Cold Spring Harbor Perspectives in Medicine
www.perspectivesinmedicine.org

Next-Generation Sequencing Strategies

Shawn E. Levy and Braden E. Boone

HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806

Correspondence: slevy@hudsonalpha.org

More than a decade ago, the term "next-generation" sequencing was coined to describe what was, at the time, revolutionary new methods to sequence RNA and DNA at a faster pace and cheaper cost than could be performed by standard bench-top protocols. Since then, the field of DNA sequencing has evolved at a rapid pace, with new breakthroughs allowing capacity to exponentially increase and cost to dramatically decrease. As genome-scale sequencing has become routine, a paradigm shift is occurring in genomics, which uses the power of high-throughput, rapid sequencing power with large-scale studies. These new approaches to genetic discovery will provide direct impact to fields such as personalized medicine, evolution, and biodiversity. This work reviews recent technology advances and methods in next-generation sequencing and highlights current large-scale sequencing efforts driving the evolution of the genomics space.

Since the fundamental discovery of the structure of DNA by Watson, Crick, and Park (Watson and Crick 1953) and the pioneering development of methods to detect the sequence of DNA bases by foundational methods such as Maxam and Gilbert (1977) and then Sanger sequencing (Sanger and Coulson 1975), 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 (Saiki et al. 1985, 1988), 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 in 2001 (Lander et al. 2001; Venter et al. 2001) and the first completed draft 3 years later (International Human Genome Sequencing Consortium 2004). Genomics has evolved at an amazing pace since the first draft of the human genome project. Dozens of next-generation sequencing (NGS) companies and technologies have been created and the corresponding field of bioinformatics has exploded as a major scientific and training discipline.

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 a highly parallel assay to measure RNA and DNA (Pease et al. 1994; Shalon et al. 1996). Between 2001 and 2006, microarrays offered the first genome-scale, parallel analysis of DNA and RNA. Beginning in 2006, second- and third-generation sequencing techniques began to emerge that permitted an unbiased means to examine billions of templates of DNA and

Editors: W. Richard McCombie, Elaine R. Mardis, James A. Knowles, and John D. McPherson
Additional Perspectives on Next-Generation Sequencing in Medicine available at www.perspectivesinmedicine.org

 doi: 10.1101/cshperspect.a025791
Cite this article as *Cold Spring Harb Perspect Med* 2019;9:a025791

RNA. Although now almost a decade old, "next-generation" sequencing remains the colloquial term 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, the genomics space has had 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 NGS instruments changing within months and chemistries and analysis algorithms changing within weeks, creating substantial challenges for both researchers and clinicians. These 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, allowing not only a maturity in the methods and instruments to develop but also a large diversity of publications, methods, and applications of sequencing technology. There are now thousands of publications with many hundreds added each year that describe the utilization of sequencing technologies. There have been a number of excellent reviews over the last several years that describe the technological landscape of sequencing (Metzker 2010; Morey et al. 2013; Reuter et al. 2015). A review of those examples and many others, when viewed in chronological order, provides an excellent history of the changing sequencing space and the amazing pace that has brought us from the first draft of the human genome 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 15 years. This review will focus on the 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 is not a reflection of a lack of its impact or importance but instead is indicative of the sheer volume of information available and the desire to highlight specific areas in this review.

SEQUENCING PLATFORMS AND CAPABILITIES

The sequencing world is a dynamic and unforgiving space, punctuated by spurts of tremendous hype and overpromise and periods of catapulting advancement. In 2011/2012, Oxford Nanopore presented at the annual Advances in Genome Biology and Technology conference that their GridION sequencing platform would sequence a human genome in 15 minutes and be commercially available by mid-2012. In the same month, Ion Torrent stated they would begin selling a machine by the end of the year that could sequence an entire human genome in a day for less than $1000 and Pacific Biosciences was approaching the 1-year anniversary of their commercial launch. Looking back from 2018, many of those announcements proved to be aggressive and development efforts proved to be far more challenging for the manufacturers. However, the past few years have seen a resurgence of platforms and innovations. In 2015, Illumina launched the HiSeq X platform, increasing sequencing capacity to levels that seemed staggering at the time. That same year, Pacific Biosciences launched their second commercial platform, Sequel, that increased data output by six- to sevenfold over the existing platform while decreasing the capital costs for the instrument by one-half. In 2018, the Sequel 2.1 chemistry was released, reporting achievements of one-half of all reads above 30,000 bases in length and 10 billion bases of data for whole-genome analysis run. Oxford Nanopore had a successful launch of their MinION sequencer and followed it with a GridION X5 (drives up to five MinION flowcells) in 2017. The PromethION (up to 48 MinION flowcells) is also now available in an early access program.

For several years, the Illumina HiSeq X system remained the highest output platform and the only sequencing technology available that could generate highly accurate data at the human genome scale at reagent costs less than $1000. In initial years, only whole genomic samples at a minimum coverage of 30× were allowed to be sequenced on the HiSeq X platform. In later years, this requirement was relaxed to include

whole-genome sequencing (WGS) of all organisms with a minimum coverage of 15×. The HiSeq X was run at a fixed read length and although the cost per base had dropped dramatically compared with NextSeq or HiSeq 2500 sequencing, library types other than WGS were unable to take advantage of the lower sequencing costs offered by the HiSeq X. In 2017, the launch of the NovaSeq 6000 platform unleashed a new era in sequencing capabilities. The platform is extremely flexible, with flow cells of differing read yields available and any type of library or sequencing length allowed. The NovaSeq 6000 is limited to a maximum of 150 nt reads but can output >20 billion paired-end reads or more than 40 billion total reads (6000 Gb) per instrument run (two S4 300 flow cells per sequencing run).

SHORT-READ AND LONG-READ SEQUENCING

The exponential increase in data output and decrease in cost per base sequenced has been primarily driven by increases in parallelization of short-read sequencing platforms such as Illumina and Ion Torrent. 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. 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 in both the literature as well as the amount of data from the platform submitted to the Sequence Read Archive (Leinonen et al. 2011) shows its efficiency and power. Illumina has been shown to be powerful for both resequencing approaches like WGS and whole-exome sequencing (WES), for read-counting applications such as RNASeq and ChIPSeq, and for structural genomic profiling such as ATAC-seq, MNase-seq, or FAIRE-seq.

Although the first draft of the human genome was completed in 2001, there are still many areas of the genome that remain poorly characterized or missing from the current assembly. This is because of 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. A number of developments have occurred that have expanded the diversity of technologies and applications available to bring light to these dark regions of the genome or to enable the efficient de novo sequencing or characterization of nontraditional species.

Two commercially available, highly parallel sequencing technologies produce long sequencing reads. Instruments by Oxford Nanopore and Pacific Biosciences both produce read lengths in the many thousands of bases per read. Both use single-molecule sequencing albeit with very different detection methods. Oxford Nanopore uses nanopores for detection as their name implies and Pacific Biosciences uses optical detection of a sequencing-by-synthesis reaction that occurs inside a zero-mode waveguide (Levene et al. 2003). The details of the chemistries used on both platforms are well reviewed elsewhere (Reuter et al. 2015).

Although Illumina has a dominant position in the sequencing space in 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 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 was recently published from the Structure Variation Analysis Group of The 1000 Genomes Project (Sudmant et al. 2015). However, appreciating the full resolution of genomic variation is only assured when a complete, reference-free, de novo assembly of a genome is possible (reviewed in detail in Chaisson et al.

2015). One means to increase resolution for assembly or structural variation is with significantly longer sequencing reads. Short-read technologies and long-read technologies have been at opposing ends of the spectra for read length and for read density. Illumina's highest output platform, the NovaSeq 6000, is limited to 150 nt reads but can output >20 billion paired-end reads or >40 billion total reads (6000 Gb) per instrument run (two S4 flow cells per sequencing run). Pacific Biosciences, the most widely proven long-read technology, produces ~400,000 reads (10 Gb) per SMRT cell instrument run but at a read length that averages >15,000 nt and can exceed 100,000 nt in length. The previous P6-C4 chemistry had a per-base error rate of ~15% but the stochastic nature of the errors allowed for highly accurate consensus sequencing, both using the circular consensus sequencing (Lou et al. 2013; Li et al. 2014) or with generating consensus reads by sequencing samples to multiples times depth on the Pacific Biosciences platform. Current reagents and protocols have consensus accuracy exceeding 99.999% (Q50) at ~45× coverage.

The Pacific Biosciences RS platform was released in 2010 and has undergone a number of iterations and chemistry revisions since. The long sequence reads of the Pacific Biosciences platform have allowed challenging areas of the genome such as the MHCI region transcripts (Chang et al. 2014; Westbrook et al. 2015) and regions of segmental duplication (Huddleston et al. 2014) to be efficiently analyzed. Studies to generate de novo assemblies have also illustrated the impact of the platform and its potential role in developing routine de novo assembly-driven rather than reference-driven analysis of human genomes (Chaisson et al. 2015; Erlich 2015). In late 2015, Pacific Biosciences announced a new platform to augment their RS-II instrument. The new instrument, named Sequel, was a significant change from the RS in both form and capabilities. The Sequel is a fraction of the size of the original RS platform and costs much less. It offers a substantial increase in read density compared with the available RS with each SMRT cell having 1 M zero-mode waveguides compared with 150,000 on the RS-II, increasing the read output by approximately seven times. In 2018, the Sequel 2.1 chemistry was released, reporting achievements of one-half of all reads above 30,000 bases in length and 10 billion bases of data for whole-genome analysis run.

The use of nanopores as a means to sequence DNA has been discussed or shown in various forms since at least 1996 (Kasianowicz et al. 1996), and the potential and challenges of nanopore sequencing and the detailed chemistry of the Oxford Nanopore platform has been well reviewed elsewhere (Branton et al. 2008; Reuter et al. 2015). The first Oxford Nanopore sequencer, the MinION, broke many barriers at its launch, as it was the first handheld DNA sequencer that did not require anything more than an active USB port to operate. It also was the lowest cost DNA sequencer to be released with an instrument price of $1000 and offered a robust online community support and sharing of information regarding use, performance, and optimization. The current MinION platform generates ~10–20 Gb of data per 48-h run, with reads up to hundreds of kilobases, and requires only a simple 10-min sample prep. The GridION X5 (drives up to five MinION flowcells) was released in 2017 and the PromethION (up to 48 MinION flowcells) is also now available in an early access program. Other technology advances include an automated sample preparation device called VolTRAX, with version 1 available in early access form and version 2 to be released in 2018. New products on the horizon include the Flongle (an adaptor for MinION for small experiments) and the SmidgION, designed to use with a smartphone in any location.

SYNTHETIC LONG READS

Generation of long sequencing reads is not limited to direct measurement using long-read technologies such as Pacific Biosciences or Oxford Nanopore. There have been a number of innovative and elegant approaches to combine biochemical and informatic approaches with short-read sequencing data to generate synthetic long reads. These methods all rely on the parti-

tioning of the genome in some manner to subhaploid concentration followed by generation of a sequencing library that can uniquely be mapped back to the subhaploid fraction. Early methods used fosmid libraries as a means to partition a genome sample (Duitama et al. 2012), 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. A number of methods have been applied with varying complexity and resolution. Several synthetic long-read technologies and methods were described in 2012. The long fragment read (LFR) method illustrated sequencing and haplotyping from 10 to 20 human cells (Peters et al. 2012). Another long fragment technology was commercialized from Steven Quake's laboratory at Stanford University as Moleculo (Kuleshov et al. 2014). 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 (Kuleshov et al. 2014). Illumina independently published an approach that targeted a 1 Mb region of the X chromosome, phasing >95% of single-nucleotide polymorphisms (SNPs) and deriving haplotype blocks of hundreds of kilobases (Kaper et al. 2013). Then, a transposase-mediated library preparation of a subhaploid fractionated genome that leverages the unique contiguity preserving activity of the Tn5 transposase (CPT-seq) was described (Adey et al. 2014) and applied to WGS (Amini et al. 2014). Whereas the LFR and Moleculo were limited to hundreds of partitions, the CPT-seq used 9216 barcode pools via combinatorial indexing. 10X Genomics commercialized another method of linked-read sequencing via the GEMCode platform and their next-generation Chromium platform (Zheng et al. 2016). The Chromium system uses a microfluidic assay to dropletize high-molecular-weight DNA into ~1,000,000 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. Many publications using Chromium technology and novel algorithms have shown their power in resolving genome variation, identify structural variations, and reconstruct haplotypes inaccessible with short-read sequencing (Collins et al. 2017; Marks et al. 2017; Xia et al. 2018).

Applications for synthetic long reads extend beyond haplotype phasing. Synthetic long-read methods have been applied to the analysis of the human microbiome with 51 additional species identified that were not observed with short-read sequencing alone. Additionally, the synthetic long reads identified extensive intraspecies variation, providing a resolution to the microbiome data that was previously unobtainable (Kuleshov et al. 2015). The Chromium platform has proven to be a huge boon in the field of evolutionary biology, with hundreds of new genomes elucidated and published at a fraction of the cost, manpower, and time previously required. Traditionally, assembly of a non-reference genome might take researchers and computational scientists several years and tens of thousands of dollars in sequencing alone. This strategy involved multiple types of libraries generated, including short-read libraries of varying insert sizes, mate-pair libraries to scaffold shorter reads, and long-read libraries to fill in repetitive regions and correct any errors arising in the short-read assembly. With the Chromium platform and the 10X Genomics Supernova assembly software, however, a researcher can often obtain a good de novo draft assembly of their species of interest for a few thousand dollars and a few weeks of analysis. Using additional short-read and long-read sequencing data in combination with linked-read Chromium sequencing data allows polishing and refinement of the genome. Recent years have seen a large number of genomes published for previously unreferenced species, including plants (Zhang et al. 2017; Hulse-Kemp et al. 2018), amphibians (Hammond et al. 2017), and mammals such as the beluga whale (Jones et al. 2017b) and sea otter (Jones et al. 2017a).

Cold Spring Harbor Perspectives in Medicine
www.perspectivesinmedicine.org

A SHIFTING PARADIGM

The ubiquitous availability of sequencing technology has not only led to an amazing array of research applications but also the rapid development of a number 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 she participated in research to drive the understanding of her cancer forward by discovering a gene fusion event that appears to drive her type of cancer. She used social media to identify others with the same cancer, and through sequencing (with the help of her father, a researcher at Rockefeller University), she identified a novel gene fusion between DNAJB1 and PRKACA, producing a fusion protein that retains kinase activity (Honeyman et al. 2014; Simon et al. 2015). This example highlights not only how powerful the ubiquitous access of sequencing has become but also how accepted it is becoming as a transformative tool in healthcare. 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 a greater and greater role in research, healthcare, and industrial experiments and the number of applications available will continue to grow with the innovation and creativity of the scientific community. Yaniv Erlich published an interesting review on sequencing from the perspective of what barriers remain to the ubiquitous use of sequencing sensors, including sequencing at home, in forensics, and in security applications (Erlich 2015). In recent years, the use of crowdsourcing and crowdfunding to drive medical research has emerged as a novel and valuable tool for new discoveries in fields such as metagenomics, educational outreach, and public health (Afshinnekoo et al. 2016). Two such research initiatives (MetaSUB and PathoMap) are discussed below.

Four years ago, the capabilities of the HiSeq X platform provided an unprecedented sequencing scale to the worldwide markets. As a result, several large-scale sequencing projects were immediately planned and launched. Significant among them for both the volume of data produced and also the transformative impact in genomics were the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2012) and the Exome Sequencing Project (Tennessen et al. 2012; Fu et al. 2013). In addition to the large-scale efforts, the available scale and platform capabilities resulted in the development of several population scale sequencing efforts. The largest effort at that time was the Genome England project, which at completion will have sequenced 100,000 genomes (Rotem et al. 2015). The Genome England project is being led by Genomics England, a company founded and owned by the United Kingdom Department of Health. Using England's National Health Service, the project will perform WGS of 100,000 samples, focusing on patients with rare diseases and their families as well as patients with common cancers. A total of about 75,000 people will be sequenced, of which about 40,000 are patients. The effort is still underway with extension of the project into 2018 (www.genomicsengland.co.uk).

Iceland has long been investing in the genetic and genomic analysis of its population, primarily through efforts supported by deCODE Genetics (now deCODE Genetics/AMGEN) (Gulcher and Stefansson 1998, 1999). These efforts have been expanded to WGS 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. 2015). Gudbjartsson et al. describes the insights gained from sequencing the whole genomes of 2636 Icelanders to a median depth of 20×. The investigators reported the discovery of 20 million SNPs as well as 1.5 million insertions or deletions. The discovered variants were annotated with respect to functional impact (gene position, functional impact, pathway, and conservation score), frequency, and density. Since this inaugural manuscript, deCODE Genetics has collected genotypic and medical information on more than 160,000 volunteers in Iceland and published many impactful manuscripts

highlighting genetic variants found to cause rare disease (Arnadottir et al. 2017; Jensson et al. 2017), predispose to various psychiatric disorders (Mullins et al. 2017; Steinberg et al. 2017), and increase risk of common diseases (Ivarsdottir et al. 2017; Kristjansson et al. 2017). 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 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 that are now driving both research and therapeutic decisions (Nik-Zainal et al. 2012; Alexandrov et al. 2013a,b; Alexandrov and Stratton 2014; Helleday et al. 2014). The results of these mutational signature studies are available in the online Catalog of Somatic Mutations in Cancer project from the Sanger Institute (Sanger Institute 2015). A collection of recent reviews profiling the current landscape of cancer genome exist in a special issue of *Current Opinion in Genetics and Development* with an editorial overview offered by Delattre and Bult (2017). Applying genomic technologies to single cancer cells has also been an active area of research (Navin 2014; Tsoucas and Yuan 2017). The robust cancer data sets from hundreds of publications as well as from consortia-based efforts such as The Cancer Genome Atlas Project (TCGA) (The Cancer Genome Atlas Consortium 2018) have yielded a powerful data set to be combined with large-scale biological models such as cancer cell line tools (Garnett and McDermott 2014). In 2018, with >11,000 tumor profiles from 33 cancer types, the TCGA Pan-Cancer Atlas project provides a comprehensive understanding of how, where, and why tumors arise in humans and is an essential resource for new treatments (Ding et al. 2018). A study by Zhang and coworkers analyzed 1120 children using WGS and WES with cancer to catalog the germline mutations present that may predispose those individuals for cancer. The most prevalently mutated genes in the patients were TP53, APC, and BRCA2 (Zhang et al. 2015). The St. Jude Lifetime Cohort Study was launched to follow childhood cancer survivors through adulthood to assess and understand the increased health risks. To date, >4000 participants have been enrolled in the Cohort, which includes sequencing of patient specimens. The initial report included profiles of secondary cancers and chronic health conditions found in childhood cancer survivors (Bhakta et al. 2017). Another large-scale cancer sequencing project is the Oncology Research Information Exchange Network (ORIEN) partnership among 18 participating cancer centers. This project profiles adult cancer patients, with clinical data and genomic data provided (germline and tumor samples are sequenced, including RNA from tumors). To date, >6500 ORIEN patients have had molecular, genetic, and clinical data made available to participating researchers and clinicians. The types of cancer profiled are widely varied, from breast and leukemia to more rare brain and melanoma cancers (www.oriencancer.org). This partnership allows oncologists an unprecedented amount of data to use in deciding on treatments for patient care.

The introduction of the NovaSeq 6000 platform has allowed sequencing capacity to exponentially explode. A sequencing center operating ten NovaSeq 6000s can sequence at least 60,000 30× genomes each year by conservative estimates, and >70,000 at maximum production. With even just a dozen sequencing centers worldwide adopting the NovaSeq platform, almost a million genomes could potentially be sequenced each year. A new research effort has been proposed by the National Institutes of Health (NIH) to leverage this sequencing capacity to further understanding of genomics and its contributions to health and disease at a true population scale. Launched in 2016, the All of Us Research Program was a key element of the Precision Medicine Initiative that aims to collect participant-provided information, including environmental, physiological, and health data and biospecimens for 1 million or more participants in the United States (allofus.nih.gov). In 2018, the NIH solicited applications for genome cen-

ters to participate in the genotyping and WGS for All of Us Research Program participants. The scale is both staggering and exciting: to sequence >700,000 genomes over 5 years. With the advent of NovaSeq sequencing, reducing both time and cost for whole-genome profiling, as well as the collaboration and coordination of multiple genome centers, this goal is well within reach and the data generated will shape medical care, risk assessment, and quality of life for generations to come.

The available sequencing power is not only being applied to humans at a grand scale. Massive projects are underway to examine our planet's smallest species. In 2013, a project called PathoMap was created to profile the New York City metagenome in the subway system. The results were provocative (Afshinnekoo et al. 2015). Approximately one-half of the sequenced DNA did not match any known organism, suggesting that the breadth of our knowledge of the metagenomics world is rather shallow. Further, the ancestry of human DNA left on subway surfaces validated U.S. Census demographic data as well as historic events such as marine bacteria found in hurricane-flooded stations. Even more intriguing was the detection of some pathogenic DNA, although the associated diseases remained undetected in the city, suggesting pathogens are a part of the normal microbiome. As a consequence of the interest generated from the PathoMap study, an international consortium was launched called the MetaSUB project (Metagenomics and Metadesign of Subways and Urban Biomes) to profile microbiomes in mass transit systems around the world (MetaSUB International Consortium 2016). It has been acknowledged that there are some risks and ethical considerations when performing large-scale microbiome and metagenome research (Shamarina et al. 2017), as well as the possibility of data skewing owing to the crowdsourcing and crowdfunding nature of these large studies (Afshinnekoo et al. 2016). Still, the study of the microbiome around us will continue to reveal additional layers of complexity in both public and personal health.

These population-scale sequencing efforts are extensions of foundational projects such as The 1000 Genomes Project, The Cancer Genome Atlas Project, and the related ENCODE Project. The 1000 Genomes Project recently published data from phase 3 of the project (Sudmant et al. 2015), describing the completion of the project and the description of >88 million variants from 2504 individuals from six populations with all variants phased into high-quality haplotypes. The last 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 consortia-based projects; some are mentioned here and others are reviewed elsewhere (see Table 2 in Reuter et al. 2015).

CONCLUSIONS

As we move through the second decade since the first draft of the human genome sequence, the evolution and innovation in the genomics field remain constant. Study design continues to increase in scope to population-scale sequencing like the All of Us Research Project while also increasing in resolution like the development and refinement of single-cell sequencing. The availability of low-cost, high-performance sequencing continues to expand the diversity of applications for genomics while the development and revision to sequencing platforms, especially in the long-read technologies, expands the horizons of the type and complexity of genome architecture that can be resolved. Population-scale 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 consortia projects bring to basic and translational research is profound. It is now inconceivable to perform any type of genomic analysis without using data from one or more consortia-based projects, be it a reference genome, a variant frequency, a sequence search, or any number of other data types that are available in the public domain.

REFERENCES

Adey A, Kitzman JO, Burton JN, Daza R, Kumar A, Christiansen L, Ronaghi M, Amini S, Gunderson KL, Steemers FJ, et al. 2014. In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res* **24:** 2041–2049.

Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, Maritz JM, Reeves D, Gandara J, Chhangawala S, et al. 2015. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Syst* **1:** 72–87.

Afshinnekoo E, Ahsanuddin S, Mason CE. 2016. Globalizing and crowdsourcing biomedical research. *Br Med Bull* **120:** 27–33.

Alexandrov LB, Stratton MR. 2014. Mutational signatures: The patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* **24:** 52–60.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al. 2013a. Signatures of mutational processes in human cancer. *Nature* **500:** 415–421.

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013b. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3:** 246–259.

Amini S, Pushkarev D, Christiansen L, Kostem E, Royce T, Turk C, Pignatelli N, Adey A, Kitzman JO, Vijayan K, et al. 2014. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat Genet* **46:** 1343–1349.

Arnadottir GA, Jensson BO, Marelsson SE, Sulem G, Oddsson A, Kristjansson RP, Benonisdottir S, Gudjonsson SA, Masson G, Thorisson GA, et al. 2017. Compound heterozygous mutations in UBA5 causing early-onset epileptic encephalopathy in two sisters. *BMC Med Genet* **18:** 103.

Bhakta N, Liu Q, Ness KK, Baassiri M, Eissa H, Yeo F, Chemaitilly W, Ehrhardt MJ, Bass J, Bishop MW, et al. 2017. The cumulative burden of surviving childhood cancer: An initial report from the St Jude Lifetime Cohort Study (SJLIFE). *Lancet* **390:** 2569–2582.

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, et al. 2008. The potential and challenges of nanopore sequencing. *Nat Biotechnol* **26:** 1146–1153.

Chaisson MJ, Wilson RK, Eichler EE. 2015. Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet* **16:** 627–640.

Chang CJ, Chen PL, Yang WS, Chao KM. 2014. A fault-tolerant method for HLA typing with PacBio data. *BMC Bioinformatics* **15:** 296.

Collins RL, Brand H, Redin CE, Hanscom C, Antolik C, Stone MR, Glessner JT, Mason T, Pregno G, Dorrani N, et al. 2017. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol* **18:** 36.

Delattre O, Bult CJ. 2017. Editorial overview: Characterizing the cancer genome: Mechanistic insights and translational opportunities. *Curr Opin Genet Dev* **42:** 78–80.

Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, Gibbs DL, Weerasinghe A, Huang KL, Tokheim C, et al. 2018. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell* **173:** 305–320.e10.

Duitama J, McEwen GK, Huebsch T, Palczewski S, Schulz S, Verstrepen K, Suk EK, Hoehe MR. 2012. Fosmid-based whole genome haplotyping of a HapMap trio child: Evaluation of single individual haplotyping techniques. *Nucleic Acids Res* **40:** 2041–2053.

Erlich Y. 2015. A vision for ubiquitous sequencing. *Genome Res* **25:** 1411–1416.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493:** 216–220.

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:** 114–119.

Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, et al. 2015. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47:** 435–444.

Gulcher J, Stefansson K. 1998. Population genomics: Laying the groundwork for genetic disease modeling and targeting. *Clin Chem Lab Med* **36:** 523–527.

Gulcher J, Stefansson K. 1999. An Icelandic saga on a centralized healthcare database and democratic decision making. *Nat Biotechnol* **17:** 620.

Hammond SA, Warren RL, Vandervalk BP, Kucuk E, Khan H, Gibb EA, Pandoh P, Kirk H, Zhao Y, Jones M, et al. 2017. The North American bullfrog draft genome provides insight into hormonal regulation of long noncoding RNA. *Nat Commun* **8:** 1433.

Helleday T, Eshtad S, Nik-Zainal S. 2014. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* **15:** 585–598.

Honeyman JN, Simon EP, Robine N, Chiaroni-Clarke R, Darcy DG, Lim II, Gleason CE, Murphy JM, Rosenberg BR, Teegan L, et al. 2014. Detection of a recurrent DNAJB1-PRKACA chimeric transcript in fibrolamellar hepatocellular carcinoma. *Science* **343:** 1010–1014.

Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY, et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* **24:** 688–696.

Hulse-Kemp AM, Maheshwari S, Stoffel K, Hill TA, Jaffe D, Williams SR, Weisenfeld N, Ramakrishnan S, Kumar V, Shah P, et al. 2018. Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library. *Hortic Res* **5:** 4.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

Ivarsdottir EV, Steinthorsdottir V, Daneshpour MS, Thorleifsson G, Sulem P, Holm H, Sigurdsson S, Hreidarsson AB, Sigurdsson G, Bjarnason R, et al. 2017. Effect of sequence variants on variance in glucose levels predicts type 2 diabetes risk and accounts for heritability. *Nat Genet* **49:** 1398–1402.

Jensson BO, Hansdottir S, Arnadottir GA, Sulem G, Kristjansson RP, Oddsson A, Benonisdottir S, Jonsson H, Helgason A, Saemundsdottir J, et al. 2017. COPA syndrome in an Icelandic family caused by a recurrent missense mutation in COPA. *BMC Med Genet* **18:** 129.

Jones SJ, Haulena M, Taylor GA, Chan S, Bilobram S, Warren RL, Hammond SA, Mungall KL, Choo C, Kirk H, et al. 2017a. The genome of the northern sea otter (*Enhydra lutris kenyoni*). *Genes (Basel)* doi: 10.3390/genes8120379.

Jones SJM, Taylor GA, Chan S, Warren RL, Hammond SA, Bilobram S, Mordecai G, Suttle CA, Miller KM, Schulze A, et al. 2017b. The genome of the beluga whale (*Delphinapterus leucas*). *Genes (Basel)* doi: 10.3390/genes8120379.

Kaper F, Swamy S, Klotzle B, Munchel S, Cottrell J, Bibikova M, Chuang HY, Kruglyak S, Ronaghi M, Eberle MA, et al. 2013. Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc Natl Acad Sci* **110:** 5552–5557.

Kasianowicz JJ, Brandin E, Branton D, Deamer DW. 1996. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci* **93:** 13770–13773.

Kristjansson RP, Benonisdottir S, Oddsson A, Galesloot TE, Thorleifsson G, Aben KK, Davidsson OB, Jonsson S, Arnadottir GA, Jensson BO, et al. 2017. Sequence variant at 4q25 near PITX2 associates with appendicitis. *Sci Rep* **7:** 3119.

Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, Kertesz M, Snyder M. 2014. Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol* **32:** 261–266.

Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M. 2015. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat Biotechnol* **34:** 64–69.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. 2011. The sequence read archive. *Nucleic Acids Res* **39:** D19–D21.

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:** 682–686.

Li Q, Li Y, Song J, Xu H, Xu J, Zhu Y, Li X, Gao H, Dong L, Qian 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:** 1041–1049.

Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, Sawyer SL. 2013. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc Natl Acad Sci* **110:** 19872–19877.

Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney JP, Fehr A, et al. 2017. Resolving the full spectrum of human genome variation using linked-reads. *BioRxiv* doi: 10.1101/230946.

Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *Proc Natl Acad Sci* **74:** 560–564.

MetaSUB International Consortium. 2016. The metagenomics and metadesign of the subways and urban biomes (MetaSUB) International Consortium inaugural meeting report. *Microbiome* **4:** 24.

Metzker ML. 2010. Sequencing technologies—The next generation. *Nat Rev Genet* **11:** 31–46.

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.

Mullins N, Ingason A, Porter H, Euesden J, Gillett A, Olafsson S, Gudbjartsson DF, Lewis CM, Sigurdsson E, Saemundsen E, et al. 2017. Reproductive fitness and genetic risk of psychiatric disorders in the general population. *Nat Commun* **8:** 15833.

Navin NE. 2014. Cancer genomics: One cell at a time. *Genome Biol* **15:** 452.

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, et al. 2012. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149:** 979–993.

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci* **91:** 5022–5026.

Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J, et al. 2012. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487:** 190–195.

Reuter JA, Spacek DV, Snyder MP. 2015. High-throughput sequencing technologies. *Mol Cell* **58:** 586–597.

Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, Bernstein BE. 2015. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* **33:** 1165–1172.

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N. 1985. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230:** 1350–1354.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239:** 487–491.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94:** 441–448.

Sanger Institute. 2015. *Catalogue of somatic mutations in cancer*. Sanger Institute, Cambridgeshire, UK.

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:** 639–645.

Shamarina D, Stoyantcheva I, Mason CE, Bibby K, Elhaik E. 2017. Communicating the promise, risks, and ethics of large-scale, open space microbiome and metagenome research. *Microbiome* **5:** 132.

Simon EP, Freije CA, Farber BA, Lalazar G, Darcy DG, Honeyman JN, Chiaroni-Clarke R, Dill BD, Molina H, Bhanot UK, et al. 2015. Transcriptomic characterization of fibrolamellar hepatocellular carcinoma. *Proc Natl Acad Sci* **112:** E5916–E5925.

Steinberg S, Gudmundsdottir S, Sveinbjornsson G, Suvisaari J, Paunio T, Torniainen-Holm M, Frigge ML, Jonsdottir GA, Huttenlocher J, Arnarsdottir S, et al. 2017. Truncating mutations in RBM12 are associated with psychosis. *Nat Genet* **49:** 1251–1254.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337:** 64–69.

The 1000 Genomes Project Consortium; Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65.

The Cancer Genome Atlas Consortium. 2018. National Human Genome Research Institute, www.genome.gov/27571644.

Tsoucas D, Yuan GC. 2017. Recent progress in single-cell cancer genomics. *Curr Opin Genet Dev* **42:** 22–32.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Watson JD, Crick FH. 1953. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171:** 737–738.

Westbrook CJ, Karl JA, Wiseman RW, Mate S, Koroleva G, Garcia K, Sanchez-Lockhart M, O'Connor DH, Palacios G. 2015. No assembly required: Full-length MHC class I allele discovery by PacBio circular consensus sequencing. *Hum Immunol* **76:** 891–896.

Xia LC, Bell JM, Wood-Bouwens C, Chen JJ, Zhang NR, Ji HP. 2018. Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acids Res* **46:** e19.

Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, Hedges D, Ma X, Zhou X, Yergeau DA, et al. 2015. Germline mutations in predisposition genes in pediatric cancer. *N Engl J Med* **373:** 2336–2346.

Zhang GQ, Liu KW, Li Z, Lohaus R, Hsiao YY, Niu SC, Wang JY, Lin YC, Xu Q, Chen LJ, et al. 2017. The *Apostasia* genome and the evolution of orchids. *Nature* **549:** 379–383.

Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, et al. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34:** 303–311.

Cold Spring Harbor Perspectives in Medicine
www.perspectivesinmedicine.org