LIKE WHAT YOU SEE?

Request more information

Chapter 9

DNA Sequencing

Outline

Objectives

9.1 List the components and the molecular reactions that occur in chain termination sequencing.
9.2 Discuss the advantages of dye primer and dye terminator sequencing.
9.3 Derive a text DNA sequence from raw sequencing data.
9.4 Describe examples of alternative sequencing methods, such as pyrosequencing and next-generation sequencing (NGS).
9.5 Show different technical approaches to NGS and the two approaches used most in clinical applications.
9.6 Describe how NGS sequencing libraries are made.
9.7 Distinguish primer and probe-based enrichment.
9.8 Define *bioinformatics,* and describe electronic systems for the communication and application of sequence information.
9.9 Recount the events of the Human Genome Project.
9.10 Explain how variant databases were developed following completion of the Human Genome Project.

DNA sequence information (the order of nucleotides in the DNA molecule) is used in the medical laboratory for a variety of purposes, including detecting mutations, typing microorganisms, identifying human haplotypes, and designating polymorphisms. Treatment strategies including targeted therapies are now selected based on the results of these techniques.[1]

DIRECT SEQUENCING

The importance of knowing the order, or sequence, of nucleotides on the DNA chain was appreciated in the earliest days of molecular analysis. Elegant genetic experiments with microorganisms indirectly detected molecular changes at the nucleotide level using phenotypic characteristics, such as nutrient requirements.

Indirect methods of investigating nucleotide sequence differences are still in use today. Without knowing the nucleotide sequence of the targeted areas, the results from many of these methods would be difficult to interpret; in fact, some methods would not be useful at all. Direct determination of the nucleotide sequence, or DNA sequencing, is the most definitive molecular method to identify genetic lesions.

Manual Sequencing

Direct determination of the order, or sequence, of nucleotides in a DNA polymer is the most specific and direct method for identifying genetic lesions (mutations) or polymorphisms, especially when looking for changes affecting only one or two nucleotides. Two types of sequencing methods were concurrently developed in the 1970s: **Maxam–Gilbert sequencing**[2] and **Sanger sequencing.**[3]

Chemical (Maxam–Gilbert) Sequencing

The Maxam–Gilbert chemical sequencing method was developed by Allan M. Maxam and Walter Gilbert. Maxam–Gilbert sequencing required a double- or single-stranded version of the DNA region to be sequenced, with one end radioactively labeled.

For sequencing, the labeled fragment, or **template,** was aliquoted into four tubes. Each aliquot was treated with a different chemical with or without high salt (Fig. 9.1). Upon addition of a strong reducing agent, such as 10% piperidine, the single-stranded DNA would break at specific nucleotides (Table 9.1).

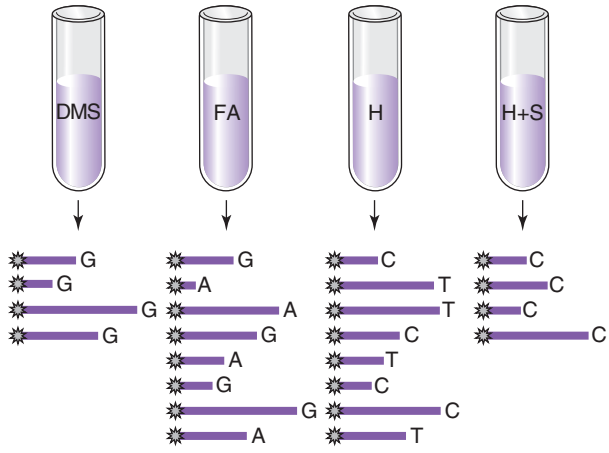

FIGURE 9.1 Chemical sequencing proceeds in four separate reactions in which the labeled DNA fragment is selectively broken at specific nucleotides. *DMS,* Dimethylsulphate; *FA,* formic acid; *H,* hydrazine; *H + S,* hydrazine + salt.

TABLE 9.1 Specific Base Reactions in Maxam–Gilbert Sequencing

Chain Breaks At:	Base Modifier	Reaction	Time (min at 25°C)
G	Dimethylsulphate	Methylates G	4
G + A	Formic acid	Protonates purines	5
T + C	Hydrazine	Splits pyrimidine rings	8
C	Hydrazine + salt	Splits only C rings	8

Advanced Concepts

To make a radioactive sequence template, (^{32}P)-ATP is added to the 5′ end of a DNA fragment, using polynucleotide kinase, or the 3′ end, using

terminal transferase plus alkaline hydrolysis to remove excess adenylic acid residues. Double-stranded fragments labeled only at one end are also produced by using restriction enzymes to cleave a labeled fragment asymmetrically, and the cleaved products are isolated by gel electrophoresis. Alternatively, denatured single strands are labeled separately, or a "sticky" end of a restriction site is filled in, incorporating radioactive nucleotides with DNA polymerase.

After the reactions, the fragments were separated by size on a denaturing polyacrylamide gel. An example of Maxam–Gilbert sequencing results is shown in Figure 9.2. The sequence was inferred from the bands on the film. The lane in which that band appeared identified the nucleotide. Bands in the purine (G + A) or pyrimidine (C + T) lane were called based on whether they were also present in the G- or C-only lanes. In that way, the sequence was read from the bottom (5′ end of the DNA molecule) to the top (3′ end of the molecule) of the gel.

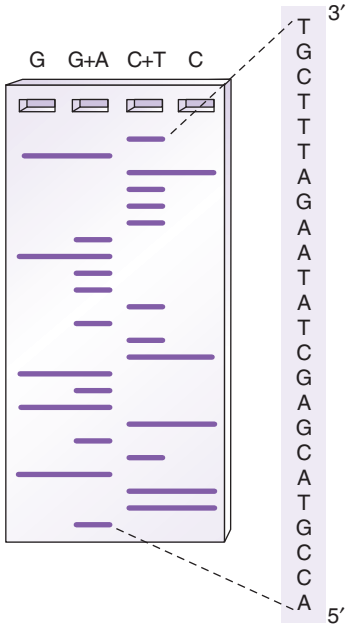

FIGURE 9.2 Products of a Maxam–Gilbert sequencing reaction. The gel is read from the bottom to the top. The size of the fragments gives the order of the nucleotides. The nucleotides are inferred from the lane in which each band appears. A or T is indicated by bands that appear in the G + A lane or C + T lane, respectively, but not in the G lane or the C lane. G is present in the G + A lane and the G lane. C is present in the C + T lane and the C lane.

Although Maxam–Gilbert sequencing was a relatively efficient way to determine short runs of sequence data, the method was not practical for high-throughput sequencing of long fragments. In addition, the hazardous chemicals hydrazine and piperidine required more elaborate precautions for use and storage. The method was therefore replaced by the **dideoxy chain termination** sequencing method for most sequencing applications.

Advanced Concepts

Polyacrylamide gels from 6% to 20% are used for sequencing. Bromophenol blue and xylene cyanol loading dyes are used to monitor the migration of the fragments. Run times range from 1 to 2 hours for short fragments (up to 50 base pairs [bp]) to 7 to 8 hours for longer fragments (more than 150 bp).

Dideoxy Chain Termination (Sanger) Sequencing

Dideoxy chain termination (Sanger) sequencing is a modification of the DNA replication process. A short, synthetic, single-stranded DNA fragment (primer) complementary to sequences just 5′ to the region of DNA to be sequenced is used for priming dideoxy sequencing reactions (Fig. 9.3). For detection of the products of the sequencing reaction, the primer is attached covalently at

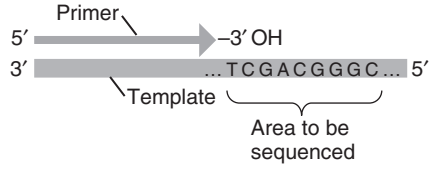

FIGURE 9.3 Manual dideoxy sequencing requires a single-stranded version of the fragment to be sequenced (template). Sequencing is primed with a short synthetic piece of DNA complementary to bases just before the region to be sequenced (primer). The sequence of the template will be determined by extension of the primer in the presence of dideoxynucleotides.

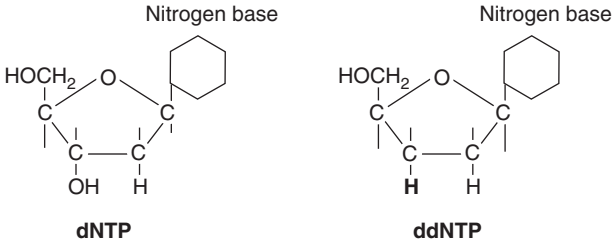

FIGURE 9.4 A dideoxynucleotide *(right)* lacks the hydroxyl group on the 3′ ribose carbon that is required for formation of a phosphodiester bond with the phosphate group of another nucleotide.

the 5′ end to a ^{32}P-labeled nucleotide or a fluorescent dye-labeled nucleotide. A previously used alternative detection strategy was to incorporate ^{32}P- or ^{35}S-labeled deoxynucleotides in the nucleotide sequencing reaction mix **(internal labeling).**

Just as in the in vivo DNA replication reaction, an in vitro DNA synthesis reaction would result in polymerization of deoxynucleotides to make full-length copies of the DNA template. For sequencing, modified **dideoxynucleotide (ddNTP)** derivatives are added to the reaction mixture. Dideoxynucleotides lack the hydroxyl group found on the 3′ ribose carbon of the deoxynucleotides (dNTPs; Fig. 9.4). DNA synthesis will stop upon incorporation of a ddNTP into the growing DNA chain (chain termination) because without the hydroxyl group at the 3′ sugar carbon, the 5′–3′ phosphodiester bond cannot be established to incorporate a subsequent nucleotide. The newly synthesized chain will terminate, therefore, with the ddNTP (Fig. 9.5).

Historical Highlights

The original dideoxy chain termination sequencing methods used in the late 1970s into the early 1980s required a single-stranded template. Templates up to a few thousand bases long were produced using M13 bacteriophage, a bacterial virus with a single-stranded DNA genome. This virus replicates by infecting *Escherichia coli,* in which the viral single-stranded circular genome is converted to a double-stranded plasmid, the **replication factor (RF).** The plasmid codes for viral gene products use the bacterial transcription and translation machinery to make new single-stranded genomes and viral proteins. To use M13 for template preparation, the RF was isolated from infected bacteria, cut with restriction enzymes, and the fragment to be sequenced was ligated into the RF. When the recombined RF was reintroduced into the host bacteria, M13 continued its life cycle producing new phages, some of which carried the inserted fragment. When the phages were spread on a lawn of host bacteria, plaques (clear spaces) of lysed bacteria formed by phage replication contained pure populations of recombinant phage. The single-stranded DNA was then isolated from the phage by picking plugs of agar from the plaques and isolating DNA from them.

Advanced Concepts

An advantage of the M13 template preparation method was that the primer that hybridizes to M13 sequences could be used to sequence any fragment ligated into the same site of the RF. Recombinant plasmids containing fragments to be sequenced include a short M13 region so that the **M13 universal primer** could still be used in some applications, even though the M13 method of template preparation is no longer practical.

For manual dideoxy sequencing, a 1:1 mixture of template and radioactively labeled primer is placed into four separate reaction tubes in sequencing buffer containing the sequencing enzyme and ingredients necessary for the polymerase activity (Fig. 9.6). Mixtures of all four dNTPs and one of the four ddNTPs are then added to each tube, with a different ddNTP in each of the four tubes.

Advanced Concepts

Polymerase chain reaction (PCR) products are currently used as sequencing templates. Residual components of the PCR reaction, especially

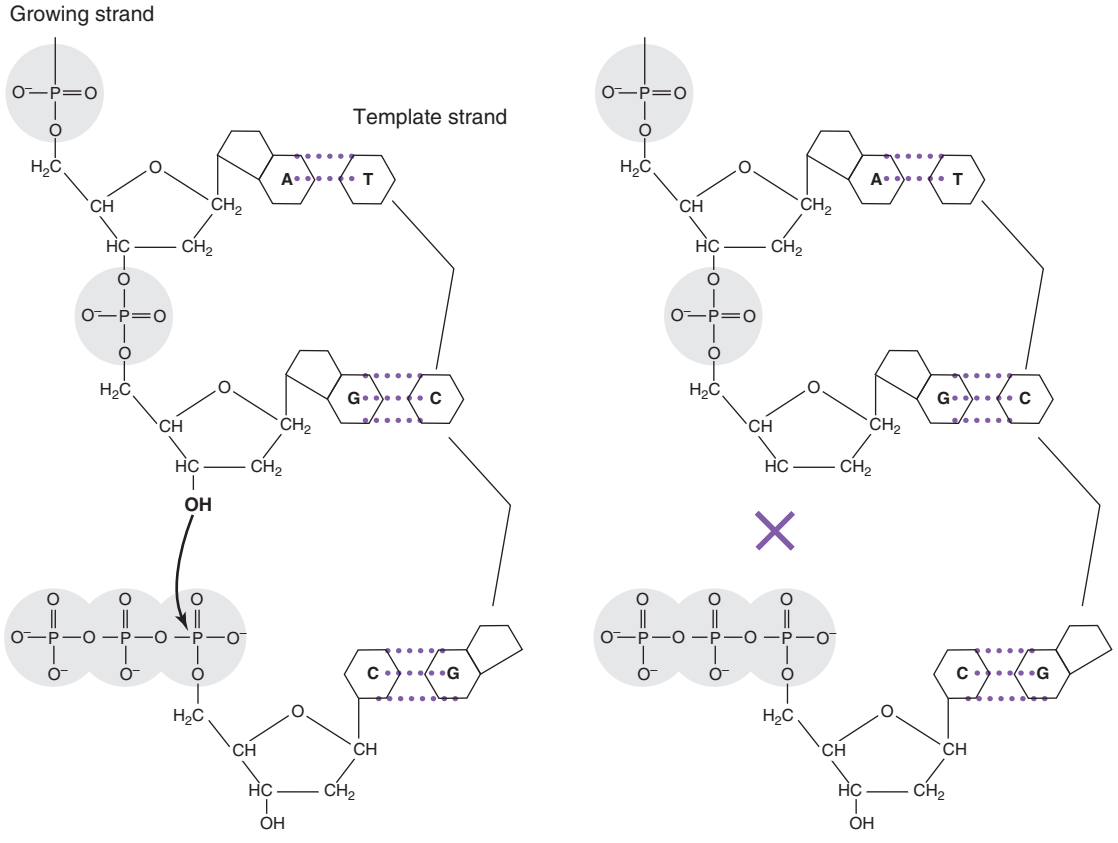

FIGURE 9.5 DNA replication *(left)* is terminated by the absence of the 3′ hydroxyl group on the dideoxyguanosine nucleotide (ddG, *right*). The resulting fragment ends in ddG.

primers and nucleotides, can interfere with the sequencing reaction and lower the quality of the sequencing ladder. PCR amplicons can be cleaned using solid-phase (column or bead) matrices, alcohol precipitation, or enzymatic digestion with alkaline phosphatase. Alternatively, amplicons can be run on an agarose gel and the bands eluted. The latter method provides not only a clean template but also confirmation of the product being sequenced. It is especially useful when the PCR reactions are not completely free of mis-primed bands or primer dimers.

The ratio of ddNTPs/dNTPs is critical for the generation of a readable sequence. If the concentration of ddNTPs is too high, polymerization will terminate too frequently early along the template. If the ddNTP concentration is too low, infrequent or no termination will occur. In the beginning days of sequencing, optimal ddNTP/dNTP ratios were determined empirically (by experimenting with different ratios). Sequencing reagent mixes have preoptimized nucleotide mixes.

With the addition of DNA polymerase enzyme to the four tubes, the reaction begins. After about 20 minutes, the reactions are terminated by addition of a stop buffer, which consists of 20 mM EDTA to chelate cations and stop enzyme activity, formamide to denature the products of the synthesis reaction, and gel loading dyes (bromophenol blue and/or xylene cyanol). All four reactions are carried out for equal times to provide consistent band intensities in all four lanes of the sequencing gel sequence.

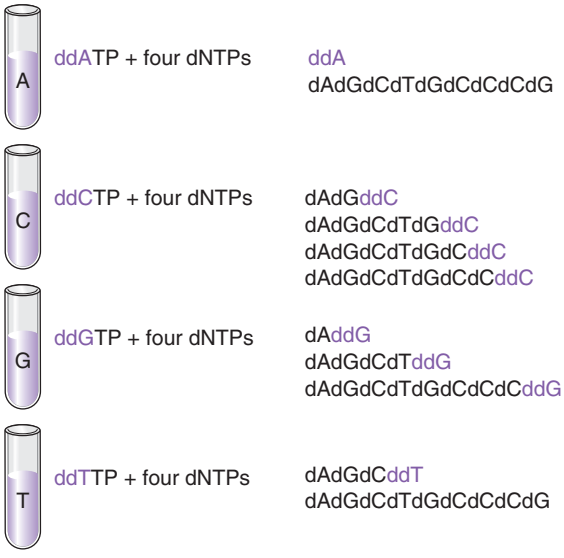

FIGURE 9.6 Components required for DNA synthesis (template, primer, enzyme, buffers, dNTPs) are mixed with a different ddNTP in each of four tubes (left). With the proper ratio of ddNTPs/dNTPs, the newly synthesized strands of DNA will terminate at each opportunity to incorporate a ddNTP. The resulting synthesis products are a series of fragments ending in either A (ddATP), C (ddCTP), G (ddGTP), or T (ddTTP). This collection of fragments is the sequencing ladder.

Advanced Concepts

Manganese (Mn^{++}) added to the sequencing reaction promotes equal incorporation of all dNTPs by the polymerase enzyme. Equal incorporation of the dNTPs makes for uniform band intensities on the sequencing gel, which eases interpretation of the sequence. Manganese increases the relative incorporation of ddNTPs as well, which will enhance the reading of the first part of the sequence by increasing intensity of the smaller bands on the gel. Modified nucleotides, deaza-dGTP and deoxyinosine triphosphate (dITP), are also added to sequencing reaction mixes to deter secondary structure in the synthesized fragments. Such additives as Mn^{++}, deaza-dGTP, and dITP are supplied in commercial sequencing buffers.

The sets of synthesized fragments are then loaded onto a denaturing polyacrylamide gel. The products of each of the four sequencing reactions are loaded into adjacent lanes, labeled A, C, G, or T, corresponding to the ddNTP in the four reaction tubes. Once the gel is dried and exposed to x-ray film, the fragment patterns are visualized by the signal on the ^{32}P-labeled primer (or incorporated deoxynucleotide). All fragments from a given tube will end in the same ddNTP; for example, all the fragments synthesized in the ddCTP tube end in C.

The four-lane gel electrophoresis pattern of the products of the four sequencing reactions is called a **sequencing ladder** (Fig. 9.7). The ladder is read to deduce the DNA sequence. From the bottom of the gel, the smallest (fastest-migrating) fragment is the one in which synthesis terminated closest to the primer. The identity of the ddNTP at a particular position is determined by the lane in which the band appears. If the smallest band is in the ddATP lane, then the first base is an A. The next larger fragment is the one that was terminated at the next position on the template. The lane that has the next larger band identifies the next nucleotide in the sequence. The sequence is thus read from the bottom (smallest, 5′-most) to the top (largest, 3′-most) fragments across or within lanes to determine the identity and order of nucleotides in the sequence.

Depending on the reagents and gel used, the number of bases per sequence read averages 300 to 400. Advances in enzyme and gel technology have increased this capability to over 500 bases per read. Sequencing reads are lengthened by loading the same ladders in intervals of 2 to 6 hours so that the larger bands are resolved with longer (e.g., 8-hour) migrations, whereas smaller bands will be resolved simultaneously in a 1- to 2-hour migration that was loaded 6 to 7 hours later.

As Sanger sequencing came into routine use, technology was improved significantly from these first manual sequencing procedures. Recombinant polymerase enzymes with in vitro removal of the exonuclease activity were faster and more processive (i.e., they stayed with the template longer, producing longer sequencing ladders). In addition, these engineered enzymes more efficiently incorporated ddNTPs and nucleotide analogs such as dITP or deaza-dGTP, which were used to deter secondary structure (internal folding and hybridization) in the template and sequencing products. Furthermore, sequencing was performed with double-stranded

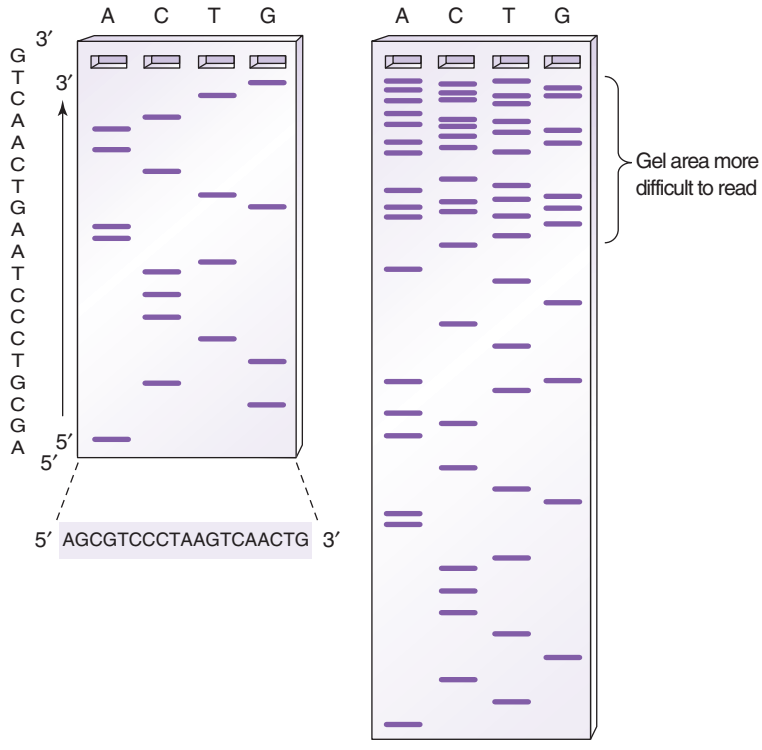

FIGURE 9.7 A sequencing ladder is read from the bottom of the gel to the top. The smallest (fastest-migrating) fragment represents the first nucleotide attached to the primer by the polymerase. Since that fragment is in lane A, from the reaction that contained ddATP *(left),* the sequence read begins with A. The next largest fragment is in lane T. The sequence, then, reads AT. The next largest fragment is in lane C, making the sequence ATC, and so forth up the gel. Larger bands on a sequencing gel can sometimes be compressed, limiting the length of sequence that can be read on a single gel run *(right).*

templates, eliminating the requirement for the preparation of single-stranded versions of the DNA to be sequenced.

Using heat-stable enzymes, the sequencing reaction took place in a thermal cycler **(cycle sequencing).** With cycle sequencing, timed manual starting and stopping of the sequencing reactions were not necessary. The labor savings in this regard increase the number of reactions that could be performed simultaneously; for example, a single operator could run 96 sequencing reactions (i.e., sequence 24 fragments) in a 96-well plate. Finally, improvements in fluorescent dye technology have led to the automation and throughput of the sequencing process and, more importantly, sequence determination.

Automated Fluorescent Sequencing

The chemistry for automated sequencing is the same as that described for manual sequencing, using double-stranded templates and cycle sequencing. Because cycle sequencing (unlike manual sequencing) does not require the sequential addition of reagents to start and stop the reaction, cycle sequencing was more easily adaptable to early high-throughput applications and automation. Universal systems combined automation of DNA isolation of the template and setup of the sequencing reactions.

Electrophoresis and reading of the sequencing ladder were also automated. A requirement for automated reading of the DNA sequence ladder is the use of fluorescent dyes instead of radioactive nucleotides to label the primers or sequencing fragments.

Advanced Concepts

Fluorescent dyes used for automated sequencing include fluorescein, rhodamine, and Bodipy (4,4-difluoro-4-bora-3a,4a-diaza-*s*-indacene) dye derivatives that are recognized by commercial detection systems.[4] Automated sequence readers excite the dyes with a laser and detect the emitted fluorescence at specific wavelengths. More advanced methods have been proposed to enhance the distinction between the dyes for more accurate determination of the sequence.[5]

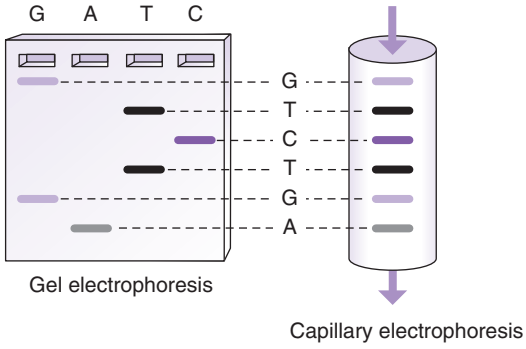

FIGURE 9.8 Instead of four gel lanes *(left)* fluorescent fragments can be run in a single gel lane or in a capillary *(right)*. Note that the sequence of nucleotides, AGTCTG, read by lane in the slab gel is read by color in the capillary.

Fluorescent dyes used for sequencing have distinct "colors," or peak wavelengths of fluorescence emission, that can be distinguished by automated sequencers. The advantage of having four distinct colors is that all four of the reaction mixes can be read in the same lane of a gel or on a capillary. Fluorescent dye color rather than lane placement will assign the fragments as ending in A, T, G, or C in the sequencing ladder (Fig. 9.8).

Approaches to Automated Sanger Sequencing

There are two approaches to automated fluorescent sequencing: **dye primer** and **dye terminator** sequencing (Fig. 9.9). The goal of both approaches is to label the fragments synthesized during the sequencing reaction according to their terminal ddNTP. Thus, fragments ending in ddATP, read as A in the sequence, will be labeled with a "green" dye; fragments ending in ddCTP, read as C in the sequence, will be labeled with a "blue" dye; fragments ending in ddGTP, read as G in the sequence, will be labeled with a "black" or "yellow" dye; and fragments ending in ddTTP, read as T in the sequence, will be labeled with a "red" dye. This facilitates reading of the sequence by the automated sequence.

In dye primer sequencing, the four different fluorescent dyes are attached to four separate aliquots of the primer. The dye molecules are attached covalently to the 5′ end of the primer during chemical synthesis, resulting in four versions of the same primer with different dye labels. The primer labeled with each "color" is added to four separate reaction tubes, one each with ddATP, ddCTP, ddGTP, or ddTTP, as shown in Figure 9.9. After addition of the remaining components of the sequencing reaction (see the previous section on manual sequencing) and of a heat-stable polymerase, the reaction is subjected to cycle sequencing in a thermal cycler. The products of the sequencing reaction are then labeled at the 5′ end, using the dye color associated with the ddNTP at the end of the fragment.

Dye terminator sequencing is performed with one of the four fluorescent dyes covalently attached to each of the ddNTPs instead of to the primer. The primer is unlabeled. A major advantage of this approach is that all four sequencing reactions are performed in the same tube (or well of a plate) instead of in four separate tubes. After addition of the rest of the reaction components and cycle sequencing, the product fragments are labeled at the 3′ end. As with dye primer sequencing, the "color" of the dye corresponds to the ddNTP that terminated the strand. Dye terminator sequencing has become the Sanger sequencing method of choice. The option of one reaction for all four nucleotides lowers the cost and labor of routine sequencing performed in many laboratories.

The Sequencing Ladder

After a sequencing reaction using fluorescent dye terminators, excess dye terminators are removed with columns or beads or by ethanol precipitation. Spin columns or bead systems bind the sequencing fragments to allow removal of residual sequencing components by rinsing with buffers. Alternately, the dye terminators are bound onto specially formulated magnetic beads, and the sequencing ladder is recovered from the supernatant as the beads are held by a magnet applied to the outside of the tube or plate.

The fragments of the sequencing ladder are completely denatured before running on a gel or capillary. Denaturing conditions (50°C to 60°C, formamide, urea denaturing gel) are maintained so that the fragments are resolved strictly according to size. Secondary structure affects migration speed and lowers the quality of the sequence. Before loading in a gel or capillary instrument, sequence ladders are cleaned, as described previously, to remove residual dye terminators, precipitated, and resuspended in formamide. The ladders are heated to 95°C to 98°C for 2 to 5 minutes and placed on ice just before loading.

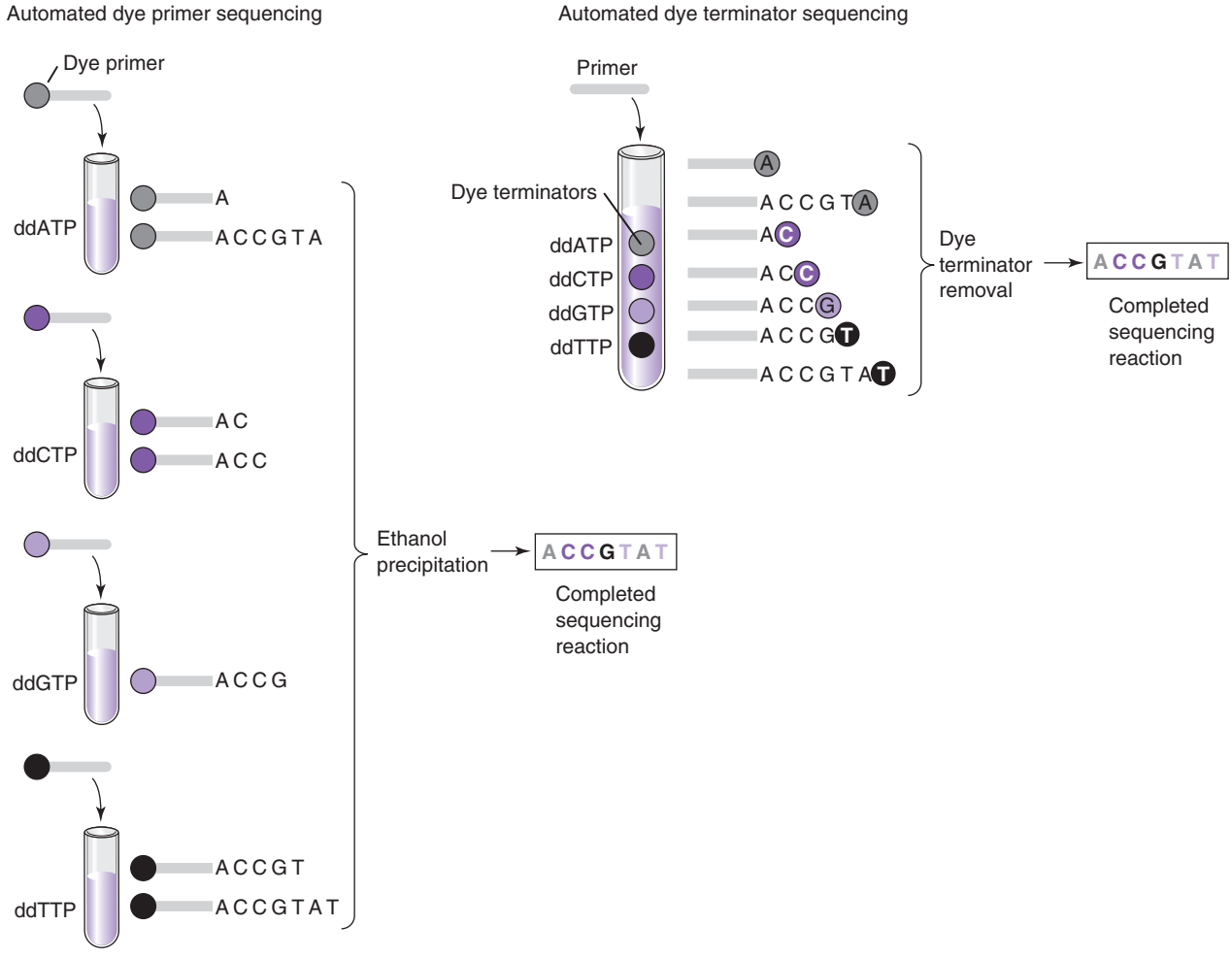

FIGURE 9.9 Fluorescent sequencing chemistries. Dye primer sequencing uses labeled primers *(left)*. The reactions take place in separate tubes and the products of all four reactions are resolved together in one lane of a gel or in a capillary. Using dye terminators *(right),* only one reaction tube is necessary because the fragments can be distinguished directly by the dideoxynucleotides on their 3′ ends.

Electrophoresis

The four sets of sequencing products in each reaction are loaded onto a single gel lane or capillary. The fluorescent dye colors, rather than lane assignment, distinguish which nucleotide is at the end of each fragment. Running all four reactions together not only increases throughput but also eliminates lane-to-lane migration variations that affect the accurate reading of the sequence. The migrating fragments pass a laser beam and a detector in the automated sequencer. The laser beam excites the dye attached to each fragment, causing the dye to emit fluorescence that is captured by the detector. The detector converts the fluorescence to an electrical signal that is imaged by computer software as a flash or peak of color.

Advanced Concepts

DNA sequences with high GC content are sometimes difficult to read due to intrastrand hybridization in the template DNA. Reagent preparations

that include 7-deaza-dGTP (7-Deaza-2′-deoxyguanosine-5′-triphosphate) or dITP instead of standard dGTP improve the resolution of bands (peaks) in regions that exhibit GC band compressions, or bunching of peaks close together so that they are not resolved, followed by several peaks running farther apart.

Fluorescent detection equipment yields results as an **electropherogram,** rather than a gel pattern. Just as the gel sequence is read from the smallest (fastest-migrating) fragments to the largest, the sequencing software reads, or "calls," the bases from the smallest (fastest-migrating) fragments that first pass the detector to the largest based on the dye emission wavelength; that is, the software calls the base by the "color" of the fluorescence of the fragment as it passes the detector. The electropherogram is a series of peaks of the four fluorescent dyes as the bands of the sequencing ladder migrate by the detector. The software assigns one of four colors—red, black, blue, or green—associated with each of the fluorescent dyes and a text letter to the peaks for ease of interpretation.

As with manual sequencing, the ratio of ddNTPs/dNTPs is key to the length of the sequence read (how much of the template sequence can be determined). Too many ddNTPs will result in a short sequence read. Too low a concentration of ddNTPs will result in loss of sequence data close to the primer but give a longer read because the sequencing enzyme will polymerize further down the template before it incorporates a ddNTP into the growing chain. The quality of the sequence (height and separation of the peaks) improves away from the primer and begins to decline at the end. At least 400 to 500 bases can be easily read with most sequencing chemistries.

Sequence Interpretation

Base calling is the process of identification of bases in a sequence by sequencing software. It is analogous to the inspection of gel bands for quality, clarity, and separation. Interpretation of sequencing data from a dye terminator reaction depends on the quality of the electropherogram, which, in turn, depends on the quality of the template, the efficiency of the sequencing reaction, and the cleanliness of the sequencing ladder. Failure to clean the sequencing ladder properly results in bright flashes of fluorescence **(dye blobs)** that obliterate parts of the sequence read (Fig. 9.10). Poor starting material results in a poor-quality sequence that cannot be read accurately (Fig. 9.11). Clear, clean sequencing ladders are read accurately by the software, and a text sequence is generated. Sequencing software also shows the certainty of each base call in the sequence. When the base call is not clear, the letter N will replace A, C, T, or G. Less-than-optimal sequences are not accurately readable by software but may be readable by an experienced operator.

Software programs can compare two sequences or test sequences with reference sequences to identify mutations or polymorphisms. Regardless of whether a sequence variant (change from a reference sequence) is found, it is important to sequence both complementary strands of DNA to confirm sequence data. This is especially critical for confirmation of mutations or polymorphisms in a sequence (Fig. 9.12). Alterations affecting a single base pair may be subtle on an electropherogram, especially if the alteration is in the heterozygous form, or mixed with the normal reference sequence. Ideally, a genetically heterozygous mutation appears as two peaks of equal height but different colors directly on top of one another, that is, at the same position in the electropherogram. The overlapping peaks should be about half the height of single base peaks. Heterozygous deletions or insertions (e.g., the *BRCA* frameshift mutations) affect all positions of the sequence downstream of the mutation (Fig. 9.13). Somatic mutations in clinical specimens are sometimes difficult to detect because they may be diluted by normal sequences that mask the somatic change.

Several software programs have been written to interpret and apply sequence data from capillary electrophoresis. Software that collects the raw data from the instrument is supplied with the electrophoresis instruments. Software that interprets, compares, or otherwise manipulates sequence data is sometimes supplied with a purchased instrument or available online. A representative sample of these applications is shown in Table 9.2. Further sequence interpretation with regard to disease association and pathogenic significance requires the use of sequence databases and clinical trial information. This information is available from public websites and institutional "data commons" collections.

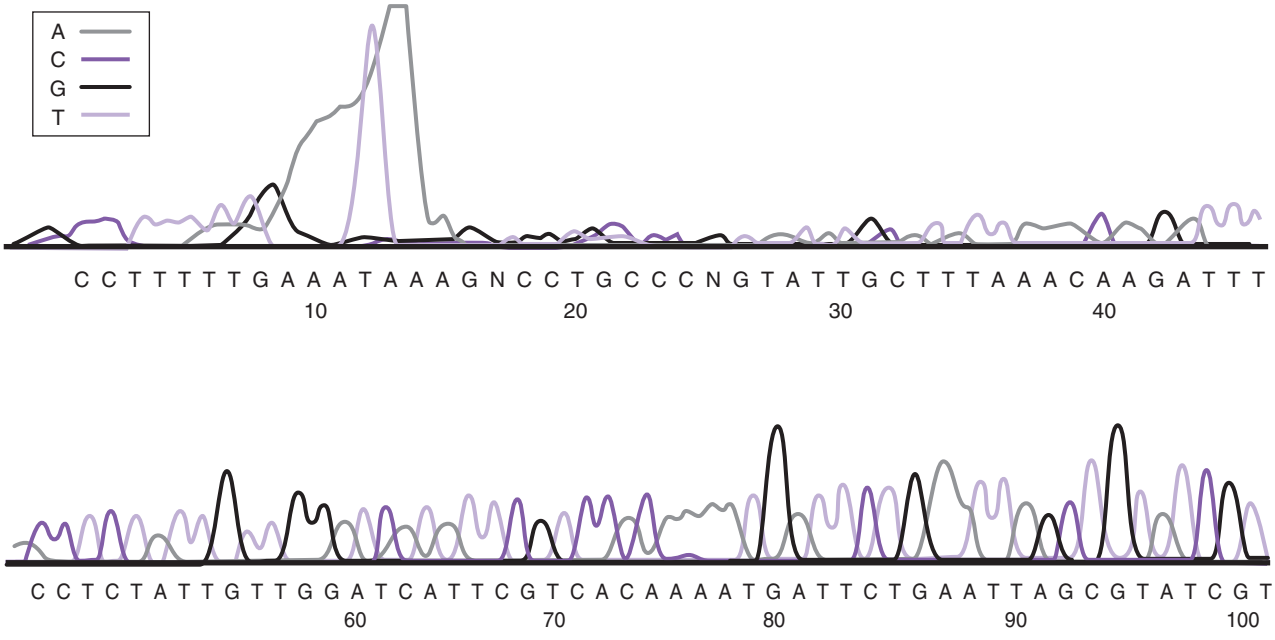

FIGURE 9.10 Electropherogram showing a dye blob at the beginning of a sequence (nucleotide positions 9 to 15). The sequence read around this area is not accurate. See Color Plate 5.

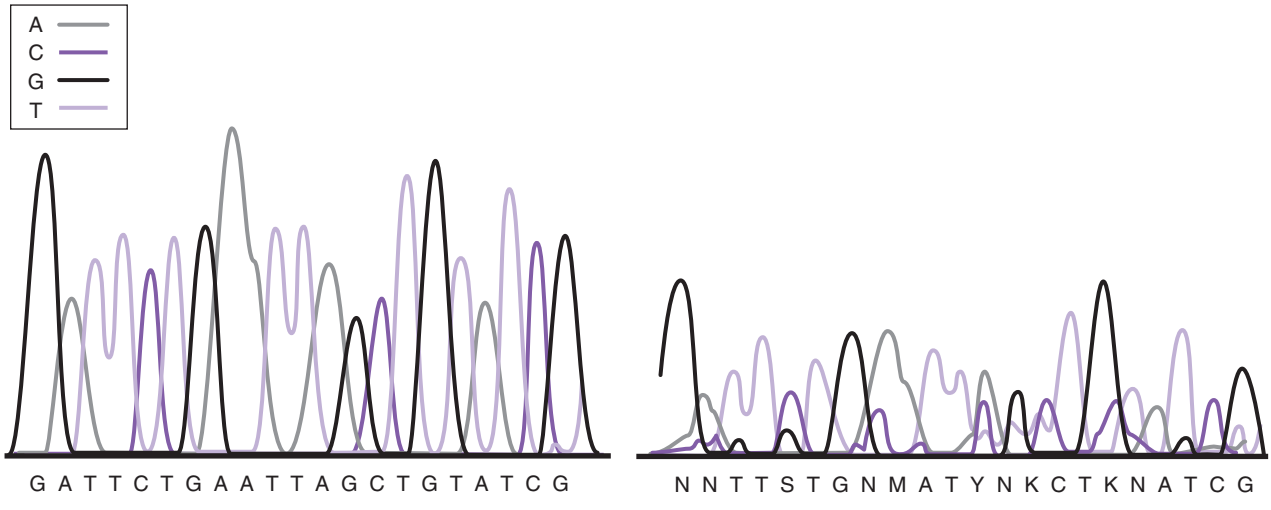

FIGURE 9.11 Examples of good sequence quality *(left)* and poor sequence quality *(right)*. Note the clean baseline on the good sequence; that is, only one color peak is present at each nucleotide position. Automatic sequence reading software will not accurately call a poor sequence. Compare the text sequences below the two scans. See Color Plate 6.

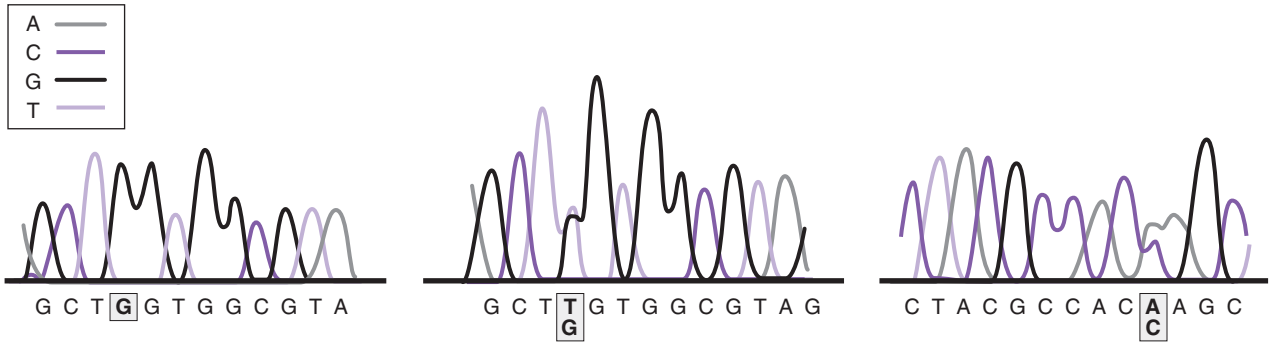

FIGURE 9.12 Sequencing of a heterozygous G to T mutation in exon 12 of the *KRAS* gene. The normal codon sequence is GGT *(left)*. The heterozygous mutation (GT, *center*) is confirmed in the reverse sequence (CA, *right*). See Color Plate 7.

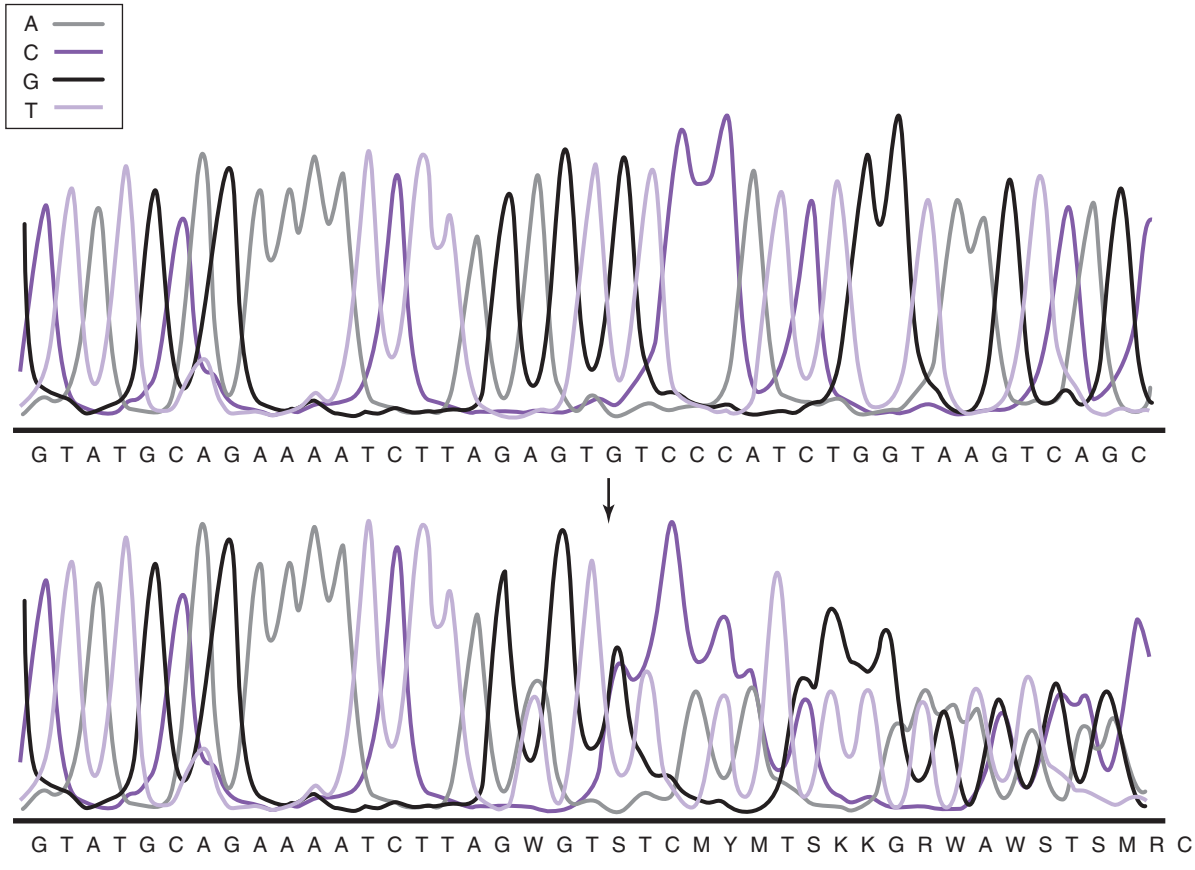

FIGURE 9.13 The 187 delAG mutation in the *BRCA*1 gene detected by Sanger sequencing. This heterozygous dinucleotide deletion is evident in the lower panel where, at the site of the mutation, two sequences are overlaid: the normal sequence and the normal sequence minus two bases. See Color Plate 8.

TABLE 9.2 Examples of Software Programs Used to Analyze and Apply Sequence Data

Software	Name	Application
BLAST	Basic Local Alignment Search Tool	Compares an input sequence with all sequences in a selected database
GRAIL	Gene Recognition and Assembly Internet Link	Finds gene-coding regions in DNA sequences
FASTA FASTQ	FAST-All derived from FAST-P (protein) and FAST-N (nucleotide) search algorithms Biological data with quality score	Rapidly aligns pairs of sequences by sequence patterns rather than individual nucleotides
Phred	Phred	Reads bases from original trace data and recalls the bases, assigning quality values to each base
Polyphred	Polyphred	Identifies single-nucleotide polymorphisms (SNPs) among the traces and assigns a rank indicating how well the trace at a site matches the expected pattern for an SNP
Phrap	Phragment Assembly Program	Uses user-supplied and internally computed data quality information to improve accuracy of assembly in the presence of repeats
TIGR Assembler	The Institute for Genomic Research	Developed by TIGR as an assembly tool to build a consensus sequence from smaller-sequence fragments
Factura	Factura	Identifies sequence features such as flanking vector sequences, restriction sites, and ambiguities
SeqScape	SeqScape	Provides mutation and SNP detection and analysis, pathogen subtyping, allele identification, and sequence confirmation
Assign	Assign	Identifies alleles for haplotyping
Matchmaker	Matchmaker	Identifies alleles for haplotyping

PYROSEQUENCING

Chain termination sequencing became the most widely used method to determine DNA sequence. Other methods were developed that yielded the same information but with less throughput capacity than the chain termination method. **Pyrosequencing** is an example of a method designed to determine a DNA sequence without having to make a sequencing ladder.[6,7] This procedure relies on the generation of light (luminescence) when nucleotides are added to a growing strand of DNA (Fig. 9.14). With this system, there are no gels, fluorescent dyes, or ddNTPs.

The pyrosequencing reaction mix consists of a single-stranded DNA template, sequencing primer, sulfurylase, and luciferase, plus the two substrates adenosine 5′ phosphosulfate (APS) and luciferin. One of the four dNTPs is added in a predetermined order to the reaction. If the nucleotide is complementary to the base in the template strand next to the 3′ end of the primer, DNA polymerase extends the primer. Pyrophosphate (PPi) is released with the formation of the phosphodiester bond between the dNTP and the primer. The PPi is converted to ATP by sulfurylase that is used to generate a luminescent signal by luciferase-catalyzed conversion of luciferin to oxyluciferin.

The process is repeated with each of the four nucleotides again added sequentially to the reaction. The generation of a signal indicates which nucleotide is the next correct base in the sequence. The results from a

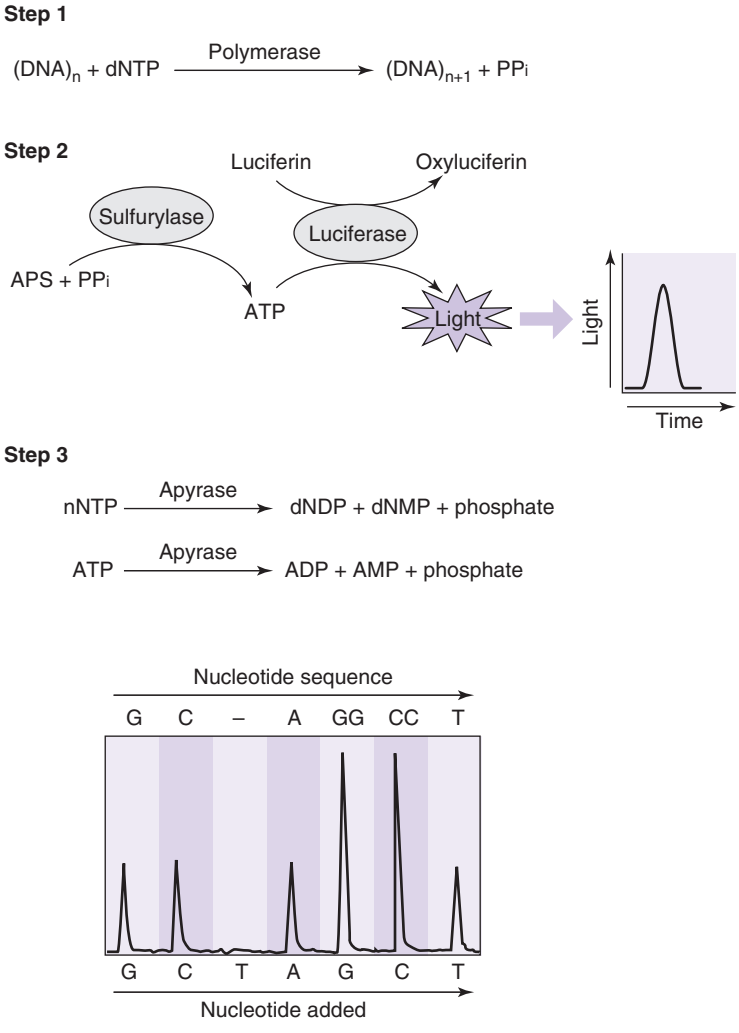

FIGURE 9.14 Pyrosequencing is the analysis of pyrophosphate (PPi) released when a nucleotide base (dNTP) is incorporated into DNA *(top left)*. The released PPi is a cofactor for ATP generation from adenosine 5′ phosphosulfate (APS). Luciferase plus ATP converts luciferin to oxyluciferin with the production of light, which is detected by a luminometer. The system is regenerated with apyrase that degrades residual free dNTP and dATP (Step 3). As nucleotides are added to the system one at a time, the sequence is determined by which of the four nucleotides generates a light signal.

pyrosequencing reaction, a **pyrogram,** consist of peaks of luminescence associated with the addition of the complementary nucleotide (Fig. 9.14). If a sequence contains a repeated nucleotide, for instance, GCAGGCCT, the results would be dG peak, dC peak, dA peak, dG peak (double height), dC peak (double height), dT peak. The nucleotide sequence is called based on the order of nucleotide bases introduced to the sequencing reaction and the peak heights.

Pyrosequencing is most useful for short- to moderate-sequence analysis. It is therefore used mostly for detection of previously known mutation or single-nucleotide polymorphism (SNP) and typing (re-sequencing) rather than for generating new sequences. It has been used for applications in mutation detection,[8] infectious disease typing,[9,10] and DNA methylation analysis.[11]

BISULFITE DNA SEQUENCING

Bisulfite DNA sequencing, or methylation-specific sequencing, is chain termination sequencing designed to detect methylated cytosine nucleotides.[12] Methylation of cytosine residues to 5-methylcytosines in DNA is an

important part of the regulation of gene expression and chromatin structure, affecting cell differentiation and diseases, including several types of cancer.

For bisulfite sequencing, 2 to 4 μg of genomic DNA is cut with restriction enzymes to facilitate denaturation. The enzymes should not cut within the region to be sequenced. The restriction digestion products are resolved on an agarose gel, and the fragments of the size of interest are purified from the gel. DNA from fixed tissue may be used directly without restriction digestion. The DNA is denatured with heat (97°C for 5 minutes) and exposed to bisulfite solution (sodium bisulfite, NaOH, and hydroquinone) for 16 to 20 hours. Buffer systems that protect DNA from bisulfite damage may be used to increase the yield of converted DNA. Overexposure to bisulfite can result in strand cleavage and loss of important regions of the DNA template. During the incubation with bisulfite, the cytosines in the reaction are deaminated, converting them to uracils, whereas the 5-methylcytosines are unchanged.

Advanced Concepts

Pyrosequencing requires a single-stranded sequencing template. Methods using streptavidin-conjugated beads have been devised to easily prepare the template. First the region of DNA to be sequenced is PCR-amplified with one of the PCR primers covalently attached to a biotin molecule. The double-stranded amplicons are then immobilized onto the beads and denatured with NaOH. After several washings to remove the non-biotinylated complementary strand (and all other reaction components), the sequencing primer is added and annealed to the pure single-stranded DNA template.

After the reaction, the treated DNA is cleaned, precipitated (or purified by adhering and washing on columns or beads), and resuspended for use as a template for PCR amplification. The primers used for amplification are altered to accommodate C to U changes in the primer-binding sites caused by the bisulfite treatment. For pyrosequencing, one primer is biotinylated for isolation of the single-stranded template.

The PCR amplicons are then sequenced by Sanger sequencing or pyrosequencing. Methylation is detected by comparing the treated sequence with the original sequence (before conversion) and noting where in the treated sequence cytosines are not changed to thymine (uracil); that is, the converted sequence will be altered relative to the reference sequence at the unmethylated C residues.

In Sanger sequencing, unmethylated cytosines will appear as red (thymine) instead of blue (cytosine) peaks on the electropherogram. In pyrosequencing, the relative light intensity of consecutive T and C additions to the reaction mix provide a quantitative degree of methylation. An example of pyrosequencing of bisulfite converted DNA is shown in Figure 9.15, where the color or height of the cytosine peaks relative to the thymine (uracil) peaks indicates the degree of methylation.

Detection methods other than sequencing have also been devised to detect DNA methylation, such as using methylation-sensitive restriction enzymes or enzymes with recognition sites generated or destroyed by the C to U changes. Other methods use PCR primers that will bind only to the converted or nonconverted sequences so that the presence or absence of PCR product indicates the methylation status. These methods, however, are not always applicable to the detection of methylation in unexplored sequences. As the role of methylation and epigenetics in human disease is increasingly recognized, bisulfite sequencing has become a popular method in the research laboratory. Clinical tests have been developed using this strategy as well.[13,14]

RNA SEQUENCING

The sequences of RNA transcripts are, for the most part, complementary to their DNA templates. Post-transcriptional processing of RNA, however, changes the RNA sequence relative to its encoding DNA. Alternative splicing and RNA editing may further modify the RNA sequence. Early methods to sequence RNA made use of ribonucleases to cut end-labeled RNA at specific nucleotides. Another approach was to infer mRNA sequence from amino acid sequence. The RNA transcript sequence can be determined from the sequencing of its complementary DNA; however, sequencing error may occur, mostly from the cDNA synthesis step.[15,16]

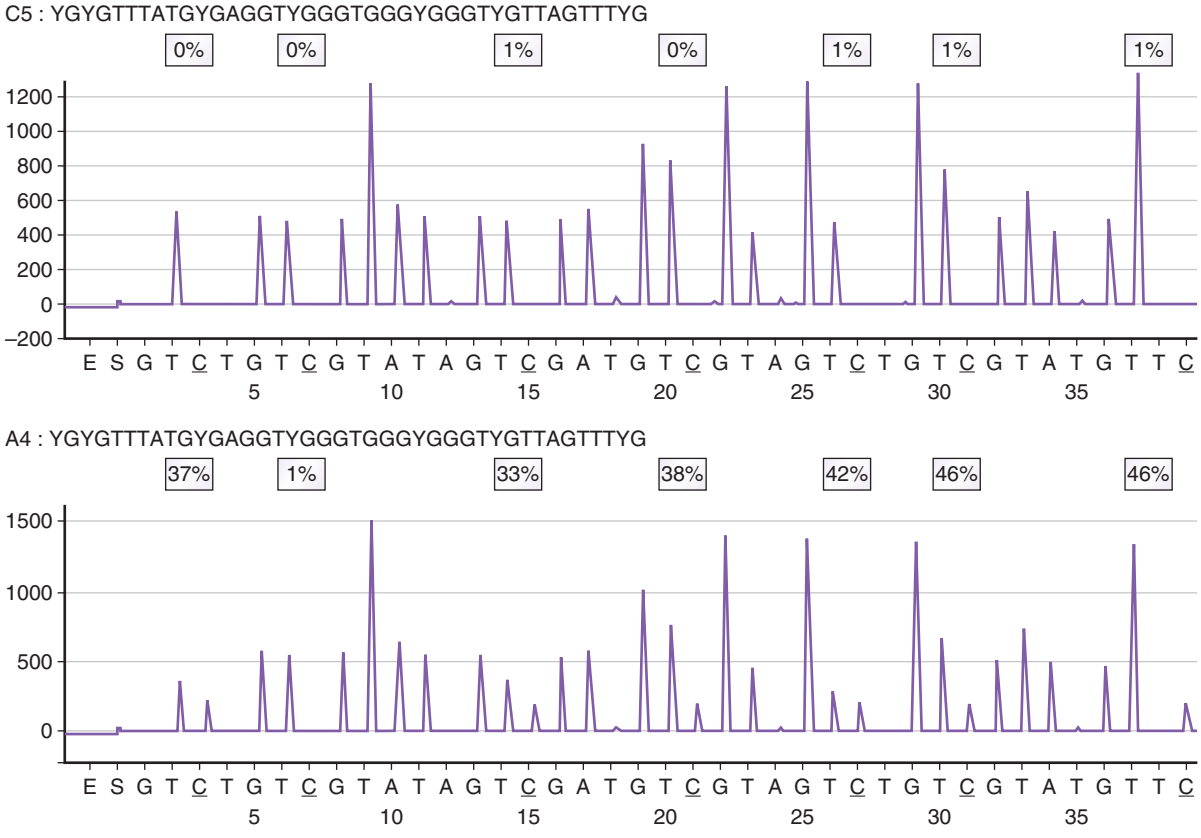

FIGURE 9.15 DNA methylation at cytosine residues detected by pyrosequencing of bisulfite-treated DNA. Exposure of a sequence to bisulfite will result in the conversion of unmethylated cytosines to uracils (T in the sequence). The pyrosequencing method will report the percent methylation that is the relative number of C to T nucleotides at each potentially methylated C position *(shaded)*. The C residues in the top panel are not methylated. All but one of the C residues in the bottom panel are methylated.

Direct sequencing of RNA has been proposed based on single-molecule sequencing technology and virtual terminator nucleotides.[17,18] In this method, mRNA is captured by immobilized polydT oligomers (Fig. 9.16). For those RNA species without polyA tails, an initial treatment with polyA polymerase is performed to add a 3′ A-tail. The 3′ ends of the captured RNA are chemically blocked to prevent extension in the sequencing step. Four reversibly dye-labeled nucleotides are then sequentially added. An image is taken, the extension inhibitors are cleaved, and alternating C, T, A, or G nucleotides are added, with imaging, cleavage, and rinsing between each nucleotide addition. After repeating this process many times (e.g., 120 cycles) the collected images are aligned and used to build the sequence from each poly(dT) anchor.

NEXT-GENERATION SEQUENCING

Data obtained from sequence analysis is best interpreted in context with population norms and variations; however, initially, few large sequence analyses were performed for multiple individuals. Furthermore, disease states involve a variety of sequence variants that can be important for diagnosis, prognosis, and treatment strategy. Although array studies were applied to this type of analysis, even the densest oligo array did not provide genomic-scale sequence data with single-base-pair resolution. Next-generation sequencing (NGS), also called massive parallel sequencing, was designed to sequence large numbers of templates carrying millions of bases simultaneously, in a run that takes a few hours. NGS

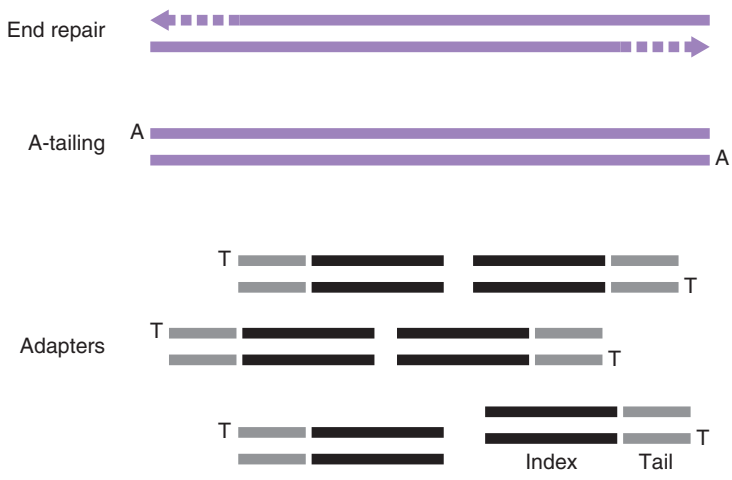

FIGURE 9.16 A next-generation sequencing library is created by the fragmentation of DNA. The fragmented DNA may have single-stranded ends that must be repaired back to double-stranded blunt ends by end repair. Addition of an A residue on the repaired blunt ends facilitates ligation of adapters carrying primer-binding sites for PCR amplification of the library.

technology has achieved gigabytes of sequencing data for a minimal cost, making genomic studies a routine component of both research and clinical analysis.

Historical Highlights

Early studies of DNA polymerase activity on an immobilized template led to the development of multiple template arrays that could be sequenced simultaneously. In 1997 high-throughput Selexa technologies were designed with capabilities of whole genome sequencing. High-throughput sequencing platforms developed in the mid-2000s resulted in a 50,000-fold drop in the cost of human genome sequencing from that of the Human Genome Project and led to the term *next-generation sequencing*. These technologies have increased in capacity and have been refined to address sequence complexity in genomes. The cost of sequencing a human genome has achieved the $1,000 cost point, which has expanded the use of sequencing analyses in the clinical laboratory.

NGS technologies include pyrosequencing, reversible dye terminator sequencing, ion-conductance sequencing, single-molecule sequencing, and sequencing by ligation. NGS requires novel methods of template preparation, such as emulsion PCR and bridge PCR, or single-molecule capabilities.[19] Powerful computer data assembly systems are required to organize the massive amounts of sequence information that are generated. These technologies can be used not only to sequence whole genomes but also to investigate populations of small genomes such as microbial diversity.[20]

Among the early challenges with massive sequencing was the integration of technologies without compromising accuracy or throughput.[21,22] These issues have been addressed with advances in bioinformatics and computer software. New challenges with system design, data accumulation and storage, clinical sensitivity, and data interpretation are being addressed, especially in dedicated sequencing facilities and commercial bioinformatics services.

NGS requires strong computer support as well as terabytes of storage space to accommodate large raw data sets. To prepare for NGS, clinical laboratories establish secure information channels and allocate space for preparation, loading, and operation of the sequencers. Interface with laboratory information systems and electronic medical records might also be arranged. Report templates are designed by the laboratory or commercial vendors and bioinformatics services.[23]

Two NGS technologies account for the majority of clinical sequencing applications: ion-conductance (pH)[24] and reversible dye terminator sequencing.[25] Both methods require the preparation of a sequencing library, sets of 100- to 500-bp-size fragments representing the regions to be sequenced. A library can represent a whole

genome or a few specific gene regions where critical variants are likely to occur.

Gene Panels

The size and application of the sequencing library depend on the selection of genes to be sequenced or gene panels. Gene panels are probe or primer sets designed to amplify specific genes, regions, or entire exomes (all protein-coding sequences).[26] NGS might also be performed to compare sequences of many organisms (rRNA genes in microbial speciation) or to detect large numbers of possible base differences in a highly polymorphic gene such as *CFTR*. Gene panels have high technical sensitivity but require knowledge of the clinical diagnosis that would justify testing particular genes.

Gene panels have been designed for disease states, such as cardiomyopathies or muscular dystrophy or cancers. These panels range from a few (less than 20) target genes to more than a thousand target genes such as those used for solid organ cancers. "Hot-spot" panels target regions of specific genes known to affect treatment response, disease state, or clinical condition. Variants in these regions are referred to as "actionable" mutations; that is, a therapeutic or medical measure might be taken as a result of the presence of a variant. Targeted panels include critical genes in particular diseases such as hematological-cancer-specific panels for lymphoid or myeloid disorders or solid-tumor-specific panels for lung, colon, breast, or other cancers. Very large panels up to 3,000 genes or more provide a large amount of information for diagnostic, prognostic, and discovery purposes. These panels, however, may produce variants of unknown significance that must be assessed by pathologists and oncologists on a patient-specific basis. With the increase in novel treatment strategies, gene variants and combinations of gene variants previously not considered actionable can become so. Whole-exome sequencing is a method of gene discovery. This more challenging approach with regard to interpretation has proven beneficial in cases of suspected inherited gene variants.[27,28] Initially, beyond the scope of clinical analysis, whole-exome and even whole-genome sequencing have been increasingly incorporated in special cases. For routine clinical laboratory work, however, small- to medium-size 15- to 500-gene panels account for the majority of sequencing procedures.

NGS Library Preparation

A collection of DNA fragments to be sequenced is a sequencing **library.** Reversible dye terminator and ion-conductance sequencing are performed on DNA fragments less than 1,000 bp in length. Genomic DNA is fragmented by a number of methods, including shearing with high-frequency acoustic energy, sonication, nebulization (forcing DNA molecules through a small opening), or enzymatic treatments. Particular methods and how they are used (e.g., pressure levels used in nebulization) produce differently sized fragments (100 to 1,000 bp). The median fragment size can be checked by gel electrophoresis or microfluidics. Starting DNA concentrations and the DNA concentration of the library is best measured by fluorometry.

Advanced Concepts

Sequencing protocols and technologies differ with respect to the amount of required input genomic DNA. The lower limits range from 10 to 50 ng of DNA. For sequencing tumor DNA from fixed tissue, 140 mm^2 tissue with at least 30% tumor is recommended. Suboptimal amounts of starting DNA will compromise sequence quality and increase the risk of PCR artifacts. Fluorometric measurement of input (and library) concentrations is recommended over spectrometry to ensure the measurement of intact DNA.

Fragmented DNA produced by enzymatic or physical methods may be used directly for whole-exome or whole-genome sequencing. The fragments will have a mixture of 5′ and 3′ overhangs, some phosphorylated. To facilitate ligation to synthetic adaptors, single-stranded fragment ends are removed or filled in with nuclease or polymerase treatment. The 5′ ends are phosphorylated. The 3′ ends can be adenylated to further enable ligation to adapters with T overhangs (Fig. 9.16).

Adapters are synthetic short dsDNA pieces carrying sequences complementary to a single primer pair. The adaptors may also contain short sequences that will identify the sample (**indexing** or **bar coding**; Fig. 9.17). This allows analysis of multiple samples in the same reaction as the sequencing software will put together sequences

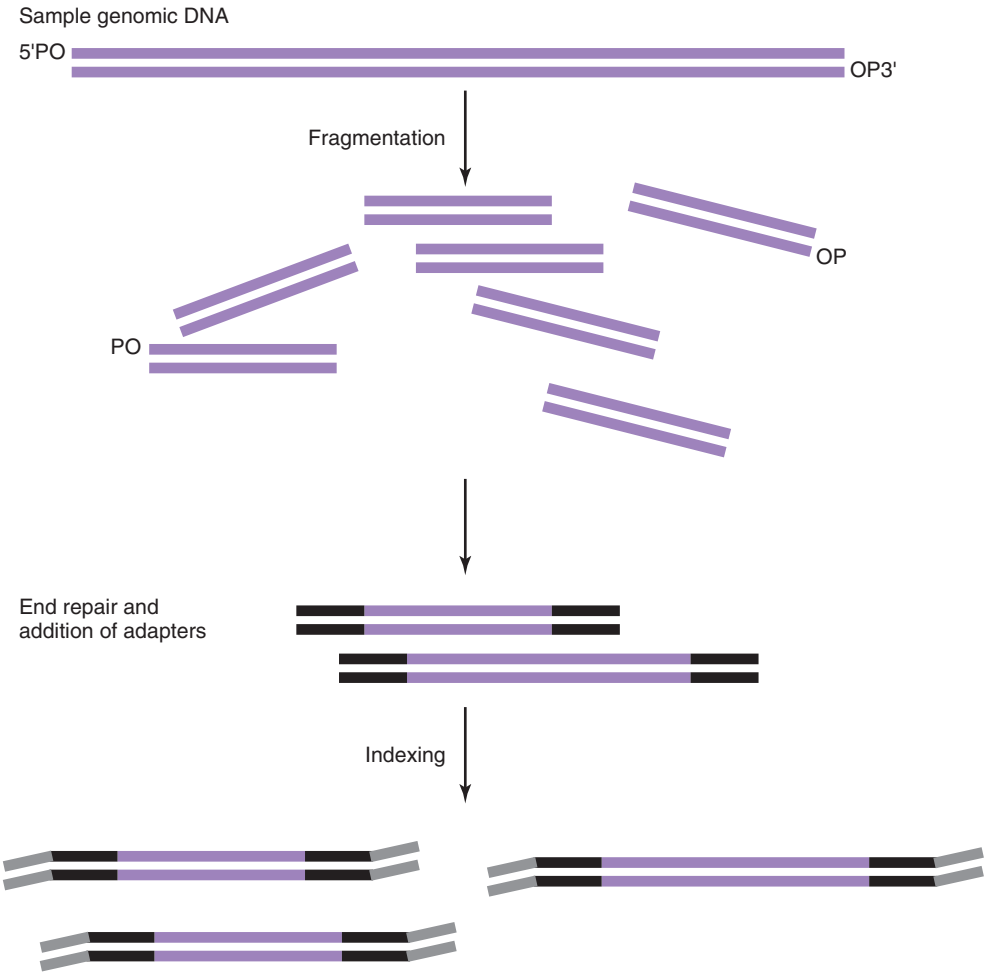

FIGURE 9.17 After fragmentation, end repair, and adapter ligation, bar codes or indexing may be performed by PCR amplification with tailed primers, or alternatively, the index sequences may be included in the adapters. Indexes are patient-specific so that multiple patient DNA can be sequenced in the same reaction and separated by their bar codes or indices after the sequencing is completed.

from fragments with the same bar code. Small genomes such as those of microbes or plasmids can be simultaneously fragmented and ligated to sequencing adapters in a single reaction tube. Reagent sets are commercially available for library preparation.

Targeted Libraries

Routine clinical sequencing of human DNA does not include the entire genome. Gene panels ranging from a few genes to whole exomes (all protein-coding regions) are employed, depending on the purpose for sequencing. The regions to be sequenced are enriched by probe hybridization or by amplification with region-specific primers.[29]

Probes are biotinylated oligonucleotides complementary to specific gene regions (Fig. 9.18). Targeted fragments to be sequenced are selected by hybridization with the biotinylated probe and captured with streptavidin-coated beads. The captured regions are ligated to adapters carrying primer-binding sites (or amplified with primer-binding sites included with short oligo probes) so that all reactions can proceed under the same amplification conditions in a single PCR reaction. Probe-based

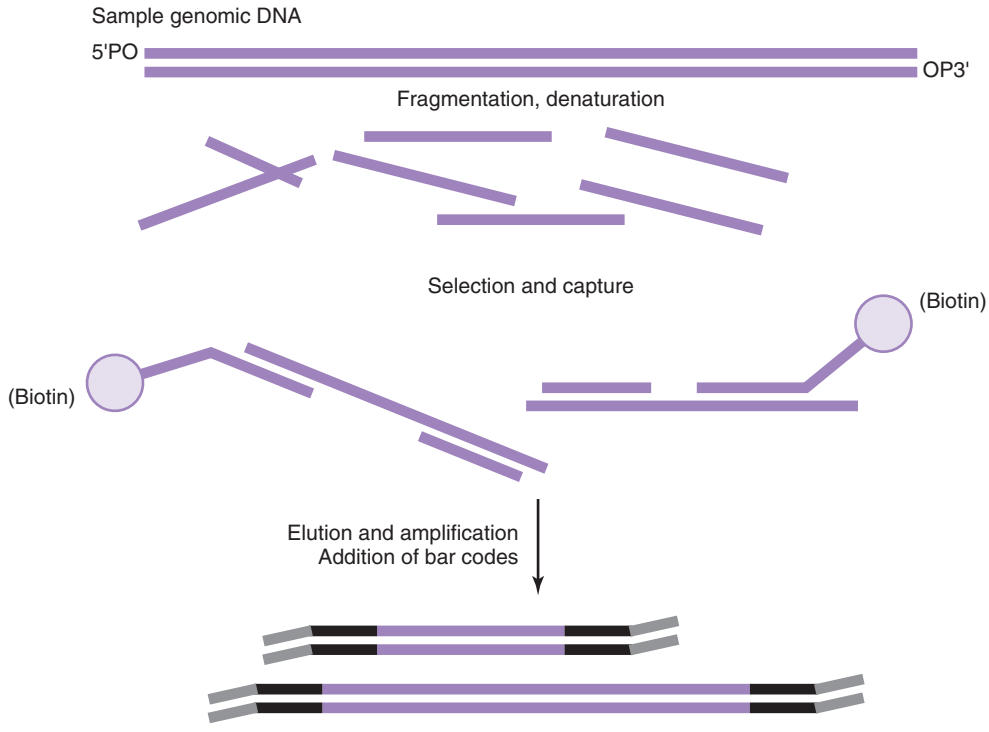

FIGURE 9.18 Targeted library preparation for NGS using probe enrichment. Fragmented DNA is denatured and hybridized to region-specific biotinylated probes. The probes are bead-captured, and the hybridized regions are amplified for sequencing. Probes may be short oligomers that can be extended across the region to be sequenced. The selected regions can then be amplified with tailed primers to add bar codes and sequencing primer-binding sites.

enrichment has the advantage of capturing sequences surrounding the region of interest and providing information from neighboring sequences. The presence of surrounding regions should be balanced because too much additional sequencing will affect the accurate sequencing of the targeted regions. The balance will depend on the average length of the DNA fragments.

Amplicon-based targeted libraries are selected by multiplex PCR with gene-specific primers tailed with binding sites for a secondary primer set (Fig. 9.19). After amplification, the secondary primers are tailed with index sequences that will identify (bar code or index) fragments from multiple samples in the same sequencing reaction and adapter sequences complementary to immobilized oligonucleotides anchored in the sequencing platform. These steps may be combined by tailing the initial multiplex PCR primers with the index and adapter sequences. Amplicon-based panel selection has the advantage of versatility and ease of use. Primer design is important, however, because sequence variations in the primer-binding sites may lower the efficiency of or even prevent amplification of particular fragments. Loss of library fragments from the sequenced regions, referred to as **allele dropout**, will cause inaccurate assessment of variant allele frequencies. Primers can be designed to produce overlapping sequences to cover less optimal regions. Paired-end or mate-pair primers produce coupled sequence fragments separated by 30 to 50 kb. By overlapping these reads, large variations not detectable in a few hundred base pairs such as translocations can be detected.

Both primer- and probe-based selections are affected by GC-rich sequencing targets. Secondary structure lowers the binding of primers and probes. GC-rich sequences also "clamp" primers in amplicon-based enrichment, lowering PCR efficiency. AT-rich regions may also be subject to poor hybridization, leading to loss of sequencing template fragments.

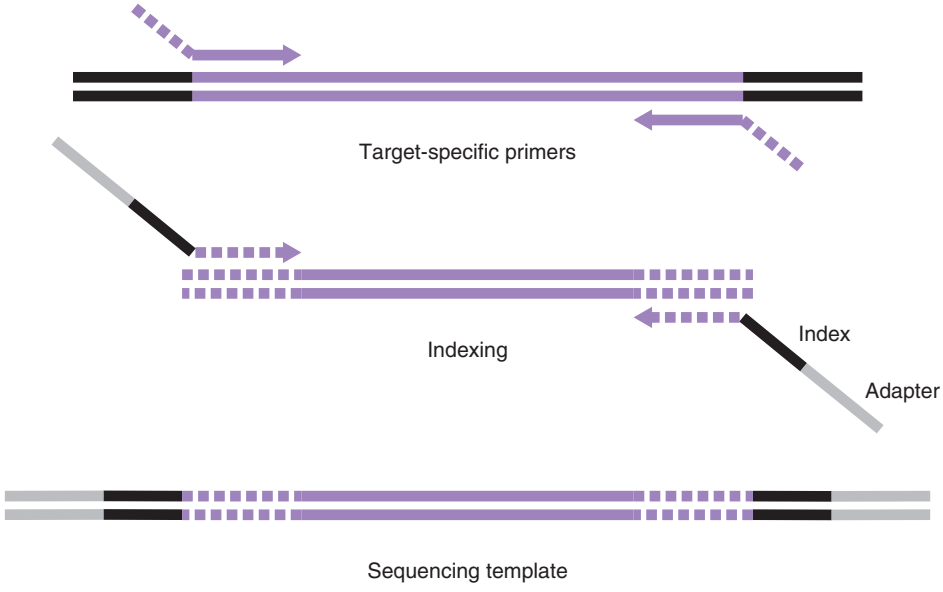

FIGURE 9.19 Targeted library preparation for NGS using amplification enrichment. Fragmented genomic DNA is end repaired and amplified with region-specific primers carrying binding sites for a single set of primers used in a second amplification. The second primer set has patient-specific index (bar-code) sequences.

Sequencing Platforms

After the introduction of NGS as a pyrosequencing technology, a variety of methods were developed for this purpose. The two most frequently used methods in clinical applications are ion-conductance and reversible dye terminator sequencing (Fig. 9.20). Both involve sequencing by synthesis and can be compared, chemically, to pyrosequencing and Sanger sequencing.

For ion-conductance sequencing, indexed libraries (gene panels) are amplified using primers immobilized on microparticles (beads) in an aqueous oil emulsion using adapters on the library fragments complementary to the immobilized primers (Fig. 9.20A). The beads carrying the amplicons (sequence templates) are placed on a solid surface (gene chip). The captured fragments are subjected to the addition of nucleotides in a predetermined order. If the nucleotide is complementary to the sequencing template, DNA polymerase will catalyze the formation of a phosphodiester bond. A hydrogen ion is released upon formation of the phosphodiester bond. The hydrogen ion will lower the pH of the reaction by a specific amount recorded by the sequencer (Fig. 9.21). This reaction occurs hundreds of thousands of times, producing sequence information from millions of sequencing panel library fragments.

In reversible dye terminator sequencing, captured or amplified fragments are hybridized to immobilized primers on a solid surface (flow cell). The fragments hybridize to the immobilized primers and are amplified by branch PCR into collections of products or **polonies** (Fig. 9.20B). Proper concentration (6 to 20 pMol) of the library DNA introduced to the flow cell will ensure that the polonies are evenly spaced on the flow cell. The polonies are sequenced in place by the sequential addition of fluorescently labeled nucleotides. If a nucleotide is complementary to the template next to the primer, DNA polymerase will extend the primer (form a phosphodiester bond). As in Sanger sequencing, each nucleotide is labeled with a specific color of fluor. An image is taken of the flow cell after each nucleotide addition (cycle), recording the presence of each added nucleotide color and location. After imaging, the fluorescent dyes are removed, and the next nucleotide is added (Fig. 9.22). Simultaneously, hundreds of thousands of polonies are sequenced in this way.

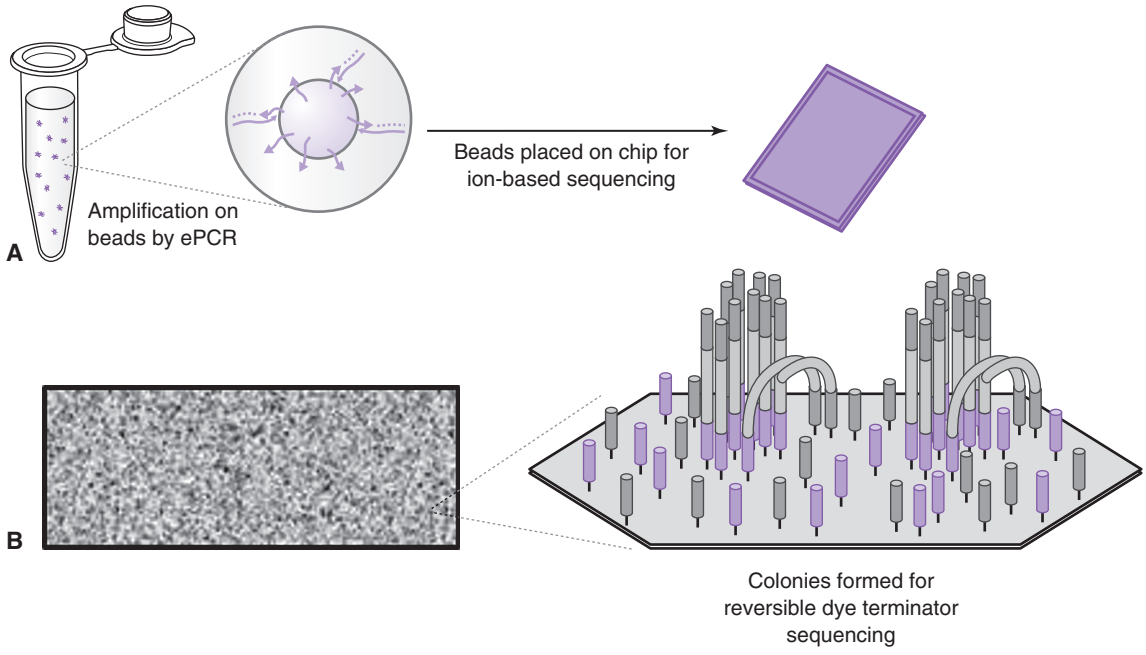

FIGURE 9.20 (A) Library amplification for ion-conductance sequencing is performed in emulsion PCR. The bar-coded libraries prepared are amplified from primer-binding sites complementary to bead-immobilized primers. At the end of the ePCR reaction, the emulsion is broken and applied to a solid surface (chip) for sequencing. (B) For reversible dye terminator sequencing, the panel is amplified by bridge PCR through primer-binding sites complementary to primers immobilized on the flow cell. Amplification in place on the solid surface produces batches or polonies of sequencing templates distributed evenly across the flow cell.

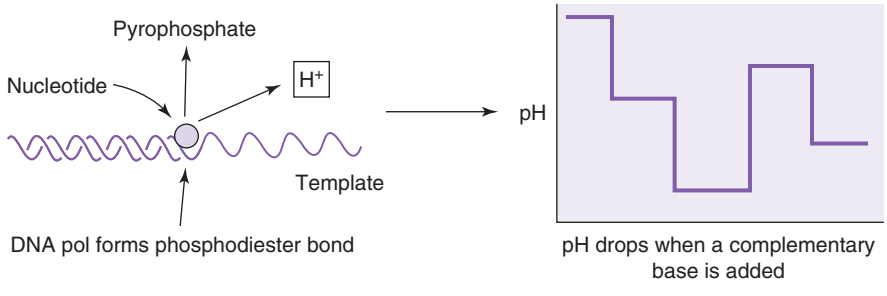

FIGURE 9.21 In ion-conductance sequencing, when the nucleotide added to the reaction is complementary to the template, it is joined to the growing chain by DNA polymerase, releasing a hydrogen ion and drop in pH identifying that nucleotide. Sequencing software converts pH changes to the nucleotide sequence.

Both sequencing platforms are accurate and efficient, with comparable performance.[30] Proper controls include a no-template sequencing control and a reference sequence control. Sequence runs take from 2.5 hours to 2 days, depending on the platform and the size of the library being sequenced.

Other sequencing platforms such as sequencing by ligation[31] and nanopore sequencing[32] are used in research applications. Sequencing by ligation uses a pool of labeled oligonucleotide DNA ligase to identify the template sequence through the known probe sequences (Fig. 9.23). Nanopore sequencing has the advantage of

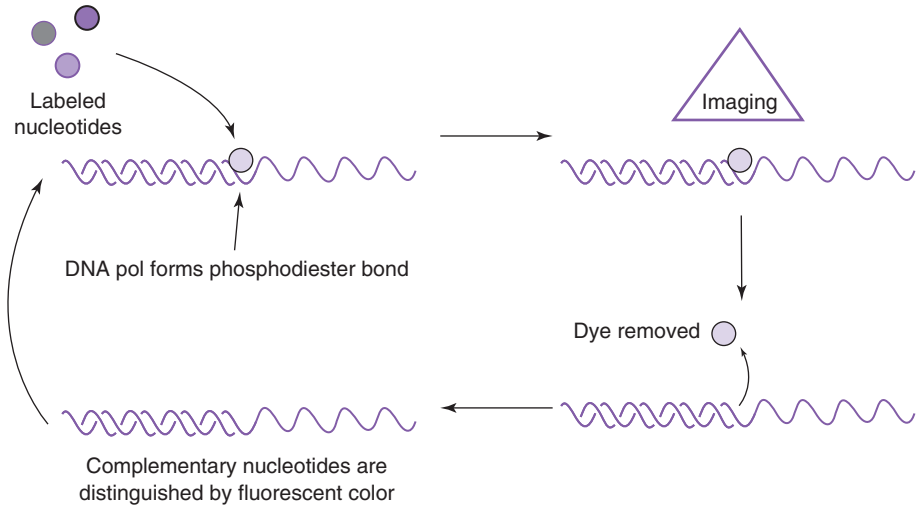

FIGURE 9.22 In reversible dye terminator sequencing, labeled nucleotides are applied to the flow cell and incorporated into growing chains by DNA polymerase at each polony location. Images are taken after rounds of fluorescent nucleotide addition; the color at each polony location indicates the next nucleotide in that sequence. Once the image is taken, the fluorescent labels are removed. Following this, another round of nucleotides is introduced.

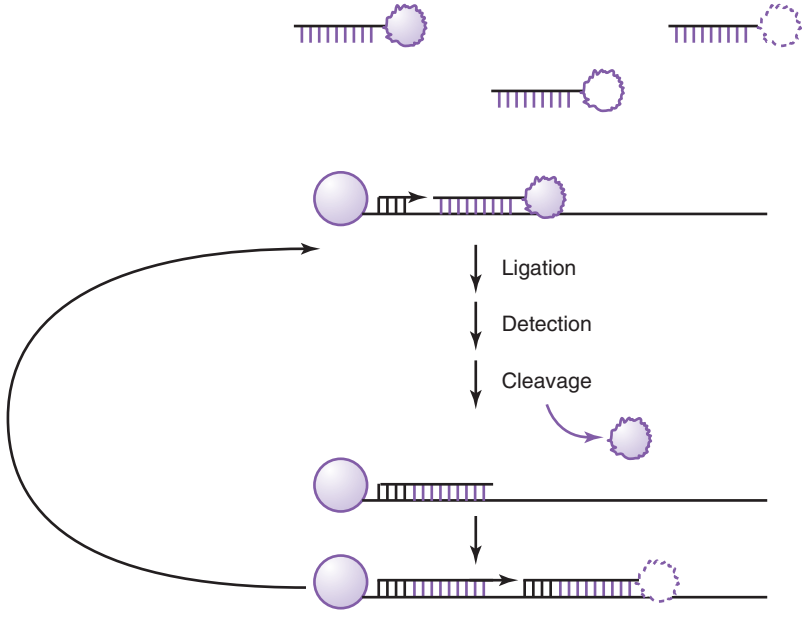

FIGURE 9.23 Sequencing by ligation uses short fluorescently labeled oligomers that hybridize in short increments if they are complementary to the DNA template. The template DNA anchored to a glass slide is flooded with fluorescent-labeled oligonucleotides. If the oligo is complementary to the template, it is ligated, and then two bases are detected at a time. The oligonucleotide is cleaved, followed by the next round of ligation. Each time, two new nucleotides are detected.

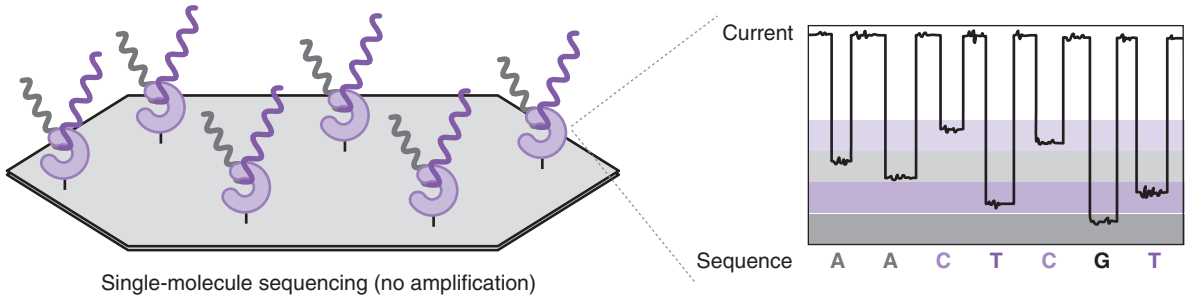

FIGURE 9.24 Long-read single-molecule sequencing uses protein ion channels through which one strand of each double-stranded DNA template is drawn. Each nucleotide passing through the pore changes the current in a characteristic way. This sequencing is rapid and does not require reassembly or short fragments for the final sequence.

not requiring fragmentation and amplification of the template DNA. One strand of long dsDNA molecules (up to 1 Mb) is drawn through protein pores. Each nucleotide is identified by a disruption in current as it passes through the pore (Fig. 9.24). This technology can also be used for direct RNA sequencing. Development of different technologies and improvement of existing technologies are actively occurring to further facilitate and widen the use of NGS.

Sequence Quality

Instrument collection and sequencing software will batch the sequences for each sample, based on the bar codes, and identify the nucleotide order in the process of base calling. Each base is assessed for quality of imaging (or conductance detection) and given a Phred score. Just as in Sanger sequencing, a Phred score of 2 to 3 (100- to 1,000-fold certainty of a correct call) is acceptable.

Each sequence is then compared to a reference sequence through read alignment. Reference sequences are considered "normal" in that there are no known significant variants; however, there is no real "normal" sequence, especially for human DNA. Variations from the reference may be the majority allele in the population, with the reference sequence carrying the minor allele. For human genome sequencing, reference genome hg19 was frequently used and reference genomes are updated periodically.[33] Reference sequences are free of known disease-related alleles, at least those found in the targeted panels.

The next step is variant identification based on comparison with the reference sequence. There are different types of variants, including single-nucleotide variants (SNVs), small insertion and/or deletion of nucleotides (indels), rearrangement of sequences (e.g., translocations), and copy-number variants (CNV; amplification or deletion of larger regions). Each of these types is handled differently by comparison software. Constitutional (genetically inherited) SNVs are identified in some programs based on a specific range of expected allele frequencies (variant allele/reference allele) for homozygosity or heterozygosity. Indels (up to 20 bp) can be identified by realignment, that is, multiple alignments (offset by one or more bases) that minimize base mismatches. Indels and even larger rearrangements can be detected by overlapping reads of paired end-primed sequences or by points of sequence diversions from 5′ and 3′ end reads (split-read analysis). Translocation breakpoints are often within introns or repetitive DNA sequences, or they contain overlaying sequence changes at the breakpoint, posing further challenges for variant identification. Optimal variant detection requires the use of the appropriate library primer design and software.

Once aligned, sequence variations from the reference (variants) are arranged in a variant call file (VCF). The VCF is a textual file that may be archived for further reference. Every variant is not of biological or clinical consequence. Some variants are synonymous or silent with regard to protein sequences. Others are common polymorphisms found in the population. Therefore, **annotation** are performed to identify critical variants.

TABLE 9.3 Annotation of Sequence Variants

Data	Description
Location of variant	Chromosome number, genomic coordinate (hg19 build)
Variant change	Reference allele, alternate allele detected in sample
Genetic state	Heterozygous, homozygous alternate
Quality of variant call	Quality/confidence score, sequencing depth at variant site (number of reads of variant and reference), probability of the reference or variant reads being balanced between + and – strands
Allele burden	The fraction of reads supporting the alternate allele (expect germline heterozygous alleles to be close to 0.5)
Variant type	The type of allele, either SNP, MNP, ins, del, or complex
Genomic position	Exon, intron, intergenic, other
Comparison to known variants	dbSNP ID, 1000 Genomes Project frequency with ethnicity, other disease-specific databases and information
Gene/coding effects	Annotated gene at the variant site, amino acid change, effect on protein sequence by the variant, algorithm scores for predicting damaging mutations

Filtering and Annotation

There are several components of annotation (Table 9.3). The confidence in the variant call is determined by sequence quality and coverage. **Coverage** is the number of times the region containing the variant is sequenced from independent fragments (read depth). Coverage is critical for confident detection of variants that are of low frequency in the sample such as somatic mutations in heterogeneous tumor tissue. Coverage of at least 500× (total of forward and reverse sequences) is recommended for detection of somatic variants.

The chromosomal and sequence location of the variant in context with the reference sequence is identified, along with the type of variant (SNP, insertion, deletion, or complex). The variant is then subjected to **filtering**. SNPs are compared to previously reported variants identified as human genome polymorphisms with the SNP rs identification number. Variants may be categorized as genetic or somatic in origin and, if genetic, as homozygous or heterozygous with the reference allele. Some variants are naturally occurring polymorphisms in particular populations. Data from the 1000 Genomes Project from major ethnic populations can be used to determine if a sequence variant is naturally present.

For gene panels and exome sequencing, variants will likely be found in gene-coding regions and adjacent intronic sequences, although intergenic areas may also be covered. The particular gene affected and the location of the variant in exon, intron, or intergenic sequences are noted. For variants found in introns, any effects of splicing are assessed. Variant effects on protein can also be estimated using algorithms such as PolyPhen and SIFT.[34] Silent variants will not change the amino acid sequence, but codon usage may have an effect on translation efficiency. Conservative amino acid substitutions or those late in the protein sequence have less effect on protein function than nonconservative mutations located early in the protein sequence. Algorithms provide scores to indicate the degree of damage to protein structure or function caused by the sequence variant. The dbNSFP database is a collection of in silico detected nonsynonymous variants.

Variants that remain after filtering may be annotated by searching in disease-specific databases, such as the Cancer Genome Atlas (TCGA), the Catalogue of Somatic Mutations in Cancer (COSMIC), My Cancer Genome, the Leiden Open (source) Variation Database (LOVD), and the Human Gene Mutation Database (HGMD). These databases and others contain population and clinical data associated with previously observed variants. The information from these databases can assist with the interpretation of the clinical effect of a variant. There are ongoing efforts to consolidate variant/disease data to ever larger and more comprehensive collections. Final reports of variants may contain information from databases, including effects on therapeutic treatments, especially targeted therapies, clinical trials, and prognosis. The clinical significance of a variant may differ with the heterogeneity of disease states as well as patient characteristics and demographics (e.g., age or gender).

Advanced Concepts

Based on professional surveys and literature reviews, a multidisciplinary group has proposed a system to categorize somatic variants in cancer.[35] It defines four tiers of variants as determined from cancer variant databases: tier I, variants with strong clinical significance; tier II, variants with potential clinical significance; tier III, variants of unknown significance (VUS); and tier IV, variants likely to be benign.

BIOINFORMATICS

Information technology has had to encompass the vast amount of data arising from the growing numbers of sequence discovery methods, especially direct sequencing and array technology. This deluge of information requires careful submission, storage, organization, and indexing of large amounts of data into databases such as those used in clinical sequencing analysis. **Bioinformatics** is the merger of biology with information technology. Part of the practice in this field is biological analysis **in silico,** that is, by computer. Bioinformatics dedicated specifically to handling sequence information is a form of **computational biology.** A list of some of the terms used in bioinformatics is shown in Table 9.4. The handling of the mountains of data being generated requires continual renewal of stored data, and a number of database programs are available for this purpose.[36,37]

Standard expression of sequence data is important for the clear communication and organized storage of sequence data. In some cases, such as in heterozygous mutations, there may be more than one base or mixed bases at the same position in the sequence. Polymorphic or heterozygous sequences are written as **consensus sequences,** or a family of sequences, with proportional representation of the polymorphic bases. The International Union of Pure and Applied Chemistry and the International Union of Biochemistry and Molecular Biology (IUB) have assigned a universal nomenclature for mixed, degenerate, or wobble bases (Table 9.5). The base designations in the IUB code are used to communicate consensus sequences and for computer input of polymorphic sequence data.

In addition to the interpretation of sequence variants, sequence information is also used in epidemiology, to speciate organisms or to find homologies within or between species. These applications involve database searches with comparisons of large regions of DNA. The **Basic Local Alignment Search Tool (BLAST)** is a system for homology searches. BLAST searches GenBank, a large database maintained by the National Center for Biotechnology Information (NCBI). Searches can be made of nucleic acid and amino acid sequences. Searches are performed by selecting a nucleotide or protein search and entering a sequence (query). Limits and parameters on the search can be added, such as the type of organisms to search (e.g., human, mouse, or other), exclusions and limits of organism or sample type, and the program. The program can optimize for highly similar sequence matches (megablast) or imperfect matches. Because sequences are directly submitted by researchers, there may be differences in the entered sequences due to the source of the sequenced material, the sequencing method, or the quality of the sequence. Selecting less-than-perfect matches will also allow cross-species matches of phylogenically conserved sequences, which can lead to the identification of important protein domains or clues to protein function.

The search will generate a number of matches or hits, with a diagram showing the alignments of the matching sequences and a color code indicating the best matches. Another section of the search results in E-values. The **E-value** (Expect value) describes the number of matches to the query by chance when searching a database of a particular size. It decreases exponentially with the quality of the match. Very low E-values (e.g., 10^{-12}) would be associated with a perfect match for a given query sequence. Further information, including the matched gene name and its organism, the source of the matched sequence and the location within that sequence, comparison of base to base or amino acid to amino acid, and plus or minus strand of the matched nucleotide sequence, are accessed by selecting the sequence or the color-coded bar in the diagram. The original submitted sequence can be accessed by selecting the sequence name.

In addition to the identification of new sequences, queries such as these are also useful for test and primer

TABLE 9.4 Bioinformatics Terminology

Term	Definition
Identity	The extent to which two sequences are the same
Alignment	Lining up two or more sequences to search for the maximal regions of identity in order to assess the extent of biological relatedness or homology
Local alignment	Alignment of some portion of two sequences
Multiple sequence alignment	Alignment of three or more sequences arranged with gaps so that common residues are aligned together
Optimal alignment	The alignment of two sequences with the best degree of identity
Conservation	Specific sequence changes (usually protein sequence) that maintain the properties of the original sequence
Similarity	The relatedness of sequences, the percent identity or conservation
Algorithm	A fixed set of commands in a computer program
Domain	A discreet portion of a protein or DNA sequence
Motif	A highly conserved short region in protein domains
Gap	A space introduced in alignment to compensate for insertions or deletions in one of the sequences being compared
Homology	Similarity attributed to descent from a common ancestor
Orthology	Homology in different species due to a common ancestral gene
Paralogy	Homology within the same species resulting from gene duplication
Query	The sequence presented for comparison with all other sequences in a selected database
Annotation	Description of functional structures, such as introns or exons in DNA or secondary structure or functional regions to protein sequences
Interface	The point of meeting between a computer and an external entity, such as an operator, a peripheral device, or a communications medium
GenBank	The genetic sequence database sponsored by the National Institutes of Health
PubMed	Search service sponsored by the National Library of Medicine that provides access to literature citations in Medline and related databases
SwissProt	Protein database sponsored by the Medical Research Council (United Kingdom)

design. Whenever a new primer or probe sequence is chosen, it is useful to query the primer or probe sequence to confirm that it belongs to the correct species and is not duplicated in multiple places in a genome. Primers and probes with multiple potential binding sites will produce mis-primes and off-target products.

Bioinformatics includes handling and updating of information for software tools and databases. The

TABLE 9.5 IUB Universal Nomenclature for Mixed Bases

Symbol	Bases	Mnemonic
A	Adenine	Adenine
C	Cytosine	Cytosine
G	Guanine	Guanine
T	Thymine	Thymine
U	Uracil	Uracil
R	A, G	puRine
Y	C, T	pYrimidine
M	A, C	aMino
K	G, T	Keto
S	C, G	Strong (3 H bonds)
W	A, T	Weak (2 H bonds)
H	A, C, T	Not G
B	C, G, T	Not A
V	A, C, G	Not T
D	A, G, T	Not C
N	A, C, G, T	aNy
X, ?	Unknown	A or C or G or T
O, -	Deletion	

accumulation of genomic and proteomic data, species and types of microorganisms based on sequences data, and variant association with disease drives the development of high-powered, reliable computer systems for storage as well as organization.

THE HUMAN GENOME PROJECT

From the first description of its double-helical structure in 1953 to the creation of the first recombinant molecule in the laboratory in 1972, DNA and the chemical nature of the arrangement of its nucleotides have attracted great interest. Gradually, this information began to accumulate, first regarding simple microorganisms and then partially in lower and higher eukaryotes. The deciphering of the human genome was a benchmark in the ongoing discovery of the molecular basis for disease and the groundwork of molecular diagnostics. In the process of solving the human DNA sequence, genomes of a variety of clinically important organisms were deciphered, advancing typing and predicting infectious disease treatment outcomes.

The first complete genome sequence of a clinically important organism was that of Epstein–Barr virus, published in 1984.[38] The 170,000-bp sequence was determined using the M13 template preparation/chain termination manual sequencing method. In 1985 and 1986, the possibility of mapping or sequencing the human genome was discussed at meetings at the University of California, Santa Cruz; Cold Spring Harbor, New York; and the Department of Energy in Santa Fe, New Mexico. The idea was controversial because of the risk that the $2 to $5 billion cost of the project might not justify the information gained, most of which would be sequences of "junk," or non-gene-coding DNA. Furthermore, there was no available technology up to the massive task. The sequencing automation and the computer power necessary to assemble the 3 billion bases of the human genome into an organized sequence of 23 chromosomes had not yet been developed.

Nevertheless, several researchers, including Walter Gilbert (of Maxam–Gilbert sequencing), Robert Sinsheimer, Leroy Hood, David Baltimore, David Botstein, Renato Dulbecco, and Charles DeLici, saw that the project was feasible because technology was rapidly advancing toward full automation of the process. In 1982, Akiyoshi Wada had proposed automated sequencing machinery and had gotten support from Hitachi Instruments. In 1987, Smith and Hood announced the first automated DNA sequencing machine.[39] Advances in the chemistry of the sequencing procedure were accompanied by advances in the biology of DNA mapping, with methods such as pulsed-field gel electrophoresis,[40,41] restriction fragment length polymorphism analysis,[42] and transcript identification. Methods were developed to clone large (500-kbp) DNA fragments in artificial chromosomes, providing long contiguous sequencing templates.[43] Finally, application of capillary electrophoresis

TABLE 9.6 Model Organisms Sequenced During the Human Genome Project

Organism	Genome Size (Mb)	Estimated Number of Genes
Epstein–Barr virus	0.17	80
Mycoplasma genitalium	0.58	470
Haemophilus influenzae	1.8	1,740
Escherichia coli K-12	4.6	4,377
E. coli O157	5.4	5,416
Saccharomyces cerevisiae	12.5	5,770
Drosophila melanogaster	180	13,000
Caenorhabditis elegans	97	19,000
Arabidopsis thaliana	90	25,000

to DNA resolution[44–46] made the sequencing procedure even more rapid and cost-efficient.

With these developments in technology, the Human Genome Project was endorsed by the National Research Council. The National Institutes of Health (NIH) established the Office of Human Genome Research with James Watson as its head. Over the next 5 years, meetings on policy, ethics, and the cost of the project resulted in a plan to complete 20 Mb of sequence of model organisms by 2005 (Table 9.6). To organize and compare the growing amount of sequence data, the BLAST and Gene Recognition and Assembly Internet Link (GRAIL) algorithms were introduced in 1990.[47,48]

For the human sequence, the decision was made to use a composite template from multiple individuals rather than a single genome from one donor. Human DNA was donated by 100 anonymous volunteers; only 10 of these genomes were sequenced. Not even the volunteers knew if their DNA was used for the project. To ensure accurate and high-quality sequencing, all regions were sequenced 5 to 10 times.

A second project started with the same goal. In 1992, Craig Venter left the NIH to start the Institute for Genomic Research (TIGR). Venter's group completed the first sequence of a free-living organism (*Haemophilus influenzae*)[49] and the sequence of the smallest free-living organism (*Mycoplasma genitalium*).[50] Venter established a new company named Celera and proposed to complete the human genome sequence in 3 years for $300 million, faster and cheaper than the NIH project. Meanwhile, Watson had resigned as head of the NIH project and was replaced by Francis Collins. In response, the Wellcome Trust doubled its support of the NIH project. The NIH moved its completion date from 2005 to 2003, with a working draft to be completed by 2001. Thus began a competitive effort on two fronts to sequence the human genome.

The two projects approached the sequencing differently (Fig. 9.25). The NIH method (hierarchical shotgun sequencing) was to start with sequences of known regions in the genome and "walk" further away into the chromosomes, always aware of where the newly generated sequences belonged in the human genome map. Venter and the researchers working with Celera—Gene Meyers, Jane Rogers, Robert Millman, John Sulston, and Todd Taylor—had a different idea. Their approach *(whole-genome shotgun sequencing)* was to start with 10 equivalents of the human genome cut into small fragments and randomly sequence the lot. Then, powerful computers would find overlapping sequences and use those to assemble the billions of bases of sequence into their proper chromosomal locations.

Initially, the Celera approach was met with skepticism. The human genome contains large amounts of repeated sequences, some of which are very difficult to sequence and even more difficult to map properly. A random sequencing method would repeatedly cover areas of the genome that are more easily sequenced and miss more difficult regions. Moreover, assembly of the whole sequence from scratch with no chromosomal landmarks would take a prohibitive amount of computer power. Nonetheless, Celera began to make headway (some alleged with the help of the publicly published sequences from the NIH), and eventually, the NIH project modified its approach to include both methods.

Over the next months, some efforts were made toward combining the two projects, but these efforts broke down over disagreements over database policy and release of completed sequences. The result of the competition was that the rough draft of the sequence was completed by

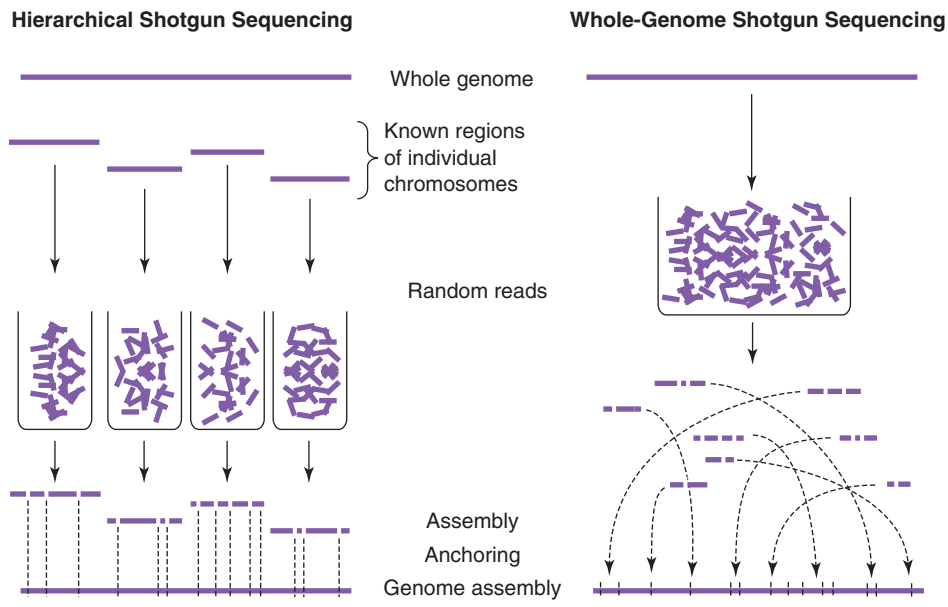

FIGURE 9.25 Comparison of two approaches for sequencing of the human genome. The hierarchical shotgun approach taken by the NIH *(left)* was to sequence from known regions so that new sequences could easily be located in the genome. The Celera whole-genome shotgun approach *(right)* was to sequence random fragments from the entire genome and then assemble the complete sequence with computers.

both projects earlier than either group had proposed, in June 2000. A joint announcement was made, and both groups published their versions of the genome, the NIH version in the journal *Nature*[51] and the Celera version in the journal *Science*.[52]

The sequence completed in 2000 was a rough draft of the genome; that is, there were still areas of missing sequence and sequences yet to be placed. Only chromosomes 21 and 22, the smallest of the chromosomes, had been fully completed. In the ensuing years, the finished sequences of each chromosome have been released (Table 9.7).

Remaining errors, gaps, and complex gene rearrangements will take years to resolve.[53] Detailed analysis of an individual genome will require sequencing of both homologs of each chromosome.[54] Even with the rough draft, interesting characteristics of the human genome were revealed. The size of the entire genome is 2.91 Gbp (2.91 billion bp). The genome was initially calculated as 54% AT and 38% GC, with 8% of the bases still to be determined. Chromosome 2 is the most GC-rich chromosome (66%), and chromosome X has the fewest GC base pairs (25%). A most surprising discovery was that the number of genes, estimated to be from 20,000 to 30,000, was much lower than expected. The average size of a human gene is 27 kbp. Chromosome 19 is the most gene-rich per unit length (23 genes/Mbp). Chromosomes 13 and Y have the fewest genes per base pair (5 genes/Mbp). Only about 2% of the sequences code for genes. Between 30% and 40% of the genome consists of repeat sequences. There is one single-base difference between two random individuals found approximately every 1,000 bases along the human DNA sequence. More detailed information, databases, references, and updated information are available at http://www.ncbi.nlm.nih.gov/.

The promise of the Human Genome Project for molecular diagnostics can be appreciated with the example of the discovery of the gene involved in cystic fibrosis. Seven years of work were required for discovery of this gene. With proper mapping information, a gene for any disease can now be found by computer, already sequenced, in a matter of minutes. Of course, all genetic diseases are not due to the malfunction of

TABLE 9.7 Completed Chromosomes

Chromosome	Completion Date
21	December 1999
22	May 2000
20	December 2001
14	January 2003
Y	June 2003
7	July 2003
6	October 2003
13	March 2004
19	March 2004
9	May 2004
9	May 2004
5	September 2004
16	December 2004
18	January 2005
X	March 2005
2	April 2005
4	April 2005
8	January 2006
11	March 2006
15	March 2006
12	March 2006
17	April 2006
3	April 2006
1	May 2006

a single gene. In fact, most diseases and normal states are driven by a combination of genes as well as by environmental influences. Without the rich information afforded by the sequence of the human genome, identification of these multicomponent diseases would be almost impossible. Ten years after the announcement of the completion of the Human Genome Project, almost 200 human genomes had been sequenced. Although the information gathered from the sequencing effort had not yielded the benefits to human health expected at the start of the project, it increased the appreciation of the vast complexity of genes and their regulation.[55]

Variant Associations With Phenotype

The Human Haplotype Mapping Project

As the Human Genome Project moved toward completion, another project was launched to further define the relationship between gene sequence and disease. This was the Human Haplotype Mapping, or HapMap, Project.[56] The goal of the project was to find blocks of sequences that are inherited together, marking particular traits and possibly disease-associated genetic lesions. The haplotype approach would reduce the number of polymorphisms required to examine the entire collection of genome/phenotype associations from the 10 million polymorphisms that exist to roughly 500,000 haplotypes. The HapMap Project revealed more than 1,000 disease-associated regions of the genome, covering commonly occurring conditions such as coronary artery disease and diabetes. With advances in technology, and the ability to generate sequence information, however, HapMap data has mostly been supplanted by higher-throughput data-gathering methods. As a result, the NCBI has planned to retire the HapMap project site because other resources, such as the 1000 Genomes Project, have become more comprehensive references for population genomics.

The 1000 Genomes Project

The 1000 Genomes Project provides a resource of structural variants in different populations.[57] The project has reconstructed the genomes of over 2,504 individuals from 26 populations by whole-genome sequencing, deep exome sequencing, and dense microarray genotyping in laboratories in the United States, United Kingdom, China, and Germany. Over 88 million variants (84.7 million SNPs, 3.6 million short insertions/deletions, and 60,000 structural variants) were verified. The resulting database includes more than 99% of single-nucleotide

variants with a frequency of greater than 1%. Data from the 1000 Genomes Project is a component of NGS variant assessment, providing more patient-specific interpretation of the clinical significance of variants. All variants from the 1000 Genomes Project are submitted to archives such as dbSNP.

The majority of HapMap SNPs are found in the 1000 Genomes Project.[58] Sites from HapMap that aren't found by the 1000 Genomes Project may be false discoveries by HapMap, the latter being based on microarray technology. Thus, there are a lot of SNPs from NGS projects that are not reported in HapMap.

The technology developed as part of the Human Genome Project made sequencing a routine method in the clinical laboratory. Small, cost-effective sequencers are available for rapid sequencing. In the clinical laboratory, sequencing is actually resequencing, or repeated analysis of the same sequence region, to detect mutations or to type microorganisms, making the task even more routine. The technology continued to develop, reducing the cost and labor of sequencing to detect multicomponent diseases or to predict predisposition to disease. Massive parallel or next-generation sequencing has supplemented and/or replaced Sanger sequencing in many clinical laboratories, and even this technology has evolved into lower-cost, user-friendly protocols. Accurate and comprehensive sequence analysis is one of the most promising areas of molecular diagnostics.

STUDY QUESTIONS

1. Read 5′ to 3′ the first 15 bases of the sequence in the gel on the right in Figure 9.7 (p. 229).

2. After an automated dye primer sequencing run, the electropherogram displays consecutive peaks of the following colors:
   red, red, black, green, green, blue, black, red, green, black, blue, blue, blue
   If the computer software displays the fluors from ddATP as green, ddCTP as blue, ddGTP as black, and ddTTP as red, what is the sequence of the region given?

3. A dideoxy sequencing electropherogram displays bright (high, wide) peaks of fluorescence, obliterating some of the sequencing peaks. What is the most likely cause of this observation? How might it be corrected?

4. In a manual sequencing reaction, the sequencing ladder on the polyacrylamide gel is very bright and readable at the bottom of the gel, but the larger (slower-migrating) fragments higher up are very faint. What is the most likely cause of this observation? How might it be corrected?

5. In an analysis of the *TP53* gene for mutations, the following sequences were produced. For each sequence, write the expected sequence of the opposite strand that would confirm the presence of the mutations detected.

   5′TATCTGTTCACTTGTGCCCT3′ (Normal)
   5′TATCTGTTCATTTGTGCCCT3′ (Homozygous substitution)
   5′TATCTGT(T/G)CACTTGTGCCCT3′ (Heterozygous substitution)
   5′TATCTGTT(C/A)(A/C)(C/T)T(T/G)(G/T)(T/G) (G/C)CC(C/T) . . . 3′ (Heterozygous deletion)

6. A sequence, TTGCTGCGCTAAA, may be methylated at one or more of the cytosine residues. After bisulfite sequencing, the following results are obtained:

   Bisulfite treated: TTGUTGCGUTAAA
   Write the sequence showing the methylated cytosines as C^{Me}.

7. In a pyrosequencing readout, the graph shows peaks of luminescence corresponding to the addition of the following nucleotides:

   dT peak, dC peak (double height), dT peak, dA peak
   What is the sequence?

8. Why is it necessary to add adenosine residues in vitro to ribosomal RNA before capture for sequencing?

9. Which of the following is next-generation sequencing?

   a. Maxam–Gilbert
   b. Tiled microarray
   c. Dideoxynucleotide chain terminator sequencing
   d. Reversible dye terminator sequencing

10. Which of the following projects would require next-generation sequencing?

   a. Mapping a mutation in the hemochromatosis gene
   b. Sequencing a viral genome
   c. Characterizing a diverse microbial population
   d. Typing a single bacterial colony

References

1. Sheikine Y, Kuo FC, Lindeman NI. Clinical and technical aspects of genomic diagnostics for precision oncology. *Journal of Clinical Oncology* 2017;35:929–933.
2. Maxam A, Gilbert W. Sequencing end-labeled DNA with base-specific chemical cleavage. *Methods in Enzymology* 1980;65: 499–560.
3. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain terminating inhibitors. *Proceedings of the National Academy of Sciences* 1977;74:5463–5467.
4. Metzker ML, Lu J, Gibbs RA. Electrophoretically uniform fluorescent dyes for automated DNA sequencing. *Science* 1996;271: 1420–1422.
5. Lewis E, Haaland WC, Nguyen F, Heller DA, Allen MJ, MacGregor RR, Berger CS, Willingham B, Burns LA, Scott GB, Kittrell C, Johnson BR, Curl RF, Metzker ML. Color-blind fluorescence detection for four-color DNA sequencing. *Proceedings of the National Academy of Sciences* 2005;102:5346–5351.
6. Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. *Science* 1998;281:363–365.
7. Harrington CT, Lin E, Olson MT, Eshleman JR. Fundamentals of pyrosequencing. *Archives of Pathology & Laboratory Medicine* 2013;137:1296–1303.
8. Insuasti-Beltran G, Gale JM, Wilson CS, Foucar K, Czuchlewski DR. Significance of MYD88 L265P mutation status in the subclassification of low-grade B-cell lymphoma/leukemia. *Archives of Pathology & Laboratory Medicine* 2015;139:1035–1041.
9. Lin S, Desmond EP. Molecular diagnosis of tuberculosis and drug resistance. *Clinical Laboratory Medicine* 2014;34:297–314.
10. Stürmer M, Reinheimer C. Description of two commercially available assays for genotyping of HIV-1. *Intervirology* 2012;55: 134–137.
11. Marsh S. Pyrosequencing applications. *Methods in Molecular Biology* 2007;373:15–24.
12. Shiraishi M, Hayatsu H. High-speed conversion of cytosine to uracil in bisulfite genomic sequencing analysis of DNA methylation. *DNA Research* 2004;11:409–415.
13. Weller M, Tabatabai G, Kästner B, Felsberg J, Steinbach JP, Wick A, Schnell O, Hau P, Herrlinger U, Sabel MC, Wirsching HG, Ketter R, Bähr O, Platten M, Tonn JC, Schlegel U, Marosi C, Goldbrunner R, Stupp R, Homicsko K, Pichler J, Nikkhah G, Meixensberger J, Vajkoczy P, Kollias S, Hüsing J, Reifenberger G, Wick W; DIRECTOR Study Group. MGMT promoter methylation is a strong prognostic biomarker for benefit from dose-intensified temozolomide rechallenge in progressive glioblastoma: the DIRECTOR trial. *Clinical Cancer Research* 2015;21:2057–2064.
14. Weller M, Stupp R, Reifenberger G, Brandes AA, van den Bent MJ, Wick W, Hegi ME. MGMT promoter methylation in malignant gliomas: ready for personalized medicine? *Nature Reviews Neurology* 2010;6:39–51.
15. Cocquet J, Chong A, Zhang G, Veitia RA. Reverse transcriptase template switching and false alternative transcripts. *Genomics* 2006;88:127–131.
16. Roberts J, Preston BD, Johnston LA, Soni A, Loeb LA, Kunkel TA. Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro. *Molecular and Cellular Biology* 1989;9:469–476.
17. Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences* 2003;100:3960–3964.
18. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM. Direct RNA sequencing. *Nature* 2009;461:814–818.
19. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 2016;17:333–351.
20. Mardis E. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* 2008;24:133–141.
21. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. *Trends in Genetics* 2008;24:142–149.
22. Fuller C, Middendorf L, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV. The challenges of sequencing by synthesis. *Nature Biotechnology* 2009;27:1013–1023.
23. Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, Temple-Smolkin RL, Voelkerding KV, Nikiforova MN. Guidelines for validation of next-generation sequencing based oncology panels. A joint consensus recommendation of the Association for Molecular Pathology and College of American Pathologists. *Journal of Molecular Diagnostics* 2017;19(3):341–365.
24. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 2011;475:348–352.

25. Ruparel H, Bi L, Li Z, Bai X, Kim DH, Turro NJ, Ju J. Design and synthesis of a 3′-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. *Proceedings of the National Academy of Sciences* 2005;102: 5932–5937.

26. Xue Y, Ankala A, Wilcox WR, Hegde MR. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. *Genetics in Medicine* 2015;17:444–451.

27. Jacob H, Abrams K, Bick DP, Brodie K, Dimmock DP, Farrell M, Geurts J, Harris J, Helbling D, Joers BJ, Kliegman R, Kowalski G, Lazar J, Margolis DA, North P, Northup J, Roquemore-Goins A, Scharer G, Shimoyama M, Strong K, Taylor B, Tsaih SW, Tschannen MR, Veith RL, Wendt-Andrae J, Wilk B, Worthey EA. Genomics in clinical practice: lessons from the front lines. *Science and Translational Medicine* 2013;5:1–5.

28. Manolio T, Fowler DM, Starita LM, Haendel MA, MacArthur DG, Biesecker LG, Worthey E, Chisholm RL, Green ED, Jacob HJ, McLeod HL, Roden D, Rodriguez LL, Williams MS, Cooper GM, Cox NJ, Herman GE, Kingsmore S, Lo C, Lutz C, MacRae CA, Nussbaum RL, Ordovas JM, Ramos EM, Robinson PN, Rubinstein WS, Seidman C, Stranger BE, Wang H, Westerfield M, Bult C. Bedside back to bench: building bridges between basic and clinical genomic research. *Cell* 2017;169:6–12.

29. Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E, Damodaran S, Bhatt D, Reeser JW, Datta J, Roychowdhury S. Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. *Human Mutation* 2015;36: 903–914.

30. Misyura M, Zhang T, Sukhai MA, Thomas M, Garg S, Kamel-Reid S, Stockley TL. Comparison of next-generation sequencing panels and platforms for detection and verification of somatic tumor variants for clinical diagnostics. *Journal of Molecular Diagnostics* 2016;18:842–850.

31. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* 2009;10:R32.

32. Lu H, Giordano F, Ning Z. Oxford nanopore MinION sequencing and genome assembly. *Bioinformatics* 2016;14:265–279.

33. Karthikeyan S, Bawa PS, Srinivasan S. hg19K: addressing a significant lacuna in hg19-based variant calling. *Molecular Genetics and Genomic Medicine* 2016;5:15–20.

34. Flanagan S, Patch AM, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet Test Mol Biomarkers* 2010;14(4):533–537.

35. Li M, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, Tsimberidou AM, Vnencak-Jones CL, Wolff DJ, Younes A, Nikiforova MN. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *Journal of Molecular Diagnostics* 2017;9:4–23.

36. Chojnacki S, Cowley A, Lee J, Foix A, Lopez R. Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Research* 2017;45(W1):W550–W553.

37. Niu S, Yang J, McDermaid A, Zhao J, Kang Y, Ma Q. Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Briefings in Bioinformatics* 2017. doi:10.1093/bib/bbx051

38. Baer R, Bankier AT, Biggin MD, Deininger PL, Farrell PJ, Gibson TJ, Hatfull G, Hudson GS, Satchwell SC, Seguin C. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. *Nature* 1984;310:207–211.

39. Hood L, Hunkapiller MW, Smith LM. Automated DNA sequencing and analysis of the human genome. *Genomics* 1987;1:201–212.

40. Schwartz D, Cantor CR. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* 1984;37: 67–75.

41. Van der Ploeg L, Schwartz DC, Cantor CR, Borst P. Antigenic variation in Trypanosoma brucei analyzed by electrophoretic separation of chromosome-sized DNA molecules. *Cell* 1984;37:77–84.

42. Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES. A genetic linkage map of the human genome. *Cell* 1987;51:319–337.

43. Riethman H, Moyzis RK, Meyne J, Burke DT, Olson MV. Cloning human telomeric DNA fragments into *Saccharomyces cerevisiae* using a yeast-artificial-chromosome vector. *Proceedings of the National Academy of Sciences* 1989;86:6240–6244.

44. Luckey J, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, Smith LM. High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Research* 1990;18:4417–4421.

45. Karger A. Separation of DNA sequencing fragments using an automated capillary electrophoresis instrument. *Electrophoresis* 1996;17:144–151.

46. Chen D, Swerdlow HP, Harke HR, Zhang JZ, Dovichi NJ. Low-cost, high-sensitivity laser-induced fluorescence detection for DNA sequencing by capillary gel electrophoresis. *Journal of Chromatography* 1991;559:237–246.

47. Altschul S, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology* 1990;215: 403–410.

48. Xu Y, Mural RJ, Uberbacher EC. Constructing gene models from accurately predicted exons: an application of dynamic programming. *Computer Applications in the Biosciences* 1994;10:613–623.

49. Fleischmann R, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM, McKenney K, Sutton GG, FitzHugh W, Fields CA, Gocayne JD, Scott JD, Shirley R, Liu L-I, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;269:496–512.

50. Fraser C, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA, Venter JC. The minimal gene complement of *Mycoplasma genitalium. Science* 1995;270:397–403.

51. Lander E, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowki J. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.

52. Venter J, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. *Science* 2001;291:1304–1351.

53. Dolgin E. The genome finishers. *Nature* 2009;462:843–845.

54. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC. The diploid genome sequence of an individual human. *PLoS Biology* 2007;5:2113–2145.

55. Hayden E. Life is complicated. *Nature* 2010;464:664–667.

56. Consortium IH. The International HapMap Project. *Nature* 2003;426:789–796.

57. Consortium TGP. A global reference for human genetic variation. *Nature* 2015;526(7571):68–74.

58. Buchanan C, Torstenson ES, Bush WS, Ritchie MD. A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *Journal of the American Medical Association* 2012;19:289–294.

LIKE WHAT YOU SEE?

Request more information