Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

Genome Sequencing Guide: An introductory toolbox to whole-genome analysis methods

Alexis Burian[*], Wufan Zhao[†], Te-Wen Lo[*], Deborah M. Thurtle-Schmidt[†]

[*] Department of Biology, Ithaca College, Ithaca, NY 14850
[†] Department of Biology, Davidson College, Davidson, NC 28035

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.

Introduction

The genome is a cellular blueprint, providing instructions for each cell in every organism. 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 (Hershey and Chase 1952). In 1953, the structure of DNA was determined (Franklin and Gosling 1953, Watson and Crick 1953, Wilkins et al. 1953), yet it was not until 1996 that the first eukaryotic genome sequence – baking and brewing yeast *Saccharomyces cerevisiae* – was published (Goffeau et al. 1996). Soon after, the first multicellular organism genome, *C. elegans (C. elegans* Sequencing Consortium 1998) was completed, prompting the race to sequence the human genome, culminating in the draft human genome sequence in 2001 (Lander et al. 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. Shortly after the publication of the human genome, a second wave of sequencing technology – called next generation sequencing – drastically decreased sequencing costs, increasing the amount of genomic and genome-scale information (Margulies et al. 2005, Shendure et al. 2005, Bentley et al. 2008, Harris et al. 2008, Lundquist et al. 2008, Wheeler et al. 2008).

Genomic sequencing methods are now widely available across disciplines, 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 now 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. Thus, this review is aimed at undergraduates and introduces sequencing technologies, analysis methods, and ethical issues associated with genome sequencing. Through reading and engaging with the anticipatory guides and discussion

© 2020 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

questions with their peers and applying these concepts to the case study included in the supplement, 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 important 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 the beginning of the field of molecular biology where each new technique builds off of previous sequencing technologies and addresses a prior shortcoming. In this section, we will review, the development of various sequencing techniques.

Sanger sequencing

Anticipatory Guides

1. *What is necessary for DNA polymerase to synthesize or add the next base?*
2. *What is nucleic acid polarity? What is the 5' end of DNA and what is the 3' end of DNA? 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 by Fred Sanger in 1977, uses the selective incorporation of dideoxynucleotides during an *in vitro* DNA replication reaction (Sanger et al. 1977) (Figure 1). During DNA replication, DNA polymerase catalyzes the synthesis of DNA by adding the next complementary nucleotide to the 3' end of the growing DNA strand. DNA polymerase requires a 3'-hydroxyl group (-OH) to be able to add the next base (Atkinson et al. 1969). Without an available 3'-OH, the polymerase cannot form a phosphodiester bond with the previous base and DNA synthesis stops. Sanger sequencing exploits this requirement. In the sequencing reaction, in addition to deoxynucleotides (**d**NTPs), which support DNA synthesis as the 3' carbon contains a hydroxyl group (Figure 1B), dideoxynucleotides (**dd**NTPs) are also present and lack the 3'-OH (Figure 1B), which prevents polymerase from adding the next base. Dideoxynucleotides are labeled with a fluorescent dye and are at a much lower concentration than the deoxynucleotides, resulting in the DNA polymerase incorporating a dNTP most of the time, supporting further synthesis. However, ddNTPs are incorporated at a low frequency, halting synthesis. This synthesis reaction results in many different DNA fragments of varying lengths complementary to the template being

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

sequenced, each ending with the fluorescently labeled ddNTP (Figure 1A). Each of the four different nucleotides are conjugated to a different dye, which emit a distinct wavelength when excited.

The sequence of the DNA molecule is determined through separating out all the newly synthesized DNA fragments by size, using capillary gel electrophoresis. The capillary (a very thin tube) contains a gel matrix (similar to agarose gel electrophoresis) that separates out DNA molecules by size with single base resolution. As in agarose gel electrophoresis, the smaller the DNA molecule the faster the molecule moves through the capillary. At the end of the capillary, a laser excites the base at the end of the chain and the fluorescent dye color is detected, allowing the sequence to be recreated by the order of the laser excitations wavelengths observed (fluorescent dyes detected) (Figure 1A).

Chip-based detection methods

Anticipatory Guides

5. *What does it mean to be heterozygous vs homozygous at a genetic locus?*
6. *Write the complementary sequence for: 5'-ATGCATCGTAT-3'*
7. *Describe the process of DNA denaturation and annealing (hybridization) during PCR.*
8. *What is a single nucleotide polymorphism (SNP)? How are SNPs related to alleles?*
9. *Why is determining more than one DNA sequence at a time beneficial?*

Sanger sequencing only allows for the sequencing of a single DNA fragment in each reaction. Although powerful – Sanger sequencing was used to sequence the human genome (Lander et al. 2001) – sequencing a single fragment at a time has limitations. Chip-based sequencing methods, called microarrays, sought to resolve this issue (Schena et al. 1995, Chee et al. 1996, Shalon et al. 1996). 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, paving the way for genomic studies. 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 of nucleic acid hybridizing or annealing (attaching through hydrogen bonding) to complementary probes on the surface of the chip (Maskos and Southern 1993). Samples are prepared by extracting nucleic acid (either DNA or RNA depending on the scientific question you want to investigate), fragmenting the nucleic acid into small pieces, denaturing the samples so that it is single stranded, and labeling the small fragments of nucleic acid with a fluorescent dye. The single-stranded samples 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. Since the specific sequence for the DNA probe in each well is known and its place on the chip documented, fluorescent signals can detect the quantity of that sequence in the original sample. The higher the fluorescent signal, the higher the quantity of that particular sequence in the sample.

To date, the main applications of DNA microarrays have been single

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

nucleotide polymorphism (SNP) detection and relative mRNA quantification. For SNP detection, each well contains a probe specific to one of two common alleles in the population of interest. An adjacent well contains DNA probes with the sequence of the other main allele in the population. Thus, the genotype of a sample can be determined by detecting the binding of the sample to these wells as the sample will only bind to the well with the exact complementary sequence: if the sample binds to one well it is homozygous for that allele, if the sample binds to the other well it is homozygous for the other allele, or if the sample binds to both wells it is heterozygous. 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 levels: mRNA samples from a control and experimental condition are extracted. These samples are reversed transcribed to generate cDNA and each sample labeled with a different fluorescent probe (generally Cy3, green, for the control and Cy5, red, for the experimental sample). Each well of the microarray contains many copies of a gene-specific probe. The two samples are washed and allowed to hybridize to the chip simultaneously, resulting in a competition experiment: If the gene is more expressed in the experimental sample and thus there is more mRNA to bind to that well, the well will appear red, for a gene more expressed in the control sample than the well will appear green, and if the gene is expressed equally in both samples than both green and red samples will bind, resulting in a yellow color when scanning. Due to advances in sequencing by synthesis (see below), microarrays are not often used for RNA quantification anymore.

Sequencing by synthesis

Anticipatory Guides

10. *Why challenges would you face if you were to sequence an entire genome with sanger sequencing?*
11. *What types of questions could you address by sequencing all the DNA in an organism?*
12. *What types of questions could you address by sequencing all the RNA in an organism?*
13. *What are fluorescent molecular dyes and why are they useful in molecular biology and chemistry?*

In the mid 1990s and early 2000s, two critical innovations brought on a fundamentally new sequencing methodology, still referred to as “next generation sequencing”, “second generation sequencing”, or more generally “sequencing by synthesis,” in which a single DNA molecule can be 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) (Ju et al. 2006, Bentley et al. 2008). Each of these modified dNTPs is labeled with a different color (similar to the ddNTPs in sanger sequencing). 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

longer fluoresces) and reverses the blocked 3' position into a 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 (Ju et al. 2006).

The other 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 to be sequenced 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" (Figure 2) (Bentley et al. 2008). 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 can proceed. For each base in the DNA strand, reversible terminator modified nucleotides are added and the attached fluorophores are excited and the chip imaged. The colored images are translated into a DNA sequence by the sequencing software, resulting in a single sequence for each cluster on the chip. Sequencing occurs for a defined number of rounds (usually between 50 – 300 bases), creating what is termed a "short read" of DNA sequence.

Third Generation Sequencing

Anticipatory Guides

14. *What are the advantages of being able to sequence longer fragments of DNA?*
15. *What are the consequences if sequencing is not accurate?*
16. *What is a processive enzyme and what is an example of a processive enzyme that uses a nucleotide substrate?*
17. *What is a protein pore in membrane bilayers?*

We are now in what is considered the third wave of DNA sequencing. Third generation sequencing technologies, called single molecule sequencing (SMRT) and nanopore sequencing, rely on sequencing single nucleotide molecules, unlike clusters of molecules in second generation sequencing, and direct detection of the DNA molecule as it is sequenced in real time (Schadt et al. 2010). Like Sanger sequencing and sequencing by synthesis, SMRT sequencing, developed by PacBio, also relies on synthesizing a new DNA strand by DNA polymerase (Eid et al. 2009). 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 - dA, dT, dG, dC - is each fluorescently labeled, emitting a distinct emission spectra (wavelength of light). The immobilized DNA polymerase begins to replicate the DNA strand and as each dNTP is added, the fluorophores are excited, and the unique emission spectra for each incorporated nucleotide is directly detected. Since each nucleotide has its own unique emission spectra, the sequence of the DNA can be easily determined based on the emission spectra detected.

A complementary third generation sequencing technology was developed that, like SMRT sequencing, relies on direct

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

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: (1) an unwinding enzyme and (2) a pore protein which allows molecules to pass through a lipid bilayer. The unwinding enzyme unwinds the double helix so that a single nucleic acid strand (DNA or RNA) passes through the pore protein (Kasianowicz et al. 1996). This pore, such as MspA, is inserted in a synthetic lipid bilayer with a variable voltage on either side of the bilayer. The most common unwinding enzymatic proteins are DNA helicase or DNA polymerase, as these proteins move nucleic acids one base at a time. As nucleotides pass through the unwinding enzyme and the pore, the mass of the nucleotide changes the current, creating a distinct change in current for each nucleotide. 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 8,000 bp as compared to sequencing by synthesis in which the longest reads are 300 bp (Jain et al. 2015). However, longer reads come at the expense of sequencing accuracy – both third generation technologies have much higher error rates than second generation sequencing (Roberts et al. 2013, Jain et al. 2015). 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.

General Discussion Questions:

i. *Why does a dideoxynucleotide not support further 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. *Why would adding all four nucleotides at the same time in sequencing by synthesis reaction result in more accurate sequencing?*
v. *Identify an example of research questions where each of these technologies would be advantageous.*
vi. *Compare and contrast the different sequencing/genome detection methods.*
vii. *For each of the sequencing methods above, enumerate the significance and limitations.*
viii. *If sequencing a new genome, why would using a combination of sequencing by synthesis (second generation sequencing) and third generation sequencing be advantageous?*

The Sequencing Pipeline

The sequencing pipeline can be best described as 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

Sample and Library Preparation

Anticipatory Guides

18. *What are challenges to sequencing many different fragments of DNA at once using sequencing by synthesis?*
19. *What are primers and why are they necessary for DNA replication?*
20. *What is cDNA and how does the sequence differ from the genomic sequence of a gene?*
21. *Why would a scientist want to sequence mRNA as opposed to genomic DNA? What is different about the resulting sequencing data and the types of questions that can be answered?*

Before nucleotides can be sequenced, the researcher must prepare the nucleic acid using traditional molecular biology techniques. Anything that results in isolation of nucleic acid – DNA or RNA – and where either determining the sequence or the amount of different specific sequences present would be illuminating can be sequenced using second and third generation sequencing technologies (example applications reviewed in: Reuter et al. 2016). Once the nucleic acid is isolated, it must be prepared for sequencing. One of the challenges to genomic sequencing methods (second and third generation sequencing) 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, and DNA polymerase can extend from that 3'-OH. Additionally, for second generation sequencing, the DNA sequences are adhered to the chip by hybridizing to a complementary DNA oligonucleotide (single stranded DNA) and a primer also binds to this sequence to support the initial amplification that must occur for cluster generation (the amplification of the signal at each spot) (Bentley et al. 2008). Thus, the necessity of a common DNA sequence on each DNA fragment for chip hybridization, cluster generation, and sequencing is at the odds with part of the innovation in second and third generation sequencing: that many different DNA pieces of unknown sequences could be 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 (Bentley et al. 2008). First, the nucleic acid sample is purified through general molecular techniques and 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. Sequencing by synthesis requires short pieces of DNA. To ensure that each piece of DNA is an appropriate length, the DNA is sheared to be 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 so that each unique DNA fragment can be replicated and sequenced simultaneously. The first step in adaptor attachment is to add a single "A" base to the 5' ends of each sequence. This single "A" hanging off the 5' end allows for the adaptors to be attached through ligation to the DNA end which have a 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

that there is enough for sequencing and extends to add the primer sequence. After this PCR step, each piece of DNA has the general adaptor and primer sequences and can now be attached to the flow cell, and primers can 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 much longer pieces can be sequenced using nanopore and SMRT sequencing technologies. Even though nanopore sequencing relies on direct detection and not sequencing by synthesis, the adaptors are necessary for the ratcheting protein to feed the nucleic acid through the pore (Jain et al. 2015).

Data analysis and bioinformatics

Anticipatory Guides

22. *What are the challenges of analyzing millions of sequences of DNA?*
23. *From the steps in sequencing by synthesis described above, identify what determines the length of the sequence returned to the user.*
24. *What is the risk of only sequencing each base one time?*
25. *What information would you like to get from the sequencing reaction for analyzing the accuracy of the sequence read?*
26. *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 – 250 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 the 5' end. A paired-end sequencing reaction sequences each fragment from both the 5' and 3' ends. The reads are returned to the user in a plain text file termed a FASTQ file (Figure 5) (Cock et al. 2009). The FASTQ file format is a repeating unit of four lines: line one is the name of the read, which follows a specific convention for the sequencing technology used and usually begins with an "@" symbol. Line two is the sequence of the read. Line three is a separator which is always a single "+" (plus) sign and makes it easier to read the file. Finally, the fourth line is the quality score line. Each base pair in the sequence on line two receives a quality score termed the Q-score that scores each base from 1-40 with 1 indicating that there is the least confidence that the base call is correct and 40 being the most confident (Bentley et al. 2008). For example, "I" is a score of 40 which translates to 99.99% accuracy for that base call or put another way, the probability that this base was incorrectly read on the sequencer is 1 in 10,000. The symbol code, which is an ASCII based code, is used so that each numerical score only takes up a single character space so that it lines up with the appropriate base. This repeating 4 lines continues for the millions of reads sequenced. For second generation sequencing, a single sequencing sample can produce over 150 million reads (sequences of 150 million different DNA fragments from the library).

Bioinformatic analysis consists of quality control analysis of the reads and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

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. A common program to analyze read quality is FastQC, which calculates the median score at each position across reads (Andrews 2010). Additionally, FastQC highlights other quality features such as if reads contain the adaptor sequence or if there are overrepresented sequences in the sample (which a researcher would most likely not expect and could indicate a problem in sample preparation). Based on quality control results, some trimming of low-quality bases may be required to ensure that only high-quality bases are being included in the analysis. Another common pre-processing step is to remove the general adaptor sequence if it is present in the read. By removing the adaptor sequence, the reads will much more reliably map 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 experiment in an organism with a sequenced genome. Thus, reads are mapped to the reference genome of interest (Figure 5). For nucleic acid samples from DNA, mapping is straight-forward (although computationally intensive) and reads are compared to the entire known genome sequence to find the place in the genome that matches the read. After all reads have been 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 RNA-seq, which sequences the mRNA of a sample and identifies gene expression, more sophisticated mapping algorithms are used that can map reads which span exon-intron boundaries, which would result in part of the read mapping in a different location than the other end of the read as compared to the genomic sequence. Once mapped, the coverage of the gene in RNA-seq samples is used to determine that gene's expression in a sample. This expression can be compared across samples to determine differentially regulated genes between different conditions. More recently, new RNA-seq mapping algorithms skip over the labor-intensive mapping portion and directly quantitate transcript levels, which has significantly decreased processing time (Patro et al. 2014, 2017, Bray et al. 2016).

General Discussion Questions:

ix. *Explain the purpose of the adaptor and primer sequences in a genomic library.*
x. *On the flow-chart diagram in Figure 4, draw the library preparation steps for fragments of DNA.*
xi. *How can you determine how many reads were sequenced from the number of lines in a FASTQ file?*
xii. *Outline the different steps needed to go from RNA-seq FASTQ files to gene expression quantitation.*
xiii. *In what applications would paired-end sequencing be desirable over single-end sequencing?*
xiv. *When looking for mutations in a sample compared to the reference genome, why would high coverage be important?*

Social Implication of Genome Sequencing

Power lies within the tools that scientists have created – power to advance medicine, research, human health and all

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

other aspects of biology. This power can be enormously beneficial to millions of people and change our lives, but researchers must consider what are the long-term consequences of this technology and see past our drive for innovation to contemplate the social and ethical ramifications. Below we illuminate some of the ethical and social consequences that arise when genetic sequencing is used for medical advancements and placed directly in the hands of consumers.

Knowledge is Power

Anticipatory Guides

27. *How will genetic testing change medical treatments?*
28. *Is the application of genetic testing limited to human diseases?*

The ease that parts of and even whole genomes can be sequenced has led to numerous discoveries linking genes to diseases (Figure 6) (Rego and Snyder 2019). Genome-wide association studies (GWAS) has facilitated linking complex genetics to differential phenotypes. GWAS identifies specific SNPs associated with diseases by comparing common sequence variants and/or genomes between healthy (unaffected individuals) to those individuals with a phenotype or interest, such as a disease (affected individuals) (Medline Plus 2020). Knowing what SNPs (genotypes) are associated with particular diseases and even more specifically which versions or types of alleles are associated with a particular disease, is the foundation of precision medicine (NRC 2011). Precision medicine allows medical professionals to choose treatments with the greatest likelihood of being effective based on understanding the genetic basis of the disease and the genetic sequence present in the particular patient that they are treating.

Sequencing technologies afford individuals the ability to obtain their genetic information independently through direct-to-consumer sequencing companies without professional medical assistance. In most cases, consumers simply provide a saliva sample via the mail which gets processed and genotyped, using a DNA microarray as described above with probes to common variants in the human population. While consumers are not informed of the sequence of their entire genome, they would learn the sequence of particular loci (or locations) of their genome known to be associated with particular diseases or phenotypes. For example, a person could find out if their genome contains a SNP associated with lactose intolerance, heart disease, or a sensitivity to caffeine. As a consumer, there are numerous choices for direct-to-consumer sequencing. A person must decide what they are hoping to learn about their genetic make-up to select the appropriate service. The number of services is continually increasing including companies that specialize in providing various genetic information such as ancestry information, information about genetic risks, and information about pharmacogenomics (how an individual metabolizes different pharmaceuticals).

Does everyone have equal access to this "power?"

Anticipatory Guides

29. *How do issues of social inequalities affect access to genetic testing and relevant treatments?*
30. *Who should be responsible for educating people about genetic testing and its implications?*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

Access to genomic technologies for the average person has traditionally been through clinical genetic testing. Sequencing by synthesis has revolutionized clinical genetic tests (tests ordered by a clinician) by allowing for 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 (Smedley et al. 2003) also plague genetic testing. Thus, access to clinical genetic testing is not equivalent across all racial and socioeconomic communities (Peters et al. 2004) due to factors such as differences in comprehensive health insurance among racial groups (Lillie-Blanton and Hoffman 2005) and mistrust of medical testing by individuals from groups historically excluded from healthcare (Peters et al. 2004). These disparities put racial minorities at increased disadvantage to reaping the benefits of clinical genetic testing.

With the rise of genomic sequencing technologies has also come a new era of direct-to-consumer genetic testing. Direct-to-consumer genetic tests are those in which a consumer can purchase the test without the involvement of a healthcare professional (Niemiec et al. 2017). These tests generally use DNA microarrays to offer services such as ancestry, health and disease risks, and lifestyle insights. Although these direct-to-consumer tests can be more affordable ($60 - $200 depending on the comprehensiveness of the test) the costs are out-of-pocket which still can be a significant barrier to access. Additionally, the health and lifestyle genetic risk factors that the direct-to-consumer report on are based on studies with inequities in representation in the genetic data. 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 (Sirugo et al. 2019). This lack of representation has consequences in 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.

Can this "power" end up in the wrong hands?

Anticipatory Guides

31. *Who might you want to keep your genetic results from?*
32. *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 their data and who has access to their data. Legislation has been enacted to protect the consumer's privacy. For example, the 21st Century Cures Act seeks to protect an individual's confidentiality when genetic information is donated to federal research purposes by removing all identifying information such as the donor's name and contact information from the genetic information (21st Century CURES Act 2016). Any information obtained from the research cannot be released to law enforcement or government agencies. In addition, the Health Insurance Portability and Accountability Act (HIPPA) will protect the privacy of genetic information if it becomes a part of one's health record

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

(HIPPA 1996). Under HIPPA, this information is protected from employers, schools, and the public. The entities that can access this information are law enforcement and health insurance.

However, health insurance companies having access to genetic information can negatively impact the individual. With the rise of genetic testing came concern about genetic discrimination. Individuals were concerned that health insurance companies could discriminate against those who tested positive for differing genetic predispositions and alter their healthcare coverage. To prevent potential discrimination, the Genetic Nondiscrimination Act (GINA) was passed in 2008 (GINA 2008) to protect individuals from health insurance companies from denying coverage and changing rates based on genetic predispositions. GINA also prohibits employers from changing the employment status based on genetic testing results. An important part of this law is that the individual cannot be showing any symptoms of the predisposition for GINA to protect them. If symptoms of the genetic difference are present, then insurance companies do have the power to alter their coverage and rates.

Direct to consumer genetic testing companies, like 23andMe, have their own privacy policies. For example, under 23andMe's privacy statement, the company guarantees that their client's data is kept private in a company protected database. Client information that is used to register for the service is stripped from the DNA sample and a random ID is assigned to the sample to further protect the privacy of the individual. 23andMe states that they will not share any client information with employers, insurance companies and law enforcements unless it is court ordered. Client information is not to be sold, leased or rented to other organizations and databases without client consent (23andMe 2020).

Discussion Questions

xv. *What human diseases are best suited for genetic testing and precision medicine?*
xvi. *What are the limitations of genetic testing?*
xvii. *Is it always beneficial for someone to know genotype(s)? Are there things that one might not want to know about your genome?*
xviii. *What direct-to-consumer service(s) do you think provides the most interesting (relevant) information?*
xix. *Who should be responsible for genetic testing costs?*
xx. *Should restrictions be placed on genetic testing? Why or why not? If yes, what sorts of restrictions are appropriate?*
xxi. *What are potential concerns regarding genetic testing (either through a medical clinic or direct-to-consumer testing)? Would these concerns stop you from getting your DNA tested? Why or why not?*
xxii. *Who should have access to genetic testing results?*
xxiii. *Is additional legislation necessary to regulate genetic testing and results? What issues would you like that legislation to address?*

References

Andrews S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. (http://www.bioinformatics.babraham

.ac.uk/projects/fastqc/).

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

Bentley DR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–9.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nature biotechnology 34: 525–7.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science 282: 2012–2018.

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SPA. 1996. Accessing genetic information with high-density DNA arrays. Science 274: 610–614.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. 2009. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research 38: 1767–1771.

Eid J, Fet al. 2009. Real-time DNA sequencing from single polymerase molecules. Science 323: 133–138.

Franklin RE, Gosling RG. 1953. Molecular Configuration in Sodium Thymonucleate. Nature 171: 740–741.

Goffeau A, et al. 1996. Life with 6000 Genes. Science 274: 546–567.

Harris TD, Bet al. 2008. Single-molecule DNA sequencing of a viral genome. Science 320: 106–109.

Hershey AD, Chase M. 1952. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. The Journal of General Physiology 36: 39–56.

Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. 2015. Improved data analysis for the MinION nanopore sequencer. Nature Methods 12: 351–356.

Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, Li X, Marma MS, Shi S, Wu J, Edwards JR, Romu A, Turro NJ. 2006. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proceedings of the National Academy of Sciences of the United States of America 103: 19635–19640.

Kasianowicz JJ, Brandin E, Branton D, Deamer DW. 1996. Characterization of individual polynucleotide molecules using a membrane channel. Proceedings of the National Academy of Sciences of the United States of America 93: 13770–3.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Lillie-Blanton M, Hoffman C. 2005. The role of health insurance coverage in reducing racial/ethnic disparities in health careHealth Affairs 24: 398–408.

Lundquist PM, et al. 2008. Parallel confocal detection of single molecules in real timeOptics Letters 33: 1026.

Margulies M, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactorsNature 437: 376–380.

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

National Research Council (US) Committee

on a Framework for Developing a New Taxonomy Disease. 2011. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. National Academies Press.

Niemiec E, Kalokairinou L, Howard HC. 2017. Current ethical and legal issues in health-related direct-to-consumer genetic testing. Personalized Medicine 14: 433–445.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 14: 417–419.

Patro R, Mount SM, Kingsford C. 2014. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature Biotechnology 32: 462–464.

Peters N, Rose A, Armstrong K. 2004. The association between race and attitudes about predictive genetic testing. Cancer Epidemiology Biomarkers and Prevention 13: 361–365.

Rego SM, Snyder MP. 2019. High throughput sequencing and assessing disease risk. Cold Spring Harbor Perspectives in Medicine 9: 1–12.

Reuter JA, Spacek D, Snyder MP. 2016. High-Throughput Sequencing Technologies. Molecular Cell 58: 586–597.

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

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74: 5463–7.

Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. Human Molecular Genetics 19: 227–240.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467–470.

Shalon D, Smith SJ, Brown PO. 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Research 6: 639–645.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309: 1728–1732.

Sirugo G, Williams SM, Tishkoff SA. 2019. The Missing Diversity in Human Genetic Studies. Cell 177: 26–31.

Smedley B., Stith A., Nelson A, eds. 2003. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. National Academies Press.

Watson JD, Crick FHC. 1953. A Structure for Deoxyribose Nucleic Acid. Nature 171: 737–738.

Wheeler DA et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872–876.

Wilkins MHF, Stokes AR, Wilson HR. 1953. Molecular structure of deoxypentose nucleic acids. Nature 171: 738–740.

United States. 1996. Health Insurance Portability and Accountability Act of 1996. Public Law 104-191.

United States. 2008. The Genetic Information Nondiscrimination Act. Public Law 110-233.

United States. 2016. 21st Century Cures Act. Public Law 114-255.

[Medline Plus] What are genome-wide

association studies? 2020. (8 November 2020; https://medlineplus.gov/genetics/understanding/genomicresearch/gwastudies/)

[23andme] Privacy Policy. 2020. (11 November 2020; www.23andme.com/about/privacy/).

Figures

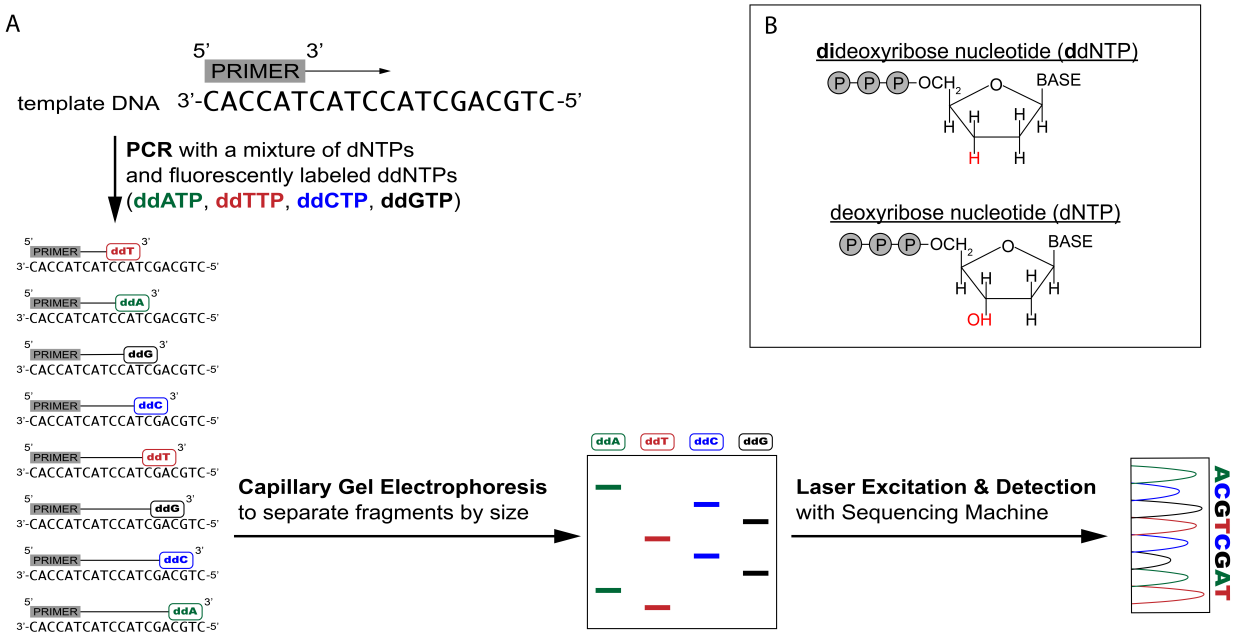

Figure 1. Chain termination sequencing. (A) Schematic of chain termination sequencing. DNA templates are amplified by DNA polymerase in a reaction containing a mixture of dNTPs and fluorescently labeled ddNTPs. Amplified fragments terminated at different lengths are separated by capillary gel electrophoresis followed by laser excitation and detection. Sequences are displayed in a chromatograph (as peaks) where each nucleotide is represented by a differently colored peak. The height of the peak indicates the confidence level at that nucleotide position. (B) Chemical structure of ddNTPs and dNTPs. The critical 3' hydroxyl group in dNTPs is highlighted in red, which is not present in ddNTPs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

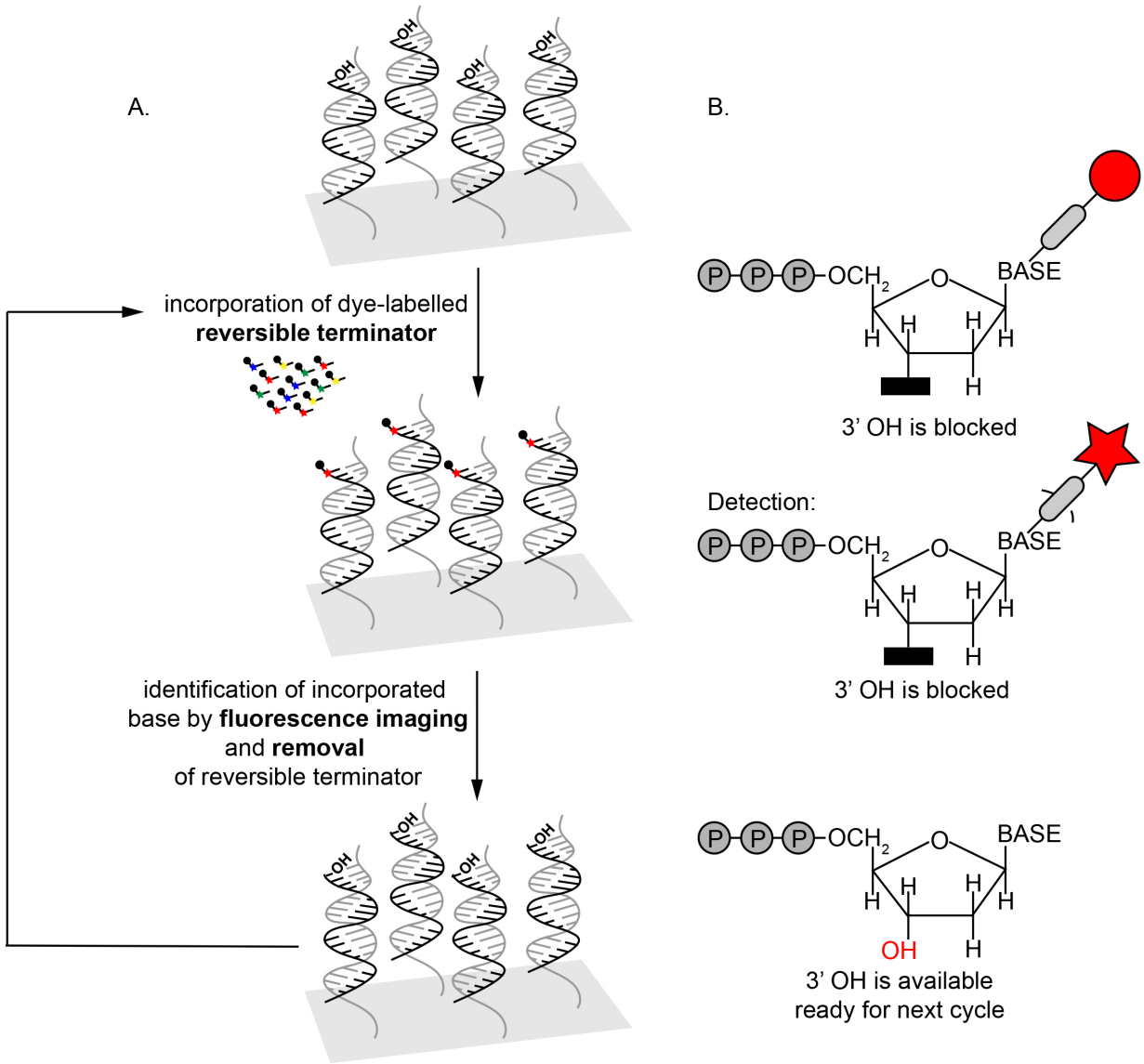

Figure 2. Sequencing by synthesis. (A) Cycle of reversible terminator incorporation, identification of incorporated base by fluorescence imaging, followed by removal of the reversible terminator. (B) Chemical structure of a nucleotide with a reversible terminator attached. The 3'-OH group is capped by 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

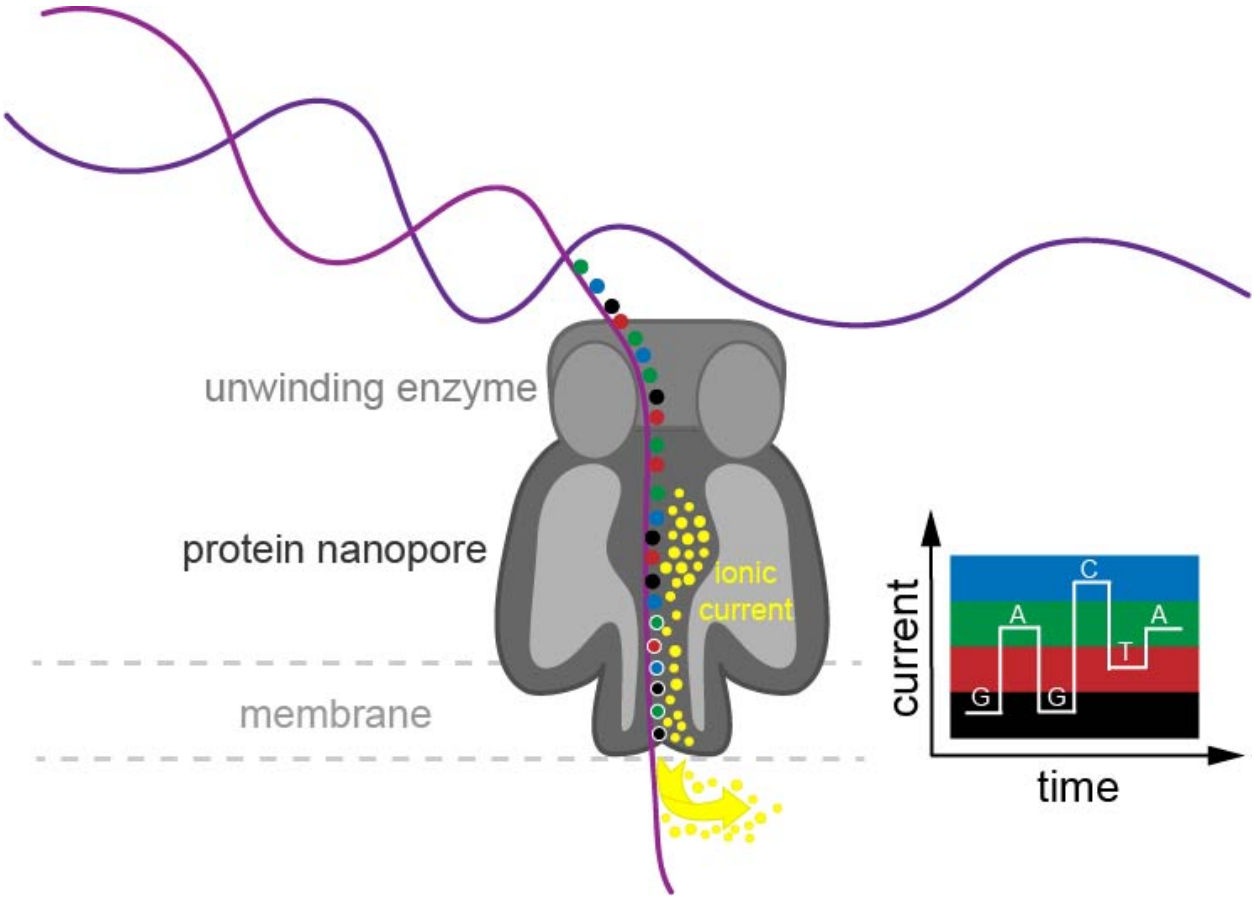

Figure 3. Nanopore sequencing. DNA double helix is unwound by unwinding enzyme and a single strand is fed through the pore inserted in a membrane. As the DNA moves through the protein nanopore, the nucleotides (colored circles) are identified by the change in ion current (yellow) across the membrane. Graph shows the identification of nucleotides in the DNA sequence based on the current measured over time.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

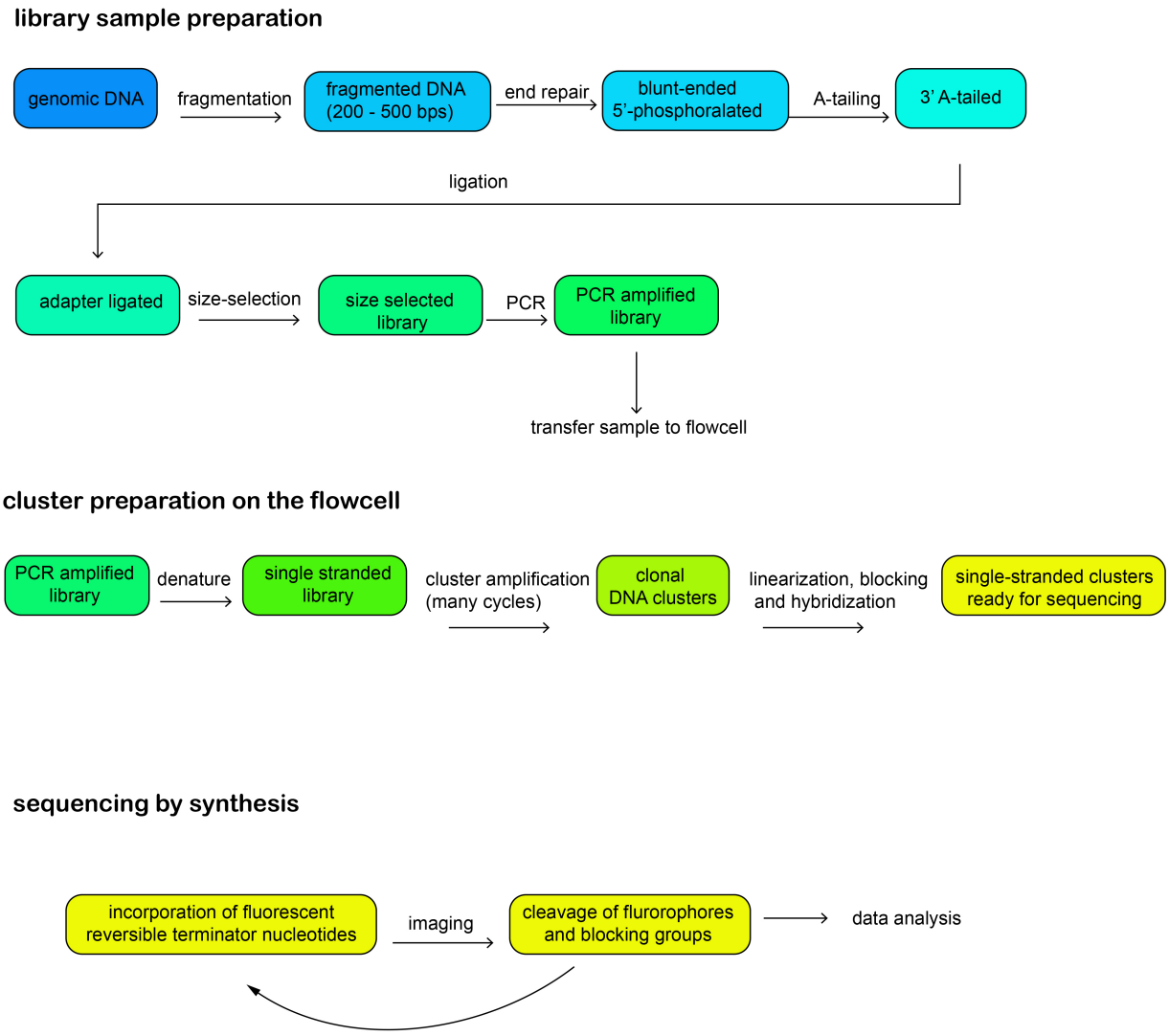

Figure 4. Sequencing by Synthesis pipeline. Genomic DNA is first fragmented into smaller templates which undergo modification, including 5'-phosphorylation and addition of 3'-A for adaptor ligation. Following size selection and PCR amplification, the library is denatured and amplified into clonal clusters that undergo linearization, blocking, and hybridization, preparing the flow cell for sequencing, using reversible terminators.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2020 doi:10.20944/preprints202011.0501.v1

Figure 5. Whole genome sequencing analysis. (A) Example four lines of each read in a FASTQ file. Components in the FASTQ file are labeled with a text box of the same color, which include the sequence ID, nucleotide sequence, and quality score. (B) Example reads mapped to a genome. An example of 1x coverage (left) and 5x coverage (right) is shown. Reads common to both the 1X and 5X examples are shown in light grey, and reads only in the 5X example are shown in dark grey.