TOPICAL REVIEW

Genome sequencing guide: An introductory toolbox to whole-genome analysis methods

Alexis N. Burian1 | Wufan Zhao2 | Te-Wen Lo1 | Deborah M. Thurtle-Schmidt2

1Department of Biology, Ithaca College, Ithaca, New York, USA
2Department of Biology, Davidson College, Davidson, North Carolina, USA

Correspondence
Deborah M. Thurtle-Schmidt, Department of Biology, Box 7118, Davidson College, Davidson, NC 28035, USA.
Email: dethurtleschmidt@davidson.edu

Funding information
Davidson College

Abstract
To fully appreciate genetics, one must understand the link between genotype (DNA sequence) and phenotype (observable characteristics). Advances in high-throughput genomic sequencing technologies and applications, so-called “-omics,” have made genetic sequencing readily available across fields in biology from applications in non-traditional study organisms to precision medicine. Thus, understanding these tools is critical for any biologist, especially those early in their career. This comprehensive review discusses the chronological development of different sequencing methods, the bioinformatics steps to analyzing this data, and social and ethical issues raised by these techniques that must be discussed and evaluated, including anticipatory guides and discussion questions for active engagement in the classroom. Additionally, the Supporting Information includes a case study to apply technical and ethical concepts from the text.

KEYWORDS
bioinformatics, ethics, genomics, sequencing

1 INTRODUCTION

Since DNA was established as the heritable material by Martha Chase and Alfred Hershey, scientists have sought to understand the structure and sequence of an organism’s genome. In 1953, the structure of DNA was determined yet it was not until 1996 that the first eukaryotic genome sequence—baking and brewing yeast Saccharomyces cerevisiae—was published. Soon after, the first multicellular organism genome, C. elegans, was completed, prompting the race to sequence the human genome, culminating in the draft human genome sequence in 2001. Sequencing the human genome was a great achievement, but with the available technologies the effort was very labor and time-intensive, prompting new advances in DNA sequencing. This second wave of sequencing technology—called next generation sequencing—drastically decreased sequencing costs, increasing the amount of genomic and genome-scale information.

Genomic sequencing methods are now widely available, providing insight into basic molecular mechanisms from evolutionary analysis to personalized medicine. Additionally, genomic technologies can be applied to any methodology or organism in which nucleic acid can be extracted, making genomic methods widely accessible.
and “-omic” techniques a staple across fields and organisms. Due to the ubiquity of these techniques, it is imperative for scientists early in their careers to understand both the power and the peril associated with genome sequencing techniques. This review introduces sequencing technologies, analysis methods, and socio-ethical issues associated with genome sequencing to undergraduates. Through reading and engaging with the anticipatory guides and discussion questions with their peers and applying these concepts to the case study included in the Supporting Information, students should achieve the following learning outcomes:

- Explain the differences between Sanger and next-generation sequencing methods.
- Compare and contrast chip-based genotyping and whole-genome sequencing methods.
- Identify advances in chemistry that enabled sequencing by synthesis.
- Outline the general pipeline for high-throughput sequencing sample preparation and data analysis.
- Illustrate how various next-generation sequencing techniques can be exploited to understand different aspects of gene expression.
- Discuss the social justice and ethical implications associated with genome sequencing techniques.

## SEQUENCING AND WHOLE-GENOME ANALYSIS METHODS

Nucleic acid sequencing techniques have evolved since their inception with each new technique building off of previous sequencing technology and addressing a prior shortcoming. In this section, we will review the development of various sequencing technologies.

### 2.1 Sanger sequencing

**Anticipatory guides**

1. What is necessary for DNA polymerase to add the next nucleotide?
2. What is nucleic acid polarity and what implications does DNA polarity have on the double helix structure?
3. How are DNA fragments separated during agarose gel electrophoresis?
4. Review the steps of DNA amplification as in PCR.

Chain termination sequencing was the first nucleic acid sequencing method and revolutionized molecular biology, resulting in the 1980 Nobel Prize. Chain termination, also called Sanger sequencing as it was developed...
In the mid 1990s and early 2000s, two critical innovations brought on a fundamentally new sequencing}

2.2 | Chip-based detection methods

Anticipatory guides

1. Define heterozygous vs homozygous at a genetic locus.
2. Write the complementary sequence for: 5'-ATGCATCGAT-3'
3. Describe the process of DNA denaturation and annealing (hybridization) during PCR.
4. What is a single nucleotide polymorphism (SNP)? How are SNPs related to alleles?
5. Why is determining more than one DNA sequence at a time beneficial?

Sanger sequencing can only sequence a single DNA fragment per reaction. Although powerful—Sanger sequencing was used to sequence the human genome\textsuperscript{7}—sequencing a single fragment at a time has limitations.

Chip-based sequencing methods, called microarrays, sought to resolve this issue.\textsuperscript{16–18} It is important to note that these methods do not actually sequence DNA but allow for the simultaneous detection of different DNA sequence variants and mRNAs at once. Generally, DNA microarray chips consist of a solid surface dotted with small wells that contain a collection of single-stranded DNA specific to a gene, allele, or genomic region called the probe. Detection of different sequences is based on denatured, single-stranded samples hybridizing (attaching through hydrogen bonding) to complementary probes on the surface of the chip.\textsuperscript{19} Samples are prepared by extracting nucleic acid, fragmenting the nucleic acid into small pieces, denaturing the samples into single strands, and labeling the small fragments of nucleic acid with a fluorescent dye. These fragments are washed across the chip and hybridized to the DNA probes in wells that are complementary to the fluorescently labeled sample. The chip is then scanned and the quantity of each sample that anneals to each well is detected based on the amount of fluorescence present. The specific sequence and location on the chip for the DNA probe in each well is known, and the fluorescent signal correlates to the quantity of that sequence in the original sample.

To date, the main applications of DNA microarrays have been single nucleotide polymorphism (SNP) detection and relative mRNA quantification. For SNP detection, two adjacent wells contain probes specific to two common alleles in the population. The genotype of the sample is determined through detection of which wells the sample binds as the sample only binds to the well with the exact complementary sequence. These DNA microarray chips are still commonly used today to genotype people, as the human genome is costly to sequence. Additionally, microarrays can be used to determine relative gene expression through differential hybridization, correlating to gene expression, of differentially labeled control and experimental samples. Due to advances in sequencing by synthesis (see below), microarrays are not often used for RNA quantification anymore.

2.3 | Sequencing by synthesis

Anticipatory guides

1. What challenges are there to sequencing an entire genome with Sanger sequencing?
2. What types of questions could you address by sequencing all the DNA or RNA in an organism?
methodology, still referred to as “next generation sequencing,” “second generation sequencing,” or more generally “sequencing by synthesis,” in which a single DNA molecule is continually sequenced. Continually sequencing the same molecule, as opposed to chain termination in Sanger sequencing, was made possible due to new chemistry termed “reversible terminator chemistry.” A nucleotide with a reversible terminator has a blocked 3'-OH, similar to a ddNTP in Sanger sequencing, but after addition of another chemical solution the blocked 3'-group is reversed to a 3'-OH, again supporting sequencing (Figure 2). Each of these modified dNTPs is labeled with a different fluorescent dye. After each base is added to the elongating DNA strand, synthesis is halted because of the blocked 3'-OH, the dye is excited, and the color of the fluorescent nucleotide is recorded. Next, a chemical solution is added which both quenches the fluorescent dye (so that it no longer fluoresces) and reverses the blocked 3'-OH, supporting the next round of sequencing. There are several reversible terminators used commercially. The most common of which are 3' blocked reversible terminators with either a 3'-ONH2, 3'-O-allyl, or 3'-O-azidomethyl.

Another key innovation for sequencing by synthesis was simultaneous sequencing of multiple DNA sequences by attaching DNA strands to a flow cell, a two-dimensional microfluidic device (which resembles a microscope slide)—very similar to a microarray used in chip-based methods described above. First, DNA fragments are attached on one end to the flow cell. Next, each DNA molecule is amplified resulting in many copies of that DNA molecule in the same spot (or cluster) on the chip, amplifying the signal (sequence of the DNA molecule)—a step called “cluster generation.” Spots on the chip have different initial DNA molecules and there can be millions of individual DNA molecules on each chip. After cluster generation, sequencing proceeds. For each base in the DNA strand, reversible terminator modified nucleotides are added and the attached fluorophores are excited and the chip imaged (Figure 2). The colored images are translated into a DNA sequence, resulting in a single sequence for each cluster on the chip. Sequencing occurs for a defined number of rounds (usually between 50–300 bases, but can be as long as 500 bases), creating what is termed a “short read” of DNA sequence. Identification of the nucleotide incorporated for each DNA fragment relies on the amplified signal from the many copies of that DNA fragment in the cluster. At times one of the DNA fragments in a cluster gets “off phase” from all other DNA fragments in the cluster by accidently

![Figure 2](image-url) Sequencing by synthesis. (a) Cycle of reversible terminator incorporation, identification of incorporated base by fluorescence imaging, followed by removal of the reversible terminator. (b) Chemical structure of a nucleotide with a reversible terminator attached. The 3'-OH group is capped by a reversible terminator (black rectangle), with a fluorophore attached to the nitrogenous base (red circle). The fluorophore is then excited (red star), and the nucleotide is recorded. Finally, the fluorophore is cleaved from the nucleotide and the 3'-OH (highlighted in red) is unblocked for the next round of sequencing.
incorporating more than one nucleotide at a time, resulting in an incorrect signal of the nucleotide incorporated for that DNA fragment. The longer the sequencing, the more likely it is that some of the DNA molecules in a cluster get “off phase,” limiting the length of the sequencing reads.\(^\text{21}\) In addition, sequencing reads should be long enough to unambiguously map to the genome thus setting the limits of read lengths for sequencing by synthesis.

2.4 | Third generation sequencing

Anticipatory guides

1. What are the advantages of being able to sequence longer fragments of DNA?
2. What are the consequences if sequencing is not accurate?
3. What is a processive enzyme and provide an example of a processive enzyme that uses a nucleotide substrate?
4. What is a protein pore in membrane bilayers?

Third generation sequencing technologies, called single molecule sequencing (SMRT) and nanopore sequencing, rely on sequencing single nucleotide molecules.\(^\text{22}\) Like Sanger sequencing and sequencing by synthesis, SMRT sequencing, developed by PacBio, also relies on synthesizing a new DNA strand by DNA polymerase.\(^\text{23}\) However, the DNA polymerase is immobilized at the bottom of a tiny well in the sequencing chip. Each well has a single piece of DNA to be sequenced and each dNTP is given a fluorescent label with a unique emission spectrum. The immobilized DNA polymerase begins to replicate the DNA strand and as each dNTP is added, the fluorophores are excited. The sequence of the DNA can then be easily determined based on the emission spectra observed, which belong to the incorporated nucleotides that were detected.

A complementary third generation sequencing technology was developed that, like SMRT sequencing, relies on direct detection of a single nucleotide molecule but does not rely on DNA synthesis. Instead, nanopore sequencing (Figure 3) uses a membrane protein complex. This protein complex consists of two proteins: (a) an unwinding enzyme and (b) a pore protein which allows molecules to pass through a lipid bilayer. The unwinding enzyme, such as polymerase or helicase, unwinds the double helix so that a single nucleic acid strand (DNA or RNA) passes through the pore protein.\(^\text{24}\) This pore protein is inserted in a synthetic lipid bilayer. A commonly used pore protein is MspA, which is a transmembrane protein found in Mycobacteria used to transport nutrients across the bacterial membrane.\(^\text{25,27,28}\) The lipid bilayer has variable voltage on either side. As nucleotides pass through the unwinding enzyme and the pore, the mass of the nucleotide creates a distinct change in current. From the specific current signature detected, the sequence of the nucleotide strand is determined.

Third generation sequencing is characterized by its ability to sequence much longer reads. Both SMRT and nanopore technologies have reported reads of at least 8000 bp as compared to sequencing by synthesis in which the longest reads are 500 bp.\(^\text{29}\) However, longer reads come at the expense of sequencing accuracy—both third generation technologies have much higher error rates than second generation sequencing.\(^\text{29,30}\) To improve sequencing accuracy, in nanopore sequencing, the two strands of DNA are ligated with a hairpin structure, thus when the DNA is denatured and passed through the pore as a single stranded molecule, both complementary strands are sequenced (Figure 3). This provides twice the sequence for one strand, helping to resolve unclear base calls. Similarly, prior to SMRT sequencing, hairpins are ligated to both ends of DNA, resulting in a circular single stranded DNA. This DNA molecule can be continually sequenced by the immobilized polymerase, resulting in better base calling due to the multiple sequencing rounds.

2.5 | General discussion questions

i. How does a dideoxynucleotide prevent elongation by DNA polymerase?

ii. What aspects of Sanger sequencing gave way to sequencing by synthesis?
iii. What aspects of DNA microarray chips gave way to sequencing by synthesis?

iv. Draw a picture of the results on a DNA microarray for a sample homozygous at a locus and for a sample heterozygous at the locus.

v. Why would adding all four nucleotides at the same time in sequencing by synthesis result in more accurate sequencing?

vi. Pose a research question appropriate for each of the technologies discussed above.

vii. Compare and contrast the different sequencing/genome detection methods.

viii. For each of the sequencing methods above, enumerate the significance and limitations.

ix. If sequencing a new genome, why would using a combination of sequencing by synthesis and third generation sequencing be advantageous?

3 | SEQUENCING PIPELINE

The sequencing pipeline is a three-step process: sample and library preparation, sequencing, followed by data analysis and bioinformatics. Above described the second step—sequencing. This section will describe the process of steps one and three.

3.1 | Sample and library preparation

Anticipatory guides

1. What are challenges to sequencing many different fragments of DNA at once using sequencing by synthesis?
2. What are primers and why are they necessary for DNA replication?
3. What is cDNA? How does cDNA sequence differ from the genomic DNA?

Before sequencing, the nucleic acid sample is isolated using traditional molecular biology techniques. Applications where determining the sequence or the amount of different specific sequences in a sample are typical applications of second and third generation sequencing technologies (reviewed in Reuter et al.) . After nucleic acid isolation, one of the challenges to genomic sequencing methods is the preparation of many millions of different sequences for sequencing at the same time (Figure 4). For sequencing by synthesis and SMRT sequencing methods (which rely on DNA polymerase) all the sequences must have at least some common sequence to which a primer can anneal. In second generation sequencing, DNA sequences are adhered to the chip by hybridizing to a complementary single-stranded DNA oligonucleotide and a primer also binds to this sequence supporting cluster generation. Thus, the necessity of a common DNA sequence on each DNA fragment for chip hybridization, cluster generation, and sequencing is at odds with the innovation that many different DNA pieces of unknown sequences are sequenced simultaneously.

To overcome this problem, the nucleic acid sample is prepared into a “library”—a collection of DNA fragments each with common sequences (adaptors) on either end (Figure 4(b)). First, if the sample is RNA, it is converted to cDNA (complementary DNA) using reverse transcriptase as DNA is much more stable than RNA and DNA polymerase requires a DNA template. Since sequencing by synthesis requires short pieces of DNA, the DNA is sheared to less than 500 bps in length. Since each fragment of DNA is unique, the same adaptors (pieces of DNA) must be added to the ends of each fragment to replicate and sequence each unique fragment simultaneously. The first step in adaptor attachment is to add a single “A” base to the 5’ ends of each sequence. This off-hanging “A” base allows the adaptors to attach through ligating to the complementary “T” overhang on the 3’ end of the adaptor. The ends of the adaptor are complementary to the end of the primer sequence, which through PCR both amplifies the library so that there is enough material for sequencing and extends to add the primer sequence. After this PCR step, the DNA can attach to the flow cell, and primers support cluster generation and subsequent sequencing by synthesis. Samples prepared for nanopore sequencing have a very similar library preparation step, adding adaptors to each of the fragments, however fragmentation is not necessary since these technologies support sequencing longer pieces. Even though nanopore sequencing relies on direct detection and not sequencing by synthesis, the adaptors are necessary to feed the nucleic acid through the pore by the ratcheting enzyme.

3.2 | Data analysis and bioinformatics

Anticipatory guides

1. From the steps in sequencing by synthesis described above, identify what determines the length of the sequence returned to the user.
2. What is the risk of only sequencing each base one time?
3. What information would be useful from the sequencing reaction for analyzing the accuracy of the base call?
4. What are intron/exon boundaries?
In high-throughput sequencing, millions of reads are sequenced. A read is the sequence of each DNA fragment and in second generation sequencing the length of the sequence is defined by the number of sequencing cycles (the number of times modified dNTPs were added and imaged), typically from 50 to 300 base pairs in 50-base-pair increments. Thus, since the DNA is typically sheared to 200–500 base pairs during the library preparation, the entire DNA fragment is not sequenced. In a typical sequencing reaction, termed single-end sequencing, the fragment is sequenced from just one end of the DNA fragment. A paired-end sequencing reaction sequences each fragment from both of the DNA ends, providing twice as much sequencing information of the same piece of DNA. The reads are returned to the user in a plain text file termed a FASTQ file (Figure 5). The FASTQ file format is a repeating unit of four lines: (a) the name of the read, which begins with an “@” symbol; (b) the sequence of the read; (c) a separator, a single “+” (plus) sign, to make the file easier to read; (d) the quality score line. Each base pair in the sequence receives a quality score termed the Q-score ranging from 1 to 40 with 1 indicating the least confidence that the base call is correct and 40 being the most confident. For example, “I” is a score of 40 which translates to 99.99% accuracy for that base call. The symbol code, which is an ASCII based code, ensures that each numerical score only takes up a single character space so that it lines up with the appropriate base. This repeating four lines continues for the millions of reads sequenced. For second generation sequencing, a single sequencing sample can produce over 150 million reads.

Bioinformatic analysis consists of quality control of the reads and then mapping the reads to the genome of interest. For quality control, the fourth line for each sequence is read in the FASTQ file, to determine if there is sufficient confidence in each base call. Based on quality control results, some trimming of low-quality bases may be required to ensure that only high-quality bases are
Another common preprocessing step is to remove the general adaptor sequence so that the reads map more reliably to the genome, which is the next step in bioinformatic analysis. Most often, second and third generation sequencing is not used to sequence a new genome from scratch but rather for analyzing and quantifying the sequences of a nucleic acid sample of interest from an organism with a sequenced reference genome by mapping the reads to this reference genome (Figure 5). For nucleic acid samples from DNA, mapping is straightforward (although computationally intensive), and reads are compared to the entire known genome to find the place that matches the read. After all reads are mapped to the genome, the amount of coverage is determined by approximating how many times each nucleotide is represented in all of the sequencing reads (Figure 5). For nucleic acid samples from RNA, mapping is more complex due to alternative splicing and the need to map reads that span exon-intron boundaries.

### 3.3 General discussion questions

i. Explain the purpose of the adaptor and primer sequences in a genomic library.

ii. On the flow-chart diagram in Figure 4, draw the library preparation steps for fragments of DNA.

iii. How can you determine how many reads were sequenced from the number of lines in a FASTQ file?

iv. Outline the different steps needed to go from RNA-seq FASTQ files to gene expression quantitation.

v. Explain how RNA-seq could be used to detect alternative splicing.

vi. In what applications would paired-end sequencing be desirable over single-end sequencing?

vii. Why is “high coverage” important when trying to identify mutations in a sample?

### 4 Social Implication of Genome Sequencing

Power lies within the genomic tools discussed above. This power can be enormously beneficial to millions of people and change lives, but researchers must consider the long-term consequences and contemplate the social and ethical ramifications. Below we illuminate some of the ethical and social consequences when genetic sequencing is used for medical advancements and placed directly in the hands of consumers.

#### 4.1 Knowledge is power

**Anticipatory guides**

1. **How will genetic testing change medical treatments?**

2. **Is the application of genetic testing limited to human diseases?**

The ease in sequencing and decrease in cost has led to numerous discoveries linking genes to diseases. Genome-wide association studies (GWAS) facilitate linking complex genetics to differential phenotypes. GWAS identifies specific SNPs associated with diseases by comparing common sequence variants and/or
genomes between unaffected individuals to those individuals with a phenotype or interest.36 Knowing what SNPs are associated with particular diseases, is the foundation of precision medicine.37 Precision medicine allows medical professionals to choose the most effective treatments based on an individual’s genetic sequence.

Sequencing technologies have led to numerous direct-to-consumer sequencing companies that allow individuals to learn about their own genetics without professional medical assistance.38 In most cases, consumers simply mail a saliva sample which is genotyped, using a DNA microarray as described above with probes to common variants in the human population. Consumers learn the sequence of their genomic loci known to be associated with different phenotypes such as lactose intolerance, heart disease, or caffeine sensitivity. As a consumer, a person must decide what they hope to learn about their genetic make-up to select the appropriate direct-to-consumer sequencing service.

4.2 | Does everyone have equal access to this “power”?

Anticipatory guides

1. How do social inequalities affect access to genetic testing and relevant treatments?
2. Who should be responsible for educating people about genetic testing and its implications?

Access to genomic technologies has traditionally been through clinical genetic testing. Sequencing by synthesis has revolutionized clinical genetic tests by allowing interrogation of multiple different genes or even the entire genome of a patient sample at once, significantly speeding up genetic testing. Results from these tests can be used both for diagnosis and to identify targeted therapies. However since clinical genetic testing is administered in a healthcare setting, social inequities which are well-documented in healthcare also plague genetic testing.39 Access to clinical genetic testing is not equivalent across all racial and socioeconomic communities40 due to factors such as differences in comprehensive health insurance among racial groups41 and mistrust of medical testing by individuals from groups historically excluded from healthcare.40 These disparities put racial minorities at increased disadvantage to reaping the benefits of clinical genetic testing.

Even though direct-to-consumer tests can be more affordable ($60–$200 depending on the comprehensiveness of the test) the costs are out-of-pocket and still represent a significant barrier to access. Additionally, the health and lifestyle genetic risk factors that direct-to-consumer report on are from studies with inequitable racial representation. The majority of genetic research databases and GWAS include mostly European-descent genomes, indicating a serious gap in “who” is being solicited to participate in genetic research.42 This lack of representation decreases the applicability of results across populations and understanding of genetic diseases in non-European populations. Thus, even if an individual has the means to access these tests, it is not a given if those results are applicable to them.

4.3 | Can this “power” end up in the wrong hands?

Anticipatory guides

1. Who might you want to keep your genetic results from?
2. How secure (private) are genetic testing results?

While there may be great benefits of having a better understanding of your genetic make-up, before signing up with a direct-to-consumer genetic test, the consumers must consider who owns and has access to their data and fully understand the companies’ privacy policies prior to submitting their sample. To help protect an individual’s privacy, legislation has been enacted to protect the consumer’s privacy. The 21st Century Cures Act seeks to protect an individual’s confidentiality when genetic information is donated to federal research purposes by removing all identifiers (i.e., donor’s name and contact information).43 Any information obtained from the research cannot be released to law enforcement or government agencies. In addition, under the Health Insurance Portability and Accountability Act (HIPPA) one’s genetic information is protected from employers, schools, and the public if it becomes a part of one’s health record.44 The entities that can access this information are law enforcement and health insurance.

With the rise of genetic testing came concern about genetic discrimination if health insurance companies had access to genetic testing results; companies could discriminate against those who tested positive for differing genetic predispositions and alter their healthcare coverage. The Genetic Nondiscrimination Act (GINA) was passed in 200845 to prevent health insurance companies from denying coverage and changing rates based on genetic predispositions. However, it only protects individuals who are not showing any symptoms of the predisposition. If symptoms of the genetic difference are present, then insurance companies could alter coverage and rates. GINA also prohibits employers from changing the employment status based on genetic testing results. Direct-to-consumer testing companies have their own
privacy policies that should be considered before using the service. These considerations highlight how a scientist should also understand how a scientific tool is being used in addition to the development of such technologies.

4.4 Discussion questions

i. **What human diseases are ideal for genetic testing and precision medicine?**

ii. **What are the limitations of genetic testing?**

iii. **Is it always beneficial for someone to know genotype(s)? Are there genotypes that one might not want to know?**

iv. **What direct-to-consumer service(s) do you think provides the most interesting (relevant) information?**

v. **Should restrictions be placed on genetic testing? Why or why not?**

vi. **Who should be responsible for genetic testing costs?**

vii. **What are potential concerns regarding genetic testing and results? What issues should that legislation address?**

viii. **Who should have access to genetic testing results?**

ix. **Should restrictions be placed on genetic testing and results? What issues should that legislation address?**

ACKNOWLEDGMENT

Funding support provided by Davidson College to D. T-S.

ORCID

Te-Wen Lo https://orcid.org/0000-0002-1231-5531
Deborah M. Thurtle-Schmidt https://orcid.org/0000-0002-9747-9585

REFERENCES

1. Hershey AD, Chase M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol. 1952;36:39–56.

2. Watson JD, Crick FHC. A structure for deoxyribose nucleic acid. Nature. 1953;171:737–8.

3. Franklin RE, Gosling RG. Molecular configuration in sodium thymonucleate. Nature. 1953;171:740–1.

4. Wilkins MHF, Stokes AR, Wilson HR. Molecular structure of deoxypentose nucleic acids. Nature. 1953;171:738–40.

5. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science. 1996;274:546–67.

6. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998;282:2012–8.

7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.

8. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376–80.

9. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005;309:1728–32.

10. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-molecule DNA sequencing of a viral genome. Science. 2008;320:106–9.

11. Lundquist PM, Zhong CF, Zhao P, Tomaney AB, Peluso PS, Dixon J, et al. Parallel confocal detection of single molecules in real time. Opt Lett. 2008;33:1026–8.

12. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452:872–6.

13. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53–9.

14. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74:5463–7.

15. Atkinson MR, Deutscher MP, Kornberg A, Russell AF, Moffatt JG. Enzymatic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2,3′-dideoxyribonucleotide. Biochemistry. 1969;8:4897–904.

16. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270:467–70.

17. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, et al. Accessing genetic information with high-density DNA arrays. Science. 1996;274:610–4.

18. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. 1996;6:639–45.

19. Maskos U, Southern EM. A novel method for the analysis of multiple sequence variants by hybridisation to oligonucleotides. Nucleic Acids Res. 1993;21:2267–8.

20. Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, et al. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci U S A. 2006;103:19635–40.

21. Tan G, Opitz L, Schlappach R, Rehbrauer H. Long fragments achieve lower base quality in Illumina paired-end sequencing. Sci Rep. 2019;9:1–7.

22. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19:2247–40.

23. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323:133–8.

24. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci U S A. 1996;93:13770–3.

25. Butler TZ, Pavlenok M, Derrington IM, Niederweis M, Gundlach JH. Single-molecule DNA detection with an engineered MspA protein nanopore. Proc Natl Acad Sci U S A. 2008;105:20647–52.
26. Reuter JA, Spacek D, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2016;58:586–97.

27. Derrington IM, Butler TZ, Collins MD, Manrao E, Pavlenok M, Niederweis M, et al. Nanopore DNA sequencing with MspA. Proc Natl Acad Sci USA. 2010;107:16060–5.

28. Manrao EA, Derrington IM, Pavlenok M, Niederweis M, Gundlach JH. Nucleotide discrimination with DNA immobilized in the MSPA nanopore. PLoS One. 2011;6:1–7.

29. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. Improved data analysis for the MinION nanopore sequencer. Nat Methods. 2015;12:351–6.

30. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. Genome Biol. 2013;14:405.

31. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research. 2010;38:1767–1771. https://doi.org/10.1093/nar/gkp1137.

32. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.

33. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.

34. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32:462–4.

35. Rego SM, Snyder MP. High throughput sequencing and assessing disease risk. Cold Spring Harb Perspect Med. 2019;9:1–12.

36. Genome-Wide Association Studies Fact Sheet. Bethesda, Maryland: National Human Genome Research Institute; 2020. https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet

37. Toward precision medicine. Washington, DC: National Academies Press; 2011.

38. Niemiec E, Kalokairinou L, Howard HC. Current ethical and legal issues in health-related direct-to-consumer genetic testing. Per Med. 2017;14:433–45.

39. Smedley BD, Stith AY, Nelson AR, editors. Unequal treatment: confronting racial and ethnic disparities in health care. Washington, DC: National Academies Press; 2003.

40. Peters N, Rose A, Armstrong K. The association between race and attitudes about predictive genetic testing. Cancer Epidemiol Biomarkers Prev. 2004;13:361–5.

41. Lillie-Blanton M, Hoffman C. The role of health insurance coverage in reducing racial/ethnic disparities in health care. Health Aff. 2005;24:398–408.

42. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. Cell. 2019;177:26–31.

43. 21st Century Cures Act. USA; 2016.

44. Health Insurance Portability and Accountability Act of 1996; 1996.

45. The Genetic Information Nondiscrimination Act; 2008.

SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Burian AN, Zhao W, Lo T-W, Thurtle-Schmidt DM. Genome sequencing guide: An introductory toolbox to whole-genome analysis methods. Biochem Mol Biol Educ. 2021; 49:815–825. https://doi.org/10.1002/bmb.21561