Chapter 8

DNA Sequencing

Investigatory approaches that study whole or partial genomes, define the field of **genomics**. The last ten years has changed scientific discovery since the entire genomic sequence for thousands of different organisms are available for analysis. In fact, sequencing the human genome has gone from a monumental task that involved multiple sequencing centers and millions of dollars into a simple task that can be done for a few thousand US dollars and a few hours. In this chapter, we will survey the methods used to sequence DNA, and in the following chapter we will consider the assembly of whole genome sequences. This chapter will explain regular chain termination sequencing, a procedure used by many to double-check plasmid assembly, and sequence shorter segments of DNA. Then the chapter will focus on **next generation sequencing (NGS)** technologies, explaining the basic procedure for some of the top selling technologies in use as of writing. Technological advancements in NGS are very fast, and therefore, the discussion will focus more on a conceptual understanding and then introduce some details and nuances of the more popular technologies.

genomics Study of the structure, function, evolution, and mapping of genomes.
next generation sequencing (NGS) The term to describe experimental techniques to simultaneously decode the order of bases for millions of genomic DNA fragments.

Molecular Biology. DOI: https://doi.org/10.1016/B978-0-12-813288-3.00008-2

1. DNA Sequencing—Overview of Chain Termination Sequencing

Chain termination sequencing, also known as **Sanger sequencing** or **dideoxy sequencing**, provides the researcher with the exact order of nucleotides for one segment of DNA. This procedure provides confirmation of the sequence for any synthetic biology plasmid construct created by a researcher, provides confirmation of mutations, insertions or deletions of a gene, and can still be used for sequencing a whole genome. Overall, the approach first involves isolating the **template DNA sample**. The typical size for the template ranges between a few hundred base pairs created using polymerase chain reaction (PCR) to tens of thousands of base pairs from a large construct made in a bacterial artificial chromosome (BAC) to whole genomes. Sources of template DNA include PCR products, plasmid DNA, genomic DNA, BACs, yeast artificial chromosomes (YACs), and/or cosmids. Only one piece of the template DNA sample, which will be referred to as the **target DNA** is sequenced during a single sequencing reaction using the Sanger method. In contrast, every piece of the template DNA is sequenced in a next generation sequencing (NGS) reaction (discussed later).

During the actual Sanger sequencing procedure, DNA polymerase copies the target region of template DNA, but the copies vary in their lengths by one nucleotide. For example, if the target DNA was 200 base pairs in length, during the sequencing reaction, a subset of the fragments would be 199 bases long, another subset would be 198 bases long, and another would be 197 bases long, continuing until only a few bases from the starting point of DNA polymerase. Although these vary in size by one nucleotide, the fragments are not created in order in the reaction. Instead, DNA polymerase randomly creates the different lengths, and the final reaction contains a mixture of all the different lengths. In order to figure out the sequence, the final nucleotide of each fragment is labeled with a fluorescent tag—one color for each of the four bases (Fig. 8.01).

The mixture of fragments are separated by size through a capillary tube filled with polymer matrix that is designed to allow small fragments travel faster than the larger fragments. In a process similar to electrophoresis, the fragments are separated based on their negative charge and the sieving properties of the polymer. The fragment movement through the tube is facilitated by high pressure and an electric current (Fig. 8.02). After the fragments travel through the polymer matrix and separate so the smallest fragments are first, a fluorescent laser excites the fluorophore at the 3′ end of the fragments causing the fluorophore to release its characteristic wavelength or color of light. The fluorescent detector records the color or wavelength, which is then recorded as a G, A, T, or C. The smallest fragment (or the shortest fragment) is first, then the next longer fragment, then the next longer fragment, the order of G, A, T, and C that are recorded are the exact same order as what they are in the template DNA (see Fig. 8.01).

Chain termination sequencing (also known as Sanger sequencing) actually involves synthesizing partial copies of a target DNA that vary in length by one nucleotide and then separating them by size.

Fluorescent detectors at the end of a capillary electrophoresis tube record the wavelength of fluorescence released from the final nucleotide and convert it into the corresponding base.

1.1. Details of the Chain Termination Method for Sequencing DNA

Now that the general idea of how Sanger or chain termination sequencing has been presented, the details of how the reaction creates the different length

chain termination sequencing Method of sequencing DNA that uses dideoxynucleotides to terminate synthesis of DNA chains. Same as dideoxy sequencing or Sanger sequencing or cycle sequencing.
dideoxy sequencing Method of sequencing DNA that uses dideoxynucleotides to terminate synthesis of DNA chains. Same as chain termination sequencing or Sanger sequencing or cycle sequencing.
Sanger sequencing Method of sequencing DNA that uses dideoxynucleotides to terminate synthesis of DNA chains. Same as chain termination sequencing or Sanger sequencing or dideoxy sequencing.
target DNA The region or defined location within a DNA sample that is to be sequenced or amplified by PCR; also known as a target sequence.
template DNA sample The sample of DNA that contains the segment of DNA or target DNA in that is to be sequenced; can be as simple as a PCR product to as complex as a whole genome.

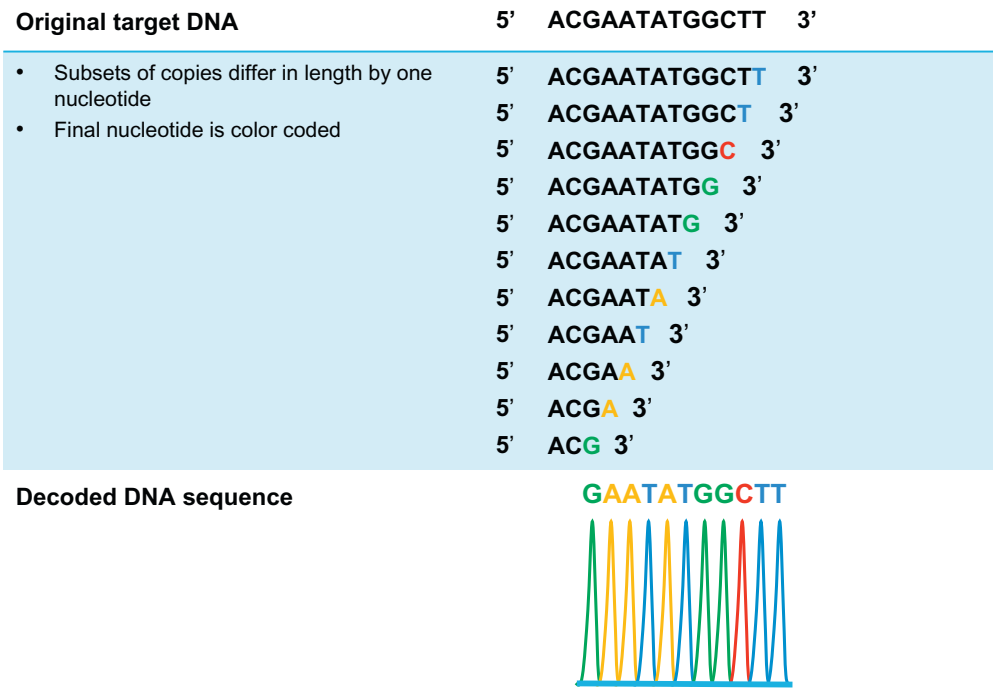


FIGURE 8.01
Sequencing—Fragments of All Possible Lengths

During chain termination or Sanger sequencing, the target DNA fragments are copied millions of times, but each copy ends at a different nucleotide position. These subsets of fragments end with a fluorescently-labeled nucleotide that reveal the identity of the final base. The final sequencing data are a series of fluorescent peaks that correspond to the original template DNA sequence.

fragments that end in a fluorescent nucleotide can be explained. Sequencing reactions need a variety of components or ingredients in order to synthesize the different length fragments. These are mixed in a small tube. First, a **sequencing primer** or short oligonucleotide that is complementary to the starting point of the target DNA is essential (Fig. 8.03). The primer supplies the 3′ hydroxyl (−OH) group that DNA polymerase uses to start assembling complementary DNA based upon the target DNA sequence (Fig. 8.04). The primer not only provides a 3′ −OH for DNA polymerase, but it also defines the start site for decoding the target DNA.

During the sequencing reaction, DNA polymerase synthesizes copies of the target DNA from a pool of **deoxyribonucleoside triphosphates (dNTPs)** for each of the four bases (dATP, dTTP, dGTP, and dCTP). These nucleotides have a hydroxyl (−OH) group attached to the 3′ carbon of the deoxyribose sugar ring. Starting at the 3′ hydroxyl of the primer, DNA polymerase connects the incoming nucleotides to the chain (Fig. 8.04). When nucleotides are joined, the phosphate group attached to the 5′-carbon of the incoming nucleotide is linked to the

deoxyribonucleoside triphosphates (dNTPs) The building blocks of DNA that are used by DNA polymerase to synthesize a complementary strand of DNA to a template or target DNA. dNTPs have a deoxyribose sugar connected to one of the four bases, adenine (A), guanine (G), cytosine (C), thymine (T) on the first carbon, and three phosphate groups attached to the 5′ carbon of the deoxyribose.
sequencing primer Short piece of single-stranded DNA that is complementary to the sequence at the beginning of the target DNA segment. It provides the essential 3' hydroxyl group for DNA polymerase to initiate DNA synthesis.

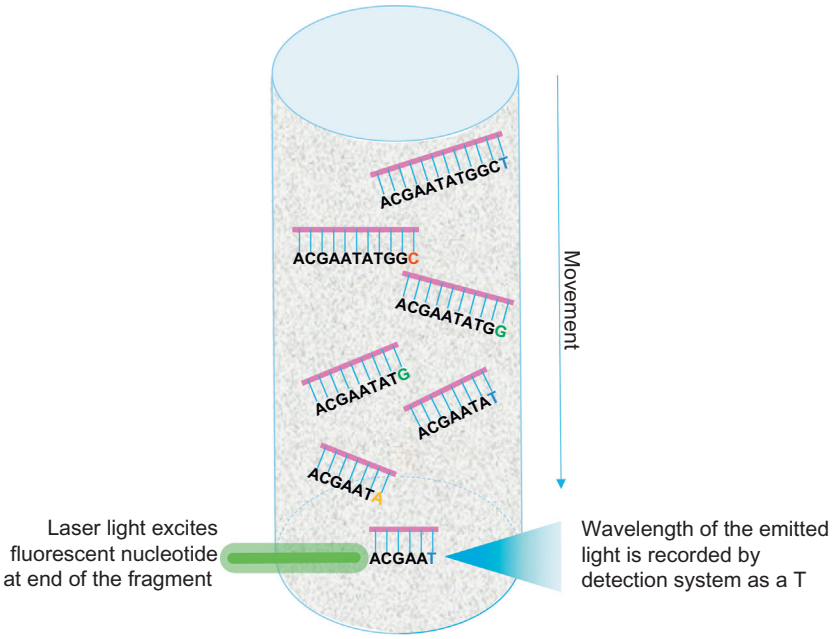


FIGURE 8.02
Separation of Fragments by Capillary Electrophoresis

Copies of the target DNA with fluorescently-labeled nucleotides separate by size in the polymer matrix of the capillary tube. The smaller fragments travel faster through the tube, and the larger fragments travel slower. As the fragments pass through the laser, the fluorophore at the 3′ end of the fragment emits the characteristic color of light for the base. In this example, the T is labeled with a fluorophore that emits blue light. The detector then records the color, and reports the color as a T. Therefore, the first decoded nucleotide for the target DNA is a T. As the next fragment moves into the laser light, the final nucleotide will emit yellow light, and this is recorded as an A. The final results are compiled as sequence files for analysis.

3′-hydroxyl group of the growing DNA chain. Or, in brief, DNA is polymerized in the 5′ to 3′ direction. During the reaction the hydrogen ion (H+) on the 3′ hydroxyl group (−OH) is released, as well as, the two outer phosphate groups from the incoming dNTP. These two outer phosphates are called **pyrophosphate** after they are released.

Sanger sequencing reactions must create copies of the template that are shorter than the original, and therefore, these reactions have an additional ingredient in the pool of dNTPs to halt DNA synthesis. There is also a mixture of **dideoxynucleoside triphosphates (ddNTPs)** for each of the four bases (ddATP, ddTTP, ddCTP, and ddGTP), which are missing the essential hydroxyl group on the 3′ carbon (Fig. 8.05). During the reaction, if DNA polymerase incorporates a ddNTP, DNA synthesis comes to a halt because the 3′−OH group is missing. There is no connection available for the next incoming nucleotide. The ratio of ddNTPs to dNTPs is established so that by chance, DNA polymerase will incorporate a ddNTP often enough to create the subsets of fragments that differ in length by one nucleotide, but not so often that none of the copies are almost full-length in comparison to the template. The term, chain termination sequencing, derives from this process. In addition to DNA polymerase, dNTPs, ddNTPs, primer, and the the template DNA sample, the sequencing reaction has buffers and ions that balance the reaction for the optimal performance for all the ingredients.

Dideoxynucleotides are used to terminate growing DNA chains and create the subsets of truncated fragments in a sequencing reaction.

dideoxynucleoside triphosphates (ddNTPs) An analog of deoxyribonucleotides triphosphate that is missing the 3′ −OH, and therefore, are not competent to have another nucleotide added onto it. ddNTPs have a dideoxyribose sugar connected to one of the four bases, adenine (A), guanine (G), cytosine (C), thymine (T) on the first carbon, and three phosphate groups attached to the 5′ carbon of the dideoxyribose.
pyrophosphate Two phosphate groups connected via a covalent bond; also called diphosphate.

FIGURE 8.03
Synthesis of DNA—Priming and Elongation

During normal DNA synthesis, DNA polymerase reads the template DNA and makes a new complementary strand of DNA. To get DNA synthesis started, a short oligonucleotide primer must anneal to the beginning of the target DNA. DNA polymerase recognizes the 3′ end of the primer and connects the 5′ end of the incoming nucleotide; hence, synthesis occurs in a 5′ to 3′ direction.

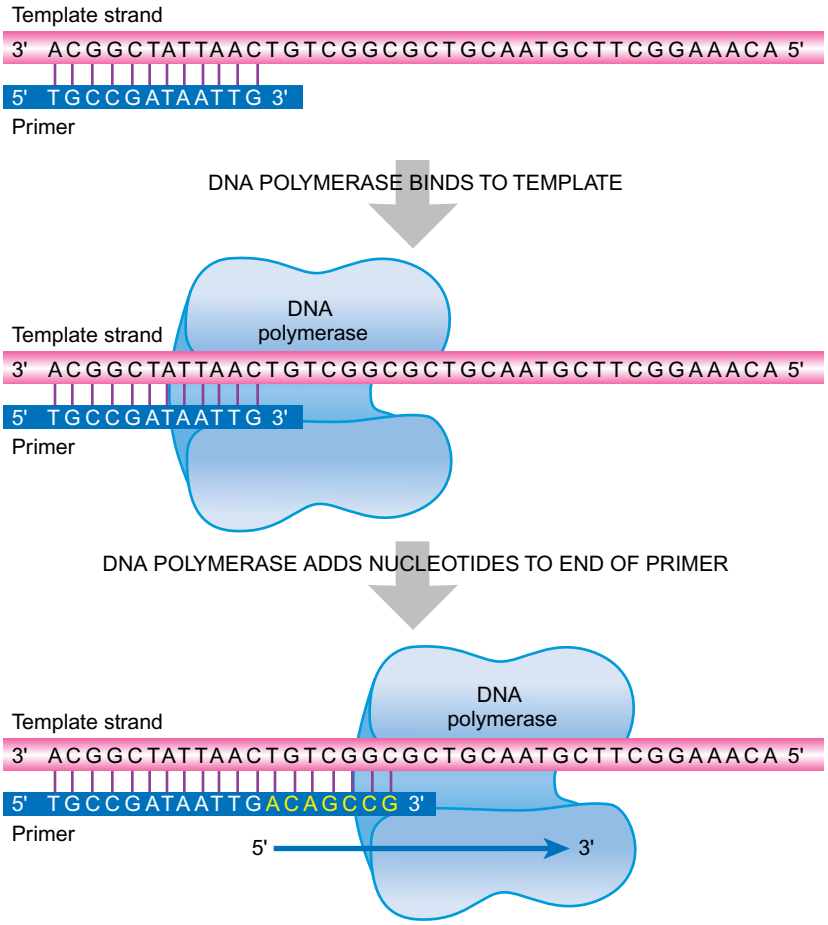
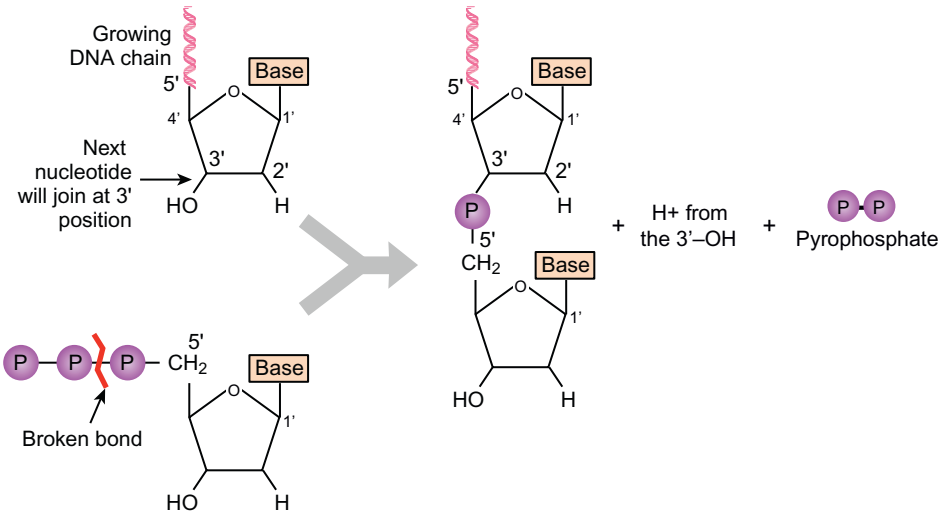


FIGURE 8.04
Synthesis of DNA—Phosphodiester Bonding

DNA polymerase links nucleotides via phosphodiester bonds. DNA polymerase breaks the bond between the first and second phosphates of the incoming deoxynucleoside triphosphate, which is then joined to the free 3′ hydroxyl of the growing DNA chain. The two outer phosphates of the dNTP (called pyrophosphate) are released from the reaction, as well as, a hydrogen ion (H +) from the 3′ —OH group.



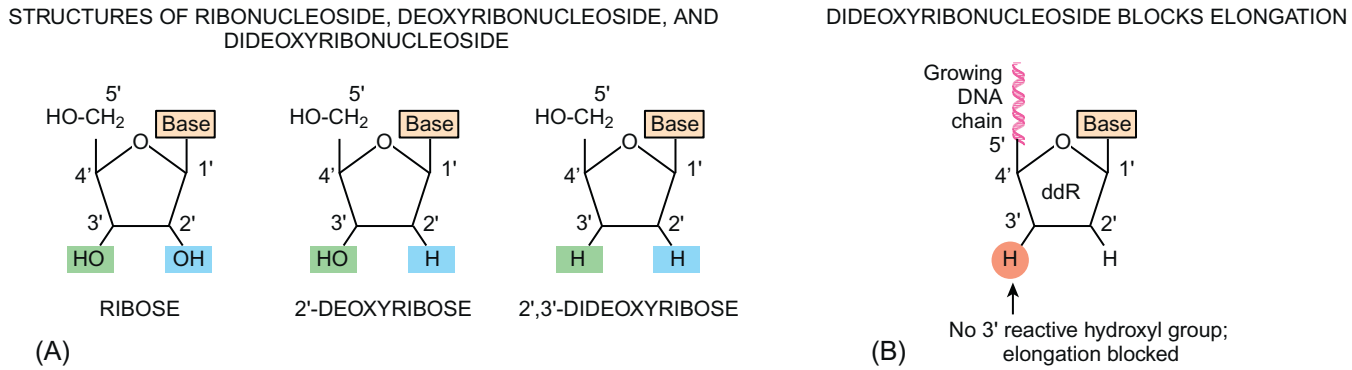


FIGURE 8.05
Dideoxyribonucleoside, Deoxyribonucleoside, and Ribonucleoside

(A) The structures of ribonucleoside, deoxyribonucleoside, and dideoxyribonucleoside differ in the number and location of hydroxyl groups on the 2′ and 3′ carbons. (B) DNA polymerase cannot add another nucleotide to a chain ending in dideoxyribose because its 3′ carbon does not have a hydroxyl group.

Another term used to describe Sanger or chain-termination sequencing is **cycle sequencing** because the reaction occurs via cycles in a thermocycler. The mixture of ingredients are placed in a thermocycler, which then heats the ingredients to around 96°C. The heat denatures the template DNA sample. Next, the temperature is lowered so the primers can anneal to the starting point of the target DNA. Finally, the temperature is adjusted so that DNA polymerase can synthesize the copies of the target DNA until a ddNTP is incorporated. Although cycling parameters may sound like PCR, sequencing reactions use only one primer, and therefore, only one strand of the template is copied. The final products are referred to as **extension products** because the reaction extends from the single primer on a single strand of the template DNA until a ddNTP is incorporated.

The final ddNTP on each of the extension products also identifies the final base added by DNA polymerase. As described above, if a ddNTP is added to a growing chain during DNA synthesis, no 3′-hydroxyl group is available for further elongation. DNA polymerase cannot add any more nucleotides. In addition to the lack of a 3' -OH group, each ddNTP has a different fluorophore, which emits a unique wavelength of light after being excited by a laser (Fig. 8.06). There is a different fluorophore for each base, and therefore, each color or wavelength of light emitted identifies the last base that was added by DNA polymerase during extension.

Dideoxyribonucleoside triphosphates are missing the hydroxyl group on both the 2′ and 3′ carbon of the sugar. Sequencing reactions include ddNTPs with a different fluorophore for each base.

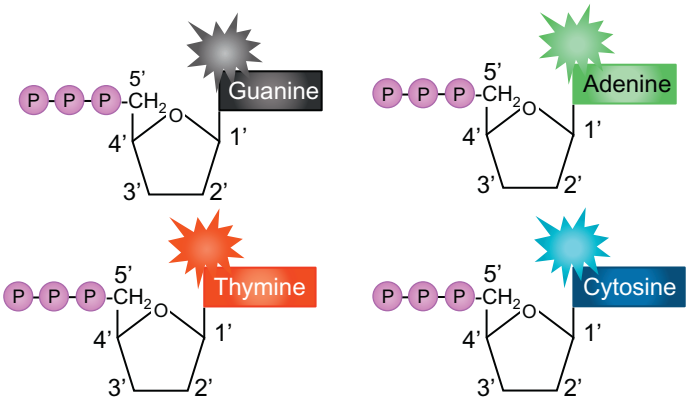
1.2. DNA Polymerases for Sequencing DNA

DNA polymerases used for sequencing elongate a primer that is annealed to a single-stranded DNA template. However, the characteristics needed for use in sequencing are more rigorous. First, the polymerase must have high processivity; that is, it must move a long way along the DNA before dissociating. Premature dissociation would give strands that ended at random before the dideoxynucleotide was incorporated. DNA polymerases for sequencing must also incorporate nucleotides rapidly and accurately. In addition, many DNA polymerases possess

cycle sequencing Method of sequencing DNA that uses dideoxynucleotides to terminate synthesis of DNA chains. Same as dideoxy sequencing or Sanger sequencing or chain termination.
extension products Copies of the template DNA created by DNA polymerase during cycle sequencing that vary in length; they originate from a single primer, and are created from one strand of the template DNA.

FIGURE 8.06
Each Base Has a Different Fluorophore

The dideoxynucleoside triphosphates for G, A, T, and C have a different fluorophore attached to their structure. Typical sequencing reactions show the color of these fluorophores as black for G, blue for C, red for T, and green for A, even though, these are not the actual colors for the fluorophores.



exonuclease activities that create sequencing errors. Exonuclease activity in the 5′ to 3′ direction is used to remove a strand of DNA ahead of the DNA synthesis point. In contrast, 3′ to 5′ exonuclease activity is used to remove incorrect bases during proofreading. Such activities can cause errors during the reaction.

In practice, no natural DNA polymerase is entirely suitable for sequencing. The first DNA polymerase used was **Klenow polymerase**, which is DNA polymerase I from *Escherichia coli* that lacks the 5′ to 3′ exonuclease domain. Klenow polymerase was originally obtained by protease digestion of purified DNA polymerase I but was later made by expression of a modified gene. Because Klenow polymerase has relatively low processivity, it can only be used to sequence around 250 bases per reaction. Another commonly used enzyme is a genetically modified DNA polymerase from bacteriophage T7. This is marketed as "**Sequenase**" and has high processivity, a rapid reaction rate, negligible exonuclease activity, and the ability to use many modified nucleotides as substrates, thus making it perfect for sequencing reactions. The modified polymerase is able to sequence more bases than Klenow polymerase. For cycle sequencing reactions, DNA polymerase from a thermophilic organism are used. *Taq* polymerase from *Thermus aquaticus* and DNA polymerase from *Thermococcus* are common sources of enzymes suitable for engineering the best attributes for sequencing.

Genetically engineered DNA polymerases with thermal stability, higher processivity and less exonuclease activity are now used for DNA sequencing.

2. Automated Sequencing

Today, the majority of cycle sequencing is done using automated machines. The reactions are carried out as described earlier, which creates millions of fragments that vary in length by one nucleotide, and end in a different fluorescent molecule for each of the four ddNTP. The sequencing instrument or machine has capillary tubes and depending on the machine, hundreds of the capillary tubes can be used simultaneously, meaning hundreds of individual sequencing reactions can be run in parallel. The machines load each individual sequencing reaction containing the different fluorescently-labeled subfragments into a single capillary tube using high voltage electricity and high pressure. The electric current acts on the DNA since the phosphate backbone imparts a negative charge, and propels the fragments into the capillary tube. These tubes are filled with a polymer matrix that helps separate the fragments by size since it inhibits the movement of the larger fragments and allows the smaller fragments to move through faster. As the fragments move

Klenow polymerase DNA polymerase I from *E. coli* that lacks the 5′ to 3′ exonuclease domain.
Sequenase Genetically modified DNA polymerase from bacteriophage T7 used for sequencing DNA.

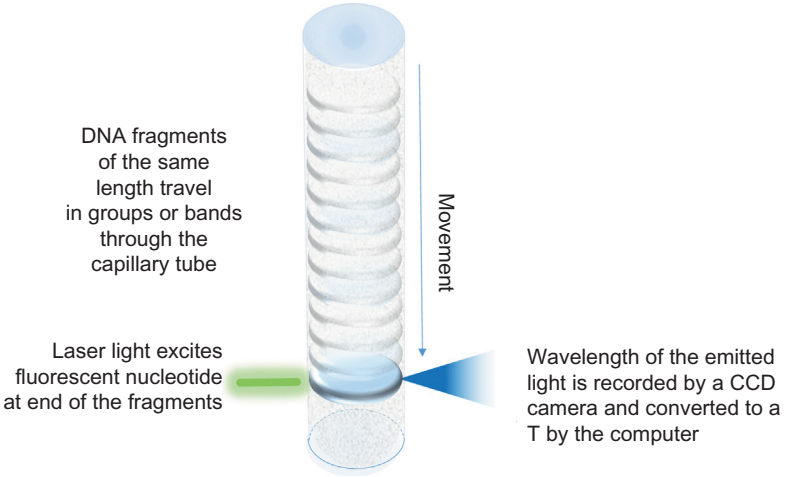


FIGURE 8.07
Separation of DNA Fragments by Capillary Electrophoresis

Groups of DNA with the same length, and therefore ending in the same fluorescently-labeled dideoxynucleotide, pass by a fluorescent laser that excites the fluorophore. The emitted wavelength of light is recorded by a CCD camera. The computer takes this data and saves the intensity of the fluorescence emission, and then converts the color into base identity.

through the matrix, they are separated in order from shortest to longest. The polymers are optimized to ensure separation of very closely sized fragments since they only differ by a single nucleotide. As the fragments move down the capillary tube and order themselves from shortest to longest, they form into groups based on size. The groups of DNA fragments that are the same size are referred to as a **band**. Each band of DNA passes the laser and detector assembly (Figs. 8.02 and 8.07). The laser beam excites the fragments by flashing the correct wavelength of light and determines which of the four bases is at the end of that group or band by the fluorescent color emitted by the ending nucleotide. A computer records the color of each group of fragments and compiles the data into actual sequence. The first bands to be recorded run right through the gel and off the end while later bands are still passing the laser. Consequently, more bases can be read from a single sequencing reaction by the continuous flow approach. On average, approximately 700 bases of sequence can be decoded from a single primer. The final results include a graph with a series of peaks that represent the intensity of fluorescent signal with the decoded sequence information along the top (as shown in Fig. 8.01).

3. Next Generation Sequencing

The key characteristic of Next Generation Sequencing (NGS) is the use of massively parallel methods. Simply put, this means that large numbers of DNA samples are sequenced side-by-side on the same apparatus. In practice this requires substantial miniaturization and highly sensitive fluorescent detectors. In fact, NGS has become so miniaturized that whole genomes can be sequenced at one time. NGS revolutionizes how researchers of all types—clinicians, ecologists, forensic biologists, agricultural scientists, and even industrial researchers—are thinking about our world. The idea of personalized medicines based upon our genetic make-up is no longer a possibility, it is fast becoming a reality. Environmentalists do not need to identify every plant and animal by traits, they can simply use a DNA sample to accurately pinpoint the genus and species of the sample. Oncologists can use NGS data to identify a drug that will effectively kill the cancer cells by finding the underlying genetic mutations that have caused the cancer to grow. NGS is being used to keep our planet healthy by understanding the genetics of various pathogens and parasites, tracking the spread of different viral diseases such as Ebola or Zika, and helping doctors

band (of DNA) Term used to describe DNA fragments of similar size that group together during electrophoretic separation.

identify and prevent the spread of antibiotic resistant bacteria. Besides advancing our understanding of the known, NGS has revolutionized our understanding of the world that is not visible to our eyes—the vast number of viruses, bacteria, sub-viral creatures that have never been seen, but now can be identified through their unique genetic codes. The use of NGS will impact you, the reader, in ways that cannot be imagined at this point!

Next generation sequencing (NGS) uses massively parallel methods where millions of sequencing reactions occur and are recorded simultaneously.

The procedure for accomplishing NGS can be generalized, and these main steps include:

1. **Genomic DNA (gDNA)** isolation;
2. NGS Library construction;
   a. gDNA fragmentation into small double-stranded pieces of approximately equal size, and
   b. Modifying the fragments so they are compatible with the sequencing platform
3. Partitioning the library fragments into separate locations on a solid surface and create clusters of identical copies;
4. Sequence each cluster; and
5. Analyze the data.

These four generalized stages of next-generation sequencing will be explained for only the two main technologies in use as of writing—Illumina sequencing by synthesis technology and the Ion Torrent sequencing technology by Thermo Fisher Scientific. The two technologies are leading the current NGS market, and provide quick and accurate sequence data. There are many different types of modifications that can be made to the basic NGS technology, such as library sifting techniques to sequence only the fragments from defined segments of the genome (as discussed later). After explaining the NGS methods of sequencing, the so-called "Third-Generation" sequencing technologies will be explained, including PacBio single-molecule real-time (SMRT) sequencing and Nanopore sequencing. These two technologies are also massively parallel sequencing techniques, but instead of working on clusters of identical DNA copies, they decode single pieces of DNA in each partition.

3.1. Illumina Sequencing by Synthesis

3.1.1 Library Preparation

In order to understand Illumina sequencing technology, our discussion will first explain how the technology can be used to sequence the whole human genome, therefore, the first step to this process is to isolate genomic DNA (gDNA) from a person. The gDNA must be pure, and contain only the DNA portion of the sample. Typical human DNA is obtained from a blood sample, but can also be obtained from scraping the inside of the cheek.

The next step of the process is to fragment the human chromosomes into small pieces of double-stranded DNA. There are two different methods of breaking up the whole human genome into small pieces. First, DNA can be sheared using **ultrasonic disruption**, or the use of sound waves that are higher than our ears can hear, passing through the liquid surrounding the gDNA. As the sound waves penetrate the liquid, small "bubbles" of trapped gases form and collapse, the force of which breaks the molecular bonds holding the DNA backbone together (Fig. 8.08). The sample is held at a constant temperature to prevent thermal degradation of the DNA fragments. After sonication, the whole chromosomes are broken into small

genomic DNA (gDNA) Sample of DNA isolated from an organism.
ultrasonic disruption Using sound waves to disrupt the molecular structure of proteins or DNA and break them into little pieces.

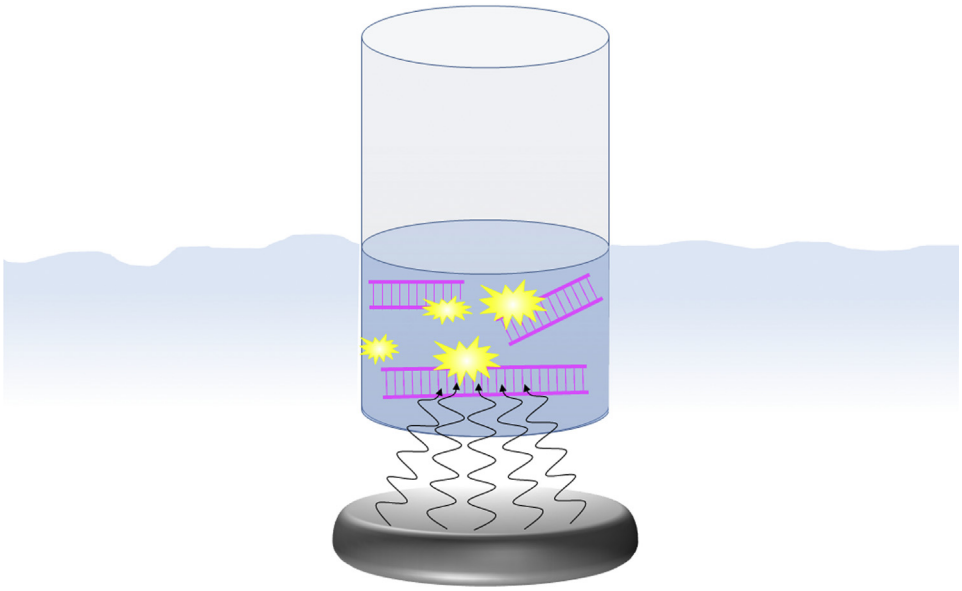


FIGURE 8.08
Ultrasonic Waves Shear Double-Stranded DNA

Ultrasonic sound waves (gray wavy lines) focused on the DNA sample (pink), create small gas bubbles in the liquid surrounding the DNA. As the bubbles collapse, the force breaks apart the covalent bonds holding the DNA backbone together. This releases smaller DNA fragments. The sheared DNA is randomly broken so every fragment ends in different locations of the genome.

pieces. The average length of these pieces will vary based on the intensity of the sound waves, length of exposure to the ultrasonic waves, and the concentration of the DNA in the liquid. The goal is to create an entire genome of fragments of roughly the same size. Note, however, that the fragments are random, and therefore, each end has a different DNA sequence.

NGS requires that some known sequence information be found on each fragment end, but each fragment of the genome generated by ultrasonic disruption has different sequences on the ends. **Adapters** are short double-stranded DNA fragments with a known sequence that are created using chemical synthesis of DNA. These are added onto each end of the whole genome fragments in order to give the sequencing primers a place to anneal, and as will be explained later, give each fragment a known sequence in order to get them to attach to the solid surface for partitioning. After breaking up the whole human genome into fragments, the ends are not even. The top strand of the double-stranded fragment is often a different length than the bottom strand. So to get the adapter to attach, the two strands of each fragment must be equal or blunt-ended, and the 5′ ends must be phosphorylated. A mixture of T4 DNA polymerase, Klenow polymerase, T4 polynucleotide kinase, and a supply of deoxyribonucleotides makes all the ends equal, and then add the 5′ phosphate groups (Fig. 8.09). In addition, Klenow polymerase has terminal transferase activity, which adds a single adenine onto the 3′ end of the genomic DNA fragments. (Note: *Taq* polymerase also has this activity, and this is used for TA cloning.) The single A overhang facilitates the binding of the adapters to the ends.

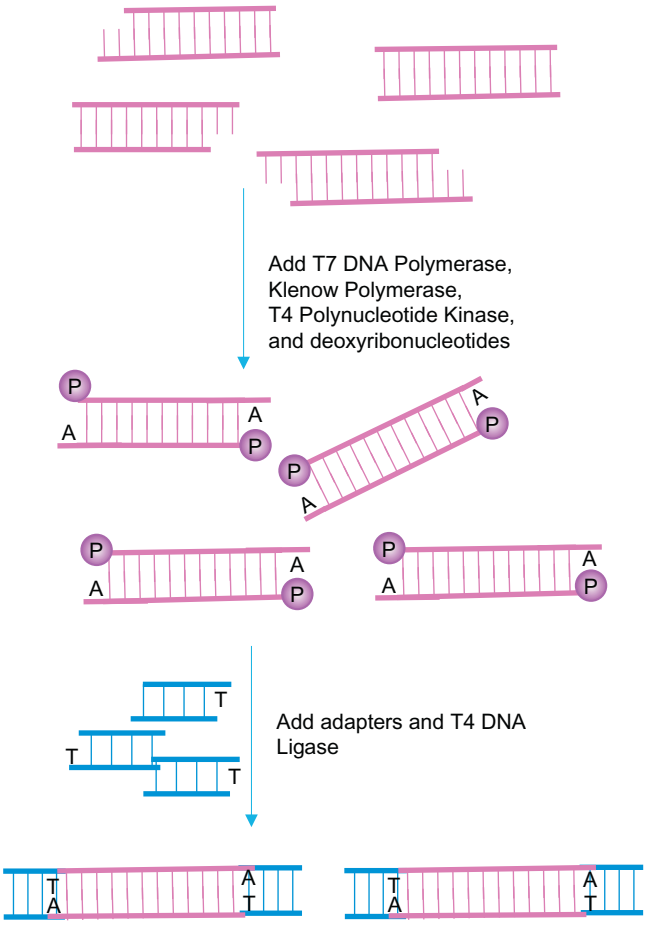
Adapters provide a known sequences onto each end of the random genomic DNA fragments.

The second method of fragmenting DNA for NGS relies upon an enzyme called Tn5 **transposase**. This is the enzyme from *E. coli* transposon, Tn5, which makes double-stranded cuts in DNA (see Chapter 25, Mobile DNA). When created by the transposon, transposase cuts its host DNA in order to move the transposon from one location to the next. Because it moves the DNA transposon, it actually binds to

adapters Short double-stranded DNA fragments with known nucleotide sequences that are added onto the ends of genomic DNA for next generation sequencing.
transposase Enzyme responsible for moving a transposon, which is a segment of DNA that can move from one location to another in a genome.

FIGURE 8.09
Adapters Are Added Onto The Ends of Genomic Fragments

After the genomic fragments are created with ultrasonic disruption, the ends are uneven. T4 DNA polymerase fills in any single-stranded regions with complementary nucleotides. Then Klenow polymerase adds a single A onto the 3′ end with its terminal transferase activity. The single A overhang facilitates the addition of the adapter, which is synthesized to have a single T overhang. Finally, DNA ligase connects the adapter to each end of the genomic DNA fragment.



Genomic DNA for next generation sequencing (NGS) is prepared by cutting genomic DNA into small pieces with endonucleases or sonication.

the two ends and inserts the transposon DNA into the cut it makes. The process for NGS is similar, but instead of having the transposon attached to the transposase, two adapters are bound to the transposase. When these transposase and adapter complexes mix with the human genomic DNA, the enzyme cuts the chromosomes into small fragments approximately 500 bases apart, and then "tags" the ends by adding the adapters (Fig. 8.10). The entire process is called **tagmentation**, which is combination of tag and fragmentation. The cuts are completely random, and the process creates a whole genome of small fragments with identical sequences on every end for the next step of NGS library preparation.

Whether added after sonication or during tagmentation, each adapter has the following key features (Fig. 8.11):

- A sequence that is complementary to the oligonucleotides that are bound to the flow cell.
- An index sequence that is unique to the genomic fragments from one sample.
- A sequence that is complementary to the sequencing primer.

Once the fragments are have adapters, they are ready for an optional PCR amplification step to increase the number of copies for each of the genomic DNA

tagmentation The method of shearing or fragmenting genomic DNA samples for next generation sequencing that employs a transposase enzyme to cut the DNA and attach the adapters.

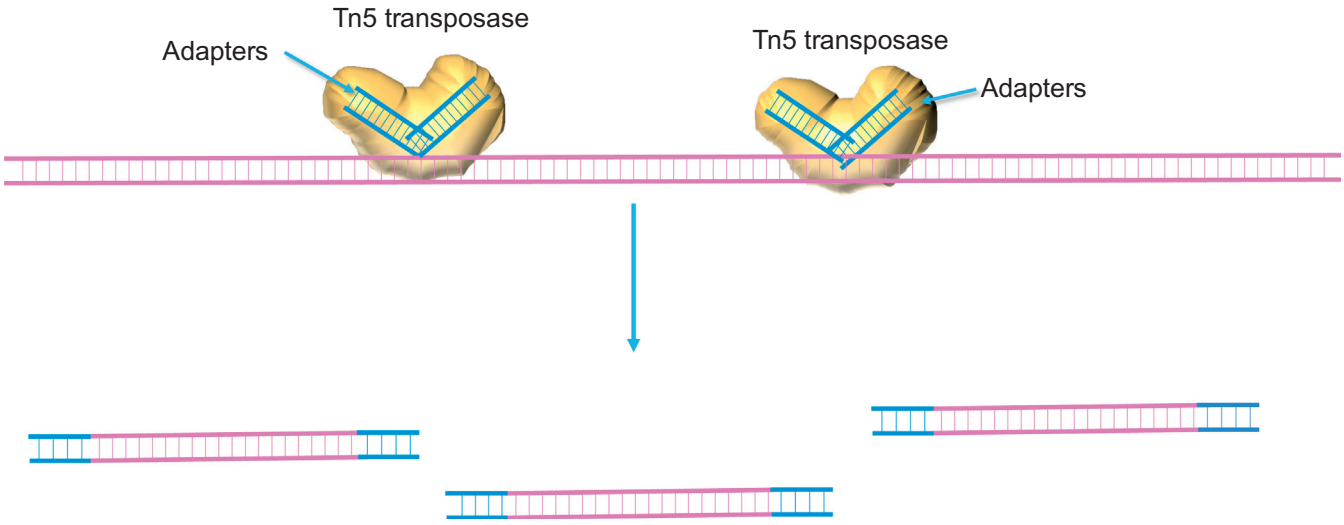


FIGURE 8.10
Tagmentation

Tn5 transposase is able to cut double-stranded DNA and insert another piece of DNA. In its natural function the other piece of DNA is a transposon (see Chapter 25: Mobile DNA), but in this reaction, the transposase inserts two adapters, one for each end of the cut. The reaction is able to fragment the DNA and add the adapters in one single step.

fragments. This step can add a unique tag or **index sequence** (also called a **barcode sequence**) if the adapter was missing this element. Index sequences are essential when genomic DNA from different organisms or different people are mixed together in one sequencing reaction. The term for mixing multiple samples into one reaction is **multiplexing**. NGS platforms can simultaneously decode so many different fragments of DNA that there is usually plenty of room for multiple samples to be mixed. This is particularly true for smaller genomes, or genomic DNAs that have been sifted for particular sequences. The amplification step is sometimes omitted for NGS libraries that have a lot of genomic DNA in the sample because it can introduce sample bias, where some pieces of the genome are amplified and others are not. If the samples are amplified, the researcher uses PCR amplification. The forward and reverse primers are tailed—with the 5′ ends of the primers containing the index sequence. The 3′ end of the PCR primers are complementary to the adapter sequences.

Index sequences found in adapters are essential to identify which genomic DNA fragment belongs to which sample when multiplexing different DNAs in one sequencing reaction.

3.1.2 Partitioning the Library and Cluster Generation

The next step of NGS is to separate or partition the library of genomic DNA fragments into discrete locations on a solid surface. Illumina sequencing platforms use a **flow cell**, which looks like a glass slide, but is actually a very complex apparatus (Fig. 8.12). There are channels that are actually microfluidic chambers that allow different liquid reagents to flow at a defined rate. In addition, each of the microfluidic chambers has a bottom surface that is coated with oligonucleotides that are complementary to the adapter ends. The original Illumina flow cells had a solid smooth

barcode sequence A unique DNA sequence that is found within the adapters added to genomic DNA fragments for next generation sequencing. The index or barcode is used to distinguish different DNA samples during multiplexing. Also called an index sequence.
flow cell A glass slide with microfluidic channels on the surface that allow sequencing reagents such as DNA polymerase, dNTPs, buffers and sequencing primers to flow over the surface. The lower surface of the channel has oligonucleotides attached to the glass that provide anchor points for the genomic DNA fragments to attach.
index sequence A unique DNA sequence that is found within the adapters added to genomic DNA fragments for next generation sequencing. The index or barcode is used to distinguish different DNA samples during multiplexing. Also called a barcode sequence.
multiplex Mixing more than one sample in the same reaction. Can be used in reference to PCR, next generation sequencing, and other genomics applications.

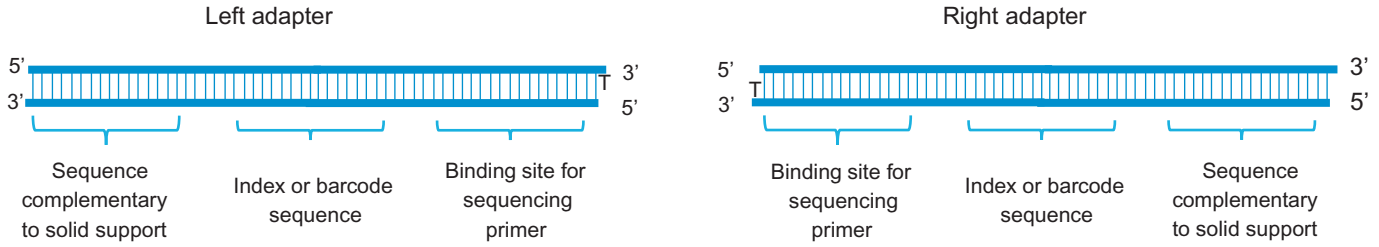
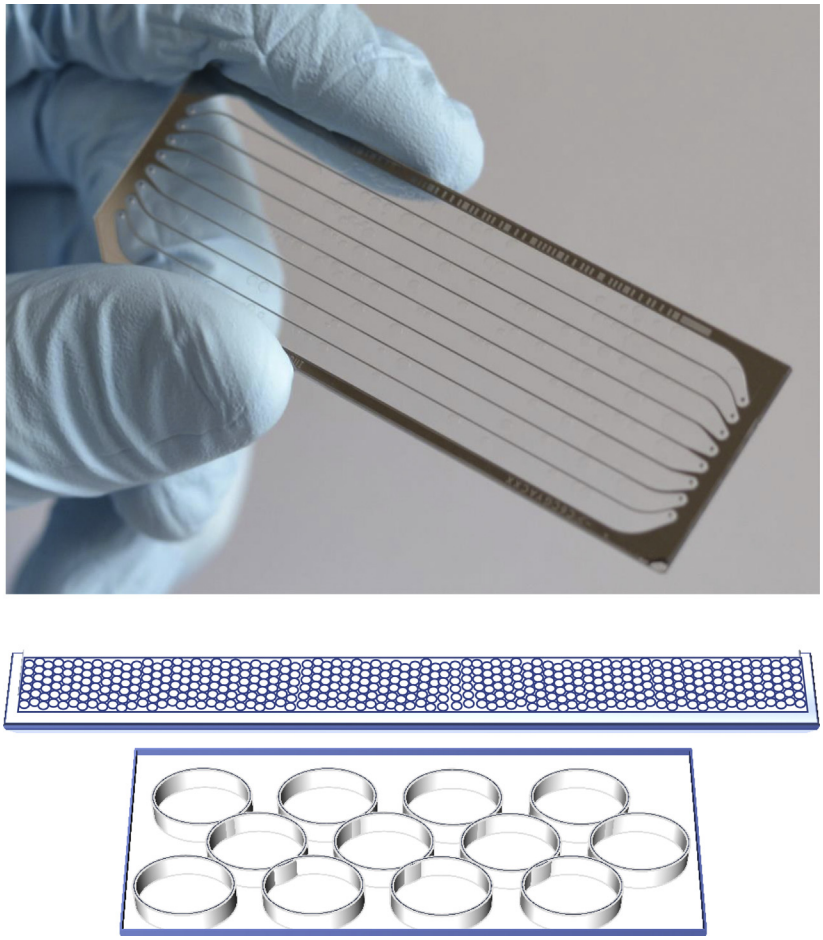


FIGURE 8.11
Adapter Features

Two different adapters are added onto the ends of genomic DNA fragments during NGS library preparation. The left adapter and right adapter have a sequence complementary to another oligonucleotide that is attached to the solid support, an index or barcode sequence, and binding site complementary to the sequencing primers. These sequences vary between left and right adapters and will vary depending upon the type of NGS sequencing methodology used.

FIGURE 8.12
Illumina Technology Partitions Genomic DNA on a Flow Cell

A picture of a flow cell that is used in the Illumina NGS sequencing platform to partition the genomic DNA fragments into discrete locations (top). Older flow cells had a smooth surface, but newer versions have a surface coated with nanowells (middle and lower diagrams) that help space out the genomic fragments and prevent overcrowding.



Flow cells partition individual genomic DNA fragments into discrete locations.

surface, but newer flow cells have nanowells that separate the different fragments into a depressions or dips in the surface. Although both types of flow cells function well, researchers using the original design must be careful to avoid overcrowding the library fragments. Overcrowded flow cells makes decoding the sequence more difficult. In order to partition the NGS library, the prepared fragments are injected into the flow cell and the temperature is adjusted so the genomic DNA adapters can anneal to their complementary oligonucleotides on the flow cell.

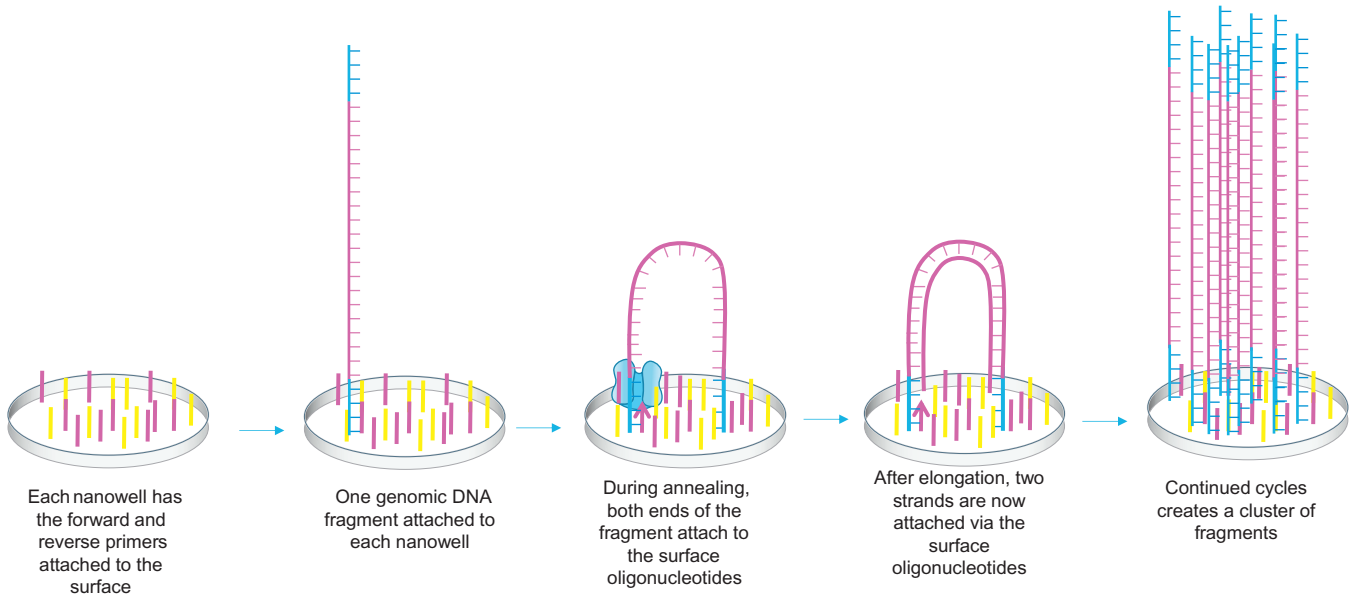


FIGURE 8.13
Bridge Amplification PCR Creates Clusters of Identical DNA Fragments

The individual genomic DNA strand attached to the flow cell does not provide ample enough signal for sequencing to occur, therefore, it must be converted into a group or cluster of identical copies via PCR. The forward and reverse primers for PCR are attached to the flow cell surface, and therefore, the genomic DNA arches over during annealing so that DNA polymerase can make a copy of the strand.

After each of the genomic DNA fragments attach to the flow cell, the next step is to create a **cluster** by PCR. This step is essential for sequencing. Cluster generation is accomplished with PCR, and the PCR primers are the oligonucleotides attached to the flow cell. The PCR amplification occurs when the genomic DNA fragment bends over so both adapters anneal to the surface via the oligonucleotides. The structure resembles a bridge, so this type of PCR has been called **bridge amplification** (Fig. 8.13). Continued cycles of denaturation, annealing, and extension/elongation temperatures convert the single copy of genomic DNA within the nanowell or position on the flat flow cell into a cluster of identical copies. (Note: An extra final denaturation leaves only one strand of the DNA attached to the flow cell for sequencing.) Each cluster represents one small segment of the whole genome, and therefore, each cluster only provides a small length of DNA sequence information. But there are tens of millions of clusters on a single flow cell, and each provides a piece of the code for the whole genome.

Each individual genomic DNA fragment is converted into a cluster of identical sequences using bridge amplification.

3.1.3 Sequencing by Synthesis

Illumina sequencing is done simultaneously for each and every cluster on a flow cell. The sequence of the genomic DNA fragments are determined by supplying a sequencing primer, DNA polymerase, and fluorescently-labeled nucleotides with each base connected to a different fluorophore so the four can be readily identified by the fluorescent detector to the flow cell (Fig. 8.14). As the ingredients move through the flow cell, the sequencing primer anneals to its complementary location on the adapter, and then DNA polymerase starts making a copy of the genomic DNA fragment using the fluorescently labeled deoxynucleotides as the building blocks. Since DNA polymerase synthesizes DNA so rapidly, the current detectors cannot record the fluorescence of

bridge amplification A type of PCR that uses forward and reverse primers that are both attached to a surface of a flow cell. During PCR, the DNA fragment bends over into an arch so that both ends of the DNA are annealed to the forward and reverse primers.
cluster (in reference to next generation sequencing) The grouping of identical genomic DNA fragment copies attached to the surface of a flow cell.

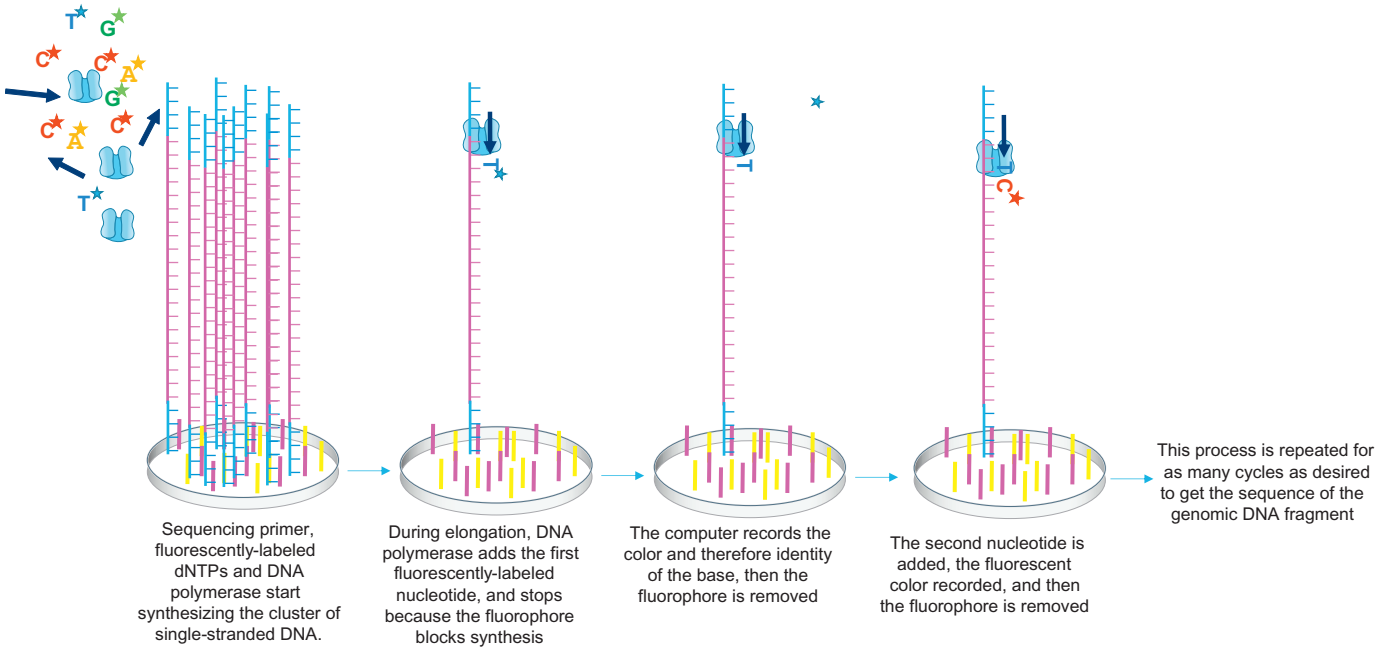


FIGURE 8.14
Sequencing by Synthesis

For each cluster of DNA on the flow cell, a supply of sequencing primers, fluorescently-labeled dNTPs, and DNA polymerase are added (left). As the temperature is adjusted the primer anneals to the DNA in the cluster, and DNA polymerase attaches. Only one DNA strand from this cluster is shown for clarity, but in reality the entire cluster of sequences are decoded simultaneously. After the first complementary nucleotide is added to the end of the primer, the fluorophore blocks DNA polymerase from adding any more nucleotides. The computer records the identity of the fluorophore, which is then removed and washed away. DNA polymerase adds the second complementary nucleotide, the identity is recorded and the fluorophore is removed. These cycles repeat to get a read from this cluster.

Illumina sequencing by synthesis creates a read from each individual cluster. Reads are generated in parallel from millions of different clusters.

every base that is added in real time, so the structure of these is designed so the fluorophore acts as a **blocking group**. The fluorophore is positioned on the nucleotide so that DNA polymerase cannot add another nucleotide until the fluorophore is removed. So as soon as DNA polymerase recognizes the 3′ hydroxyl group (3′ −OH) of the sequencing primer, it finds the complementary fluorescently-labeled nucleotide to the genomic DNA and connects it to the end of the primer. The fluorescence of the added nucleotide is recorded by the computer. Once the fluorophores for every cluster are recorded, then the fluorophore is removed, which allows DNA polymerase to add another complementary fluorescently-labeled nucleotide. The fluorescence of this nucleotide is recorded for every cluster, and then this fluorophore is removed. DNA polymerase adds the third complementary fluorescently-labeled nucleotide, and the fluorescence is recorded. The process continues in each cluster, and the computer records the identity of the added nucleotide via the color of its fluorescence. The final sequence for each cluster is called a **read**. Since there are tens of millions of clusters in a single flow cell, there are actually tens of millions of reads generated in these sequencing by synthesis reactions.

Illumina sequencing reads the different segments of the genomic DNA fragment at different points. There are actually multiple reads from one cluster. First, the flow cell is flooded with DNA polymerase, fluorescently-labeled deoxynucleotides, and a sequencing primer that anneals to the left side of the genomic DNA, and this combination decodes the genomic DNA from the 5′ side. This is the first read or read 1.

blocking group A functional group added to nucleotides to inhibit the addition of another molecule, such as another nucleotide by DNA polymerase.
read (in reference to next generation sequencing) The final sequence determined by decoding the order of nucleotides in a single cluster of identical genomic DNA fragments.

These ingredients or reagents are then washed from the flow cell, and a mixture of DNA polymerase, fluorescently-labeled deoxynucleotides, and a primer complementary to the right adapter are added. This decodes the index or barcode found in the right adapter. After the information is obtained from this second read or read 2, the ingredients are washed from the flow cell, and a new mixture of DNA polymerase, fluorescently-labeled oligonucleotides, and a third sequencing primer are added. These produce read 3, which decodes the left side adapter index or barcode. Finally, read 4 sequences the genomic DNA from the opposite side. In summary, there are four reads for each cluster that are recorded by the computer, and there are tens of millions of clusters for each flow cell. The ability to process all this information computationally with the decreased cost of computers and increased capacity for storage has been instrumental for the development of NGS, and the ability to decode a whole genome.

Illumina sequencing technology uses reversible dye terminators to record the nucleotide as a G, A, T, or C. The dye termination is reversible, and after removal, DNA polymerase can add another nucleotide.

3.1.4 Analyzing NGS Data

The first step of analyzing the data obtained from the sequencing by synthesis reactions is to combine the different reads from each cluster. Two of the four reads decoded the index sequence on both adapters. If the flow cell is simply loaded with one genomic DNA sample, the information from the adapter index or barcode sequences is less important. The index or barcodes are more important when samples are multiplexed, or when more than one genomic DNA sequence are mixed together. Then the adapter reads are used to sort each cluster into the different samples.

Reads from the genomic DNA can be aligned into continuous sequence information by either comparison to a previously sequenced genome, or by comparing one read to another looking for overlapping sequences. The first method is most common method since so many organisms have had their genomes decoded already. In the case of the human genome, the **reference genome** is available for viewing by anyone at Genome Reference Consortium Website (https://www.ncbi.nlm.nih.gov/grc). Each of the reads is individually aligned with the reference which is stored as a binary alignment mapping (BAM) file. Special software packages use the BAM file to see if any of the reads have differences from the reference genome, called a **variant**. The most important part is to determine the biological significance for the variant, which can be done with combination of experiments and mining other databases for other researchers' results.

Biological variants can be found by comparing the sequence of one genome to the reference genome.

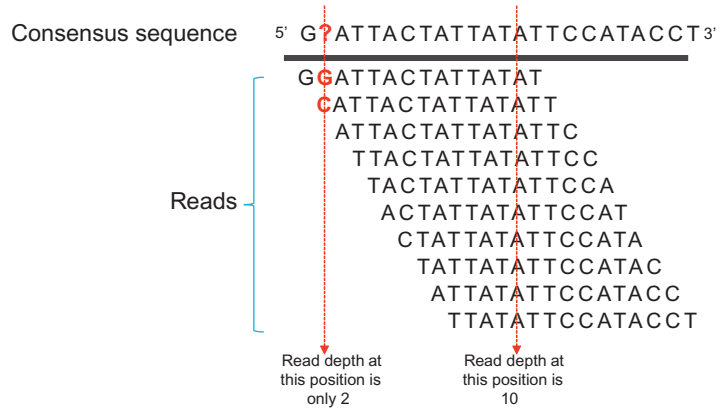
In contrast, ***de novo* sequencing** or sequencing DNA from a species that does not have a reference genome uses alignment of overlapping sequences to order the reads. Computer algorithms look for matching sequences among the millions of reads, aligning the reads into as long of a sequence as possible. The goal of the sequencing alignments is to put together as many reads into one contiguous length of sequence information, called a **contig**. Contig length is another important NGS parameter, and the goal of any sequence analysis is to have one contig for each chromosome, although in reality, the repetitive regions often make this impossible. Many reads may have minor discrepancies from each other so after the alignment, the computer usually picks the base that is most common at each position to create the **consensus sequence** (Chapter 9: Genomics and Systems Biology provides more details on the creation of a contig and consensus sequence).

The human genome has 3.2 billion base pairs of DNA sequence, a length of DNA that is actually about six feet. Every cell has the DNA genome packaged into a nucleus (except for red blood cells). If someone would print out their genetic

consensus sequence Idealized base sequence consisting of the bases most often found at each position.
contig A length of decoded DNA sequence that is continuous and has no gaps.
***de novo* sequencing** Decoding the order of nucleotide bases in a genome from an organism that has never been sequenced.
reference genome The consensus genetic code for an organism that has been determined in previous sequencing experiments and used for comparison.
variant A nucleotide sequence found in the genomic DNA sample that differs from the reference genome, which can include single nucleotide polymorphisms (SNPs), insertions or deletions (indels), or chromosomal aberrations.

FIGURE 8.15
Read Depth

The number of reads that overlap varies for each position of the genome during next generation sequencing. The greater number of reads, the more confidence in the identity of that particular base. In the position that has a read depth of two, the base could either be a G or C.



code using a standard sized font such as New Courier 12 point and using 1 inch margins, the total number of G's, A's, T's, and C's per page would be 3008. So, it would take 106,383 pages to print an entire human genome. When a sample of human genomic DNA is isolated for next-generation sequencing, there are many copies of the genome that are mixed together. The multiple copies of the genome ensure that there are multiple genomic DNA fragments from each and every chromosome in the final genomic DNA library. Because the shearing is random, the ends of the fragments are rarely exactly the same. The term average read depth refers how many reads were decoded on average across the whole genome. Typically, good read depth for a whole genome should be 20–30X. Assuming 30X coverage, the number of pages of G's, A's, T's, and C's now is 3,191,490 pages of print. These numbers are used to explain why NGS is so reliant on computing power! This type of data analysis requires large amounts of storage and computer capacity. In fact, one genome sequencing data can consume 80–90 Gigabytes of computer storage.

De novo sequencing of a genome relies on finding overlapping sequences among all the reads to form contigs.

Genome coverage is not equal throughout the whole human genome. Each of the 3.2 billion bases of the human genome will vary as to how many cluster reads decoded that position. The number of reads that cover a location is called **read depth**. The coverage for a whole genome is an average of the read depth. If there is a mistake in one read, and the read depth is two, it can be difficult to ascertain which of the two possible bases is correct. On the other hand, if there is a mistake in one read, and the read depth is 30, it is easy to identify which of the two bases is correct since the majority rules. The more reads, the more confidence in the sequence (Fig. 8.15).

Read depth for the different nucleotide positions in a genome vary based on how many genomic DNA fragments were decoded from the area of the chromosome. Some areas will have no reads, and others may have multiple. An average read depth of 20–30X is considered sufficient for whole genome sequencing.

Although the decoding power of NGS is unprecedented, there are still gaps or regions of the human genome that are missing sequence information. The reason stems from the repetitive nature of the human genome. When each cluster decodes between 100 and 200 bases of information and the read falls into a repeated area, it is impossible to figure out the exact number of repeats (see Fig. 8.16). There are new third-generation sequencing technologies (as discussed later) that are working to fill in these gaps.

3.2. Ion Torrent Sequencing Technology

3.2.1 Library Preparation and Partitioning of the Library Fragments

Although the Ion Torrent sequencing platform differs dramatically in its sequencing method, library preparation for this sequencing and Illumina based technology is similar. Ion Torrent technology also sequences small DNA fragments that have adapters added onto each end. So library preparation requires physical or enzymatic breakage of a genomic DNA sequence into millions of tiny double-stranded pieces,

read depth The number of reads that were overlapping in their decoding of one specific location in the genome during next generation sequencing.

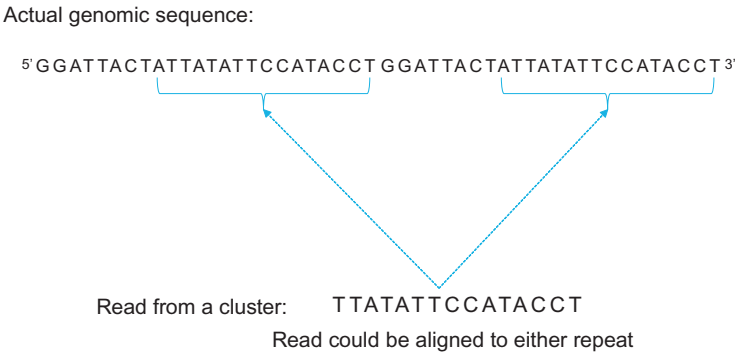


FIGURE 8.16
Aligning Reads From Repeated Regions

When the read is derived from a repeated region, it is difficult to determine where it should be aligned. This read could be aligned to either location in the genome.

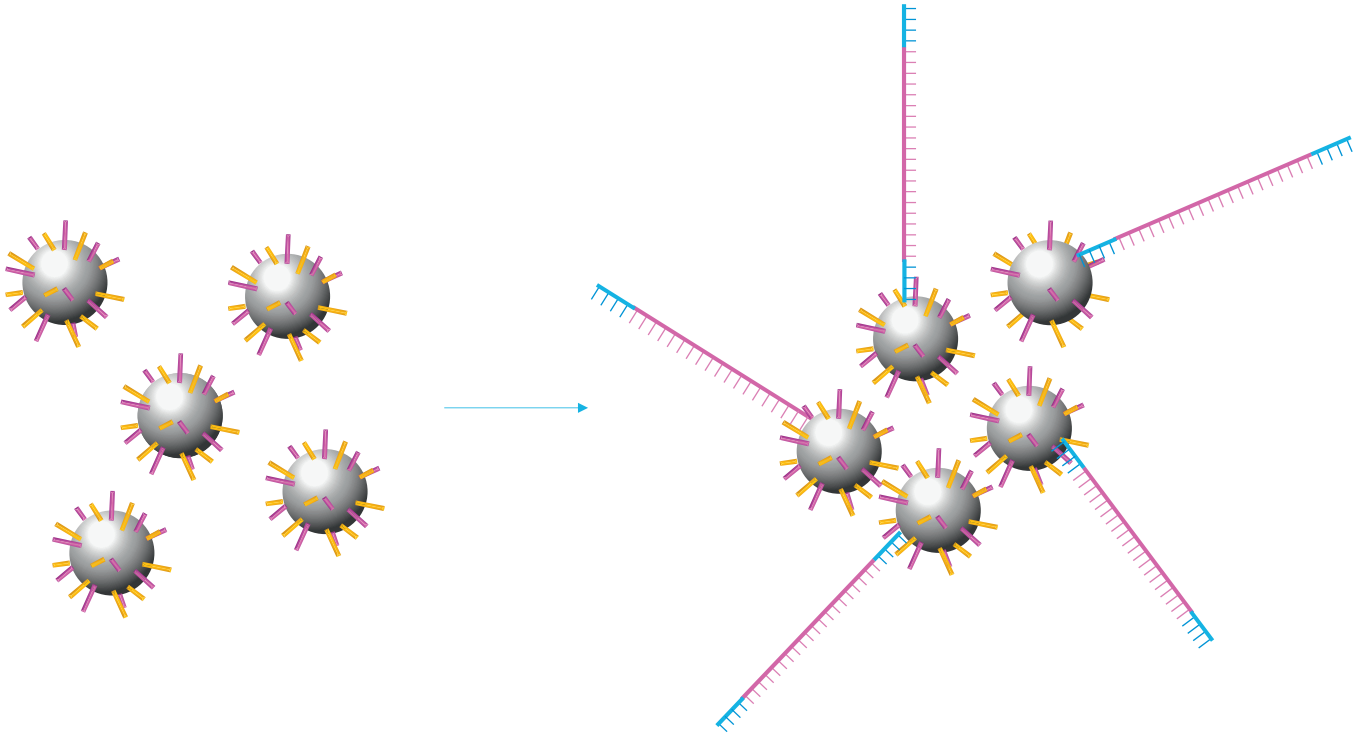


FIGURE 8.17
Ion Torrent NGS Platforms Use Beads to Partition Genomic DNA Fragments

Microbeads coated with oligonucleotides (pink and yellow) complementary to the adapters are combined with genomic DNA library fragments so that one fragment binds to each bead.

the size of which is in the 300–800 base pairs in range. After establishing the fragments, each piece of the genome must have adapters added onto each end. As before, the adapters are used to create a known DNA sequence on each end. The sequence has binding sites for the different sequencing primers, PCR primers, and also for the binding of the genomic DNA fragment to the various partitions or wells.

After library construction, the next step of Ion Torrent sequencing is the partitioning each individual genomic fragment into a separate location. This step differs from Illumina sequencing. The DNA pieces are not attached to a flow cell, but instead, the fragments are attached to tiny microbeads that are coated with a complementary oligonucleotide to the fragment's adapter sequence. Annealing the fragment to the bead is controlled so that there is only one DNA fragment for each bead, essentially separating the fragments so that there is one bead for each original piece of the genome (Fig. 8.17).

Ion Torrent sequencing platforms separate individual genomic DNA fragments by attaching to beads that are partitioned into individual water droplets in an oil and water emulsion.

FIGURE 8.18
Emulsions Partition Each Bead Into a Separate Water Droplet

When the beads are mixed with water and oil, each bead is encapsulated in a separate water droplet that prevents the PCR reactions from cross contaminating other beads.

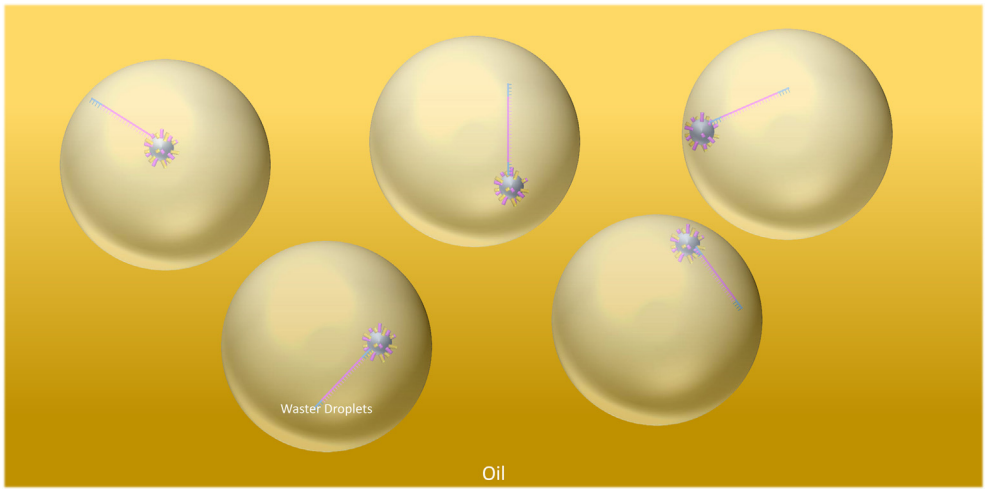
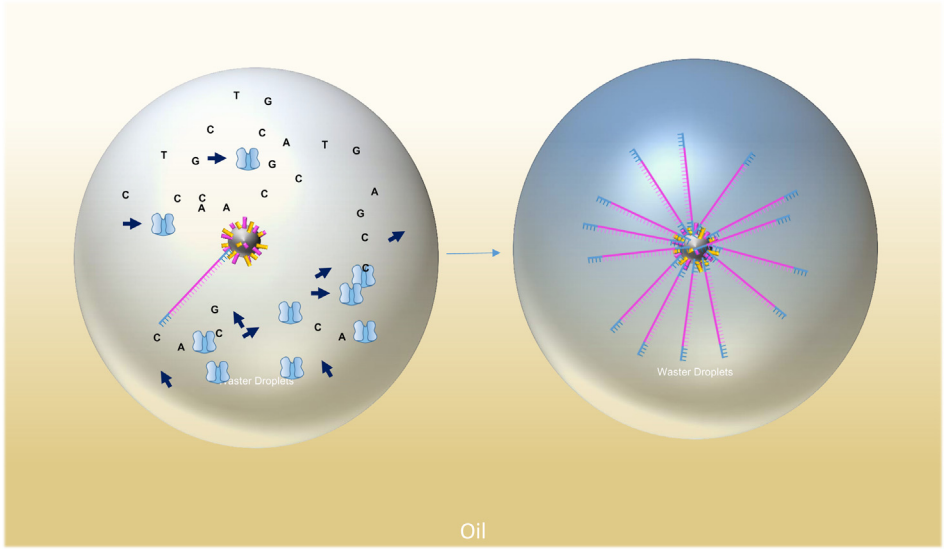


FIGURE 8.19
Emulsion PCR

DNA polymerase, PCR primers, and dNTPs are added to the emulsion in order to synthesize thousands of copies of the original DNA fragment.



As was outlined for the Illumina partitioning, one fragment of DNA cannot be assessed for its sequence, and therefore, the individual fragment is amplified into millions of copies using PCR. Since the beads are not really physically separated at this point, simply adding PCR reagents could cause cross contamination of one fragment onto another bead. To solve this problem, **emulsion PCR** is used to "physically" separate the beads and amplify the genomic DNA. The beads are separated by encapsulating the bead in a small drop of liquid suspended in a solution of oil, that is, the beads are found in the water droplets of an emulsion (Fig. 8.18). The PCR reagents such as thermostable DNA polymerase, dNTPs, and PCR primers complementary to the adapter are then mixed into the emulsion, and since they are water soluble, they partition into the water droplets also. The surrounding oil prevents the diffusion of the PCR products from one bead to the next. After each cycle of PCR, the copies of the original fragment anneal to complementary oligonucleotides that are attached to the beads. After multiple PCR cycles, each bead becomes coated with thousands of copies of the original genomic DNA fragment (Fig. 8.19).

emulsion PCR PCR reactions that occur in water droplets that are surrounded by oil, which essentially makes each droplet a separate PCR reaction vessel.

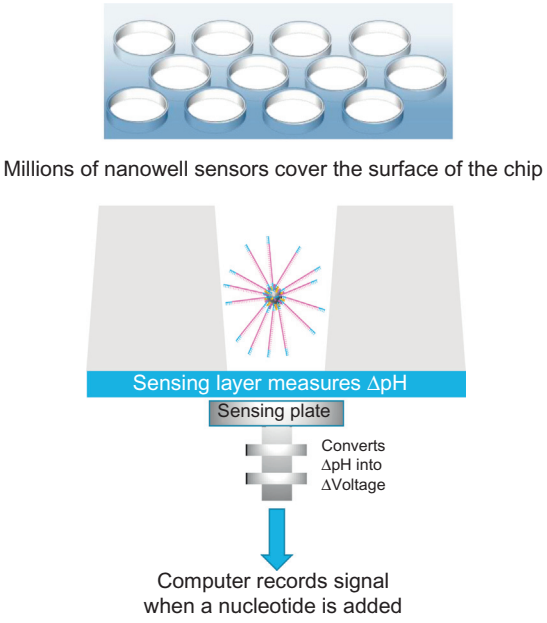


FIGURE 8.20
Ion Torrent Semi-Conductor Chip

The surface of the semi-conductor chip has millions of nanowells (top) that each has their own individual sensor for pH underneath (bottom). The pH sensor data is converted into a change of voltage, which is then sent to the computer.

3.2.2 Sequencing by Synthesis on the Ion Torrent Platform

After emulsion PCR, the oil is removed and each of the beads is partitioned into separate microwells that are found in a dense array on a flat surface called the semi-conductor chip (Fig. 8.20). The top surface of the chip has many small wells, just big enough to fit one bead bathed in a small amount of liquid. After the beads are separated into their wells, DNA polymerase and a sequencing primer are added. The temperature is adjusted so the sequencing primer anneals to the adapter region of the genomic DNA fragment. Notice that no dNTPs are present, so DNA polymerase has no nucleotides to add onto the primer at this point, but will sit poised to add a nucleotide complementary to the genomic DNA when they become available. Remember that there are thousands of copies of the genomic DNA on the bead, and each copy has a primer and DNA polymerase waiting for action.

Semi-conductor chips for Ion Torrent platforms have miniature pH meters that measure the release of protons (H +) during phosphodiester bond formation. This pH change is converted into a voltage change, and then recorded as a 1 or 0 by the computer. A 1 means the nucleotide was added, and a 0 means the nucleotide was not added.

Sequencing begins when a single type of dNTP is flooded over the surface of the chip. For example, dGTP is added to the entire surface of the chip and spreads across the millions of wells. If there is a well that contains a genomic DNA fragment with a C for the nucleotide next to the end of the primer, then the dGTP is added on to the primer by DNA polymerase (Fig. 8.21). If the genomic DNA fragment does not have a C next to the primer, then the dGTP is not added. When the dGTP is added, a hydrogen ion (H +) is released, which changes the pH of the liquid in the well (see Fig. 8.04). Underneath each of the wells or holes is a sensor that records the pH of the liquid in the well, and converts the change into a voltage change. And below the pH meter is the semiconductor that converts voltage changes into a 1 or 0, which the computer records. The wells that added the dGTP have the first decoded nucleotide of the sequences, whereas, other wells do not have anything recorded for the first nucleotide. Any excess dGTP is removed from the chip after the data is recorded.

Next, a second type of dNTP is flooded over the chip. As the second nucleotide, such as dATP, flows into the wells, those with DNA polymerase poised next to T will use the dATP. In other wells, if DNA polymerase is poised next to any of the other bases, nothing happens. The pH of the wells where dATP is added changes, the sensor records the change, and finally the computer records the sequence in those wells only. Notice that in some wells, the genomic DNA could have the sequence CT, so there would already two nucleotides decoded, but in other wells,

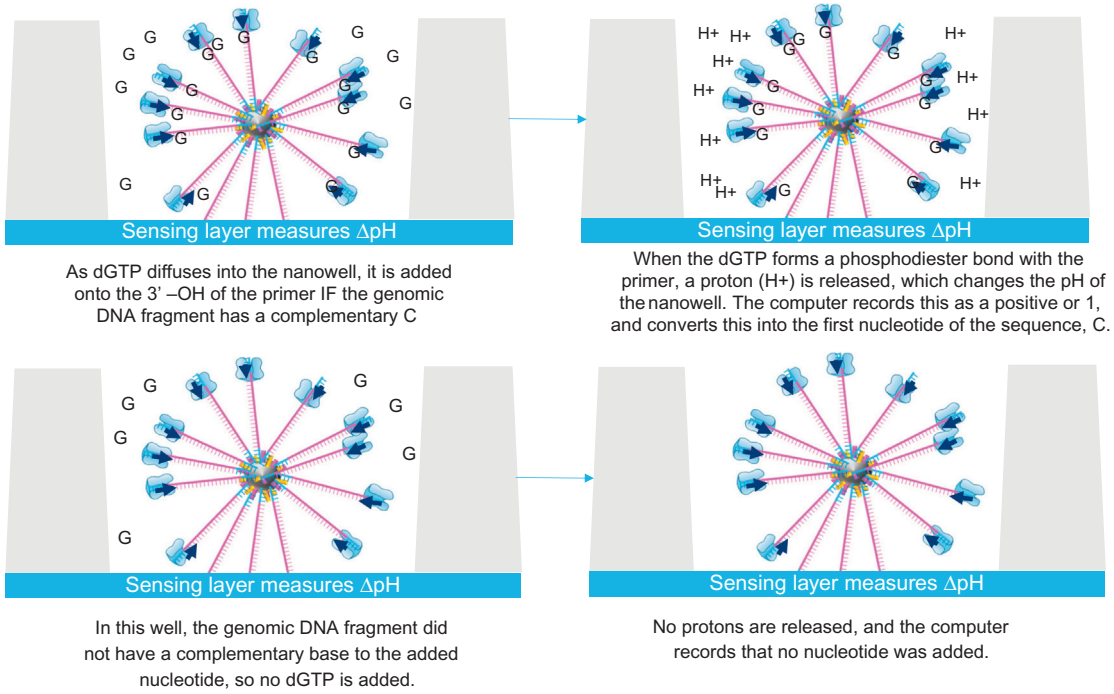


FIGURE 8.21
Sequencing by Synthesis in Ion Torrent Platforms

DNA polymerase, PCR primers, and one of the four dNTPs are washed over the surface of the semi-conductor chip. If the genomic DNA fragment has the complementary base next to the 3′ end of the sequencing primer, DNA polymerase will incorporate the single dNTP. In the top left, dGTP was incorporated. As the phosphodiester bond forms, a proton ($H+$) is released, which changes the pH of the nanowell (top right). The pH change is converted to a change in voltage, which is recorded as 1. In contrast, when the genomic DNA fragment does not have the complementary base to the added dGTP (bottom row), then the pH does not change. The change is voltage is recorded as a 0.

neither the dGTP or dATP were complementary, so no sequence information exists. Again, the excess dATP is washed away.

In a similar manner, a third type of dNTP (dCTP for example) is added to the nanowells of the semiconductor chip. In any well that DNA polymerase is poised next to a G, the dCTP is added onto the 3′ end, $H+$ is released, and the pH changes are recorded and converted to a voltage increase, which in turn is recorded. After addition of the final dNTP (dTTP for example), every one of the wells should have at minimum one nucleotide decoded. At this point, the process begins again, where dATP, dCTP, dGTP, and dTTP are added separately. If DNA polymerase adds the base, then the data is recorded. Because this sequencing technology relies on pH changes, which are the exact same for all four nucleotides, the bases must be added separately so the computer can ascertain what base is added (Fig. 8.22).

This method of sequencing is much faster than the Illumina-based method that uses fluorophores that act as blocking groups, but its weakness is when the genomic DNA has multiple identical bases in a row. For example, if the genomic DNA has two T nucleotides in a row, and the wells are flooded with dATP, then DNA polymerase adds two dATP since there is nothing to stop synthesis. The pH of the well will change, but since double the $H+$ are released, then there will be double the voltage recorded (Fig. 8.23). The technology does fine for two or three identical bases in a row, but when there are greater numbers identical nucleotides, then the pH and voltage measures do not vary as much, and the recorded pH/voltage changes can be misinterpreted, leading to sequencing errors.

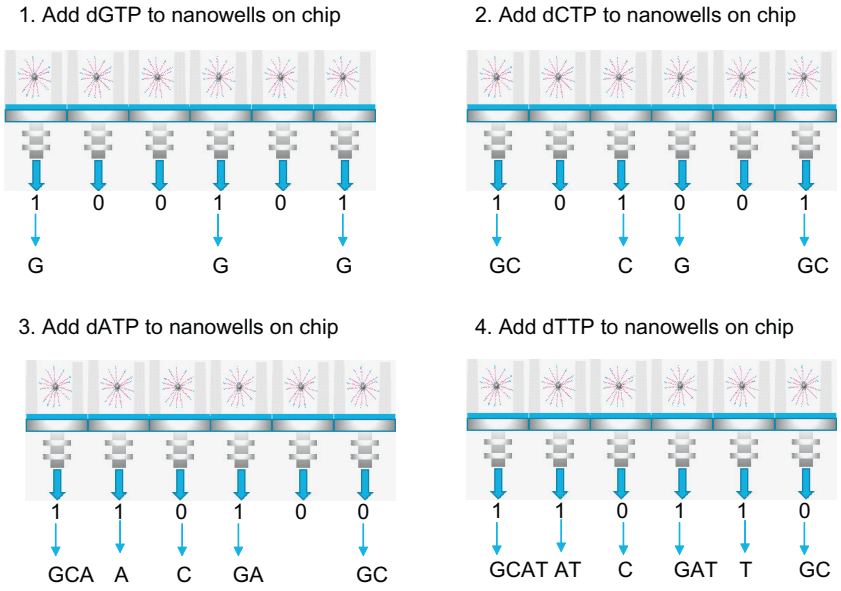
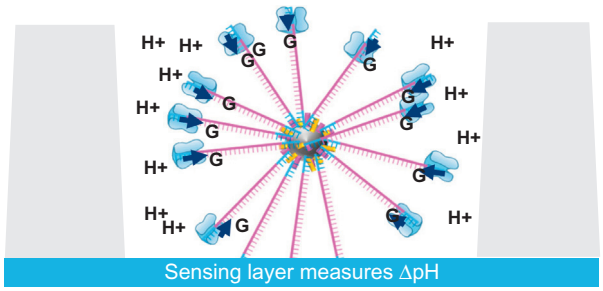
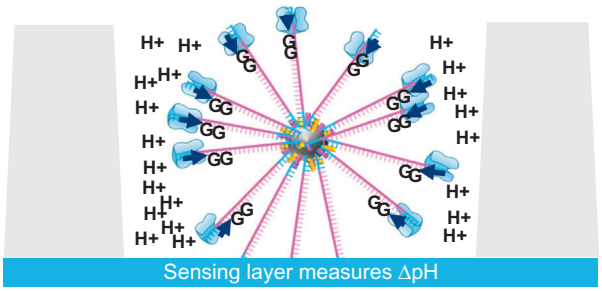


FIGURE 8.22
Sequencing Reads

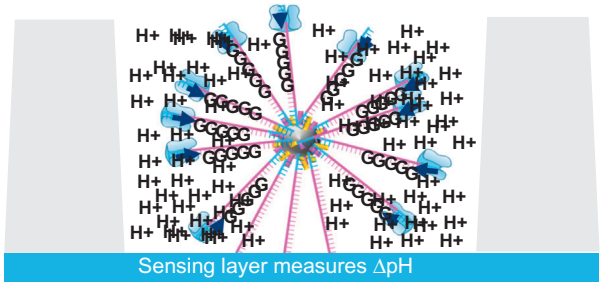
As the different nucleotides are added one by one, the computer records whether or not the pH changed in each of the millions of nanowells. If the pH changes, the computer adds a that nucleotide base onto the sequence read for that well. If the nucleotide is not added, then no information is added. By the end of the first four nucleotide additions, at least one base should be recorded for each well.



When the dGTP forms a phosphodiester bond with the primer, a proton (H+) is released, which changes the pH of the nanowell. The computer records this as a positive or 1, and converts this into the first nucleotide of the sequence, C.



When 2 dGTPs are added, double the protons are released, and the pH is also twice the normal amount.



If the genomic DNA has a long stretch of repeated bases, the change in pH can be difficult to distinguish the exact number of nucleotides that were added.

FIGURE 8.23
Multiple Repeated Bases Can be Misinterpreted

Since Ion Torrent relies on the detection of pH differences due to nucleotide additions, when the genomic DNA fragment has multiple repeated bases, the pH differences can be difficult to determine how many bases are added.

4. Targeted Sequencing

The discussions of the various sequencing technologies have centered on the use of whole genomes, and NGS has the power to analyze this much data. In many uses though, sequencing the whole genome can cost too much, provide too much information, and take too much time to analyze. Instead of whole genome sequencing (WGS), clinicians and researchers often focus their NGS analysis on a subset of the whole genome. In **targeted sequencing**, a series of genes or regions of interest are isolated or enriched from a whole genomic DNA sample before sequencing using next generation technologies. The real advantage to enriching the sample of genomic DNA is coverage because the chosen regions are more defined. The more reads obtained for a specific set of genes, the better the results.

Targeted sequencing has many applications for understanding cancers, inherited genetic diseases, effects of the environment on our genome, and even understanding biodiversity of our planet. For example, the human genome has many regions that do not code for proteins, and therefore, are less likely to harbor the mistakes that cause a particular genetic disease. The regions that do code for proteins, all the exons, are more likely to have the deleterious mistakes. Therefore, instead of sequencing the whole genome, many researchers and clinicians are sequencing all the exons, or the **exome**, called **whole exome sequencing**. Other subsets of genes that are being studied in targeted sequencing include groups of cancer causing genes, which can find the root genetic mistakes that have led to the development of that cancer. Some examples include sequencing genes associated with leukemia, melanomas, myelomas, and many others. There are targeted sequencing experiments that focus on subsets of genes associated with genetic diseases including cardiomyopathies and neuromuscular disorders. In addition, many companies offer custom designs that are designed by the researcher for their specific interest.

Targeted NGS sequencing analyzes a subset of genes or regions of interest from a whole genome.

Two different methods are currently being used to create a targeted DNA sequencing library. The first method uses highly multiplexed PCR reactions to amplify the regions of the genome that are of interest (Fig. 8.24). This procedure starts with a sample of whole genomic DNA. The genomic DNA is mixed with a set of PCR primers that amplify the target genes. Some of these PCR reactions have up to 24,000 forward and reverse primer pairs in a single reaction. Each primer pair amplifies one region of interest from the genome. The PCR reaction then is performed as usual with the entire set of PCR primers. The final PCR reaction is then cleaned up to remove the excess PCR primers, and then collection of PCR amplicons are attached to the adapters that are compatible to the NGS platform. The final set of PCR amplicons are sequenced and compared to the reference genome as in any other NGS reaction.

Multiplex PCR can amplify specific regions of a genome for targeted sequencing.

The second method of targeted sequencing uses biotinylated oligonucleotide probes that have sequences complementary to the chosen targeted genes or exons. The first step of the targeted sequencing procedure is to isolate whole genome DNA, fragment the whole chromosomes into small pieces, and then add adapters onto the ends to make them compatible with the sequencing platform. The complete set of fragments is then mixed and annealed to the biotinylated probes (Fig. 8.25). The set of probes, called a **panel**, can have from just a few different probes to over hundreds of thousands. The sequence of each probe is unique so it binds to a different location of the selected target genes. For example, a panel that is used to isolate all the exons from the human genome contains more than 400,000 different probes, every single one chemically synthesized to have sequences complementary to a

exome The subset of a whole genome that has the DNA with only the protein coding information (exons).
panel (in reference to next generation sequencing) A set of individual oligonucleotide probes that are labeled with a biotin group on the 5′ end. Each oligo probe has a sequence complementary to a gene or region of interest in the genome, and is used to isolate these genomic regions from the whole genomic DNA sample for targeted sequencing.
targeted sequencing Isolating a series of genes or regions of interest from a whole genomic DNA sample before sequencing using next generation technologies.
whole exome sequencing A targeted sequencing strategy that only decodes the genomic DNA fragments that have exons.

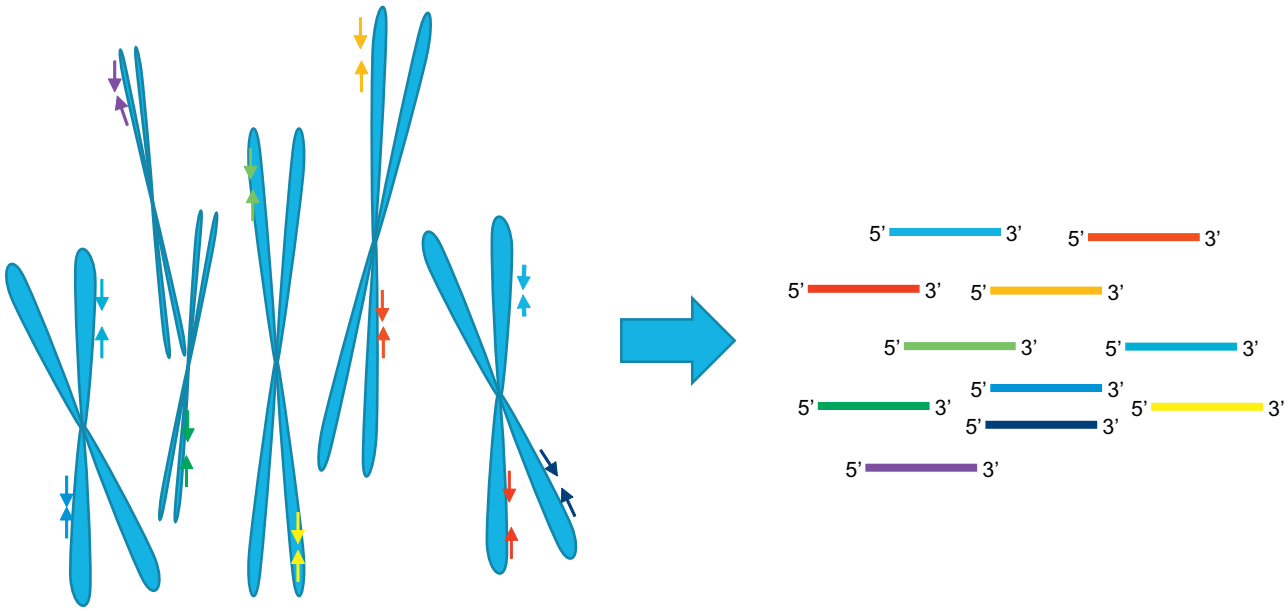


FIGURE 8.24
Multiplex PCR Creates a Targeted Sequencing Library

Simultaneously amplifying many different regions of a genome using different PCR primer sets creates a series of PCR amplicons for just the genes or regions of interest in the genome.

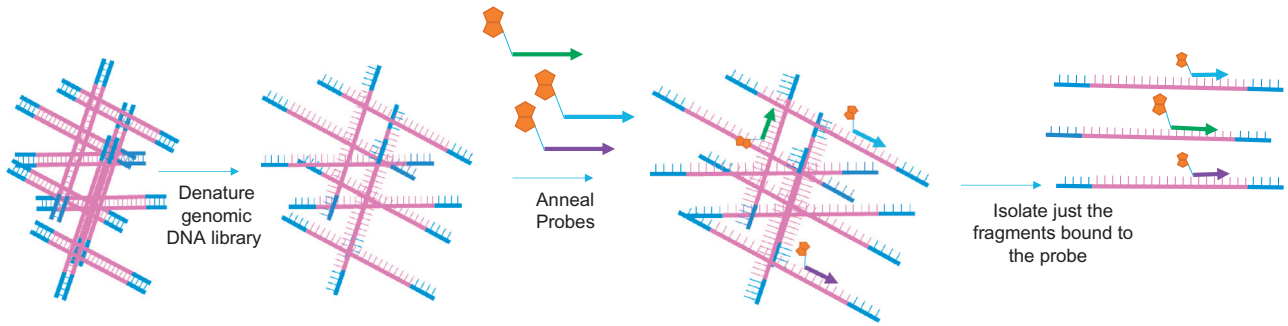


FIGURE 8.25
Biotinylated Oligonucleotide Probes Create a Targeted Sequencing Library

Genomic DNA fragments with the genes or regions of interest can be isolated using biotinylated oligonucleotide probes. The sequence of the probe is complementary to the targeted genes or regions of interest. Once the probe binds to the genomic fragment, they can be isolated by binding to magnetic beads coated with streptavidin (not shown).

different exon in the human genome. Although this is a ton of different probes, not every piece of the original DNA sample will anneal to a probe.

After annealing the probes to the genomic DNA fragments, the probe-bound fragments must be separated from the remaining genomic DNA fragments. That is where the biotinylated part of the probe becomes important. As described in Chapter 5, Manipulation of Nucleic Acids, biotin groups bind tightly to another molecule called strepavidin, therefore, the sample of genomic DNA bound to these biotinylated probes is mixed with magnetic beads coated with streptavidin. All the DNA pieces bound to the biotinylated probes stick to the beads, and are easily separated from the remaining genomic DNA using a magnet. After cleaning up the non-specifically bound genomic DNA, the fragments of interest are removed from the magnet by heating them. This causes the probe and DNA fragment to denature, releasing the DNA fragments for NGS analysis.

Panels of biotinylated oligonucleotide probes can be used to isolate specific sequences of interest from a whole genome NGS library.

Targeted sequencing is changing the medical field rapidly. If a researcher wanted to ascertain what gene was defective in a series of different patients that had the same cancer, they could isolate DNA from the tumor biopsy and a blood sample to provide the normal sequence and cancer sequence from each patient. Then by using targeted sequencing with a different barcode or index for each of the samples, the whole series of patients' DNA could be sequenced at one time. The results give both specific individual data, as well as, comparative data from one patient to the next. These types of experiments are leading to major discoveries on the types of mutations that are found in different cancers, and some correlations are being made between the type of mutation and the prognosis for the patient. Another surprising outcome of this research is finding better drugs to target cancer. For many years, cancer drugs were given based on the location of origin for the cancer. Although this works in general, the type of cancer can vary even if it originates in the same spot. Targeted sequencing can identify and categorize cancers based on the genetic profiles, and not on their location of origin, which allows the doctor to treat the cancer with a drug that works against the cancer found in each patient, and not the typical cancer found in a specific organ.

5. Third-Generation Sequencing

The reason the following two sequencing technologies are considered a third generation is their ability to decode single copies of a genomic DNA fragment. Unlike the previous next generation method that requires multiple copies of each genomic DNA fragment (clusters in Illumina and coated beads in Ion Torrent), these newer sequencing methods have established methods that decode one single strand of DNA.

5.1. Nanopore Detectors for DNA

Nanotechnology is based on microscopic machinery that operates at the level of single molecules. The first example of a single-molecule sequencing method is called **nanopore sequencing**. As the name implies, the genomic DNA passes through a nanopore, or small opening in a membrane that is approximately 1 nm in diameter. The diameter is so small that it only permits one single strand of DNA to pass through at a time. As the single-stranded DNA transits the pore, a detector records how the current changes from the channel. Every base has a slightly different structure, and therefore each base blocks the current differently. The advantages of nanopore technology are its high speed and its ability to handle long DNA molecules. In addition, many nanopores can be assembled into a very small region, and many long fragments of DNA sequence can be determined simultaneously.

Nanopore detectors permit a single DNA strand to pass through a tiny pore and sequence it as it passes.

A practical **nanopore detector** consists of a channel in a membrane that separates two aqueous compartments. When a voltage is applied across the membrane, ions flow through the open channel. Since DNA is negatively charged, the DNA is pulled through the nanopore to the positive side. The DNA molecules enter the pore and are pulled through in extended conformation, one at a time (Fig. 8.26). During the time the channel is occupied by the DNA, the normal ionic current is reduced. The amount of reduction depends on the base sequence (G > C > T > A); therefore, a computer can measure the current and decipher the sequence based on the differences.

Nanopore detectors use either a pore created by a protein such as alpha-hemolysin from *Staphylococcus* or a solid state nanopore that is a synthetic membrane with fabricated pores created with ion or electron beams. Surprisingly, the precision of the protein pores is much greater than the drilled holes in a synthetic membrane. Extensive research on protein structure and function allows the protein pores to be changed in structure by Angstroms. These subtle protein changes can

nanopore detector Detector that allows a single strand of DNA through a molecular pore and records its characteristics as it passes through.
nanopore sequencing Determining the order of bases for a single-strand of DNA as it passes through a small pore or channel in a membrane.

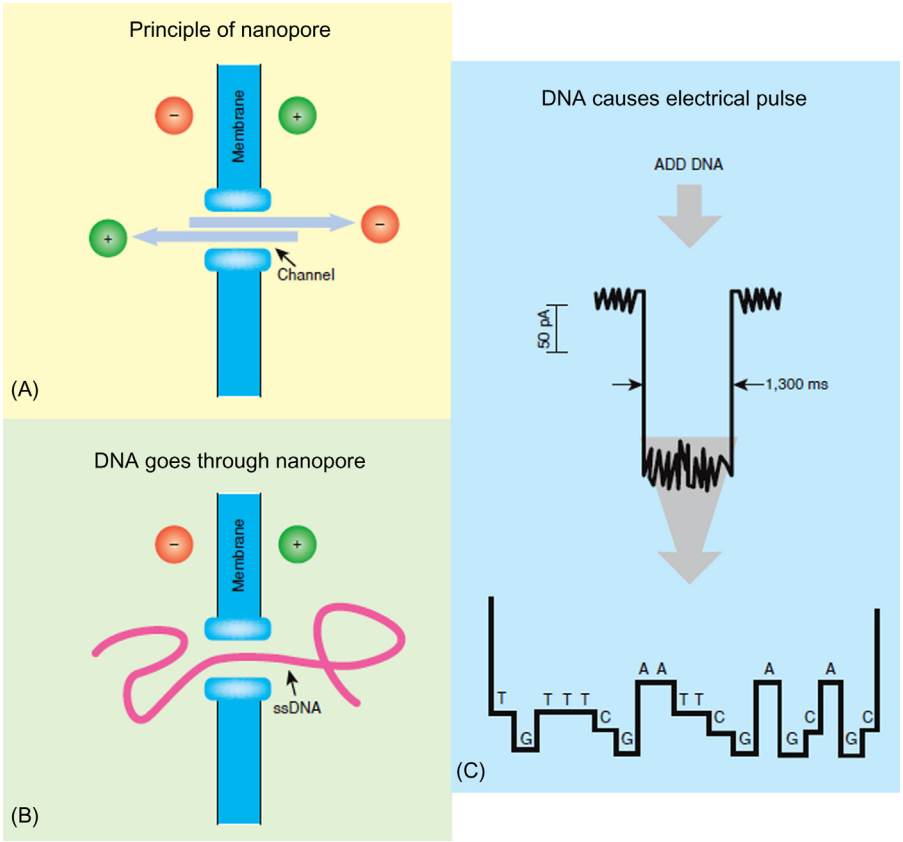


FIGURE 8.26
Principle of the Nanopore Detector

(A) Nanopores are small openings in a membrane that only allow one molecule through at a time. The nanopore membrane separates two compartments of different charge. (B) Since there is a charge separation between compartments, negatively charged molecules like DNA can pass through the pore in an extended conformation. (C) While the DNA is passing through the pore, a detector measures how much the current, due to normal ion flow, is reduced. Since each base alters the current by different amounts, the detector can determine the sequence as the DNA passes through the pore.

alter how fast the DNA travels through the pore, and proteins also allow various attachments that can make sure only the desired DNA is allowed to pass through the channel. For example, the mouth of the alpha-hemolysin channel is about 2.5 nm wide—roughly 10 atomic diameters. Double-stranded DNA can enter the pore mouth, but toward the middle, the channel narrows to less than 2 nm, which prevents dsDNA from going any further. The dsDNA remains stuck until the strands separate, allowing single-stranded DNA to pass through the length of the pore.

Current nanopore detectors are small enough to be carried or large enough to decode multiple genomes simultaneously. The small devices connect to a regular computer via a typical USB (Universal Serial Bus) port so they are easily used in the field as well as in the lab. As the data are processed, the resulting sequence is listed in real-time. In addition, the length of the DNA fragments for sequencing does not have to be small. At this point in time, the technology still suffers from inaccurate sequencing errors, but for re-sequencing or for quick analysis, this technology provides accurate enough data.

5.2. Long Reads From SMRT Sequencing

Pacific Biosciences **SMRT** (single-molecule real-time) sequencing (Fig. 8.27) uses **zero-mode waveguides** (ZMWs) or nanocontainers that are so small only a single piece of template DNA can occupy the space. As in several other sequencing methods, DNA polymerase extends a growing chain by adding nucleotides tagged with four alternative fluorescent dyes. Incoming nucleotides emit a flash of light as they

SMRT sequencing (for single-molecule real-time) Third-generation sequencing method that identifies the nucleotide added onto a growing strand of DNA using small nanowells that only allow enough light to penetrate so one DNA polymerase and one DNA template are visible. Since the fluorophores identify the nucleotide base by being bound to pyrophosphate, the sequence is read in real time, and the template can be tens of thousands of bases in length.
zero-mode waveguides (ZMWs) Small nanosized metal cylindrical wells that reduce background light so that only a very small portion of the cylinder can be visualized for fluorescent light flashes.

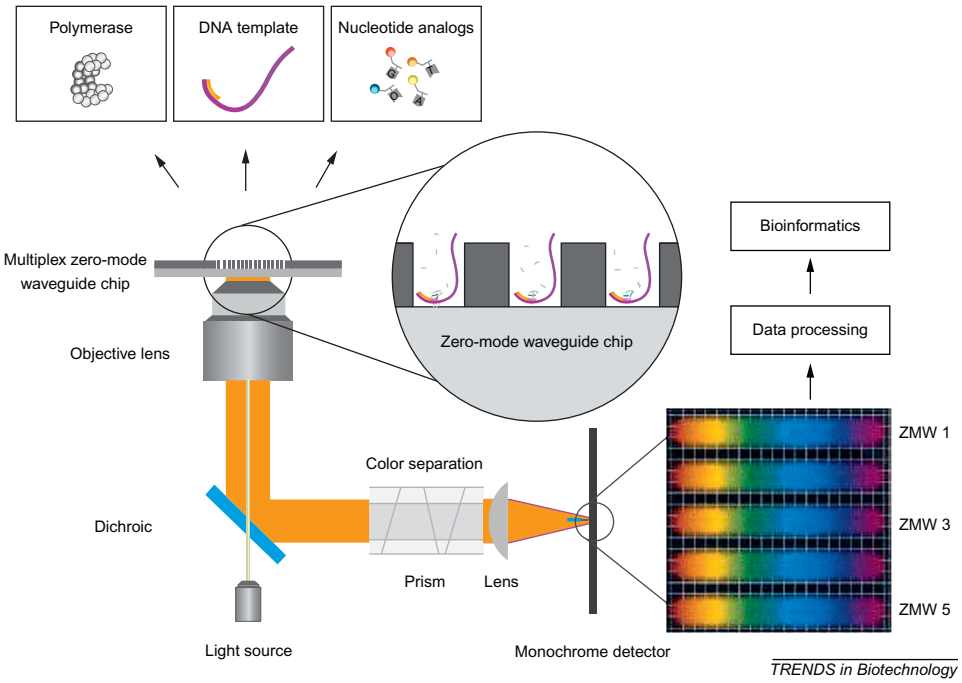


FIGURE 8.27
Principle of SMRT Sequencing

A schematic illustration representing the highly parallel optic system used in single-molecule real-time (SMRT) DNA sequencing technology (Pacific Biosciences). This method uses SMRT chips that contain thousands of zero-mode waveguides (ZMWs). The presence of a fluorescent dye within the detection volume (within the ZMWs) indicates nucleotide incorporation. This leads to a light flash that is separated into a spatial array, from which the identity of the incorporated base can be determined. *(Credit: Figure reproduced from the document 'Pacific Biosciences Technology Backgrounder' (dated 2/2/2008), with expressed permission of Pacific Biosciences.)*

SMRT sequencing has zero-mode waveguides that permit the visualization of a single DNA polymerase adding a nucleotide onto the growing DNA. The phosphate molecules are labeled with a fluorescent tag that emits light upon release from the DNA.

are linked in place. The fluorescent tag is then washed away to allow the next cycle. The sequence of colors reveals the order of the bases.

Two novel features are critical. The reactions are carried out inside nanocontainers—that is, within hollow metal cylindrical wells 20 nm across called ZMWs. Their small size reduces background light enough for individual flashes from each single reacting nucleotide to be detected. Several thousand ZMWs are assembled onto a single chip. The second breakthrough is in attaching the fluorescent tag. Instead of linking it to the part of the incoming nucleotide that will be incorporated into the growing chain, it is attached to the pyrophosphate group that is discarded (see Fig. 8.04). Thus, the DNA does not accumulate tags. Instead each extension reaction gives a brief burst of color. The fluorophore is washed away, and then the next cycle begins. The length of the genomic DNA piece that can be sequenced with this technology is huge in comparison to the next generation sequencing reads. A typical read for SMRT sequencing is 20,000 bases, whereas, the Illumina reads are only between 100 and 200 bases. SMRT sequencing is becoming a method to resequence repetitive genomes, and is currently being used to decode the final gaps of the human genome.

6. DNA Microarrays for Sequence Analysis

DNA chips were developed to allow automated side-by-side analysis of multiple DNA sequences. In practice, the simultaneous analysis of thousands of DNA

DNA chip Chip used to simultaneously detect and identify many short DNA fragments by DNA-DNA hybridization. Also known as DNA array or oligonucleotide array detector.

sequences is possible. The first chip was introduced by a company called Affymetrix in California in the early 1990s. Since then, DNA chips have been used for a variety of purposes including sequencing, detection of mutations, and gene expression analysis. DNA chips all rely on hybridization between single-stranded DNA permanently attached to the chip and DNA (or RNA) in solution. Many different DNA molecules are attached to a single chip forming an array of spots on a solid support (the chip). The DNA or RNA to be analyzed must be labeled, usually with fluorescent dyes. Hybridization at each spot is scanned and the signals are analyzed by appropriate software to generate colorful data arrays. Two major variants of the DNA chip exist. Earlier chips mostly used short oligonucleotides. However, it is also possible to attach full-length cDNA molecules. Prefabricated cDNA or oligonucleotides may be attached to the chip. Alternatively, oligonucleotides may be synthesized directly onto the surface of the chip by a modification of the phosphoramidite method described in Chapter 5, Manipulation of Nuclecic Acids. Modern arrays may have 100,000 or more oligonucleotides mounted on a single chip.

DNA arrays are used for a variety of purposes, including sequencing. Large numbers of probes are bound to the chip in a grid-like pattern, and then, hybridization with labeled target DNA occurs on the chip surface.

Key Concepts

- Chain termination sequencing requires that the original template DNA be copied such that the copies vary in size by one nucleotide.
- Dideoxynucleotides are missing the 3′ −OH group, and therefore, DNA polymerase is unable to add another nucleotide onto the strand of DNA, even if there is a template sequence.
- Chain termination sequencing involves mixing a single-stranded template DNA, deoxynucleotides, fluorescently-labeled dideoxynucleotides, and DNA polymerase. The reaction creates millions of copies of the template DNA that terminate at various positions due to the incorporation of the fluorescently labeled dideoxynucleotides.
- The different sized fragments generated during chain termination (Sanger sequencing) are separated by length using capillary electrophoresis. At one end of the capillary tube, a fluorescent light excites the final fluorescently-labeled dideoxynucleotide, and a detector records the wavelength of emitted light that is then converted into sequence based on the different fluorophores.
- During phosphodiester bond formation, a deoxynucleoside triphosphate is added to the 3′ −OH group of the DNA chain. The reaction releases a H + or hydrogen ion (which is recorded by Ion Torrent sequencing) and a pyrophosphate (which is recorded by SMRT sequencing).
- DNA polymerases have been genetically modified to have higher processivity, more fidelity, and less exonuclease activity so as to provide better, longer, and more accurate fragment subsets of the template.
- NGS assembles the sequence from a large number of individual genomic DNA fragments in parallel using miniaturized chips or beads and picoliter amounts of reagents. The template DNA in NGS is usually entire genomic DNA that has been sheared into small fragments using either tagmentation or ultrasonic disruption.
- Adapters are added to each genomic DNA fragment so the fragments have known sequence on the ends. Adapters have binding sites for PCR primers, sequencing primers, barcodes or index sequences, and a sequence complementary to the solid surface for separating or partitioning in NGS.
- Illumina sequencing partitions the individual genomic DNA fragments on a flow cell, and then creates a cluster of identical DNAs for sequencing using bridge amplification.

- During sequencing by synthesis in Illumina sequencers, DNA polymerase adds nucleotides one at a time because reversible fluorescent dye terminators block the addition of another nucleotide until they are removed.
- Ion Torrent sequencing partitions individual genomic DNA by first attaching the fragment to a bead and then uses emulsion PCR to create the thousands of copies.
- During sequencing by synthesis in Ion Torrent systems, each individual base is added one-at-a-time because it only detects the change in pH that occurs during the $H+$ release during phosphodiester bond formation.
- An individual read is the decoded sequence for each Illumina cluster or each Ion Torrent bead. These technologies have short read lengths of less than 200 bases.
- Read depth and average read depth are two important parameters for NGS analysis, and can affect the accuracy of the final consensus sequence.
- Targeted sequencing creates a subset of genomic DNA fragments for NGS. These subsets can include whole exomes (all the exons) or genes associated with a particular disease or cancer.
- Targeted sequencing libraries can be created with highly multiplexed PCR reactions or using a panel of biotinylated oligonucleotide probes to isolate the genomic DNA fragments of interest from a whole genome.
- Third-generation sequencing focuses on single DNA molecule analysis. SMRT sequencing uses ZMWs that are so small that light cannot penetrate into the well beyond the size of a single DNA polymerase and template DNA complex. As a nucleotide is added by DNA polymerase, a fluorescently-labeled pyrophosphate flashes a wave length of light that identifies the nucleotide as a G, A, T, or C. A computer records these flashes and assembles the data into sequence.
- Another third-generation sequencing method uses synthetic or biological nanopores, which are just big enough for a single strand of DNA to pass through. Since each individual nucleotide reduces the voltage across the membrane to a different extent, the actual sequence is determined by measuring the current across the membrane.

Review Questions

1. Define the term genomics.
2. What is the chain termination method for sequencing DNA and how does it work?
3. How does a dideoxynucleotide terminate a growing DNA chain? What key functional group is missing on dideoxynucleotides versus regular deoxynucleotides?
4. During the phosphodiester bond formation, what are the reactants? What are the products?
5. How are DNA fragments separated and detected after chain termination sequencing?
6. What is the advantage of capillary separation in Sanger sequencing?
7. What is the advantage of Sequenase over DNA polymerase and Klenow polymerase?
8. During library preparation for next generation sequencing, what are two methods of fragmenting the genomic DNA?
9. How does ultrasonic disruption break apart DNA?
10. How does tagmentation break apart DNA?
11. What are adapters? What types of sequences are found within the adapters?
12. What is the key part of the adapter that allows samples to be multiplexed?
13. How does chain termination sequencing and Illumina sequencing differ? How are they the same?
14. What is bridge amplification and why is it important?
15. Define read and read depth.
16. What is *de novo* sequencing? How does it compare to analyzing sequencing data with a reference genome?

17. What specific product of the phosphodiester bond formation is measured in Ion Torrent sequencing? Why do Ion Torrent sequencing platforms add each individual nucleotide separately?
18. Describe two methods of creating a targeted sequencing library for next generation sequencing.
19. What is meant by third-generation sequencing?
20. What is the basis for nanopore technology?
21. How can DNA be sequenced by nanopore detectors? What is the advantage of nanopore technology?
22. What two properties of single-molecule real-time sequencing (SMRT) are novel to this third-generation sequencing technique?

Further Reading

Corlett, R.T., 2017. A bigger toolbox: biotechnology in biodiversity conservation. Trends Biotechnol. 35 (1), 55–65.

Gilad, Y., 2009. Characterizing natural variation using next-generation sequencing technologies. Trends Genet. 25 (10), 463–471.

Hyman, D.M., et al., 2017. Implementing genome-driven oncology. Cell 168 (4), 584–599.

Kilpinen, H., Barrett, J.C., 2013. How next-generation sequencing is transforming complex disease genetics. Trends Genet. 29 (1), 23–30.

Koboldt, D.C., et al., 2013. The next-generation sequencing revolution and its impact on genomics. Cell 155 (1), 27–38.

Liu, X.S., Mardis, E.R., 2017. Applications of immunogenomics to cancer. Cell 168 (4), 600–612.

Lu, H., et al., 2016. Oxford nanopore MinION sequencing and genome assembly. Genom. Proteom. Bioinform. 14 (5), 265–279.

Mardis, E.R., 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. 24 (3), 133–141.

Pogrebniak, K.L., Curtis, C., 2018. Harnessing tumor evolution to circumvent resistance. Trends Genet. 34 (8), 639–651.

Roh, S.W., 2010. Comparing microarrays and next-generation sequencing technologies for microbial ecology research. Trends Biotechnol. 28 (6), 291–299.

van Dijk, E.L., et al., 2018. The third revolution in sequencing technology. Trends Genet. Available from: https://doi.org/10.1016/J.TIG.2018.05.008.

van Dijk, E.L., 2014. Ten years of next-generation sequencing technology. Trends Genet. 30 (9), 418–426.

Zou, Z., et al., 2017. Technologies for analysis of circulating tumour DNA: progress and promise. TrAC Trends Anal. Chem. 97, 36–49.