1

An Introduction to Next-Generation Biological Platforms

VIRGINIA MOHLERE, WENTING WANG, AND GANIRAJU MANYAM

1.1 Introduction

When Sanger and Coulson first described a reliable, efficient method for DNA sequencing in 1975 (Sanger and Coulson, 1975), they made possible the full sequencing of both genes and entire genomes. Although the method was resource-intensive, many institutions invested in the necessary equipment, and Sanger sequencing remained the standard for the next 30 years.

Refinement of the process increased read lengths from around 25 to almost 750 base pairs (Schadt et al., 2010, fig. 1). Although this greatly increased efficiency and reliability, the Sanger method still required not only large equipment but also significant human investment, as the process requires the work of several people. This prompted researchers and companies such as Applied Biosystems to seek improved sequencing techniques and instruments. Starting in the late 2000s, new instruments came on the market that, although they actually decreased read length, lessened run time and could be operated more easily with fewer human resources (Schadt et al., 2010).

Despite discoveries that have illuminated new therapeutic targets, clarified the role of specific mutations in clinical response, and yielded new methods for diagnosis and predicting prognosis (Chin et al., 2011), the initial promise of genomic data has largely remained unfulfilled so far. The difficulties are numerous. The functional consequences of individual mutations are not always clear. In fact, it is often logistically challenging to determine which discovered mutations make a critical contribution to disease and which are due merely to genetic instability and confer little functional effect.

In part, these difficulties lie in the methods used to acquire data. Microarray plates started to replace the labor-intensive Sanger method in the mid-1990s (Schena et al., 1995). These plates consist of many small wells that contain probe sets (e.g., up to 54,000 on the Affymetrix GeneChip

[www.affymetrix.com]), or stacks of bases. The target sequence is fluorescently labeled and washed onto a chip; levels of matching sequences are then analyzed by a laser, and the signal from laser indicates the amount of gene expression. Depending on how the data are measured and then analyzed, several metrics can be determined, including the concentration of a particular gene's mRNA transcript at a discrete point in time; differences in expression of the same gene among many samples; or differences in phenotype, reaction to a particular treatment, or prognosis that arise from differences in expression levels among samples (McGee and Chen, 2005).

The ability to place large numbers of probes on one chip, and later the availability of standard commercial microarray chips, greatly decreased the cost of expression assays. They are not, however, without their drawbacks. For example, to construct the probe sets on the microarray, the genome of the organism studied must be well characterized. Also, microarray data are obtained from sequences hybridized to the probes stuck to the plate, and this process can introduce errors, not only because of unreliable probes but also because of cross-hybridization of imperfectly matching target sequences. Methods that require samples to be amplified by polymerase chain reaction (PCR) might introduce unavoidable errors not in the original sample, and these are not easy to determine. Also, because microarray data are gathered by measuring the fluorescence signal, both very rare and very common signals (those that are very faint and those that are very bright) near the detection limits of the assay at either end cannot be measured accurately (McCormick et al., 2011).

To overcome these limitations, research has continued to find more efficient ways to quantify biomolecular data. This has given rise to next-generation sequencing (NGS), also called high-throughput sequencing. These methods measure single molecules of DNA or RNA using methods, such as nanopores, described later in this chapter. Such technologies aim to overcome the limitations of previous methods by generating millions of short reads to provide detailed views of cellular activity at nucleotide resolution. "Short," in this case, means that sequences that are generally read are 18–25 nt long. This length serves two purposes: first, it is easier and cheaper to gather shorter sequences; second, many small DNA and RNA elements are known to be within this size range, so they will be captured at this length (McCormick et al., 2011). These reads are then assembled into longer sequences.

However, using short sequences runs the risk that each read might map to more than one site in a given genome. To ensure that the reads are generated with good quality, many copies are run with slightly overlapping ends. The number of repeats required to ensure correct mapping is called "coverage," and

experience has indicated that the convergence between accuracy and efficiency occurs at about 28–30× coverage (McCarthy et al., 2012).

The direct assessment enabled by NGS will not only reduce some kinds of introduced sequencing errors by methods such as PCR, but also provide information about catalysis and DNA processing that might otherwise be masked by interim amplification steps (Schadt et al., 2010). Importantly, NGS techniques can quantify the abundance of molecules based on the read count, or so-called digital signal, in contrast to the "analog signal" measured by array techniques. Future enhancements of these methods also hold the potential to increase read length into thousands of bases and to decrease the time to results to mere hours – both of which would also decrease the overall cost.

Each platform of NGS data – whole genome, miRNA, methylation, and so forth – represents a different kind of data and is quantified differently. One of the goals of NGS is to combine many platforms. The goal of this volume is to provide integrated models that can assess large sets of diverse biological data and still provide meaningful results.

The major NGS platforms include the following:

- The epigenome: changes in transcription that do not affect the original DNA strand, such as methylation and histone changes
- The genome: the entire DNA sequence
- The exome: only genes transcribed by RNA
- The transcriptome: RNA-based platforms and those assessing proteins

Examples and detailed descriptions of some platforms are presented in the following sections. Despite the differences in the type of data produced by each of these platforms, given that they are all single-molecule–based, the form of the data are often largely the same (McCormick et al., 2011). These NGS data are often described in four levels. Level 1 is the raw data file. Level 2 data have been processed and normalized – that is, images have been converted to "reads," or sequence fragments. Poor-quality signals have been removed, and sequences have been mapped by aligning them to a reference sequence. Level 3 data have been interpreted, and level 4 data have been summarized. Unfortunately, there are few standards for processing high-throughput data, which of course leads to the risk of false comparisons if similar data have been analyzed and interpreted differently (Martens et al., 2011). One of the purposes of this volume is to suggest analyses that might lead to such standardization.

Like previous sequencing methods, NGS has its own caveats. For example, nanopore technologies sometimes result in a nucleotide becoming stuck as it passes through the pore and thus being counted by the scanner more than once

(Schadt et al., 2010). In other methods, a reagent may not bind to every target sequence, decreasing the signal strength. All NGS techniques produce short sequence reads that might map to more than one sequence of the reference genome, and multiple sequencing runs are needed to minimize this effect. This need for multiple runs increases the time and cost and remains a limitation of NGS. Additionally, NGS technologies result in enormous data sets that require a substantial investment in data storage and both computational and human effort to manage and analyze the information and derive meaningful results (Chin et al., 2011). These challenges drive the need for the analytical methods described in this volume.

1.2 The Biology of Gene Silencing

Gene silencing describes the process of inhibitory gene expression regulation at various levels: the genome, epigenome, and transcriptome. DNA regulatory elements and transcription factors control gene expression at the genome level. DNA methylation inhibits gene expression via epigenome, whereas RNA interference is used to repress gene expression in the transcriptome. This section describes the molecular biology of repression through DNA methylation and RNA inference, as these processes are often used to elaborate other high-throughput methodologies in this volume.

1.2.1 DNA Methylation

DNA methylation is a normal biological process that plays an important role in the regulation of gene transcription. DNA methylation is an epigenetic change – one that affects gene expression but not the gene sequences. Epigenetic changes are often very stable (long-lasting) and can be inherited; a particular site can even be methylated in one cell and unmethylated in another (Das and Singal, 2004; Jabbari and Bernardi, 2004; Krueger et al., 2012). However, these changes are also reversible, making them attractive targets for therapy. Epigenetic changes have been found in many diseases and development processes, including cancer, viral infection, and developmental abnormalities such as X-inactivation (Das and Singal, 2004).

The DNA methylation process is a chemical change that adds a methyl group (CH_3) to the carbon 5 position of a cytosine pyrimidine ring or to the 6 position of an adenine purine ring. These mostly occur in the cytosine sequence identified by 5′CG3′. This is called the "CpG dinucleotide," because most CpG sites (in which a cytosine-C is located next to a guanine-G in the series of bases) are separated by one phosphate (p). This designation differentiates the

CpG – in which C and G are side by side – from the CG base pair (Lander et al., 2001). Among the 16 possible nucleotide combinations, the CpG dinucleotide should occur around 6% of the time, but its rate of occurrence is only a fraction of that expected rate (5%–10% of it) (Antequera and Bird, 1993). This low frequency is thought to occur because cytosine, when it is methylated, mutates easily, and the mutations are often identified and repaired (Daura-Oller et al., 2009). Thus CpG islands tend to cluster in unmethylated regions of the genome. On average, these occur about every 100 bp (Antequera and Bird, 1993; Cross, 1995).

DNA methylation is powered by enzymes called DNA methyltransferases. At present, three families of DNA methyltransferases have been described in mammals. During embryonic development, DNA methyltransferases and mechanical regulators (e.g., methylation centers) strictly control methylation, which ensures that genes are expressed or silenced to drive correct cell differentiation (Laird, 2010).

The outcome of DNA methylation depends on its location: methylation in the promotor region of a gene always leads to decreased expression. In contrast, methylation in the transcribed region can have various effects (Laird, 2010). The actual mechanisms of the repression elicited by DNA methylation can involve either interfering with the binding sites of specific transcription factors (e.g., nuclear factor–κB, a protein found in almost all cell types) or direct binding to proteins that prevent transcription. Some types of cancer show characteristic patterns of DNA methylation disruption. Aberrant DNA methylation that contributes to cancer development falls into two broad categories: hypomethylation and hypermethylation.

Hypomethylation has been found in numerous types of solid tumors, such as hepatocellular, cervical, and prostate cancer. It has also been noted in some forms of cancer affecting the blood-forming elements. The level of hypomethylation often increases with later progression of disease. Congenital hypomethylation is characterized by facial abnormalities, immunodeficiency, and instability of chromatin, the bundle of DNA and protein inside a cell nucleus. A decreased methylation rate is thought to enable the expression of some oncogenes, such as H-RAS, which is associated with bladder cancer and other types of cancer (Parikh et al., 2007; Kompier et al., 2010).

Far more common is hypermethylation. There are several pathways that protect against "runaway" methylation – chromatin blocking DNA methyltransferase, demethylation triggers in the cell, the timing of replication, and even transcription itself (Clark and Melki, 2002). These protective measures can be overcome, however, usually as a result of gene mutation. Genes known to be susceptible to changes that result in hypermethylation are involved in

regulating the cell cycle, DNA repair, drug resistance, angiogenesis (the formation of blood vessels), and metastasis – in other words, ubiquitous genes with critical functions (Das and Singal, 2004).

Different cancer types frequently show hypermethylation in type-specific genes, such as steroid receptor and cell adhesion genes in breast cancer (Yang et al., 2001). Hypermethylated genes have been discovered in association with leukemia, lung cancer (for which more than 40 are known; Tsou et al., 2002), and prostate cancer, among others. Ongoing research indicates that hypermethylation is associated with a broad range of disease characteristics and may be useful in predicting disease outcomes. Methylation is an active enough branch of research that a number of methods have been developed to study it. Some of the earliest methods were based on gel blotting and Sanger methods. Later, array-based techniques were created using methylation probes on chips. This allowed multiplexing of samples and brought methylation studies into the high-throughput era. These methylation-specific probes can now be used with NGS instruments for true single-molecule sequencing (see later; Laird, 2010).

1.2.2 RNA Interference

RNA interference (RNAi) is a process of gene silencing that occurs after gene transcription. The identification of RNAi has greatly advanced the study of gene function, and the mechanics of the process are being investigated for their therapeutic potential. Long strands of double-stranded (ds) RNA complementary to specific mRNA were found, first in plants and then artificially in mammalian cells, to silence genes via the action of very short segments (Fire et al., 1998; Elbashir et al., 2001). That the process occurs among plants, fungi, and animals indicates that RNAi is an ancient feature of gene regulation (Bagasra and Prilliman, 2004). RNAi is thought to be a natural protection system against virus-mediated gene expression and mutation (Malone and Hannon, 2009).

Broadly speaking, RNAi occurs in two steps. dsRNA is cut by the Dicer enzyme into short components between 21 and 25 nt in length, each of which has a 5′ phosphate group and 3′ overhangs of about 2 nt. The strand that is complementary to the mRNA target is called the guide strand, and the other is the passenger strand. The resulting fragments are then delivered by Dicer to the RNA-induced silencing complex (RISC), a mix of enzymes that further processes the fragment, separates the guide strand from the passenger strand, and directs the guide strand to bind with the mRNA target. This binding stops gene transcription, "silencing" the expression of the gene. When RNAi was first described, it was hoped that the process would prove to be a powerful therapeutic tool. However, this has not proved to be the case, for reasons

discussed later. It remains, however, a highly useful method for performing genetic manipulations and studying gene function (Berger and Randall, 2010).

Several molecules are known to associate with RISC and trigger RNAi: small interfering RNA (siRNA), microRNA (miRNA), and piwi-interacting RNA (piRNA) (Malone and Hannon, 2009; Sakurai et al., 2011). Each of these silences gene expression in a different way. piRNA is a relatively newly discovered small RNA about which little is known (Esteller, 2011). The features of siRNA and miRNA are described next.

siRNA

Small interfering RNA (siRNA) was first described in 1999 and was subsequently found to be about 21 nt long (Hamilton and Baulcome, 1999). siRNA, a product of RNA interference, plays an important role in gene silencing. siRNAs are produced from a dsRNA that has been cleaved by Dicer. These approximately 21-nt siRNA fragments, still in double-stranded form, are bound to RNAi nuclease (part of RISC). This complex is then catalyzed, the guide and passenger strands are split from one another, and the resulting siRNA is ferried by RISC to its target string (Bagasra and Prilliman, 2004). The siRNA sequences match perfectly (are homologous) to the target sequences. This suggests that siRNAs would be strong agents of gene repression, but this has not been demonstrated. The siRNA molecule is negatively charged, which contributes to the molecules being subject to breakdown by nucleases, its clearance by the kidneys, and "off-target silencing," or the silencing of genes other than the target. Off-target silencing occurs when the central region (usually 2–8 nt long) matches sequences in more than one gene (Berger and Randall, 2010). There is also evidence that siRNA inhibition does not last past transcription, making its gene-silencing effects short-lived (with a half-life of only minutes). These shortcomings are difficult to address. Some studies have shown that chemically modifying the sugar regions of siRNA molecules can reduce off-target silencing in individual sequences; however, work remains to find a standardized method to solve the problem. Such modifications also increase the stability of siRNA in serum, delaying its breakdown (Jackson et al., 2006; Watts et al., 2008; Gao et al., 2011).

Other limitations of siRNAs in therapeutic use concern the mode of siRNA delivery to cells and the breakdown of siRNA by the immune system. The most widely studied method of insertion of siRNA sequences uses a viral vector. This can stimulate the immune system of the cell, which then degrades the siRNA and prevents gene silencing (Gao et al., 2011). New delivery systems are under investigation to try to bypass immune stimulation. One such method uses liposomes, or cellular components coated in lipids, which can encase drugs

or other molecules and cross cell membranes without stimulating immunity (Guo et al., 2010; Gao et al., 2011). However, the use of liposomes can result in other forms of toxicity, such as cell contraction and inhibited mitosis, so more research is needed (Stewart et al., 1992). The relatively new field of nanotechnology holds promise in enabling efficient siRNA delivery systems, such as nanospheres, carbon-fiber nanotubes, and magnetized nanocrystals (Katas et al., 2009; Ladeira et al., 2010; Lee et al., 2010; Wang et al., 2010). Work is also being done that attaches peptides specific to certain receptors to siRNA molecules to improve their specificity and increase the half-life of the siRNA (Dassie et al., 2009; Guo et al., 2010). Despite the unexpected difficulties in using siRNA in the clinic, it remains an active area of research. Larger data sets and more effective algorithms to predict siRNA activity are anticipated to provide the keys to these challenges.

miRNA

Like siRNA, microRNA (miRNA) is a small molecule, usually about 22–24 nt in size. miRNA is a post-transcriptional regulator that acts to repress the translation of a protein, degrade messenger RNA (mRNA), or silence a gene. So far, approximately 15,000 miRNAs are known (Ladomery et al., 2011). miRNAs have been found in animals, plants, and viruses and are ubiquitous among all animals with bilateral symmetry, which proves the importance and antiquity of these molecules in gene regulation (Chen, 2010). The process of miRNA formation is different from that of siRNA. When a palindromic sequence of bases occurs (often in the 3′ untranslated region [UTR]), the molecule can fold up and stick to another, creating a stem, with the bases between the palindrome regions making a loop at the top (a "hairpin" shape). The stem is then cut from the RNA strand, and the Dicer enzyme attaches to the stem and carries the miRNA to RISC, as in siRNA processing. However, unlike siRNA, the unfolded (or "mature") miRNA does not accomplish gene silencing by attaching to coding sequences. Instead, it turns off genes through one or more of the following mechanisms: (1) promoting mRNA decay, (2) inhibiting protein translation, or (3) directing mRNA to move to parts of the cell where it will be broken down (Cannell et al., 2008; Bartel, 2009; Beezhold et al., 2010). These mechanisms are poorly understood and are the subject of much research.

The function of miRNAs is another active research topic. miRNAs are not intrinsically harmful; they are known to participate in many different cellular processes, including stem cell development, cell differentiation, cell cycle regulation, apoptosis, and transformation (either normal or malignant). All of these processes require the switching off or fine tuning of the expression of

specific genes at specific times, so the overall role of miRNAS is incredibly complicated. Because the same miRNA can bind to different sections of mRNA to inhibit transcription, the same miRNA can target quite a number of genes. Some evidence has shown that where the miRNA attaches depends on that gene's promoter (Beezhold et al., 2010).

Because miRNAs are noncoding genes, their expression can be regulated by transcription factors (proteins): miRNAs affect gene expression, but their own expression in turn can be influenced. For example, the p53 tumor suppressor protein, one of the most frequently studied proteins in cancer research, has an apparent effect on miRNA. When p53 is mutated, tumor suppression is lowered. However, mutated p53 has been shown to hinder the activity of tumor-suppressing miRNAs, thereby strengthening its tumorigenic action (Beezhold et al., 2010; Ladomery et al., 2011).

1.3 High-Throughput Profiling

The high-throughput methods described in this section represent a "middle path" between older technologies and NGS. They result in large data sets but do not produce biological resolution at the single-molecule level. However, the data analysis challenges are similar to those for NGS.

1.3.1 Molecular Inversion Probe Arrays

Molecular inversion probes (MIPs) are used mainly to identify and analyze single-nucleotide polymorphisms (SNPs) – that is, when a DNA sequence differs from the biological norm by only one nucleotide. These microarrays are used to analyze single strands of DNA. First described by Chowdhary et al. (1994), a MIP is constructed of oligonucleotide probes for two segments of DNA complementary to sequences flanking a particular target, connected by a "linker sequence" (which can include a barcode for easy identification), for a total probe length of about 120 nt (Ji and Welch, 2009). When the target sequence is found, the complementary strand attaches to that sequence and the linkers join together, making a loop with the target DNA sequence – a single base pair in the case of an SNP – in the middle. One way to imagine this is by remembering the name Chowdhary et al. originally gave the assay: the "padlock probe" (Figure 1.1). The assay is broken down into three parts: hybridizing, in which the complementary strand is created; circularizing, in which the "padlocks" are set into place; and amplifying, in which the probe sequences are amplified to enhance the signal. The probes can then be counted using high-throughput sequencing methods. Because these probes "lock away" the sequence of interest, only that specific sequence is captured by the assay,

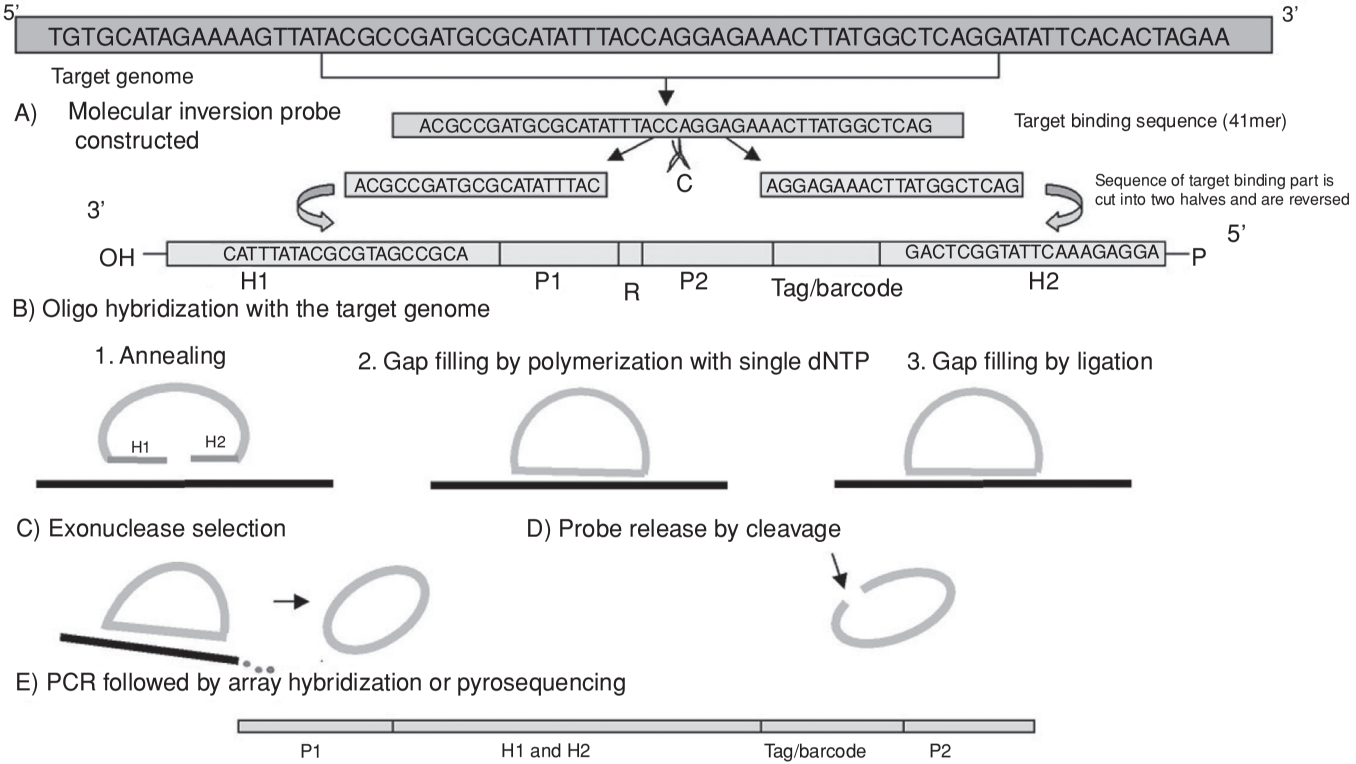

Figure 1.1 Molecular inversion probe assay. Reprinted under a Creative Commons license from Thiyagarajan et al. PathogenMIPer: a tool for the design of molecular inversion probes to detect multiple pathogens. *BMC Bioinformatics* 2006;7:500.