make the Oxford Nanopore platform unique. Sequencing is performed on template DNA by measuring the translocation of the DNA through a protein pore. The current MinION platform generates approximately 100 Mb of data per 16-hour run, with an average read length of approximately 6 kb (10). Oxford Nanopore has announced that a next-generation version of the MinION will be released in 2016, along with the higher-output PromethION and an automated sample preparation device called Voltrax that

will connect to either platform (58). Oxford Nanopore has reported advancements that will allow its platforms to break the gigabase and terabase barriers per run and has announced a pay-as-you-go pricing system for the new platforms. It will be interesting to see whether these changes have any fundamental impact on the genomics field.

During the early access program, several studies illustrated the capabilities of the Oxford Nanopore platform, including combining nanopore data with Illumina data to produce a hybrid assembly (42, 97). Library preparation methods that support sequence capture and sample indexing have been described (57), and there has been rapid development of analysis algorithms for long-read sequencing data. These include scaffolding methods for assembly of draft genomes (126), tools for real-time visualization and analysis of MinION data (15), and error corrections to drastically improve read accuracy (55, 114).

Long-read sequencing has also been robustly applied to sequencing full-length transcripts with both the Pacific Biosciences and Oxford Nanopore platforms. Pacific Biosciences launched its SMRT Analysis 2.2 software with support for the Iso-seq method of analyzing full-length transcripts and gene isoforms, without the requirement of assembly. Several studies have utilized this approach, including analyses of human pluripotent stem cells (11), allele-specific transcription (116), and alternative splicing (118). It will be interesting to see how the transcriptional analysis field evolves as long-read sequencing becomes more available and cost effective, particularly in light of a study that raised questions about errors in short-read data and transcriptional analysis (99). Although this paper is thought provoking, it raises as many questions about analysis techniques and algorithms as it does about sequencing technology. As more options and more transcriptional data from long-read sequencing become available, a more thorough evaluation will be possible to determine the balance between the lower read output (and therefore lower dynamic range) of long-read technologies compared with the lower alignment accuracy and relative insensitivities for all splice variants of short-read technologies.

SYNTHETIC LONG READS

The generation of long sequencing reads is not limited to direct measurement using long-read technologies such as the Pacific Biosciences and Oxford Nanopore platforms; several innovative and elegant approaches have combined biochemical and informatic approaches with short-read sequencing data to generate synthetic long reads. These methods all rely on partitioning the genome to a subhaploid concentration and then generating a sequencing library that can be uniquely mapped back to the subhaploid fraction. Early methods used fosmid libraries to partition a genome sample (28), whereas later methods have relied on the use of a diversity of synthetic sequences added as barcodes in a manner that allows differentiation of hundreds to hundreds of thousands of sequences based on the barcodes. Several methods have been applied with varying complexity and resolution. Several synthetic long-read technologies and methods were described in 2012. The long-fragment-read method illustrated sequencing and haplotyping from 10–20 human cells (92). Another long-fragment technology was commercialized from Stephen Quake's laboratory at Stanford University as Moleculo (63). Moleculo was acquired by Illumina and is now available in a kit form. The Moleculo technology was used to phase a human genome to greater than 99% completeness with N50 phase blocks in the 400–500-kb range (63). The N50 statistic refers to the contig length produced following data analysis and is similar to the mean or median length of the resulting contigs. The formal definition of N50 is the length for which the collection of all contigs of that length or longer contains at least half of the sum of the lengths of all contigs, and for which the collection of all contigs of that length or shorter also contains at least half of the sum of the lengths of all contigs. Illumina independently published an approach that targeted

a 1-Mb region of the X chromosome, phasing more than 95% of SNPs and deriving haplotype blocks of hundreds of kilobases (56).

More recently, a transposase-mediated library preparation of a subhaploid fractionated genome that leverages the unique contiguity-preserving activity of the Tn5 transposase (CPT-seq) has been described (3) and applied to whole-genome sequencing (8). Whereas the long-fragment-read and Moleculo technologies were limited to hundreds of partitions, CPT-seq uses 9,216 barcode pools via combinatorial indexing. 10X Genomics recently commercialized another method via their GemCode platform, extending the number of subhaploid partitions that can be resolved to 750,000. GemCode uses a microfluidic assay to dropletize high-molecular-weight DNA into approximately 100,000 droplets and combines each droplet with a dissolvable bead known as a Gem. Each Gem contains oligonucleotides with a single barcode sequence that is introduced in the sequencing library preparation method. The unique barcode is used following the sequencing data generation to partition the sequencing reads and provide phasing and structural variation analysis. In publications describing results with the GemCode platform, the example data released show more than 99% of SNPs phased and haplotype block sizes of more than 12 Mb.

Applications for synthetic long reads extend beyond haplotype phasing. Kuleshov et al. (62) used synthetic long-read methods to analyze the human microbiome and identified 51 additional species that were not observed with short-read sequencing alone. Additionally, these synthetic long reads revealed extensive intraspecies variation, providing a resolution to the microbiome data that was previously unobtainable.

TOWARD A REFERENCE-FREE ANALYSIS

Many new methods, technologies, and algorithms have emerged that can provide routine, efficient synthesis of very long fragments of DNA, either as synthetic reads or through direct sequencing. These methods promise to dramatically increase our understanding of almost all fields of biology, including genomes, epigenomes, transcriptomes, epitranscriptomes, and metagenomes. Nearly all of these methods rely on an analysis method that is rooted in the initial alignment of the sequencing data to a linear reference genome. Representing genomic information in tracks in a uniform manner that allows flexible, extendable track types is critical for the continuing advancement of genomics (46, 102).

Diploid organisms receive two sets of chromosomes, one from each parent. The sequences of these chromosomes are highly similar but differ at all positions that show heterozygous variation. These variations can be at the single-nucleotide level or at the insertion/deletion and chromosomal rearrangement/translocation levels. Representing the diploid genome as a single linear sequence requires removing much of this variation and using additional reporting or representation methods beyond the linear representations. Complicating the issue is that any single sample may have genomic content that is not represented in the reference, regardless of the representation method. In other words, any sequencing reads that are unique to an individual and are not part of the reference may be missed, and any single haploid reference is insufficient to fully represent genomic diversity (21). The most recent release of the human genome (build 38, GRCh38) contains so-called alternate loci, defined as "multiple representations for regions that are too complex to be represented by a single path" (40). These are regions, often of modest size, that are found in some genomes but are not part of the reference representation. Alignment software is being revised to accommodate these alternative representations and allow for more comprehensive mapping and flexibility. The nearly ubiquitously used alignment program Burrows-Wheeler Aligner (BWA) has detailed capabilities for dealing with these alternate loci (68). A new alternate loci–tolerant aligner (SRPRISM) and companion reference-guided assembly tool (ARGO) have recently been described

as part of a study to assemble a single-haplotype genome derived from a hydatidiform mole (112). The assembly of the haploid DNA from the hydatidiform mole has added to the curation framework available for genome annotation and assembly and provided at least one accurate allelic representation across loci with a complex genomic architecture. These data facilitate the understanding of the genomic architecture and diversity in complex regions, including regions such as the immunoglobulin heavy-chain locus (127).

Although the best way to robustly use and represent multitrack reference information as a step toward a reference-free assembly remains an open and active area of study, techniques such as graph-based representation remain interesting potential solutions. Nearly all assembly programs use graph representations and graph algorithms to assemble reads into a genome representation, and using graphs to represent DNA sequences has a long history; for example, string graphs (82) and De Bruijn graphs (23) are used in this context. As discussed in detail by Church et al. (21), the newly formed Global Alliance for Genomics and Health (**http://genomicsandhealth.org**) is leading an effort to formalize data structures for graph-based reference assemblies. These efforts will require developing infrastructure and analysis tools to support these new structures and achieve widespread adoption across the biological and clinical research communities. Although these efforts will likely take many months to years for full adoption, the steady progress and direction of advancing beyond a linear haploid reference represent an important step.

UBIQUITOUS USE OF SEQUENCING

As mentioned above, the sheer volume of publications and applications enabled by sequencing approaches has grown at a staggering rate. The ubiquitous availability of sequencing technology has led not only to an amazing array of research applications, but also to the rapid development of laboratories offering sequencing for clinical testing purposes. There are also a growing number of inspiring stories of how sequencing has led to transformative results for patient care. For example, Elana Simon was diagnosed with a rare form of liver cancer (fibrolamellar hepatocellular carcinoma) and participated in research that revealed a gene fusion event that appears to drive her type of cancer. Elana's father is a researcher at Rockefeller University, and while doing research for a high school internship, Elana used social media to identify others with the same rare cancer, leading to a cohort of 15 tumor/normal pairs for analysis, including her own sample. Sequencing of the cohort revealed that all of the tumors contained a novel gene fusion between *DNAJB1* and *PRKACA*, producing a fusion protein that retains kinase activity. Elana was a coauthor on two publications describing this work (51, 110). This example highlights not only how powerful the ubiquitous access to sequencing tools has become, but also how accepted they are becoming as a transformative tool in health care. Although not all studies examining rare diseases or cohorts of tumors with sequencing technologies will have as compelling an outcome, there is no doubt that sequencing will play an increasing role in research, health care, and industrial experiments and that the number of available applications will continue to grow with the innovation and creativity of the scientific community. Erlich (31) has published an interesting review of sequencing that focuses on the barriers that remain to the ubiquitous use of sequencing sensors, including sequencing at home, in forensics, and in security applications.

The application of genomic tools to human cancer has been an exceptionally active area of research and development since the earliest days of the field. Sequencing applications, both genome-wide and targeted, are revealing complex mutational signatures associated with different types of cancers and revealing complexities and molecular signatures of cancer that are now driving both research and therapeutic decisions (5–7, 49, 87). The results of these mutational signature studies are available in the online Catalogue of Somatic Mutations in Cancer project from the Wellcome

Trust Sanger Institute (**http://cancer.sanger.ac.uk/cosmic**). Among several excellent reviews that have discussed progress in understanding the cancer genome landscape are articles by Offit (88), Vogelstein et al. (124), and Wheeler & Wang (131). Applying genomic technologies to single cancer cells has also been an active area of research, as reviewed by Navin (83). The robust cancer data sets from hundreds of publications as well as from consortium-based efforts such as the Cancer Genome Atlas have yielded a powerful data set that can be combined with large-scale biological models such as cancer cell line tools (38). This combination has the potential to completely transform how cancer patients are stratified for treatment. Additionally, a recent study by Zhang et al. (136) used whole-genome and whole-exome sequencing to analyze 1,120 children with cancer and catalog germline mutations that may predispose those individuals to cancer. The most prevalently mutated genes in the patients were *TP53*, *APC*, and *BRCA2*. Although many questions remain regarding how cancer cells metastasize and avoid detection by the host immune system, the last decade of cancer genomics has transformed the field of cancer research (131).

Decreases in the cost of sequencing, increases in instrument output, and advances in sample preparation continue to drive the field of single-cell genomics forward. Single-cell studies have recently examined transcriptomes to better understand single-cell physiology (24, 39), DNA methylation (34, 111), ChIP-seq (100), and genome sequencing (37). These and related studies are developing data sets to provide genomic resolution at the single-cell level, opening opportunities to appreciate the variability in single-cell physiology as these cells function as parts of more complex organ systems and organisms. Now that many of the initial challenges of analyzing single-cell amounts of RNA and DNA have been overcome, the next step is appropriately powering studies to examine enough individual cells to develop data sets that are sufficiently broad and deep to accurately resolve the cellular dynamics. Recent work by Macosko et al. (74) demonstrated a highly parallel approach for single-cell analysis termed Drop-seq that is not limited by the number of available wells in standard laboratory formats. Drop-seq utilizes a microfluidic partitioning method not unlike the methods described earlier for partitioning DNA into subhaploid amounts for phasing studies. Rather than partitioning DNA, this method partitions individual cells into droplets and associates a unique barcode with the RNA from that individual cell. The original Drop-seq study analyzed mRNA transcripts from 44,808 mouse retinal cells simultaneously while retaining each transcript's cell of origin, which allowed the authors to identify 39 transcriptionally distinct cell populations from the mouse retina (74). This parallelization of single-cell genomics provides a foundation for examining tissue or organ physiology not by isolating RNA or DNA in bulk from the total tissue, but rather by tagging RNA or DNA from individual cells via molecular barcodes and having the flexibility to examine population expression or DNA signatures and parse data to the single cell.

DATA SHARING, STANDARDS DEVELOPMENT, AND CLINICAL APPLICATIONS

The initiation of consortiums to develop widely available standards and samples for RNA [the External RNA Control Consortium (81)] and DNA [the Genome in a Bottle Consortium (137)] brings a stable sample set for comparing and developing validation and technical standards. Data consortiums such as the Exome Aggregation Consortium (32) and the St. Jude Pediatric Cancer Genome Project (136) provide unprecedented value to the scientific community. Although the technology and market will continue to evolve at a rapid pace, the maturity of the platforms and the availability of public reference data sets and biochemical standards allow for rigorous testing and development that have opened the door to the clinical application of sequencing in a targeted

manner (such as sequencing specific panels of genes), in the broader exome (covering annotated protein-coding regions), and finally in the entire genome.

Sequencing all or some of the protein-coding regions of the genome has been an effective and efficient method to characterize the ~1–2% of the genome annotated to encode proteins. The earliest genome selection methods were performed by either microarray selection (4, 50, 89) or multiplex amplification (93). These developments were quickly followed by the use of oligonucleotides to target regions of interest, which quickly became the dominant selection method for large-scale partitioning of the human genome (41). Since these foundational efforts, several commercialized technologies for oligonucleotide-based sequence capture have been developed, with platforms from Agilent Technologies, Illumina, and Roche NimbleGen being the most dominant over the last five years.

Whole-exome sequencing quickly became an efficient and accurate tool to examine the protein-coding regions of the genome, with numerous papers illustrating its power in rare disease diagnosis (9, 20, 73, 85, 86, 98) and clinical impact (133). Although effective and efficient, whole-exome sequencing determines a causative variant in only approximately 25% of cases (135). The diagnostic rate may appear low, but these results should be considered in the context of limited power to appreciate multivariant effects and the possible impact of variations outside the exonic regions, such as deep intronic or regulatory variants that could play a role. One of the more striking features of exome studies that have examined rare disease (135) and more common disorders, such as autism (84) and sporadic schizophrenia (134), is the exceptional rate of de novo mutation observed. These observations are profoundly changing our perception of these diseases (61, 122) and providing novel frameworks for the analysis of diseases with extensive locus heterogeneity (105).

The cost of DNA sequencing has steadily decreased since the introduction of next-generation sequencing, regardless of the platform type or technology (130). More than two years after the announcement of the Illumina HiSeq X platform, overall sequencing costs have largely stabilized, and the availability of sequencing on the HiSeq X platform has become widely available, with more than 20 HiSeq X sites around the world. The wide availability of low-cost human genome sequencing has resulted in a broader use of whole-genome sequencing for the study of human variation and disease (132) but with some important and notable considerations for the clinical use of sequencing in relatively healthy individuals (25). In an exploratory study using 12 volunteer adults, Dewey et al. (25) reported that whole-genome sequencing had incomplete coverage of inherited disease genes, low reproducibility of detection of genetic variation with the highest potential clinical effects, and uncertainty about clinically reportable findings. The uncertainties and concerns highlighted in this study illustrate the importance of rigorous quality control, rigorous technical standards, and the use of high-quality data to produce genome data. In the data produced by Dewey et al. (25), 9–17% of genes associated with inherited disease or annotated as important by the American College of Medical Genetics and Genomics were inadequately or inconsistently covered, particularly for insertion/deletion variants, and 4 of the 12 individuals had mutations in genes annotated as disease-causing without showing the presence of disease, indicating that those mutations had an uncertain effect, had a lesser significance than originally believed, or were sequencing errors. The authors concluded that the practical burden of reportable genetic findings from genome sequencing will vary considerably based on laboratory or institutional sequencing expertise, reliance on pathogenicity classifications in mutation databases, and access to and methods for evaluation of published evidence.

As clinical applications for whole-genome or whole-exome sequencing become more common, it will become increasingly important to carefully evaluate the technical performance of the capture tools available from commercial vendors. Patwardhan et al. (90) evaluated the differential performance of four commercially available exome capture reagents and compared them

with an augmented exome strategy that enhanced coverage over medically relevant genes. The authors reported superior variant sensitivities in the enhanced regions compared with traditional exome sequencing or whole-genome sequencing. Of a set of 56 genes that were recommended for return of secondary findings by the American College of Medical Genetics and Genomics, very few had greater than 95% of the disease-associated variants covered to at least 30× when using a 31.5× coverage whole-genome data set from the Sequence Read Archive (under accession PRJNA289286). The authors concluded that clinicians should carefully consider the analytical performance of any platform before determining the most appropriate reagent to use for a specific study in order to avoid false negative results.

In contrast to the study by Patwardhan et al. (90), other recent studies have found whole-genome sequencing to be superior to whole-exome sequencing in terms of overall variant sensitivity and a lack of bias because there is no selection procedure for whole-genome libraries. Lelieveld et al. (66) found that whole-exome libraries required two to three times more coverage to achieve similar variant sensitivities compared with whole-genome libraries. This increase in needed read depth for whole-exome sequencing helps to normalize the cost differential between the two methods. Importantly, Lelieveld et al. (66) did not observe any significant differences in the ability of exome or whole-genome sequencing methods to completely cover 2,759 clinically relevant genes. In a study similar to the work by Lelieveld et al. (66), Meynert et al. (79) compared polymorphism detection sensitivity and systematic biases using a set of tissue samples that underwent both whole-genome and whole-exome sequencing. They found that exome sequencing required a mean depth of 40× to reach a 95% detection sensitivity, whereas the same sample analyzed with whole-genome sequencing needed only a mean depth of 14× for the same sensitivity. They also reported greater uniformity of coverage and reduced bias in the detection of nonreference alleles in whole-genome data compared with whole-exome data.

CONCLUSIONS

As we move through the second decade since the first draft of the human genome sequence, the genomics field continues to advance rapidly. Study designs have continued to increase in scope (including population-scale sequencing efforts such as the 100,000 Genomes Project) and resolution (through the development of techniques such as Drop-seq for single-cell sequencing). The availability of low-cost, high-performance sequencing continues to expand the diversity of researchers and applications of genomics, while the development and revision of sequencing platforms (especially long-read technologies) expand the horizons of the type and complexity of genome and transcriptome architecture that can be resolved. The advancement of sequencing technologies by commercial companies and the development of applications for those technologies by the scientific community will continue to be a robust symbiotic relationship.

For the first time in several years, there are indications and potential for new technologies to challenge the sequencing status quo. Pacific Biosciences and Oxford Nanopore will challenge each other with their new platforms—Pacific Biosciences with Sequel, and Oxford Nanopore with PromethION. Together, these long-read technologies have an opportunity to challenge Illumina owing to the limitations of short-read technologies for analyzing structural variation and haplotype phasing as well as transcript splicing variation. That said, short-read sequencing will be bolstered by its ease of use and massive output as well as the development of companion technologies, such as the 10X Genomics GemCode platform or CPT-seq for the generation of synthetic long sequencing reads that drastically improve phasing and structural variation analysis. The debate will continue on what platforms will ultimately be the most useful in the clinic and in the lab as well as what capabilities will be necessary for a platform to provide the next inflection

point in resolution to fundamentally add to our understanding of how the sequence of a genome is transformed into the complexities of life through the biological process.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors acknowledge and thank all of our colleagues at the HudsonAlpha Institute for Biotechnology, with a special thanks to the dedicated staff of the Genomic Services Laboratory for their efforts in testing, optimizing, and supporting a diverse array of genomic and bioinformatic methods and technologies.

LITERATURE CITED

1. 1000 Genomes Proj. Consort. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65
2. Acevedo A, Andino R. 2014. Library preparation for highly accurate population sequencing of RNA viruses. *Nat. Protoc.* 9:1760–69
3. Adey A, Kitzman JO, Burton JN, Daza R, Kumar A, et al. 2014. In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* 24:2041–49
4. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* 4:903–5
5. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, et al. 2013. Signatures of mutational processes in human cancer. *Nature* 500:415–21
6. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* 3:246–59
7. Alexandrov LB, Stratton MR. 2014. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* 24:52–60
8. Amini S, Pushkarev D, Christiansen L, Kostem E, Royce T, et al. 2014. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* 46:1343–49
9. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, et al. 2010. Clinical assessment incorporating a personal genome. *Lancet* 375:1525–35
10. Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, et al. 2015. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.* 33:296–300
11. Au KF, Sebastiano V. 2014. The transcriptome of human pluripotent stem cells. *Curr. Opin. Genet. Dev.* 28:71–77
12. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59
13. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, et al. 2008. The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* 26:1146–53
14. Burghel GJ, Hurst CD, Watson CM, Chambers PA, Dickinson H, et al. 2015. Towards a next-generation sequencing diagnostic service for tumour genotyping: a comparison of panels and platforms. *Biomed. Res. Int.* 2015:478017
15. Cao MD, Ganesamoorthy D, Cooper MA, Coin LJM. 2016. Realtime analysis and visualization of MinION sequencing data with npReader. *Bioinformatics* 32:764–66
16. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. 2012. Pacific Biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genom.* 13:375

17. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 517:608–11
18. Chaisson MJ, Wilson RK, Eichler EE. 2015. Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.* 16:627–40
19. Chang CJ, Chen PL, Yang WS, Chao KM. 2014. A fault-tolerant method for HLA typing with PacBio data. *BMC Bioinform.* 15:296
20. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, et al. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *PNAS* 106:19096–101
21. Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, et al. 2015. Extending reference assembly models. *Genome Biol.* 16:13
22. Church GM, Gao Y, Kosuri S. 2012. Next-generation digital information storage in DNA. *Science* 337:1628
23. Compeau PE, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* 29:987–91
24. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, et al. 2015. A survey of human brain transcriptome diversity at the single cell level. *PNAS* 112:7285–90
25. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, et al. 2014. Clinical interpretation and implications of whole-genome sequencing. *JAMA* 311:1035–45
26. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36:e105
27. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327:78–81
28. Duitama J, McEwen GK, Huebsch T, Palczewski S, Schulz S, et al. 2012. Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of single individual haplotyping techniques. *Nucleic Acids Res.* 40:2041–53
29. Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–38
30. English AC, Richards S, Han Y, Wang M, Vee V, et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLOS ONE* 7:e47768
31. Erlich Y. 2015. A vision for ubiquitous sequencing. *Genome Res.* 25:1411–16
32. Exome Aggregation Consort., Lek M, Karczewski KJ, Minikel EV, Samocha KE, et al. 2015. Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv. doi: 10.1101/030338
33. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, et al. 2012. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* 30:1232–39
34. Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schonegger A, et al. 2015. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep.* 10:1386–97
35. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, et al. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7:461–65
36. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493:216–20
37. Fu Y, Li C, Lu S, Zhou W, Tang F, et al. 2015. Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *PNAS* 112:11923–28
38. Garnett MJ, McDermott U. 2014. The evolving role of cancer cell line-based screens to define the impact of cancer genomes on drug response. *Curr. Opin. Genet. Dev.* 24:114–19
39. Gaublomme JT, Yosef N, Lee Y, Gertner RS, Yang LV, et al. 2015. Single-cell genomics unveils critical regulators of Th17 cell pathogenicity. *Cell* 163:1400–12
40. Genome Ref. Consort. 2015. *Human genome overview: information concerning the continuing improvement of the human genome.* **http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human**
41. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* 27:182–89

42. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* 25:1750–56
43. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, et al. 2015. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* 47:435–44
44. Gulcher J, Stefansson K. 1998. Population genomics: laying the groundwork for genetic disease modeling and targeting. *Clin. Chem. Lab. Med.* 36:523–27
45. Gulcher J, Stefansson K. 1999. An Icelandic saga on a centralized healthcare database and democratic decision making. *Nat. Biotechnol.* 17:620
46. Gundersen S, Kalas M, Abul O, Frigessi A, Hovig E, Sandve GK. 2011. Identifying elemental genomic track types and representing them uniformly. *BMC Bioinform.* 12:494
47. Guo J, Xu N, Li Z, Zhang S, Wu J, et al. 2008. Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *PNAS* 105:9145–50
48. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, et al. 2008. Single-molecule DNA sequencing of a viral genome. *Science* 320:106–9
49. Helleday T, Eshtad S, Nik-Zainal S. 2014. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* 15:585–98
50. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* 39:1522–27
51. Honeyman JN, Simon EP, Robine N, Chiaroni-Clarke R, Darcy DG, et al. 2014. Detection of a recurrent *DNAJB1-PRKACA* chimeric transcript in fibrolamellar hepatocellular carcinoma. *Science* 343:1010–14
52. Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* 24:688–96
53. Illumina. 2014. *Sequencing methods review: a review of publications featuring Illumina® Technology*. Publ. No. 073-2014-001, Illumina, San Diego, CA. **http://www.illumina.com/techniques/sequencing/ngs-library-prep/library-prep-methods.html**
54. Int. Hum. Genome Seq. Consort. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–45
55. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. 2015. Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* 12:351–56
56. Kaper F, Swamy S, Klotzle B, Munchel S, Cottrell J, et al. 2013. Whole-genome haplotyping by dilution, amplification, and sequencing. *PNAS* 110:5552–57
57. Karamitros T, Magiorkinis G. 2015. A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. *Nucleic Acids Res.* 43:e152
58. Karow J. 2015. Oxford Nanopore outlines specs for new sequencers, automated sample prep system, pay-as-go pricing. *Genome Web*, May 15. **http://www.genomeweb.com/sequencing- technology/oxford-nanopore-outlines-specs-new-sequencers-automated-sample-prep-system-pay**
59. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. 1996. Characterization of individual polynucleotide molecules using a membrane channel. *PNAS* 93:13770–73
60. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, et al. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30:693–700
61. Ku CS, Polychronakos C, Tan EK, Naidoo N, Pawitan Y, et al. 2013. A new paradigm emerges from the study of de novo mutations in the context of neurodevelopmental disease. *Mol. Psychiatry* 18:141–53
62. Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M. 2016. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat. Biotechnol.* 34:64–69
63. Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, et al. 2014. Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.* 32:261–66
64. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
65. Leinonen R, Sugawara H, Shumway M (Int. Nucleotide Seq. Database Consort.). 2011. The sequence read archive. *Nucleic Acids Res.* 39:D19–21

66. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. 2015. Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. *Hum. Mutat.* 36:815–22
67. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299:682–86
68. Li H. 2014. On the graphical representation of sequences. *Heng Li's Blog*, July 25 **http://lh3.github.io/2014/07/25/on-the-graphical-representation-of-sequences**
69. Li Q, Li Y, Song J, Xu H, Xu J, et al. 2014. High-accuracy de novo assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy. *New Phytol.* 204:1041–49
70. Liu L, Li Y, Li S, Hu N, He Y, et al. 2012. Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012:251364
71. Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, et al. 2013. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res.* 23:121–28
72. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, et al. 2013. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *PNAS* 110:19872–77
73. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, et al. 2010. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* 362:1181–91
74. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161:1202–14
75. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–80
76. Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *PNAS* 74:560–64
77. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, et al. 2011. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLOS ONE* 6:e22751
78. Metzker ML. 2010. Sequencing technologies—the next generation. *Nat. Rev. Genet.* 11:31–46
79. Meynert AM, Ansari M, FitzPatrick DR, Taylor MS. 2014. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinform.* 15:247
80. Morey M, Fernandez-Marmiesse A, Castineiras D, Fraga JM, Couce ML, Cocho JA. 2013. A glimpse into past, present, and future DNA sequencing. *Mol. Genet. Metab.* 110:3–24
81. Munro SA, Lund SP, Pine PS, Binder H, Clevert DA, et al. 2014. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nat. Commun.* 5:5125
82. Myers EW. 2005. The fragment assembly string graph. *Bioinform.* 21(Suppl. 2):ii79–85
83. Navin NE. 2014. Cancer genomics: one cell at a time. *Genome Biol.* 15:452
84. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485:242–45
85. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. 2010. Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.* 42:30–35
86. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461:272–76
87. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, et al. 2012. Mutational processes molding the genomes of 21 breast cancers. *Cell* 149:979–93
88. Offit K. 2014. Decade in review—genomics: a decade of discovery in cancer genomics. *Nat. Rev. Clin. Oncol.* 11:632–34
89. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. 2007. Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* 4:907–9
90. Patwardhan A, Harris J, Leng N, Bartha G, Church DM, et al. 2015. Achieving high-sensitivity for clinical applications using augmented exome sequencing. *Genome Med.* 7:71
91. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *PNAS* 91:5022–26
92. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, et al. 2012. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 487:190–95

93. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, et al. 2007. Multiplex amplification of large sets of human exons. *Nat. Methods* 4:931–36
94. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, et al. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genom.* 13:341
95. Quick J, Quinlan AR, Loman NJ. 2014. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *GigaScience* 3:22
96. Reuter JA, Spacek DV, Snyder MP. 2015. High-throughput sequencing technologies. *Mol. Cell* 58: 586–97
97. Risse J, Thomson M, Patrick S, Blakely G, Koutsovoulos G, et al. 2015. A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing data. *GigaScience* 4:60
98. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328:636–39
99. Robert C, Watson M. 2015. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol.* 16:177
100. Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, et al. 2015. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* 33:1165–72
101. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475:348–52
102. Rydbeck H, Sandve GK, Ferkingstad E, Simovski B, Rye M, Hovig E. 2015. ClusTrack: feature extraction and similarity measures for clustering of genome-wide data sets. *PLOS ONE* 10:e0123261
103. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, et al. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239:487–91
104. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, et al. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230:1350–54
105. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, et al. 2014. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* 46:944–50
106. Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94:441–48
107. Shalon D, Smith SJ, Brown PO. 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6:639–45
108. Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31:1009–14
109. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728–32
110. Simon EP, Freije CA, Farber BA, Lalazar G, Darcy DG, et al. 2015. Transcriptomic characterization of fibrolamellar hepatocellular carcinoma. *PNAS* 112:E5916–25
111. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, et al. 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* 11:817–20
112. Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, et al. 2014. Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res.* 24:2066–76
113. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526:75–81
114. Szalay T, Golovchenko JA. 2015. De novo sequencing and variant calling with nanopores using PoreSeq. *Nat. Biotechnol.* 33:1087–91
115. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337:64–69
116. Tilgner H, Grubert F, Sharon D, Snyder MP. 2014. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *PNAS* 111:9869–74
117. Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 38:e159
118. Treutlein B, Gokce O, Quake SR, Sudhof TC. 2014. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *PNAS* 111:E1291–99

119. Trujillano D, Weiss ME, Koster J, Papachristos EB, Werber M, et al. 2015. Validation of a semiconductor next-generation sequencing assay for the clinical genetic screening of CFTR. *Mol. Genet. Genom. Med.* 3:396–403
120. Uemura S, Aitken CE, Korlach J, Flusberg BA, Turner SW, Puglisi JD. 2010. Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature* 464:1012–17
121. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, et al. 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18:1051–63
122. Veltman JA, Brunner HG. 2012. De novo mutations in human genetic disease. *Nat. Rev. Genet.* 13:565–75
123. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
124. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr., Kinzler KW. 2013. Cancer genome landscapes. *Science* 339:1546–58
125. Wang Y, Yang Q, Wang Z. 2014. The evolution of nanopore sequencing. *Front. Genet.* 5:449
126. Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, et al. 2015. LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience* 4:35
127. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, et al. 2013. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am. J. Hum. Genet.* 92:530–46
128. Watson JD, Crick FH. 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 171:737–38
129. Westbrook CJ, Karl JA, Wiseman RW, Mate S, Koroleva G, et al. 2015. No assembly required: full-length MHC class I allele discovery by PacBio circular consensus sequencing. *Hum. Immunol.* 76:891–96
130. Wetterstrand K. 2016. *DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP)*. **http://www.genome.gov/sequencingcosts**
131. Wheeler DA, Wang L. 2013. From human genome to cancer genome: the first decade. *Genome Res.* 23:1054–62
132. Willig LK, Petrikin JE, Smith LD, Saunders CJ, Thiffault I, et al. 2015. Whole-genome sequencing for identification of Mendelian disorders in critically ill infants: a retrospective analysis of diagnostic and clinical findings. *Lancet Respir. Med.* 3:377–87
133. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, et al. 2011. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.* 13:255–62
134. Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, et al. 2011. Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat. Genet.* 43:864–68
135. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, et al. 2013. Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N. Engl. J. Med.* 369:1502–11
136. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, et al. 2015. Germline mutations in predisposition genes in pediatric cancer. *N. Engl. J. Med.* 373:2336–46
137. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, et al. 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* 32:246–51