CHAPTER

3

Genomic Technologies*

OUTLINE

3.1 ADVANCES IN GENOMICS

Advances in genomics have broadened the scope of many already existing techniques from the gene scale to the genome scale with a concomitant drop in cost; DNA-sequencing and gene-expression-measurement technologies being the greatest beneficiaries. Genomics has two broad aspects: structural and functional. Structural genomics attempts to study the three-dimensional (3D) structure of proteins encoded by a genome. Therefore, the structural genomics approach requires the knowledge of the genome sequence, which is integrated with experimental and modeling data to predict the 3D structure of proteins. As the name implies, functional genomics aims to study gene (and protein) functions and interactions. Thus, functional genomics focuses on processes, such as transcription, translation, and protein−protein interaction. In reality, structural and functional aspects of genomics have overlaps simply because they both require knowledge of the genome sequence.

With the advancement of genomics, traditional molecular biology techniques—such as cloning, nucleic acid amplification, sequencing, mutagenesis, mutation detection, gene and protein interaction and expression studies—have been significantly improved in terms of their efficiency, cost, and high-throughput nature. Of these techniques, DNA-sequencing and gene-expression technologies have been revolutionized the most, and the scope of these techniques has been improved from the gene scale to the genome scale.

3.2 FROM SANGER SEQUENCING TO PYROSEQUENCING

Genome sequencing is the most direct method of detecting mutations, such as single nucleotide

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

polymorphisms (SNPs) and copy number variations (CNVs). The development of the dideoxy method of DNA sequencing was a major step forward for the science of molecular biology. The dideoxy method of DNA sequencing was published by Sanger and colleagues in 1977.[1] The technique is based on the **chain-termination** principle—that is, when DNA polymerase elongates the DNA chain, the incorporation of a dideoxynucleotide causes the termination of further chain elongation. This technique is not discussed any further because it is now the subject of textbooks. About 20 years after the development of Sanger's dideoxy sequencing, Pal Nyren introduced the **pyrosequencing** technique.[2] The pyrosequencing technique paved the way for the development and commercialization of large-scale, high-throughput, massively parallel sequencing technology, popularly referred to as **next-generation sequencing** or **next-gen sequencing** (**NGS**) **technology**.

3.3 PYROSEQUENCING, MUTATION DETECTION, AND SNP GENOTYPING

Pyrosequencing is based on the **sequencing by synthesis** principle. When DNA polymerase elongates the DNA chain, pyrophosphates are released. Each released pyrophosphate triggers a series of reactions that generates a detectable quantum of light. Therefore, pyrosequencing enables real-time detection of the sequence of a gene. Consequently, this technique is useful in the rapid detection of point mutations in the sequence and in SNP genotyping, including genotyping of microbes.

The DNA template that needs to be sequenced is first amplified by polymerase chain reaction (PCR). The amplicon (double-stranded amplified fragment) length is usually less than 200 bp for efficient pyrosequencing, but could be longer. While the number of cycles in regular PCR is around 30, the number of cycles in PCR for pyrosequencing is around 50. This is to ensure that the primers and the free nucleotides are utilized as much as possible. One of the two PCR primers is biotinylated at the 5′-end. The PCR amplicon containing a biotinylated end is captured on streptavidin-coated sepharose beads, denatured by alkali, and purified prior to pyrosequencing. The biotinylated strand is used as the template for pyrosequencing. A pyrosequencing primer (the third primer) is added to the purified biotinylated PCR strand and pyrosequencing is carried out.

Pyrosequencing is conducted in 96-well plates. During this process, the sequencing primer is first allowed to anneal with the DNA template in the presence of four enzymes—DNA polymerase, ATP sulfurylase, luciferase, and apyrase—and two substrates—adenosine 5′-phosphosulfate (APS) and luciferin—but without the deoxynucleotide triphosphates (dNTPs). Then, individual dNTPs are added to the reaction sequentially in a fixed order, which is programmed before the run. Out of the four dNTPs, only dATP is replaced by deoxyadenosine alpha-thio triphosphate (dATPαS). If the added dNTP is complementary to the base in the template strand, it is incorporated by the DNA polymerase and a pyrophosphate (PP_i) is released. ATP sulfurylase uses this PP_i and APS to generate ATP. The ATP is utilized by luciferase to oxidize luciferin into oxyluciferin with the concomitant emission of light, which is recorded by a charge-coupled device (CCD) camera in the form of a peak. Because of the stoichiometry of the reaction, the peak height is directly proportional to the number of nucleotides incorporated in tandem. Thus, if two of the same bases are incorporated back to back, the peak height becomes double, and so on. If the injected dNTP is not complementary to the template base, no signal is produced. Unutilized dNTPs are degraded by apyrase. The apyrase reaction is very important to keep the background noise level low. The readout of the pyrosequencing is called a **pyrogram** (Figure 3.1).

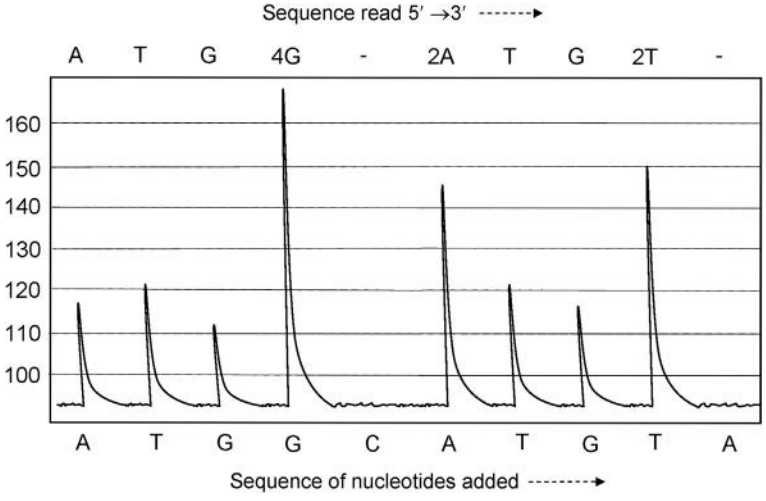

FIGURE 3.1 **A hypothetical pyrogram showing the sequence determination.** The peak height is proportional to the number of contiguous bases. There are four "G"s, two "A"s and two "T"s in this sequence. No peak was found at C in the middle and at A at the far right. The sequence for this window is ATGGGGGAATGTT.

By comparing the pyrogram of the query DNA (sample) with that of the wild-type DNA (reference), SNPs can be detected. The algorithm involves statistical analysis for significance. The enzymatic reactions of pyrosequencing are:

1. $DNA_n + dNTP \rightarrow DNA_{n+1} + PP_i$ (catalyzed by DNA polymerase)
2. $PP_i + APS \rightarrow ATP$ (catalyzed by ATP sulfurylase)
3. $ATP + luciferin + O_2 \rightarrow oxyluciferin + light\ quanta$ (catalyzed by luciferase)
4. Unincorporated $dNTP \rightarrow dNMP + 2\ P_i$ (catalyzed by apyrase)

3.4 NEXT-GENERATION SEQUENCING PLATFORMS

Next-generation sequencing (NGS) is high-throughput, massively parallel sequencing. NGS is also referred to as **second-generation** sequencing technology (the first generation being the original sequencing techniques of Sanger, and Maxam and Gilbert). The proposed cost of the first human genome sequencing was $3 billion ($3000 million). The sequencing of the genome of Dr J. Craig Venter reportedly cost $100 million, whereas the sequencing of the genome of Dr James Watson cost less than $1 million.[3] It is obvious that since the turn of the millennium, there has been a tremendous improvement in sequencing technology in terms of automation, high-throughput nature, and lowering the cost. The ultimate dream is to bring the sequencing cost down to $1000 per genome so that the genome of an individual can be sequenced for the purpose of personalized medicine and personalized nutrition.

Essentially, all NGS platforms discussed below utilize the following steps: DNA (sequencing) library preparation, immobilization of library fragments on a solid support, amplification of the fragments, massively parallel sequencing of the fragments, and computer-aided assembly of the sequence[a]. In this process, each nucleotide base incorporated is detected by a "wash-and-scan" method; millions of reactions are imaged per run to achieve the massively parallel sequencing; each read length is short. A DNA-sequencing library for use in NGS platforms is a collection of surface-anchored single-stranded fragments. The preparation of the sequencing library is a crucial step. *Therefore, the NGS technology does not need the DNA fragments to be cloned for sequencing*. Three popular NGS platforms discussed below are **Roche 454**, **Illumina Solexa**, and **ABI SOLiD**. All these technologies directly read the sequence of individual fragments without the need for cloning the fragments.

3.4.1 Roche 454

Roche 454 was the first NGS platform, introduced in the market in 2005. It is a high-throughput, large-scale, parallel pyrosequencing system. The 454 GS-FLX + system can sequence roughly 0.7 gigabases ($1\ Gb = 10^9$ bases) of DNA per run; the run time being 23 hours.[4] The coverage is $10\times$[b]. By 2013, the average read length was 700−800 bases. *These numbers are arbitrary because they keep improving with time*.

The 454 NGS platform represents a single-molecule improvement to standard pyrosequencing. In this technique, the sequencing library is amplified via **emulsion-PCR** (**em-PCR**), while **pyrosequencing chemistry** is used for sequencing the fragments. In em-PCR, a single DNA template molecule is clonally amplified in an oil/water emulsion (Figure 3.2). In brief, the technique comprises the following steps: (1) DNA-sequencing library preparation (DNA fragmentation + adapter ligation), (2) one fragment−one bead complex formation, (3) fragment amplification by em-PCR, (4) purification, and (5) sequencing by synthesis.

The process begins with shattering of a large DNA molecule, such as genomic DNA, into approximately 800−1000-bp-long fragments. These double-stranded DNA (dsDNA) fragments are blunt ended (polished) and end ligated with universal adapters (A and B). These adapters provide priming sequences for both amplification and sequencing. The A/B-adapter-ligated dsDNA fragments are selected using streptavidin−biotin purification discussed before, denatured into single strands, and combined with an excess of micrometer-sized DNA capture beads or in a 1:1 DNA/bead ratio (but not an excess of DNA, in order to ensure generation of monoclonal beads). The surface of these beads carries oligonucleotides complementary to the adapter sequences on the fragment library. Next, the DNA

[a]If a genome is resequenced, the fragment assembly can be performed with the aid of the reference genome, called **reference assembly**. If a genome is sequenced for the first time, its assembly is called **de novo assembly**.

[b]Coverage denotes the number of times a genome (or a target sequence) has been sequenced. Thus, a $10\times$ coverage for a sequenced genome means that the entire genome has been sequenced 10 times over. So, the higher the coverage, the greater is the depth of sequencing (hence the term **deep sequencing**). A high coverage ensures that the base calling is accurate. Coverage (C) = [read length (L) × number of reads (N)]/G (haploid genome length). Thus, if a target sequence of 5000 bp is assembled from 100 reads with an average read length of 300 nucleotides, the coverage is $(300 \times 100)/5000 = 6\times$. Intuitively, a $6\times$ sequence coverage for the genome appears to mean that each base of the genome has been read 6 times over, but in reality that may not be the case because some parts of the genome of higher eukaryotes are not easily amenable to sequencing, such as intronic sequences and highly repeated sequences.

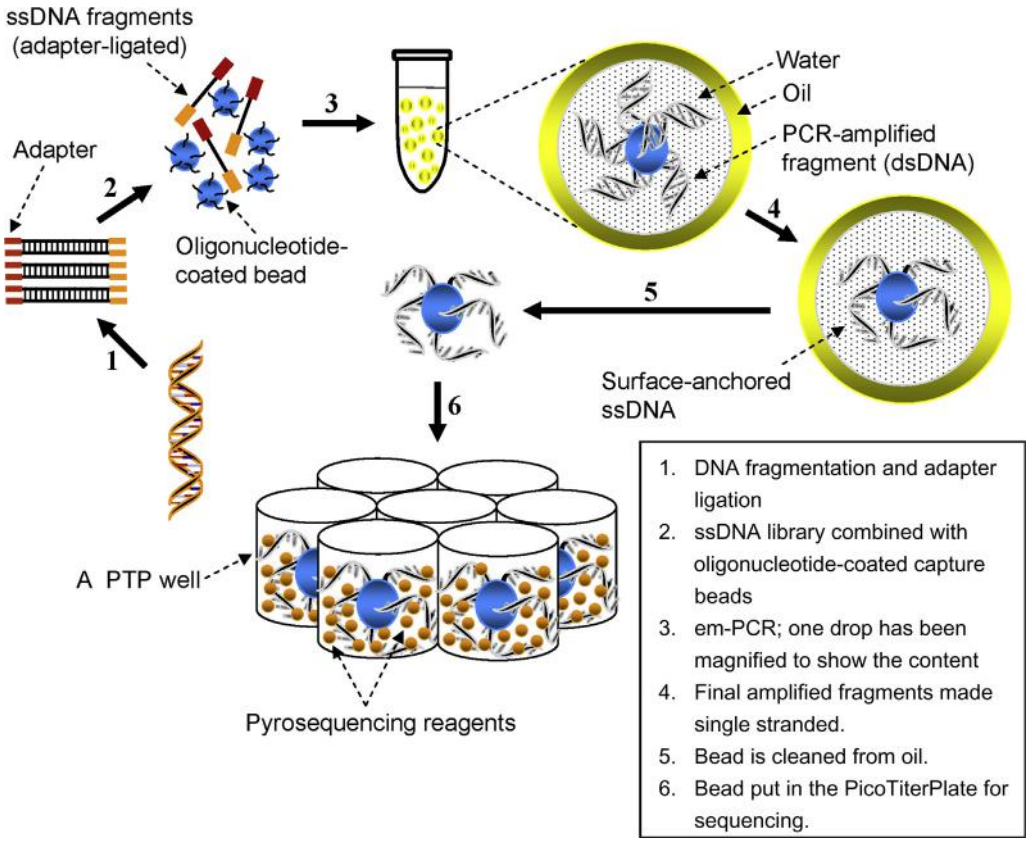

FIGURE 3.2 **Principles of 454 sequencing.** A DNA sequencing library is prepared by ligating adapters to end-polished DNA fragments. Single-stranded (ss) fragments are combined with DNA capture beads containing oligonucleotides complementary to the adapters. The DNA fragments, beads, and PCR reagents are combined within an aqueous mixture, mixed with synthetic oil, and vigorously shaken, which results in the formation of water-in-oil emulsion droplets. Typically, most droplets contain only one bead and one DNA fragment each. The DNA fragment is amplified in emulsion-PCR (em-PCR). The PCR products are purified, denatured, and sequenced in a picotiter plate (PTP) using pyrosequencing chemistry.

fragments, beads, and PCR reagents are combined within an aqueous mixture, which is then mixed with synthetic oil and vigorously shaken. The shaking results in the formation of water-in-oil emulsion droplets (micro-reactors). Typically, most droplets contain only one bead and one DNA fragment each, surrounded by the aqueous layer, which, in turn, is surrounded by the oil layer. The DNA fragment in each droplet is PCR amplified into clonally amplified copies. This PCR process is called **emulsion-PCR** (**em-PCR**). Thus, each bead will bear on its surface PCR products that have been amplified from a single molecule from the template library; these beads are therefore called monoclonal beads. In these bead-immobilized amplicons, the hybridized strand is washed away leaving the beads with surface-anchored single strands.

Next, the beads are screened from the oil and cleaned. The amplified DNA sequencing library, thus generated, is then loaded onto a picotiter plate (PTP) for pyrosequencing. The PTP contains 1.6 million wells; each well is approximately 44 μm in diameter and 75 picoliters in volume.[5,6] Each well can accommodate only a single capture bead. The pyrosequencing reaction mix is also packed into these wells. The PTP is loaded onto an automated pyrosequencing platform, such as the Roche 454 GS-FLX + system, and the DNA fragments are subjected to high-throughput parallel pyrosequencing. The beads that do not contain DNA are eliminated, and the beads that hold more than one type of DNA fragment (polyclonal beads) will be readily filtered out during sequencing signal processing.

3.4.2 Illumina Solexa

Solexa was founded in 1998 in the UK to develop high-throughput sequencing using fluorescently labeled nucleotides and a **sequencing-by-synthesis** approach,

like 454. However, while 454 employs pyrosequencing chemistry for sequencing, Solexa employs fluorescent **reversible terminator chemistry**[c]. The first Solexa sequencer (Genome Analyzer) was introduced in 2006, and could sequence 1 Gb in a single run. In 2007, Illumina acquired Solexa, and by 2011 this sequencing capability had increased to 600 Gb in a single run.[7] The coverage is 30×. By 2013, the run time in the HiSeq 2000/2500 platform was 11 days (regular mode) or 2 days (rapid run mode), and the average read length was ~100 bases.[4] *As indicated earlier, these numbers are arbitrary because they keep improving with time*. The main steps in the Solexa technology are the following: (1) DNA-sequencing library preparation (DNA fragmentation + adapter ligation), (2) addition to flow-cell channels, (3) bridge amplification, (4) cluster generation, and (5) sequencing by synthesis.

For DNA-sequencing library preparation, long DNA is randomly fragmented by ultrasonication; fragments are blunt ended and adapter ligated at both ends. The adapter-ligated fragments are size selected for a length of 250−350 bp, and subjected to small-cycle (10−15 cycles) PCR to increase the yield, which is verified by gel analysis. The desired fragment size pool is isolated and used as the source of the DNA-sequencing library. The dsDNA fragments are denatured and added to the flow-cell channels. The flow-cell channels already contain surface-anchored oligonucleotide primers that immobilize these single-stranded fragments by hybridizing to the adapters. The next step is cluster generation. First, the immobilized fragments are subjected to standard PCR amplification so that many copies of the original fragment are produced and localized in a tight cluster. The double-stranded PCR products in the cluster are denatured and the original strands (hybridized to the surface-anchored primers providing the template for amplification) are washed away leaving the newly synthesized strands, which are now surface anchored. These surface-anchored single strands flip over to hybridize with their nearest surface-anchored primers, forming a bridge-like appearance. Polymerase in the PCR mix extends the hybridized primer, forming a double-stranded bridge. This process of PCR amplification is called **bridge amplification**. When the double-stranded bridge is denatured, two single-stranded molecules are obtained, each of which is now surface anchored. The bridge amplification PCR cycles are repeated to obtain dense clusters of amplified single-stranded products. In this way, several million dense clusters are generated in each channel of the flow cell. These initial clusters have both forward and reverse strand clusters. Next, the reverse strands are cleaved and washed away, leaving the forward strand clusters (Figure 3.3).

The strands are then sequenced using sequencing primers. The first sequencing cycle is initiated by adding all four fluorescently labeled reversible terminator bases (each base contains a different fluorophore), sequencing primers, and DNA polymerase to the flow cell. The polymerase can perform only single base extension; thus, only the base complementary to the template strand is incorporated and the extension stops because of the blocked 3′-end of the added base. Next, the unincorporated bases are removed and the added base is subjected to laser excitation. Following laser excitation, the emitted fluorescence is captured by a CCD camera. Thus, the first base is imaged. The first base of each fragment is similarly recorded and imaged. Then the fluorophore and the terminal 3′-OH end block of the first base are chemically removed, allowing the second cycle to take place. In a similar fashion, the second base added is imaged for all fragments. The cycle is repeated to determine the sequence of bases in each fragment, one base at time. The sequence is assembled by computer software using a reference genome (**reference assembly**). If there is no reference genome and the sequence is new, the sequence assembly is done by the **de novo assembly** method. To score SNPs, the sequence obtained is aligned and compared to a reference (e.g. reference genome) and sequence differences are identified.

3.4.3 ABI SOLiD

Applied Biosystems commercialized its SOLiD platform in 2008. The acronym SOLiD stands for **sequencing by oligonucleotide ligation and detection**. Unlike the 454 and Solexa platforms that use a sequencing-by-synthesis approach, the SOLiD platform uses a **sequencing-by-ligation** approach, and employs sequencing-by-ligation chemistry for sequencing.

Most recent SOLiD platforms, such as the SOLiD 4 system, produce 80−100 Gb of usable DNA data per

[c]In reversible terminator chemistry, each of the four types of dNTPs is labeled with a unique removable fluorophore at the base. Additionally, the 3′-OH end is chemically blocked, but the 5′-PO_4 end is free. After the fluorophore-conjugated dNTP is incorporated by DNA polymerase into the DNA chain, the fluorescence image of the fluorophore is captured using laser excitation. Next, the fluorophore and the 3′-OH block are chemically removed. The resulting 3′-OH end of the newly incorporated dNTP is ready to accept the next incoming nucleotide. This cycle is repeated.

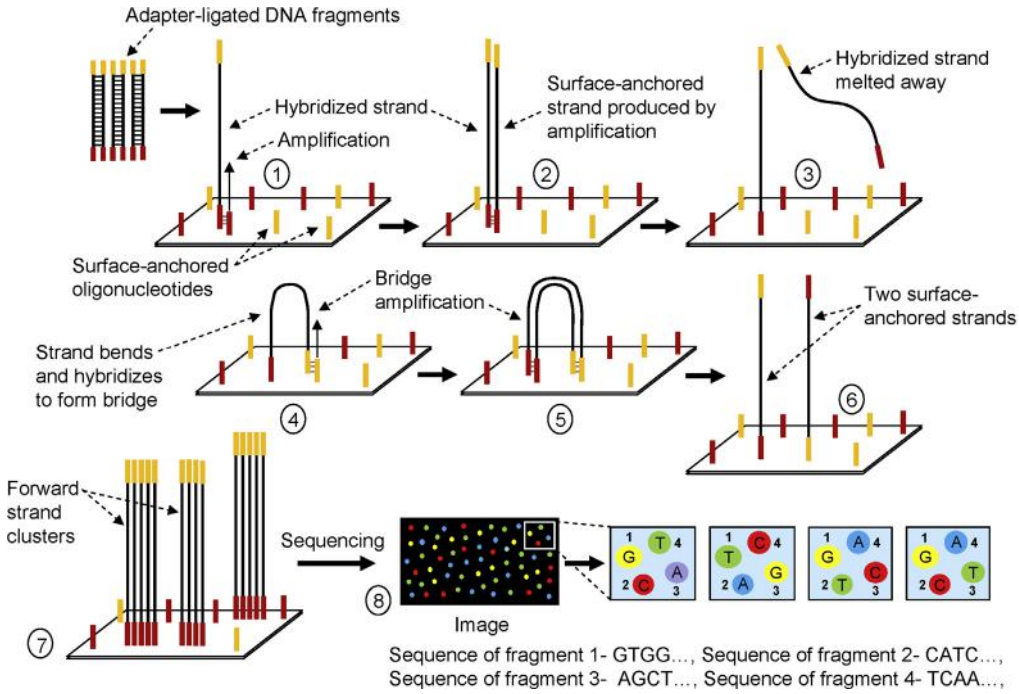

FIGURE 3.3 **Principles of Illumina Solexa sequencing.** The DNA-sequencing library is prepared by ligating adapters to the end-polished DNA fragments. The single-stranded fragments are allowed to hybridize with surface-anchored oligonucleotides that are complementary to the adapters. Initial PCR amplification of the strands followed by bridge (PCR) amplification results in the generation of single-stranded clusters. The strands are then sequenced using fluorescent reversible terminator chemistry (see text for details).

run.[8] The coverage is 30×. By 2013, the average read length of SOLiD sequencing was ~50 bases. *As indicated above, these numbers are arbitrary because they keep improving with time*. In brief, the technique comprises the following steps: (1) DNA-sequencing library preparation (DNA fragmentation + adapter ligation), (2) one fragment−one bead complex formation, (3) fragment amplification by em-PCR, (4) purification, (5) bead immobilization on glass slide, and (6) sequencing by ligation.

The sequencing library preparation for SOLiD sequencing involves shearing of large DNA molecules into 400−600-bp fragments. The fragments are end repaired, adapter ligated, and immobilized on paramagnetic beads. The dilution and anchoring process ensures that only one template per location is tethered. The fragments on the beads are amplified by em-PCR, the beads with extended templates are separated out from undesired beads, the extended templates on the beads are 3′-end modified, and then the beads are immobilized on a glass slide.

The sequencing-by-ligation chemistry utilizes a di-base (two-base) query system for interrogating the sequence and a fluorescent dye for detection. This is also known as **two-base encoding**. The system uses four fluorescent dyes to interrogate all sixteen (4^2) possible two-base combinations. This system utilizes a number of probes; each probe is eight nucleotides (nt) long (8-mer), in which the first two bases at the 5′-end represent the unique two-base combination, and the fluorophore is at the 3′-end. The process begins when a sequencing primer is allowed to hybridize with the universal adapter. Next, a probe that contains the two-base combination complementary to the two bases immediately 3′ to the adapter hybridizes. The base pairing results in the ligation of the 8-mer to the sequencing primer, thereby extending the sequencing primer. The ligation step is followed by fluorescence detection and base calling. Next, a regeneration step removes three 3′ bases from the ligated 8-mer (including the fluorescent group). This prepares the extended primer for another round of ligation. This process is repeated until a specific read length is achieved. Then this extended hybridized sequence is melted away, and the process is repeated with new 8-mers (primer reset) (Figure 3.4).

There are even fully automated benchtop versions of these sequencing instruments available, such as the 454 GS Junior of Roche, MiSeq of Illumina, and Ion Personal Genome Machine and Ion Proton, both of Life Technologies (discussed below).

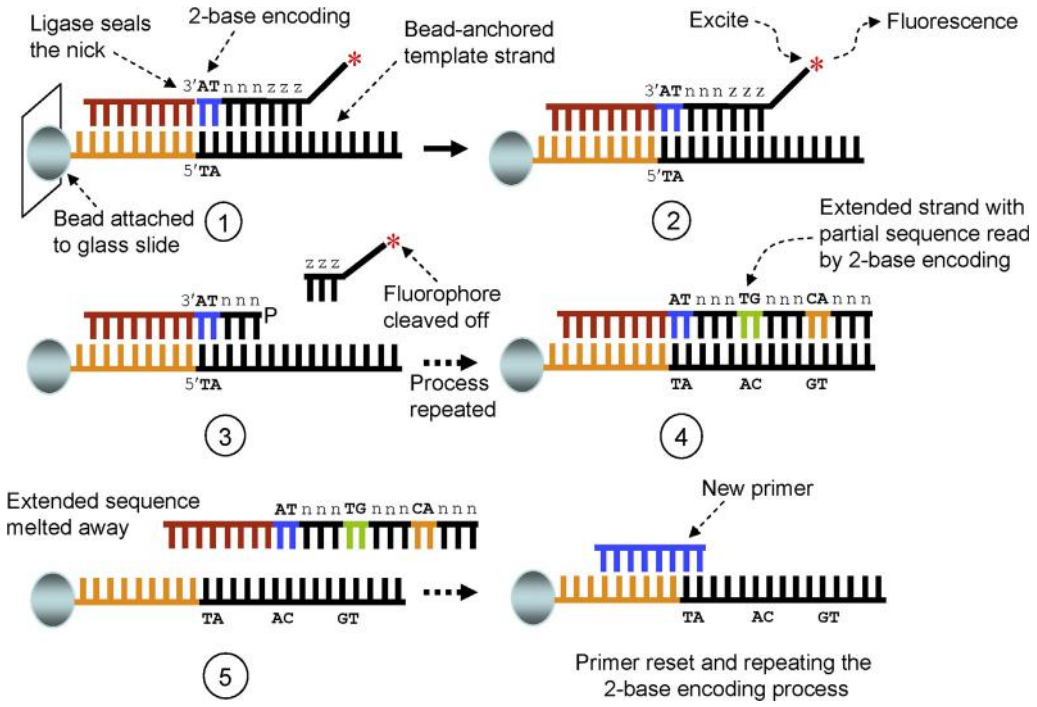

FIGURE 3.4 **Principles of SOLiD sequencing.** The DNA-sequencing library is prepared by ligating adapters to the end-polished DNA fragments, and immobilized on paramagnetic beads. The dilution and anchoring process ensures that only one template per location is tethered. The fragments on the beads are amplified by em-PCR, the extended templates on the beads are 3′-end modified, and the beads are immobilized on a glass slide. The sequencing-by-ligation chemistry utilizes a two-base encoding query system for interrogating the sequence and a fluorescent dye for detection (see text for details).

3.5 NEXT-NEXT-GENERATION SEQUENCING TECHNOLOGY

The invention of DNA sequencing technology was pioneered by Fred Sanger in the UK, and by Alan Maxam and Walter Gilbert in the USA. Sanger's dideoxy-chain-termination method ultimately became the sequencing method of choice because it was technically easier to perform and could be scaled up. These methods are popularly referred to as **first-generation sequencing technology**. The read lengths of these methods are typically 600−800 bp, but could be longer. **The original human genome sequencing project largely relied on the automated and scaled-up version of first-generation sequencing technology**. The main drawbacks of first-generation sequencing technology are the slow progress, because only a small amount of DNA could be sequenced per unit time (low throughput), and high cost (cost per base sequenced).

The introduction of **second-generation sequencing technology** (also known as **next-generation sequencing technology**), three popular platforms of which are discussed above, was an attempt to solve the two major problems of first-generation sequencing technology—that is, to introduce high-throughput sequencing technology for a lower cost of sequencing. However, the second-generation sequencing technology platforms have their own technical problems; for example, a PCR-generated DNA-sequencing library may have PCR-introduced bias and errors, fluorescent nucleotide labeling is not fully efficient, exonucleases are inefficient with labeled nucleotides, detection of single-molecule fluorescence has a high error rate because of the inherent noise in a fluorescence-driven base call, and the same strand can not be "re-read." The noise is due to the fact that the base addition is $< 100\%$ efficient; as a result, as the number of incorporation cycles increases, the population of molecules becomes asynchronous, which results in errors in sequencing read. Although the very high-throughput nature of these methods tends to alleviate some of these problems, the future goal is to develop next-next-generation sequencing technology that will be more efficient and free from the technical problems encountered in second-generation sequencing technology.

Next-next-generation sequencing technology[9] is **third-generation sequencing technology**, although the boundary between the second-generation and third-generation technologies may not be distinct. Ideal desired features of the true third-generation sequencing

technology will probably include the following: single-molecule sequencing technology, no PCR amplification, less complex sample preparation, no pausing of sequencing after each base incorporation (hence increase in sequencing rate), increased read length, and decreased cost. Some of the currently available sequencing technologies that are at the border between the current second-generation and the futuristic third-generation include Life Technologies' **Ion Torrent** semiconductor sequencer that employs a sequencing-by-synthesis approach and uses **pH change** (from the released hydrogen ion during the polymerization of nucleotides) to detect nucleotide incorporation, and **Helicose's** Genetic Analysis Platform that employs a sequencing-by-synthesis approach of a **single molecule** using a defined primer and works by imaging individual DNA molecules as they are extended. The Ion Torrent workflow involves generation of the sequencing library, amplification of the library fragments onto proprietary Ion Sphere particles by em-PCR, deposition of the Ion Sphere particles coated with template in the Ion chip, and sequencing. The average read length is up to 200 bases.

The only truly third-generation sequencing approach so far introduced seems to be the **single-molecule real-time** (**SMRT**) sequencing technology developed by Pacific Biosciences (PacBio). It employs a sequencing-by-synthesis approach and allows for direct observation of the synthesis of a single strand of DNA by DNA polymerase in real time. The SMRT technology of PacBio utilizes what is called a **zero-mode waveguide** (ZMW). A ZMW is a hole, tens of nanometers in diameter, fabricated in a 100-nm metal film deposited on a glass substrate. An active polymerase is immobilized at the bottom of each ZMW chamber. The ZMW, being so small, prevents visible laser light from passing entirely through it; the laser exponentially decays as it enters the ZMW. Because of this property, a laser passed through the glass into the ZMW only illuminates the bottom 30 nm of the ZMW chamber. Nucleotides are allowed to diffuse into the ZMW chamber; each base is labeled with a different fluorescent dye. The incorporated base can be recognized based on the fluorescence emission, which happens within the illuminated section of the nanochamber, and the synthesis of a single DNA molecule is directly recorded.[10] In this method, the same DNA molecule can be resequenced by creating a circular DNA template and separating the newly synthesized DNA strand from the template. In the PacBio *RS* platform, the average read length is about 3000 bases and the run time is very short, about 20 min[4] Various other approaches are being tested, such as **transmission electron microscopy** to directly image single DNA molecules, and a **nanopore-based** single-molecule sequencing approach. The sequencing community has been eagerly waiting to get their hands on third-generation sequencing technology.

3.6 HIGH-DENSITY OLIGONUCLEOTIDE-PROBE-BASED ARRAY TO INVESTIGATE GENOME EXPRESSION

Microarray and global gene-expression profiling is a crucial genomic technology. The term **microarray** is often used synonymously with DNA microarray and high-throughput gene-expression measurement. However, it can also be used in the context of expression profiling of proteins, carbohydrates, and tissues. The current discussion on microarray will focus on gene expression. Gene-expression microarray is a nucleic-acid-hybridization-based technique. Studies on nucleic-acid hybridization were pioneered independently by Paul Doty and Sol Spiegelman and their colleagues. The DNA−RNA hybridization principles were utilized to develop a number of widely used techniques to study gene expression, such as in situ hybridization, Northern blot, and solution hybridization.[11] These techniques mostly measure the expression of a single gene in multiple tissues and at multiple time points. Before the advent of genomics, a number of techniques were also developed to analyze differential gene-expression profiles, involving a large number of samples, multiple target sequences (a large number of transcripts), and many tissues at the same time; for example, ribonuclease (RNAse) protection assay (RPA), subtractive hybridization, differential display, serial analysis of gene expression (SAGE), and branched DNA (bDNA) signal amplification technique.[12]

However, global gene-expression profiling was revolutionized with the advent of the microarray. In 1996, Affymetrix commercialized its oligonucleotide-based DNA chip under the proprietary name GeneChip®. A microarray can be either a complementary DNA (cDNA) microarray or an oligonucleotide microarray. Currently, high-density oligonucleotide microarray is the method of choice. In an oligonucleotide microarray, an array of oligonucleotide probes (usually 20−80-mer) are synthesized either on-chip (on the platform) or by conventional synthesis followed by immobilization on the platform. An example of on-chip synthesis of oligonucleotides is the photolithographic technique, which is used by Affymetrix (Figure 3.5A). Another related technology uses an ink jet to spray oligonucleotide probes on the microarray. The fabrication of an oligonucleotide array is carried out by high-speed robotics. These robots rely on pins or needles to transfer the sample from a reservoir to the platform. The

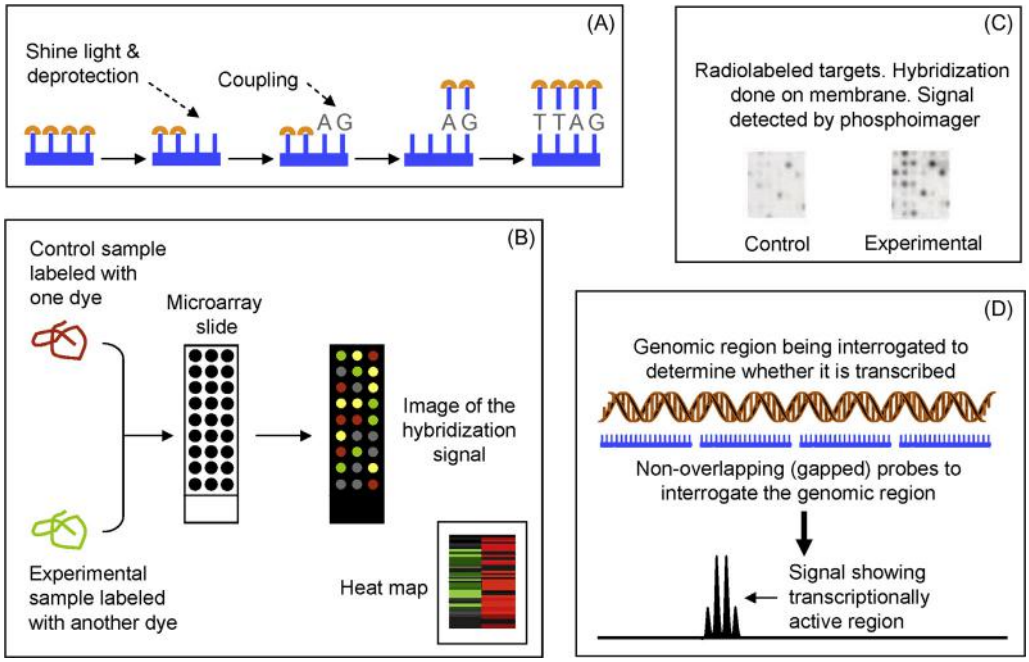

FIGURE 3.5 **High-density oligonucleotide-based array.** (A) Microarray fabrication by photolithographic synthesis, which involves repeated cycles of targeted deprotection, coupling, and protection of the coupled bases. (B) Microarray using fluorescent-dye-labeled targets and competitive hybridization of the two probes on the same array slide. The inset shows what a heat map could look like. (C) Microarray using radiolabeled targets. (D) Use of tiling array to identify a genomic region that was previously not known to be transcriptionally active.

pin diameter and shape, solution viscosity, and platform characteristics determine the volume transferred and how far the solution will spread. The number of spots on the microarray can vary between a few thousand to 30,000 on a 25×75-mm slide, each spot representing the product of a specific gene, and is generated by depositing between 1 and 10 nl ($1\ \text{nl} = 10^{-3}\ \mu\text{l}$) of PCR product representing that specific gene, usually at concentration of 100–500 μg/ml. The spot diameter can be between 75 and 200 μm, and the distance between spots is about 200 μm.[11] In a cDNA microarray format, customized cDNA probes are immobilized on a solid surface (glass or nylon membrane). The DNA fragments can be PCR amplified or be library clones. Thus, the array density is lower than in DNA chip, and the spotted cDNAs are longer than oligonucleotide probes.

To detect gene expression, the microarray is hybridized with the labeled **target**, which is the reverse-transcribed copy of the mRNA. The mRNA-derived cDNA is labeled, in most cases by fluorescent dyes, such as Cy3 and Cy5. Purified poly(A)$^+$ mRNA is usually recommended as the starting material for improving the signal/noise ratio—that is, for increased sensitivity and low background. Hybridization spots containing fluorescent dyes are detected by laser scanning of the microarray. The laser scanner is hooked to a confocal microscope and a CCD camera. The fluorescent tags are excited by the laser, while the microscope and the camera work together to create a digital image of the array. The results are then analyzed using special analysis software (Figure 3.5B).

For cDNA microarrays spotted on nylon membrane, the target cDNA population is radioactively labeled. Radiolabeled hybridization spots can be detected and analyzed by a phosphoimager (Figure 3.5C). Differences in the expression of specific sequences can be further validated using other conventional methods, such as Northern blot, reverse transcriptase-polymerase chain reaction (RT-PCR), RNAse protection assay, or bDNA assay.

Microarray data can be transformed into a colored graphical representation, the so-called **heat map** (Figure 3.5B inset). In the heat map, increased expression is displayed by the intensity of a certain color (such as red), whereas decreased expression is displayed with another color (such as green), and a third color (black, the absence of other colors) may represent no changes in expression pattern.

3.6.1 Tiling Array as a Versatile Tool to Interrogate the Whole Genome

A tiling array is an oligonucleotide-based whole-genome microarray, and has proved to be very useful for whole-genome functional analysis beyond simple

gene-expression profiling. Because the tiling array is a variation of the microarray, it is conducted in the same way as a regular expression microarray, the main difference being the probe design. Tiling arrays probe for known contiguous sequences, such as a genomic region whose expression is not known. The resolution power of tiling arrays depends on the probe design—that is, whether the probes are spaced apart (gapped) or overlapping.

Whole-genome tiling arrays can be used for the interrogation of genomic regions for transcription, antisense transcription, and alternative splicing; interrogation of transcription-factor-binding sites and genomic polymorphism, and mapping of genomic methylation sites; and comparative genomic hybridization (CGH).[13,14] Figure 3.5D shows just one application of the tiling array, how a tiling array can be used to detect a region of the genome that was not previously known to be transcriptionally active. Tiling arrays designed to detect SNPs utilize overlapping probes so that every base is interrogated for mutation. The number of oligoprobes used in a whole-genome tiling array can be many millions. For example, in order to comprehensively identify coding sequences in the human genome, Bertone et al.[15] used genome tiling arrays by designing about 52 million oligoprobes (36-nt long) positioned every 46 nt, on average. These probes cover 1.5 Gb of nonrepetitive genomic DNA, both sense and antisense strands.

Tiling array platforms are designed and fabricated in the same way as the regular expression microarray platforms described above.

3.7 GENOME-WIDE MUTAGENESIS, GENOME EDITING, AND INTERFERENCE OF GENOME EXPRESSION

The best way to study the function of a gene is to silence its expression and analyze the resulting phenotype. The principal method of silencing the expression of a gene is **gene targeting** (**gene knockout**) by homologous recombination in embryonic stem (ES) cells. Using homologous recombination, a specific genetic locus can be disrupted (knockout) or replaced with another functional open reading frame (ORF) (knock-in) in ES cells of mice. By replacing the endogenous mouse gene with a human ortholog, a humanized mouse model can also be produced. The targeting construct contains an expression cassette that is flanked by two long stretches of genomic DNA. These two stretches of genomic DNA, called **homology arms**, have the same sequence as that of the genomic DNA flanking the target locus. Thus, the homology arms facilitate recombination and integration of the construct into the locus, thereby disrupting the endogenous ORF (Figure 3.6). The gene-targeting technique is limited to the generation of mouse models because it requires knowledge of the ES cells in which the targeting is done to mutate the gene. Currently, the biology of mouse ES cells is well understood. As a result, gene knockout models are mouse models, and this technique cannot be routinely performed in other animal models.

The only organism where systematic targeting of a vast number (96%) of the annotated ORFs has been

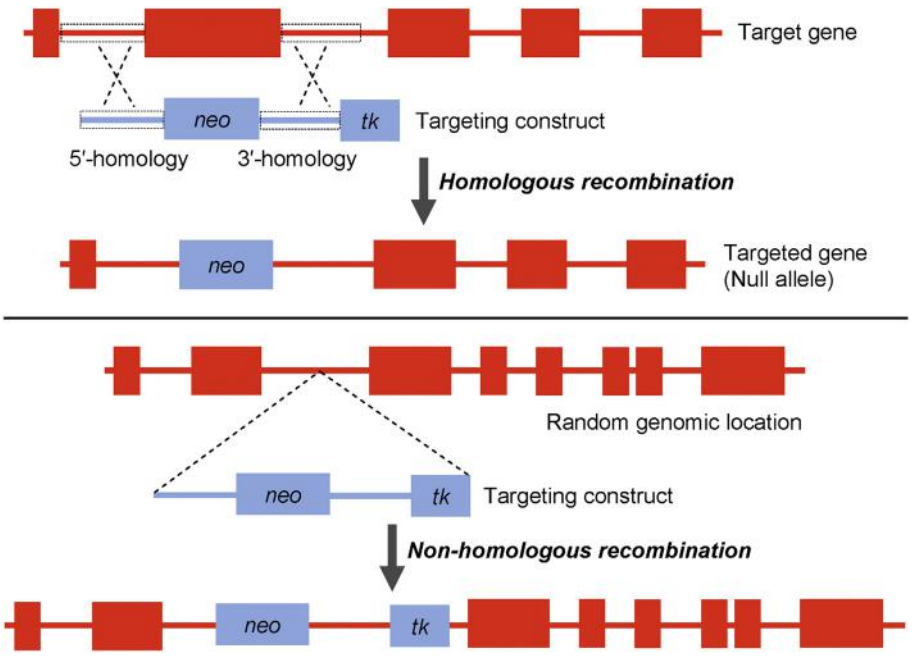

FIGURE 3.6 **Gene targeting.** The upper panel shows the generation of a null allele through gene targeting. The targeting construct is integrated through homologous recombination, which has a low frequency. In homologous recombination, the thymidine kinase (*tk*) gene, which is a negative selection marker, is not integrated. Only the *neo* gene, which is the positive selection marker, is integrated through legitimate recombination. The lower panel shows the random integration of the entire targeting construct by nonhomologous recombination, which has a higher frequency than homologous recombination.

achieved is yeast (*Saccharomyces cerevisiae*).[16] Each ORF was precisely targeted and replaced by mitotic recombination with the *KanMX* targeting cassette. The *KanMX* gene (which confers kanamycin resistance) in each cassette is flanked by yeast sequence that facilitates recombination and integration of the cassette in the yeast genome; in addition to the yeast sequence, the *KanMX* gene is also flanked by two distinct 20-nt sequences that serve as molecular barcodes to uniquely identify each deletion mutant.

Such an achievement could be a reality even for the mouse a few years from now. The International Knockout Mouse Consortium (IKMC) has been working to mutate all protein-coding genes in the mouse using a combination of gene trapping and gene targeting in C57BL/6 mouse ES cells.[17] **Gene trapping** is an insertional mutagenesis technique that randomly generates ES cells with well-characterized mutations. A gene-trap vector construct, called the **trap cassette**, contains a promoterless reporter cassette (such as *lacZ*). There is an upstream splice acceptor site and a downstream poly(A) sequence in the trap cassette. The splice acceptor sequence is not bypassed by the RNA-splicing machinery. The trap cassette reporter is used to identify the ES cells where the gene-trap construct is integrated. The gene-trap construct can be electroporated into the ES cells, or delivered using a retroviral vector. In some ES cells, the construct will be correctly integrated in an intron to produce incorrect splicing of the target gene, such that all exons downstream of the insertion site are not expressed. The endogenous functional promoter of the target gene will drive transcription producing fusion transcripts. The fusion protein translated from the fusion transcript provides a means of rapid identification of the disrupted gene. The targeted gene is identified by sequencing of the transcribed product. Figure 3.7 shows the gene-trap technique.

The limitations of classical gene targeting could soon be overcome by **zinc-finger nuclease** (**ZFN**) or **TAL effector nuclease** (**TALEN**) technology. A Zn finger is a small protein structural motif that has a Zn ion in a coordination complex with either four cysteines (Cys_4) or two cysteines and two histidines (Cys_2His_2) to stabilize the so-called finger-like fold (Figure 3.8 inset). A large class of transcription factors containing a Zn finger bind to the major groove of DNA through their Zn-finger DNA-binding domains; each domain actually recognizes a specific trinucleotide sequence in the DNA. A ZFN is an engineered synthetic protein that consists of an engineered Zn-finger DNA-binding domain fused to the cleavage domain of the *Fok*I restriction endonuclease. *Fok*I is a type IIS restriction endonuclease. Type IIS restriction endonucleases cleave the DNA outside of the recognition sequence, to one side. *Fok*I recognizes an asymmetric nucleotide sequence and cleaves one strand 9 nt downstream and the other strand 13 nt upstream of the recognition site, as follows: 5′-GGATG$(N)_9$▼-3′/ 3′-CCTAC$(N)_{13}$▲-5′. The *Fok*I cleavage domain induces double-strand breaks (DSBs) in specific DNA sequences, which triggers DNA repair. Eukaryotic cells repair

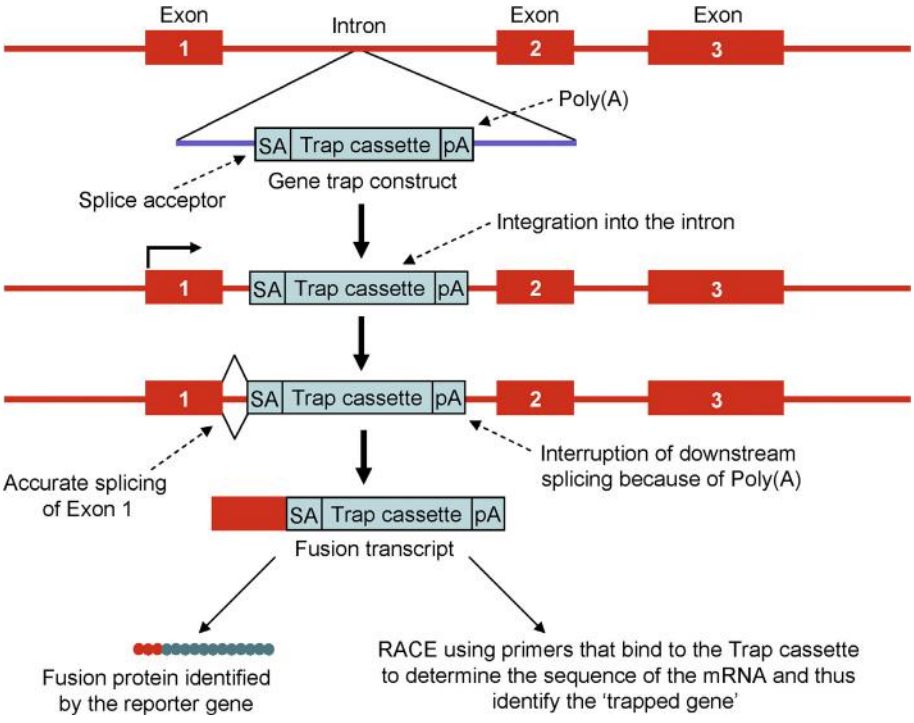

FIGURE 3.7 **Gene trapping is an insertional mutagenesis technique.** Random insertion of the trap cassette in the genome generates ES cells with well-characterized mutations. The trap cassette reporter is used to identify the ES cells where the gene-trap construct is integrated. Rapid amplification of cDNA ends (RACE) using trap-cassette-specific primers is employed to identify the trapped genes in the ES cells. Where the construct is correctly integrated into an intron, this produces incorrect splicing of the target gene, such that all exons downstream of the insertion site are not expressed.

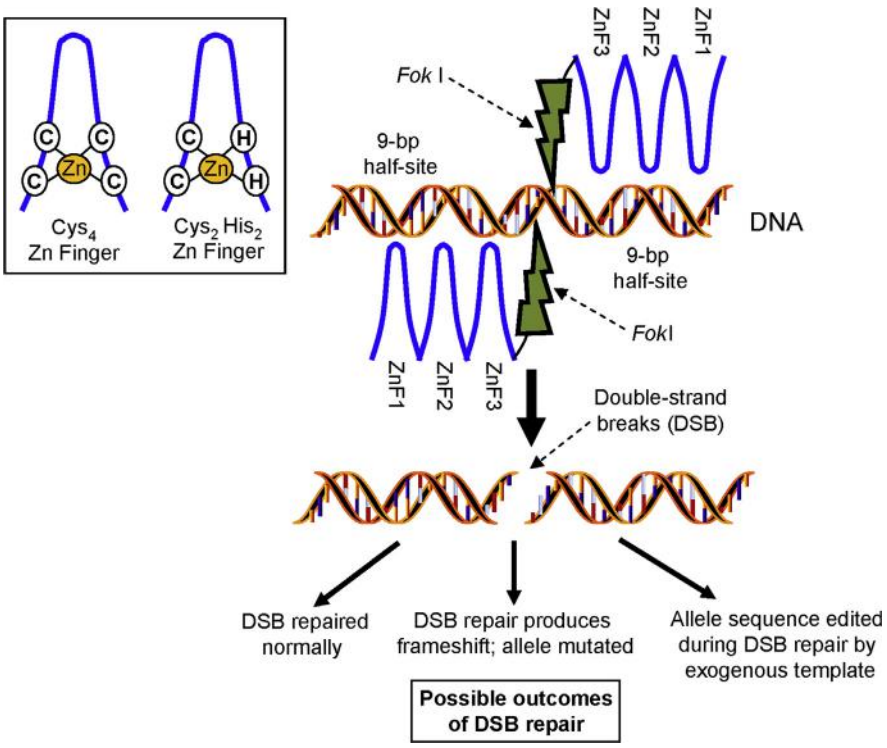

FIGURE 3.8 **Gene and genome manipulation using Zn-finger nuclease.** The figure shows a pair of ZFNs bound to their target site. Three Zn-finger domains are marked ZnF1, 2, and 3. Each three-finger array binds to a 9-bp half-site and is associated with a *Fok*I nuclease domain. A ZFN pair cleaves its target site within the variable-length spacer sequence between the half-sites. There are three possible outcomes of the DSB repair. The inset shows two types of Zn-finger motifs, a Cys_4 and a Cys_2His_2 motif.

DSBs using homology-directed repair (HDR) or non-homologous end-joining (NHEJ) pathways, and these repair pathways can be utilized to edit the genome. For example, by providing template (homologous) donor DNA along with ZFNs for HDR, information encoded on the introduced template can be used to repair the DSB, and in that process some nucleotides can be changed (gene editing including correction), or it is even possible to add a new gene at the site of the break. The NHEJ repair pathway ligates the two broken ends, with occasional small insertions or deletions at the site of the break, resulting in frame-shift and disruption of the target gene (Figure 3.8). Thus, the genome-editing function of ZFNs is based on the introduction of site-specific DNA DSBs into the locus of interest. By fusing *Fok*I to different types of Zn fingers that recognize different trinucleotide sequences, the ZFNs can be targeted to different parts of the genome for desired genome editing. ZFN technology has been successfully used to manipulate the genomes of many plant and animal species.

One of the major achievements of ZFN technology has been the generation of gene knockout models in species other than mice, which was not possible using the standard gene-targeting technique. By microinjection of ZFNs designed to target an integrated reporter and two endogenous rat genes, immunoglobulin M (*IgM*) and *Rab38*, in a one-cell rat embryo, successful gene targeting was reported. A high frequency of animals had 25 to 100% disruption at the target loci and these mutations were faithfully and efficiently transmitted through the germline. Transcription-activator-like effector nuclease (TALEN) technology is similar to ZFN technology. The main difference is in the DNA-targeting protein, which is the TAL effector (TALE) protein. The TALE protein can be fused to *Fok*I to generate the TALEN. Unlike ZFN and TALEN that are protein-guided genome editing tools, CRISPR-Cas system is a RNA-guided genome editing tool. CRISPR stands for Clustered Regularly Interspaced Short Palindromic Repeats, and Cas is CRISPR-associated nuclease. Target recognition by Cas nuclease requires a ''seed sequence'' within CRISPR RNA (crRNA) that acts as a guide to Cas. Thus, almost any DNA sequence can be targeted by redesigning the crRNA seed sequence. In prokaryotes, the CRISPR-Cas system acts as RNA interference (RNAi, discussed in the following section) based immune system to defend against invading viral DNA because the short crRNAs that guide the recognition of targets for degradation are produced by the processing of a long transcript.[18]

RNA interference (RNAi) is another way of **knocking down** (instead of **knocking out**) genome expression and studying the phenotype. In *Caenorhabditis elegans*, the effect of silencing gene expression on a large scale has been studied by multiple groups, who were able to study about a third of the predicted genes. Using a reusable RNAi library of 16,757 bacterial clones, Kamath et al.[19] were able to knock down the expression of about 86% of the 19,427 predicted genes. Each bacterial strain in the library was capable of expressing dsRNA

designed to correspond to a single gene. Mutant phenotypes for 1722 genes were identified; about two-thirds of these were not previously associated with a phenotype. Such genome-wide RNAi analysis has also been accomplished in *Drosophila*.[20] The authors applied an RNAi screen of 19,470 dsRNAs in cultured cells to characterize the function of nearly 91% of predicted *Drosophila* genes in cell growth and viability. Interestingly, the authors found 438 dsRNAs that identified essential genes, among which 80% lacked mutant alleles.

3.8 SPECIAL TOPIC: OPTICAL MAPPING OF DNA

Michael L. Kotewicz, Ph.D., Office of Applied Research and Safety Assessment, CFSAN, FDA

3.8.1 Introduction

In chromosomes, which range from 1–6 million bp in bacteria to 100 million bp in humans, what graphic software tools allow one to locate and distinguish details as small as single nucleotide polymorphisms, mid-sized chromosomal changes (10,000–200,000 bp), and inversions across millions of base pairs? No graphic tool, to date, performs ideally at both these extremes. One software tool well suited for the fine-scale mapping of nucleotides and detailed chromosome alignments is **Mauve**.[21] Mauve and the updated **progressiveMauve** are extremely powerful desktop graphic tools for aligning chromosomes and defining both homologous genome segments and single-nucleotide differences. At the opposite scale, the graphic software in **MapSolver™** was designed to work with optical maps of chromosome restriction fragments and *in silico* sequence-based maps of reference bacterial chromosomes. MapSolver's strengths are its easy graphic ability to ramp up and down thousands and millions of base pairs and to detail differences in aligned optical maps and reference *in silico* chromosome maps.

Optical maps are physical maps assembled from overlapping restriction-fragment maps of long chromosomal pieces, and they represent a sample of the sequence across the complete chromosome. For each restriction fragment, the cut site at the beginning of the fragment and the cut site at the end score the presence of these sequence pairs; for example, a *Bam*HI map scores GGATCC pair sets in the chromosome as well as measuring the nucleotide distance between those sequence pairs. The map could be considered a digital chromosome. Within the limits of fragment size measurements, 1–2%, where sets of fragments in a new isolate's optical map align to fragments from a reference sequenced genome, there is a direct correlation of the map fragments with the reference sequences and genes in those fragments. The alignment scores represent the strength of the correlation of map and sequence, where the limit of detection for differences such as insertions and deletions is 1–5 kb in the optical map. The optical mapping software optimally presents a simple graphic, best suited to detect, measure, and display chromosome differences from about 5000 to millions of bp. Differences created by events such as close-proximity multiple prophage insertions can span 300,000 bp and complex multiple inversions can span several million bp. In contrast, Mauve is like a street map, detailed to seeing single nucleotide addresses, and just as one would not use a street map to find continents on a globe, Mauve is not quite as well suited for rapidly determining and viewing these larger chromosomal differences; something that optical maps do extremely well. What optical maps lose in terms of resolution and nucleotide detail, they make up for in ease of use and perspective. It is worth testing a set of alignments in both software packages and comparing the advantages and limitations of each for examining chromosome differences (Figure 3.9). Mauve gives a sequence-based segmental view of compared chromosomes, while MapSolver™ gives a difference-based alignment of restriction fragments. For maps, the sequence information is correlated, albeit indirectly, with sequences in aligned reference fragments.

3.8.2 Optical Maps

Optical maps are physical maps generated from long chromosomal DNA preparations attached and restriction digested on surfaces. For a number of reasons—including G/C content, average fragment size generated for a given genome, and overall number of cuts—optical maps are usually generated using six-base-cutter restriction enzymes, such as *Bam*HI (GGATCC) or *Nco*I (CCATGG), although there is some flexibility in enzyme of choice. In addition to displaying the physical DNA maps, MapSolver™ software is used to generate reference *in silico* maps from sequence data. These annotated reference genomes are used to define the differences found in comparative alignments with optical maps.

There is an additional use for MapSolver™: higher resolution mini-maps, usually generated on shorter DNA sequences ranging from 5000 to 1 million bp using more-frequently cutting restriction enzymes, such as four-base-recognition enzymes. These mini-maps are useful in several regards. One is for comparative genomic studies determining the structures of chromosomal variations. The other is for the rapid display of sequencing misassemblies. Initially, mini-maps were conceived as allowing a more detailed map to be constructed by sub-cutting sites within larger fragments of *in silico*

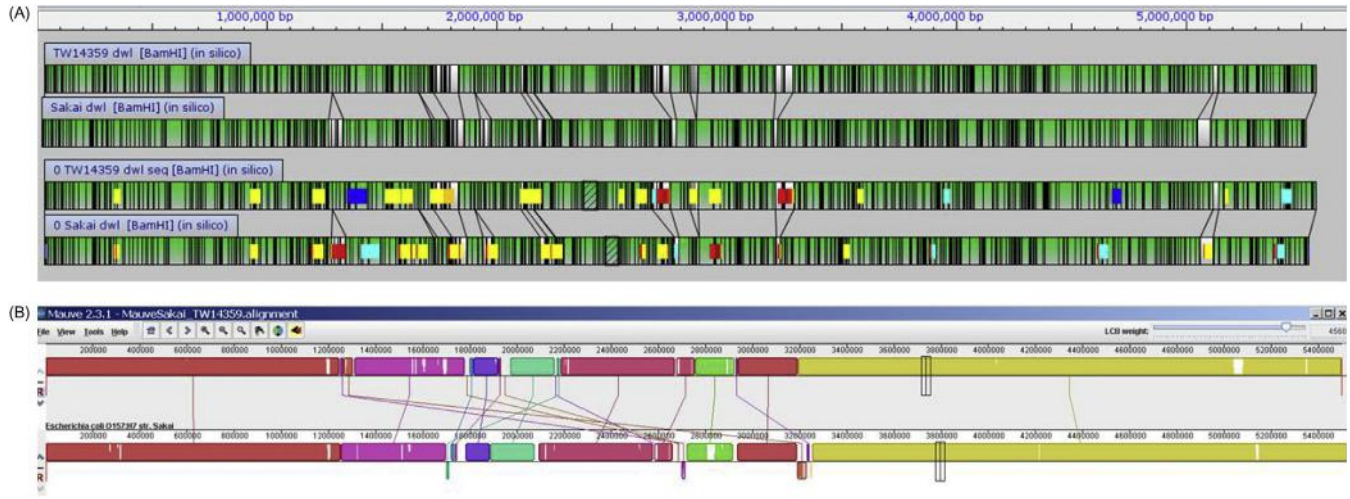

FIGURE 3.9 **The alignment of the *in silico* optical maps of two related strains of *E. coli* O157:H7: TW14359 from the 2006 US spinach-associated outbreak, and Sakai, the Japanese outbreak associated with sprouts.** (A) Two pairs of aligned maps using MapSolver™, the non-aligned regions of the chromosomes are white, aligned regions are green; in the lower aligned pair, regions of interest have been "painted" from the sequence-based annotations. Prophages are yellow/orange, prophages carrying the Shiga toxin genes are red, and pathogenicity islands are blue. (B) Mauve alignment of the same two sequenced chromosomes, where similarly colored sections reflect sequence matches; note white streaks within colored boxes, indicating short unaligned sequences within larger aligned sequence blocks.

*Bam*HI (GGATCC) maps with *Sau*3AI (GATC) fragmentation. Dr. David Lacher has refined mini-mapping in our laboratory. He noted that *Sau*3AI produces a much more heterogeneous mixture of large and small fragments, and that other four-base cutters such as *Hha*I (GCGC) and even other six-base cutters such as *Hpa*I (GTTAAC) provide a more evenly distributed, higher density set of fragments in these *in silico* mini-maps, especially for *E. coli*. For example, the six *Bam*HI fragments for the 112-kb TW14359 *yehV* prophage region produce a *Sau*3AI mini-map with 344 fragments; *Hha*I produces a much more homogenous set of 725 fragments that yields better coverage of differences. The *Hha*I mini-map of the *yehV* region of TW14359 and Sakai (Figure 3.10) shows the detail of two 1.3-kb insertion/deletions (indels) in the left flanking chromosomal DNA outside the prophages. The mini-map clearly shows two distinctive differences within the two *yehV* prophages, but in addition the mini-map details another 1.3-kb indel, a 12.6-kb region containing Shiga toxin genes in Sakai, and a quite different, unaligned 14.5-kb set of fragments, hence different sequence, in TW14359. The remaining 28 mini-map fragments (7.0 kb) are homologous in the two prophages, delineating the variant Shiga toxin region within otherwise homologous regions.

Optical mapping is a also a corroborative technology for sequencing; it is independent of amplification technologies, and importantly, mistakes in DNA assemblies are readily identified, notably across ribosomal RNA and repeated conserved regions of multiple prophages.[22] It is also a complementary and refining technology for traditional low-resolution pulsed-field gel electrophoresis (PFGE) analysis, the gold standard for bacterial epidemiological identification.[23] A contiguous 600-fragment map locates chromosomal markers, and it greatly exceeds the 40-fragment resolution of PFGE. Most importantly, the optical maps define the contiguous relationships of all the fragments, while PFGE gives no direct band correlation with chromosomal position. Optical mapping accurately identifies both large fragments not resolved by PFGE and small fragments not detected by PFGE (Figure 3.11).

Optical maps are fundamentally shorthand representations of the sequences of chromosomes generated by mapping restriction-enzyme cut sites; they are reflections of whole-chromosome sequences. For a typical bacterial chromosome of 4−6 Mbp, six-base-recognition restriction enzymes such as *Bam*HI (GGATCC) or *Nco*I (CCATGG)—for *Escherichia coli* and *Salmonella enterica* isolates—generate a map with 400−600 contiguous restriction fragments. Changes in genome sequences ablate or create cut sites, creating restriction fragment length polymorphisms (RFLPs). More importantly, differences in chromosomes between related strains generate changes in the sizes and distribution of fragments that light up in aligned maps. Optical mapping allows the rapid construction of ordered restriction fragment maps for chromosomes that can be as small as 150−200-kb bacterial plasmids, but optical maps are optimally suited for detecting differences in chromosomes of bacteria which range from 1−10 million bp. Overall, the 5-Mbp chromosomes of bacteria can be sized to within 10−20 kb, an accuracy of about 0.1 to 0.3%.[24] Whereas single nucleotide polymorphisms are

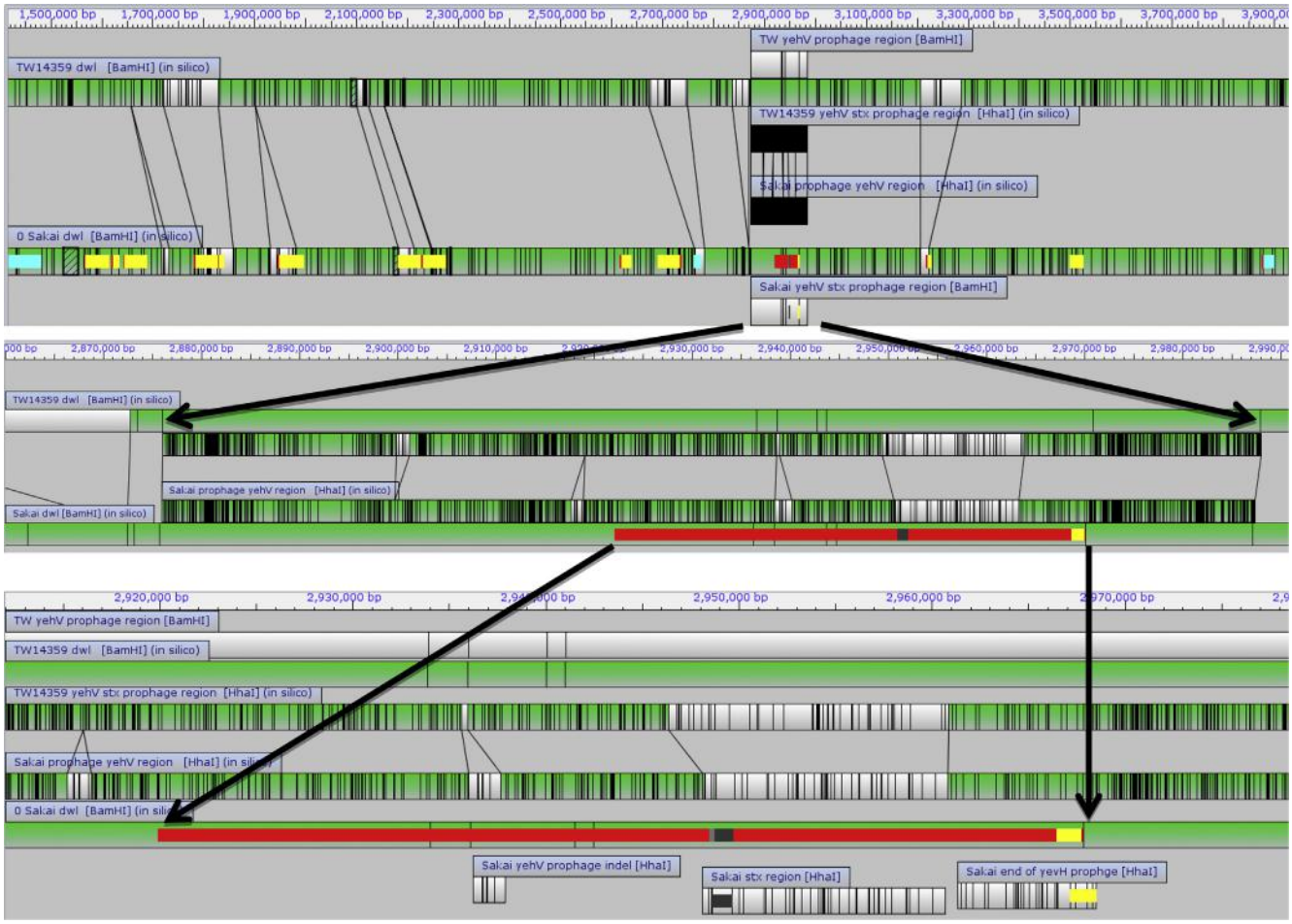

FIGURE 3.10 **Mini-maps: six-base cutter *Bam*HI (GGATCC) versus four-base cutter *Hha*I (GCGC).** Three successively enlarged MapSolver™ views of the *yehV* prophages.

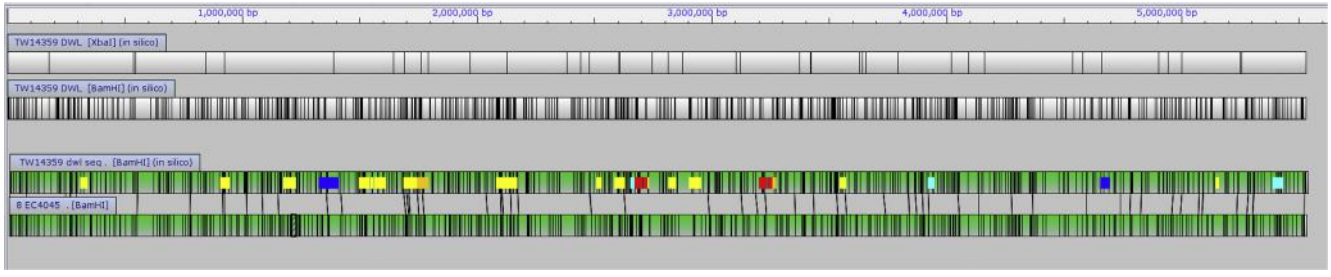

FIGURE 3.11 **Optical limit of detection.** Upper two unaligned maps: *Xba*I (42 fragments) versus *Bam*HI (642 fragments) *in silico* TW14359 maps; lower two maps: aligned painted *in silico* (642 fragments) versus optical map (529 fragments) of spinach-outbreak strain, isolates TW14359 and EC4045. A total of 113 fragment differences are in small fragments, 21 to 1000 bp, at the optical limit of detection.

crucial for differentiating highly clonal *Salmonella* isolates, *Escherichia coli* strains, particularly pathogens such as *E. coli* O157:H7 isolates, differ by prophages and insertions and deletions.[25]

There are two other related technologies for determining the structure of chromosomes with comparable mid to long molecule resolution, one involving fluidic separation of large DNA molecules from Pathogenetix, Woburn, MA, and the other involving nanochannel fluidic chips that spread out confined native long genome fragments labeled at restriction-enzyme-nicked sites with fluorescent tags, from BioNano, San Diego, CA. This discussion is focused on optical mapping using hardware (the Argus mapping station) and software (MapSolver™)

for comparative genomics, from OpGen, Gaithersburg, MD.

There are a number of other nucleotide-based software packages for looking at long regions of DNA molecules, including DNASTAR's Lasergene and a number of retired (2007) Genetics Computer Group (GCG) available within the European Molecular Biology **Open** Software Suite, an open-source software analysis package at EMBOSS (http://helix.nih.gov/Applications/emboss.html). Software packages from next-generation sequencing companies are continually upgrading and although designed and useful for examining sequence contigs (consensus regions of DNA derived from sets of overlapping DNA segments) and assemblies and although not necessarily optimized for comparative genomics, they are moving in that direction.

3.8.3 Overview; Making an Optical Map

Optical maps have been generated for a wide range of bacterial species involved in industrial microbiology, clinical illnesses, and food-borne bacterial outbreaks, as well as for larger chromosomes from fungal and mammalian sources. For a bacterium, the optical map of its chromosome is generated by growing up cells from an isolate or a set of isolates and gently lysing them to release high-molecular-weight DNA (Figure 3.12A). The DNA molecules are loaded into carefully designed microfluidic channels (Figure 3.12B, in this case 40 channels in a 2×2 cm area. DNA molecules attach by charge interactions with the derivatized glass surface and distribute as long linear individual molecules onto the surface (Figure 3.12C).

The attached molecules are digested with an appropriate restriction enzyme and the DNA is stained with

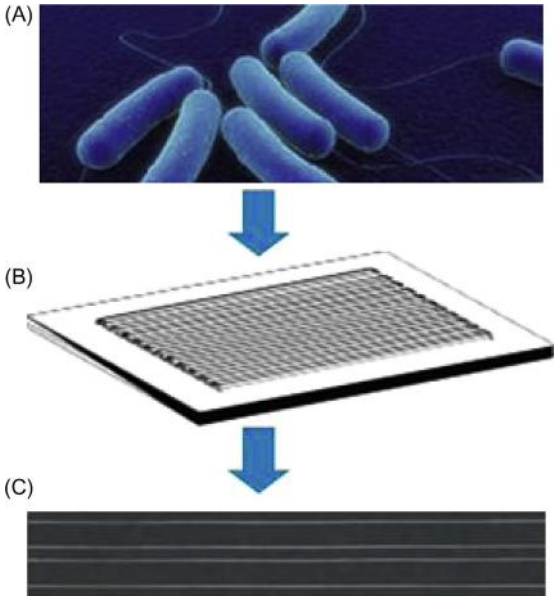

FIGURE 3.12 **Preparation of high-molecular-weight DNA.** (A) Bacterial cells prior to lysis; (B) forty-microfluidic-chamber device on coverslip; (C) DNA in one channel attached to derivatized glass surface.

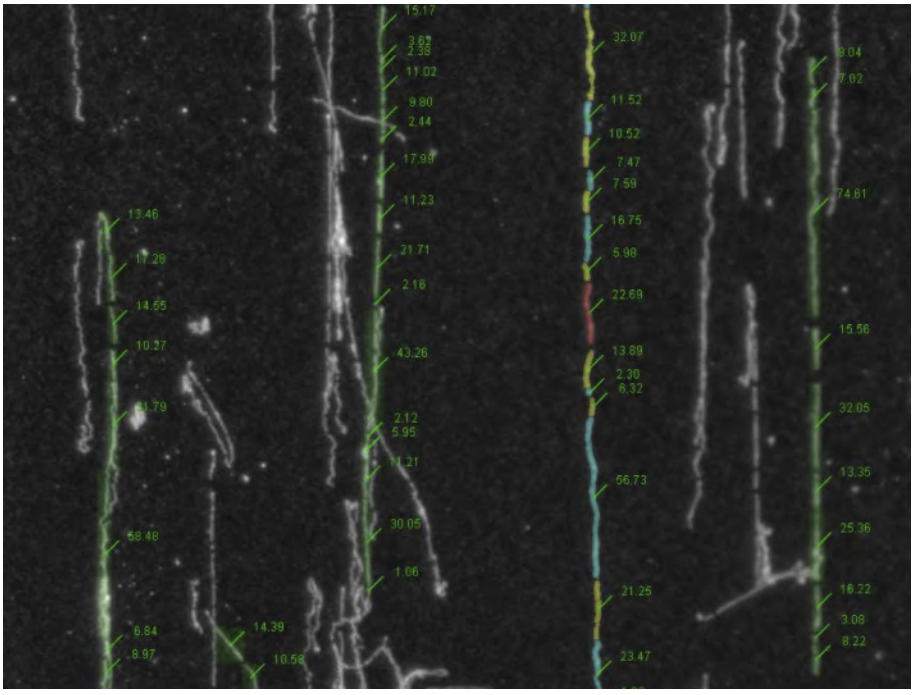

FIGURE 3.13 **Restriction-digested DNA attached to the cover slip surface as seen under the Argus microscope and assembly platform.** The image is from an assembly data set; molecular weights of fragments are indicated. The multicolored strand to the right of the figure center line is a molecule from the assembled map, for examination of details. Note the extent of linearity or wiggle in each restriction fragment and the gap sizes. These are some of the quality control parameters used to judge data sets.

the fluorescent dye JOJO-1. The salt conditions of the wash after staining cause the DNA to constrict a slight amount such that a small measurable gap is created at the cut sites, but the restriction fragments remain attached to the surface. Automated software is used to measure the sizes and positions of contiguous restriction fragments along thousands of chromosome-fragment molecules. Depending on the size of the genome of the organism being mapped, molecules are collected, each containing 10 to 100 contiguous restriction fragments. For example, 2000 to 50,000 molecules are usually collected for analysis for a 1 to 6-million-bp bacterial chromosome. The attached, digested DNA fragments range from 250 to 400 kb; some are as large as 1.5 Mbp. **The limit of detection of fragments is about 500 bp** (Figure 3.13).

The data from these thousands of molecules are assembled into complete genomic maps by overlapping same-sized fragment runs, similar to the assembly of overlapping DNA sequencing runs (Figure 3.14A). In these assemblies, the minimum coverage for each fragment is 30× (Figure 3.14B and C). The completed assemblies are usually oriented to a defined start reference or origin and scaled to a reference sequence.

3.8.4 Conclusions

Optical mapping provides information on the genome that cannot be obtained from PFGE profiles and a perspective very different from comparing whole-chromosome sequences. Optical mapping is a powerful tool for studying structural genomics because it provides a bird's eye view of chromosomal morphology and architecture. Consequently, optical mapping can be used to visualize and compare different genomes, such as genomes of related species/strains, as well as genomes of pathogenic and nonpathogenic strains within a bacterial species. Optical mapping can also be used to study the same genome in different states.

Since some of the first publications in 1993, optical mapping has been developed and extended from sizing restriction fragments on bacteriophage lambda and bacterial artificial chromosome (BAC) clones (48,500 to 150,000 bp), to scaffolding larger chromosomes such as those in *Candida albicans* (8 chromosomes, 16 Mbp),[26] *Plasmodium falciparum* (14 chromosomes, 23.3 Mbp),[27] rice (24 chromosomes, 389 Mbp),[28] maize (20 chromosomes, 2300 Mbp),[29] mouse (40 chromosomes, 2500 Mbp),[30] humans (46 chromosomes, 3000 Mbp),[31] and most recently the goat genome (60 chromosomes, 2900 Mbp).[32,33]

With its mid-range resolution and graphic flexibilities, optical mapping is ideal for the examination of whole

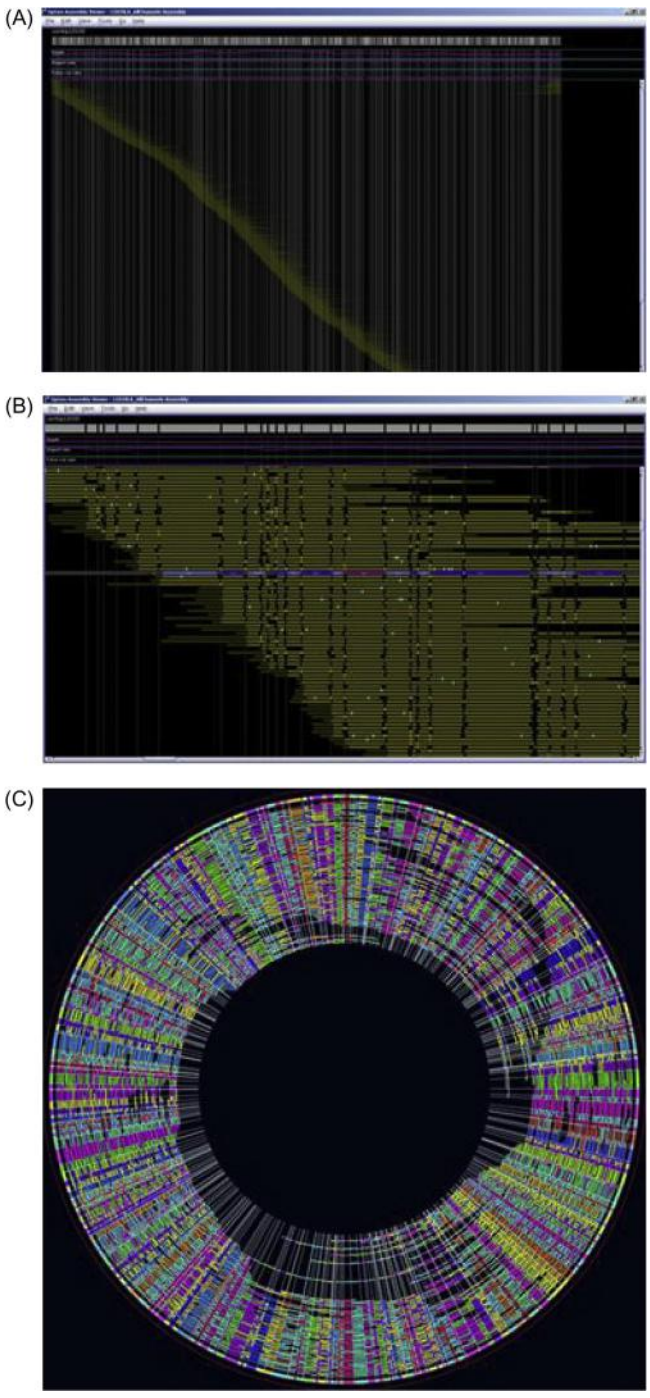

FIGURE 3.14 **Assembly of collected molecules.** (A) In a typical matching of an alignment of 1500 to 50,000 molecules, overlapping restriction fragments grow the chromosome until ends cease growth, or for circular chromosomes, until overlap to previous fragment sets occurs. (B) An enlargement of the overlapping molecule assembly. (C) A graphic representation of the like-colored fragments assembling, in this case into a circular chromosome. In all cases, a criterion of a minimum 30 molecules representation for each restriction fragment is set. More often, there are hundreds of fragments present for many assemblies, adding to the statistical reliability of fragment-size determinations.

chromosomes extending from viruses to humans, for independently validating sequence assemblies, for scaffolding higher-order 10–100-Mbp chromosome sequence contigs, and for rapidly detecting differences between the chromosomes of outbreak strains of bacteria.[34–36]

References

1. Sanger F, et al. *Proc Natl Acad Sci USA* 1977;**74**:5463–7.
2. Ronaghi M, et al. *Anal Biochem* 1996;**242**:84–9.
3. Wheeler DA, et al. *Nature* 2008;**452**:872–6.
4. Loman NJ, et al. *Nat Rev Microbiol* 2012;**10**:599–606.
5. 454 Life Sciences Corporation. *How is genome sequencing done?* Available online at: <http://www.454.com/downloads/news-events/how-genome-sequencing-is-done_FINAL.pdf>.
6. Mardis ER. *Annu Rev Genomics Hum Genet* 2008;**9**:387–402.
7. Illumina. *Hist Illumina Seq* 2013.. Available online at: <http://www.illumina.com/technology/solexa_technology.ilmn>.
8. Applied Biosystems. *Applied biosytems SOLiD 4 system*; 2010. Available online at: <http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_078637.pdf>.
9. Schadt EE, et al. *Hum Mol Genet* 2010;**19**(Review Issue 2):R227–40.
10. Pacific Biosciences. *SMRT Technol* 2013. Available online at: <http://www.pacificbiosciences.com/products/smrt-technology/>.
11. Choudhuri S. *J Biochem Mol Toxicol* 2004;**18**:171–9.
12. Choudhuri S. *Toxicol Mech Methods* 2006;**16**:137–59.
13. Mockler TC, et al. *Genomics* 2005;**85**:1–15.
14. Yazaki J, et al. *Curr Opin Plant Biol* 2007;**10**:1–9.
15. Bertone P, et al. *Science* 2004;**306**:2242–6.
16. Giaever G, et al. *Nature* 2002;**418**:387–91.
17. Skarnes WC, et al. *Nature* 2011;**474**:337–42.
18. Gaj T, et al. *Trends Biotechnol* 2013;**31**:397–405.
19. Kamath RS, et al. *Nature* 2003;**421**:231–7.
20. Boutros M, et al. *Science* 2004;**303**:832–5.
21. Darling AE, et al. *PLoS ONE* 2010;**5**:e11147.
22. Latreille P, et al. *BMC Genomics* 2007;**8**:321.
23. Ribot EM, et al. *Foodborne Pathog Dis* 2006;**3**:59–67.
24. Kotewicz ML, et al. *Microbiology* 2007;**153**:1720–33.
25. Kudva IT, et al. *J Bacteriol* 2002;**184**:1873–9.
26. van het Hoog M, et al. *Genome Biol* 2007;**8**:R52.
27. Riley MC, et al. *Malar J* 2011;**10**:252.
28. Zhou S, et al. *BMC Genomics* 2007;**8**:278.
29. Zhou S, et al. *PLoS Genet* 2009;**5**:e1000711.
30. Church DM, et al. *PLoS Biol* 2009;**7**:e1000112.
31. Teague B, et al. *Proc Natl Acad Sci USA* 2010;**107**:10848–53.
32. Dong Y, et al. *Nat Biotechnol* 2012;**31**:135–41.
33. Mak HC. *Nat Biotechnol* 2013;**31**:123.
34. Zhou S, et al. *J Bacteriol* 2004;**186**:7773–82.
35. Chen Q, et al. *Microbiology* 2006;**152**:1041–54.
36. Kotewicz ML, et al. *Microbiology* 2008;**154**:3518–28.