CHAPTER

1

Overview of Technical Aspects and Chemistries of Next-Generation Sequencing

Ian S. Hagemann

Departments of Pathology and Immunology and of Obstetrics and Gynecology, Washington University School of Medicine, St. Louis, MO, USA

OUTLINE

CLINICAL MOLECULAR TESTING: FINER AND FINER RESOLUTION

Progress in applying genetic knowledge to clinical medicine has always been tightly linked to the nature of the genetic information that was available for individual patients.

Classical cytogenetics provides pan-genomic information at the level of whole chromosomes and sub-chromosomal structures on the scale of megabases. The availability of clinical cytogenetics made it possible to establish genotype–phenotype correlations for major developmental disabilities, including +21 in Down syndrome, the "fragile" X site in Fragile X syndrome, monosomy X in Turner syndrome, and the frequent occurrence of trisomies, particularly +13, +17, and +14, in spontaneous abortions.

Over time, new experimental techniques have allowed knowledge to be accumulated at finer and finer levels of resolution, such that genotype–phenotype correlations are now routinely established at the single-nucleotide level. Thus it is now well known that germline *F5* p.R506Q mutation is responsible for the factor V Leiden phenotype [1] and that loss of imprinting at the *SNRPN* locus is responsible for Prader–Willi syndrome [1], to cite examples of two different types of molecular lesions. Clinical advances have been closely paralleled by progress in research testing, since the underlying technologies tend to be similar.

Historically, much clinical molecular testing has taken an indirect approach to determining gene sequences. Although the sequence was fundamentally the analyte of interest, indirect approaches such as restriction fragment length polymorphism (RFLP) analysis, allele-specific polymerase chain reaction (PCR), multiplex ligation-dependent probe amplification (MLPA), and invader chemistry assays have proven easier to implement in the clinical laboratory—easier and more cost-effective to standardize, to perform, and to interpret [2].

Technological advances in the past two decades have begun to change this paradigm by vastly facilitating the acquisition of gene sequence data. Famously, the human genome project required an investment of 10 years and about 10 billion dollars to determine the genomic sequence of a single reference individual. While the technology used for that project was innovative at the time, the effort and cost were clearly monumental and the project could never have been translated directly into a clinical testing modality. Fundamental technical advances, broadly described as next-generation sequencing (NGS), have lowered the cost and difficulty of genomic sequencing by orders of magnitude, so that it is now practical to consider implementing these methods for clinical testing.

The first section of this book is a survey of the technologies used for NGS today. The present chapter focuses on the lowest-level building blocks of NGS: the chemical and technological basis of the methods used to convert nucleic acids into sequence. Subsequent chapters deal with methods for selecting the molecules to be sequenced (whole genome, exome, or gene panels) as well as different approaches for enriching the reagent pool for these molecules (capture and amplification) (Chapters 2–4). The section closes with a chapter on emerging "third-generation" methods, which promise to eventually allow single-molecule sequencing (Chapter 5), as well as a chapter on RNA-based methods which allow NGS technology to be used for expression profiling (Chapter 6).

SANGER SEQUENCING

Chemistry of Sanger Sequencing, Electrophoresis, Detection

In Sanger sequencing [3], DNA polymerase is used to synthesize numerous copies of the sequence of interest in a single primer extension step, using single-stranded DNA as a template. Chain-terminating 2′,3′-dideoxynucleotide triphosphates (ddNTPs) are spiked into the reaction. At each nucleotide incorporation event, there is chance that a ddNTP will be added in place of a dNTP, in which case, in the absence of a 3′ hydroxyl group, the growing DNA chain will be terminated. The endpoint of the reaction is therefore a collection of DNA molecules of varying lengths, each terminated by a dideoxynucleotide [4].

The original Sanger sequencing method consists of two steps. In the "labeling and termination" step, primer extension is performed in four parallel reactions, each reaction containing a different ddNTP in addition to [α-^{35}S]dATP and dNTPs. A "chase" step is then performed with abundant unlabeled dNTPs. Any molecules that have not incorporated a ddNTP will be extended so that they do not interfere with detection. The products are then separated by polyacrylamide gel electrophoresis in four parallel lanes representing ddA, ddT, ddC, and ddG terminators. The DNA sequence is read off of an autoradiograph of the resulting gel by calling peaks in each of the four lanes (Figure 1.1A).

Historically, Sanger sequencing employed the Klenow fragment of *Escherichia coli* DNA polymerase I. The Klenow fragment has 5′→3′ polymerase and 3′→5′ exonuclease activity, but lacks 5′→3′ exonuclease activity [5], thus preventing degradation of desired DNA polymerase products. Klenow fragment is only moderately processive and discriminates against incorporation of ddNTPs, a tendency which can be reduced by including Mn^{2+} in the reaction [6]. Sequenase, which was also commonly used, is a modified T7 DNA polymerase with enhanced processivity over Klenow fragment, a high elongation rate, decreased exonuclease activity, and minimal discrimination between dNTPs and ddNTPs [6,7].

Several variants of Sanger sequencing have been developed. In one of these, thermal cycle sequencing, 20−30 denaturation−annealing−extension cycles are carried out, so that small numbers of template molecules can be repeatedly utilized; since only a single sequencing primer is present, the result is linear amplification of the signal, rather than exponential amplification as would be the case in a PCR [4,8]. The high-temperature steps present in thermal cycle sequencing protocols have the advantage of melting double-stranded templates and disrupting secondary structures that may form in the template. A high-temperature polymerase, such as *Taq*, is required. *Taq* polymerase discriminates against ddNTPs, requiring adjustment of the relative concentration of dNTPs and ddNTPs in these reactions. Native *Taq* polymerase also possesses undesirable 5′→3′ exonuclease activity, but this has been engineered out of commercially available recombinant *Taq* [4].

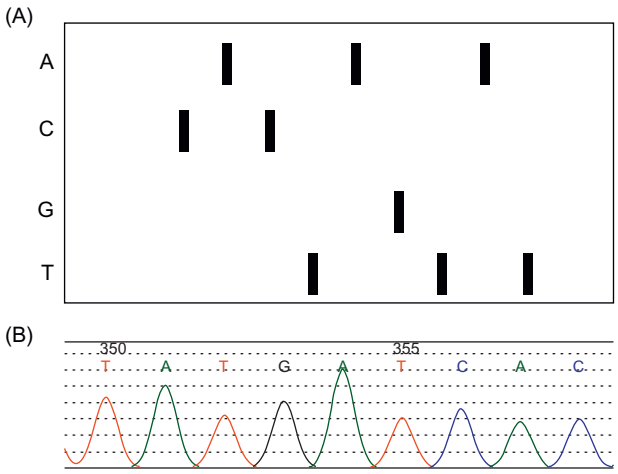

FIGURE 1.1 **Sanger sequencing.** (A) Mockup of the results of gel electrophoresis for Sanger sequencing of the DNA molecule 5′-TATGATCAC-3′. The sequence can be read from right to left on the gel. (B) Electropherogram for Sanger sequencing of the same molecule. The results are read from left to right.

Other variant approaches consist of different detection methods:

- When radioisotope detection was in use, the original [α-^{32}P]dATP protocol was modified to allow use of [α-^{33}P]dATP and [α-^{35}S]dATP, lower-energy emitters producing sharper bands on the autoradiogram [9].
- Chemiluminescent detection was also reported using biotinylated primers, streptavidin, and biotinylated alkaline phosphatase [10].
- 5′-end labeling of the primer ensures that only authentic primer elongation products will be detected, thus reducing the effect of nicks in template molecules serving as priming sites [4].

Modern Sanger sequencing (automated fluorescent sequencing, dye-terminator sequencing) uses fluorescently labeled ddNTPs that allow the amplification step to be performed in a single reaction. The product of the reaction is a mixture of single-stranded DNA fragments of various lengths, each tagged at one end with a fluorophore indicating the identity of the 3′ nucleotide. The reaction is separated by capillary electrophoresis. Continuous recording of four-color fluorescence intensity at the end of the capillary results in an electropherogram (Figure 1.1B) that can be interpreted by base-calling software, such as Mutation Surveyor (SoftGenetics LLC, State College, PA).

Clinical Sanger sequencing today uses the fluorescent dye-terminator method and is accomplished with commercially available kits. The BigDye family of products (Applied Biosystems (ABI)/Life Technologies) is commonly used. BigDye v3.1 is recommended by the vendor for most applications, including when long read lengths are desired. The older BigDye v1.1 chemistry remains useful for specialty applications, specifically for cases in which bases close to the sequencing primer are of greatest interest [11]. The ABI PRISM dGTP BigDye Terminator v3.0 kit may be useful for difficult templates, such as those with high GC content [12]. These kits are optimized for readout on ABI capillary electrophoresis platforms, such as the ABI 31*xx* and 37*xx* series Genetic Analyzers. These instruments vary in features, particularly in the number of capillaries which ranges from 1 (310 Genetic Analyzer) to 96 (3730*xl* DNA Analyzer), and in the available modes (Table 1.1).

Applications in Clinical Genomics

Sanger sequencing is a "first-generation" DNA sequencing method. Despite the advantages of next-generation sequencing techniques, where throughput is orders of magnitude higher, Sanger sequencing retains an essential place in clinical genomics for at least two specific purposes.

First, Sanger sequencing serves as an orthogonal method for confirming sequence variants identified by NGS. When validating clinical NGS tests, reference materials sequenced by Sanger approaches provide ground truth against which the NGS assay can be benchmarked. These materials may include well-characterized publicly available reagents, such as cell lines studied in the HapMap project, or archival clinical samples previously tested by Sanger methods.

TABLE 1.1 Estimated Sequencing Throughput of ABI Genetic Analyzers, as Reported by the Vendor [13]

Instrument	Capillaries	Sample Capacity	Mode	Length of Read*	Runs/Day	Output/Day
310	1	96 tubes	Standard	600	9	5 kb
			Rapid	425	38	15 kb
3130*xl*	16	96- or 384-well plate	Long read	950	8	121 kb
			Ultra rapid	500	41	328 kb
3730*xl*	96	96- or 384-well plate	Extra long read	900	8	691 kb
			Standard	700	24	1.6 Mb
			Rapid	550	40	2.1 Mb
			Resequencing	400	72	2.8 Mb

*Length of read is reported for 98.5% base-calling accuracy with fewer than 2% N's.

As an orthogonal method, Sanger sequencing provides a means to confirm variants identified by NGS. It would be impractical to Sanger-confirm every variant, given the large number of primers, reactions, and interpretations that would be required. However, there may be instances where the veracity of a specific variant is in doubt; e.g., called variants that are biologically implausible or otherwise suspected of being spurious. Sanger sequencing is the easiest method to resolve these uncertainties and is therefore an invaluable protocol in any clinical genomics laboratory.

Second, Sanger sequencing provides a means to "patch" the coverage of regions that are poorly covered by NGS. In targeted NGS testing, there may be regions that are resistant to sequencing, due to poor capture, amplification, or other idiosyncrasies. These regions are often rich in GC content. One approach to restoring coverage of these areas is to increase the quantity of input DNA, but the quantity available may be limited. It may be possible to redesign the amplification step or capture reagents, or otherwise troubleshoot the NGS technology. However, a very practical approach, when the area to be backfilled is small, is to use Sanger sequencing to span the regions poorly covered by NGS.

When Sanger sequencing is used for backfilling NGS data, the NGS and Sanger data must be integrated together for purposes of analysis and reporting, which represents a challenge since these data are obtained by different methods and do not have a one-to-one correspondence to one another. Analyses that are natural for NGS data may be difficult to map onto data obtained by Sanger. For example, measures of sequence quality that are meaningful for NGS are not applicable to Sanger; the concept of depth of coverage can only be indirectly applied to Sanger data; allele frequencies are indirectly and imprecisely ascertained in Sanger sequence from peak heights rather than read counts; and Sanger data do not have paired ends. While NGS may potentially be validated to allow meaningful variant calling from a single nonreference read, the sensitivity of Sanger sequencing has a floor of approximately 20%: variants with a lower allele frequency may be indistinguishable from noise or sequencing errors (discussed below). Thus the performance of an NGS assay may be altered in areas of Sanger patching, and these deviations in performance must be documented and/or disclaimed.

Technical Constraints

Read Length and Input Requirements

Read lengths achieved with Sanger sequencing are on the order of 700−1000 bp per reaction [12]. Thus, a small number of Sanger reactions may be sufficient to cover one or two failed exons in a targeted NGS panel. Required input for Sanger sequencing varies by protocol and by type of template, but as a rule of thumb for double-stranded linear DNA, 10 ng of template per 100 bp of template length gives satisfactory results. Paradoxically, excessive template quantity results in short usable sequence length (i.e., short sequence reads).

Input DNA must consist of a relatively pure population of sequences. Each molecule to which the sequencing primer hybridizes will contribute to the electropherogram: the final electropherogram will be a superposition of all of the input molecules. Sequence diversity at a small number of positions (e.g., a heterozygous single-nucleotide variant (SNV) or deletion of a few nucleotides) will be resolvable by human readers or by

analysis software. More complex diversity within the input DNA will be very difficult to resolve and/or may be indistinguishable from sequencing errors.

Pooled Input DNA Puts a Limit on Sensitivity

Unlike NGS technologies, in which each sequence read originates from sequencing of a single molecule of DNA, the results of Sanger sequencing represent the pooled characteristics of all of the template molecules. This presents no difficulty if the template is a homogeneous population. However, clinical samples may be heterogeneous in at least two ways.

Genomic DNA represents a pool of the patient's two haplotypes, so positions at which the patient is heterozygous will result in an ambiguous call if some form of heterozygote analysis is not specifically enabled. Mutation Surveyor (SoftGenetics LLC, State College, PA) is one Sanger analysis package with the ability to deconvolute heterozygous SNVs and insertion−deletion variants (indels) [14].

In cancer samples, DNA extracted from bulk tumor tissue is intrinsically a mixture of nontumor stroma and of any subclones present within the tumor, so alleles may be present at less than 50% variant allele frequency (VAF). Mitochondrial heteroplasmy is another scenario where variants of low VAF may be clinically relevant. Variant bases with low allele frequency appear in electropherograms as low peaks which may be indistinguishable from baseline noise. The sensitivity of Sanger sequencing must therefore be validated in each laboratory, but is usually cited as being in the neighborhood of 20% [14]. Variant alleles below this frequency may truly be present within the specimen and are faithfully identified by NGS but cannot reliably be confirmed by Sanger sequencing.

A related issue is that Sanger sequencing is not phase-resolved. The two copies of each gene carried by an euploid cell population are averaged together in Sanger sequencing, and variants on one chromosome cannot be differentiated from variants on the other. This limitation is problematic if more than one pathogenic variant is detected in a given gene: variants in *cis* would imply retention of one functional copy, while variants in *trans* would mean that both copies are mutated. The lack of phase resolution is also problematic if Sanger data are to be used to determine the patient's diplotype for complex alleles, as is the case for HLA typing or drug-metabolism genes (e.g., *CYP2D6*, *CYP2D19*). At these highly polymorphic loci, multiple positions often need to be assayed with preservation of phase data in order to assign haplotypes unambiguously based on sequence. It is, however, usually possible to use databases of known haplotypes, combined with data describing the probability of each haplotype in the patient's ethnic group, to ascertain the most probable diplotype. External data cannot be leveraged for phase resolution in the case of somatic variants, which by definition are unique to the patient and are not segregating in the population as discrete haplotypes.

CYCLIC ARRAY SEQUENCING

Many of the currently available next-generation (NGS) approaches have been described as cyclic array sequencing platforms, because they involve dispersal of target sequences across the surface of a two-dimensional array, followed by sequencing of those targets [9]. The resulting short sequence reads can be reassembled *de novo* or, much more commonly in clinical applications, aligned to a reference genome.

NGS has been shown to have adequate sensitivity and specificity for clinical testing. Tested against the gold standard of Sanger sequencing, an NGS cardiomyopathy panel consisting of 48 genes showed perfect analytical sensitivity and specificity [15]. A sensorineural hearing loss panel, OtoSeq, was similarly shown to have 100% analytical sensitivity and 99.997% analytical specificity [16]. NGS tests designed to detect somatic variants in cancer have also been validated as having clinical sensitivity and specificity exceeding 99% and adequate for clinical use [17,18].

NGS workflows involve [1] obtaining the nucleic acid of interest; [2] preparing a sequencing library, which may involve enrichment of target sequences; and [3] carrying out the sequencing on the chosen platform. Many platforms have been developed and have been reviewed elsewhere [9,19], but this discussion will be limited to a focused review of those platforms that, thanks to suitable cost and technical parameters, have found a place in clinical genomic testing. Because of constant evolution in the marketplace, technical comparisons between platforms have a limited life span: new instruments, reagent kits, and protocols appear constantly, with major implications for assay performance [20].

ILLUMINA SEQUENCING

Sequencers and reagents developed by Solexa and further commercialized by Illumina, Inc. (San Diego, CA) have come to be one of the most frequently used platforms in clinical genomics, thanks to their versatility and favorable cost/speed/throughput trade-offs. The distinguishing feature of Illumina sequencing is that prepared libraries are hybridized to the two-dimensional surface of a flow cell, then subjected to "bridge amplification" that results in the creation of a localized cluster (a PCR colony, or polony) of about 2000 identical library fragments within a diameter of ~1 μm, and across a single lane of a flow cell there can be over 37 million individual amplified clusters [21] (Figure 1.2). These fragments are sequenced in place by successive incorporation of fluorescently labeled, reversibly terminating nucleotides (known as sequencing by synthesis). After each incorporation step, the surface of the flow cell is imaged by a charge-coupled device (CCD) to query each position for the identity of the most recently incorporated nucleotide. Successive cycles of deprotection, incorporation, and imaging result in a series of large image files that are subsequently analyzed to determine the sequence at each polony.

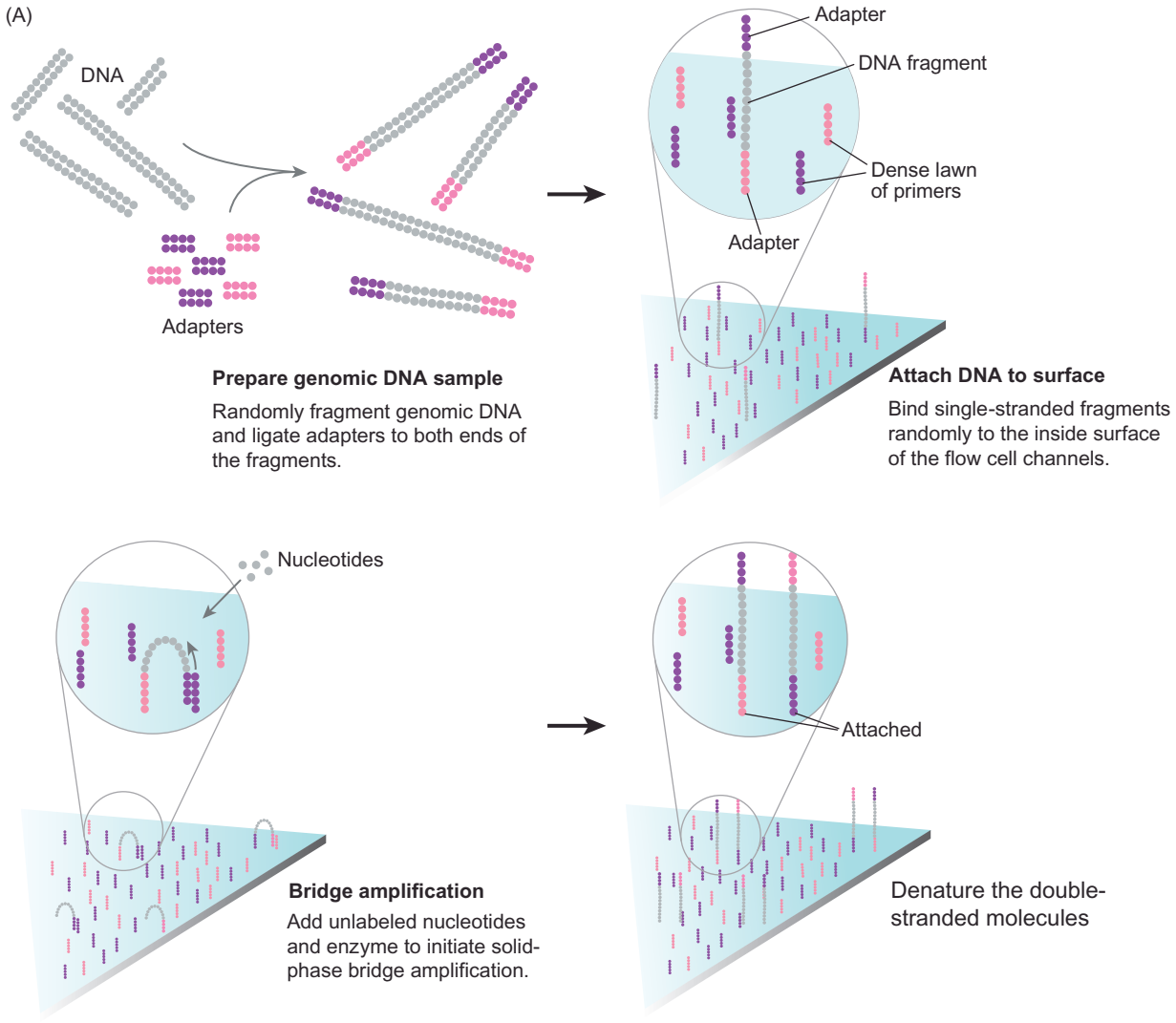

FIGURE 1.2 **Illumina sequencing.** (A) Initial steps in library hybridization and loading on Illumina flow cell. (B) First and subsequent cycles of sequencing by synthesis. *Reprinted with permission from* Annual Review of Genomics and Human Genetics *by ANNUAL REVIEWS, copyright 2008.*

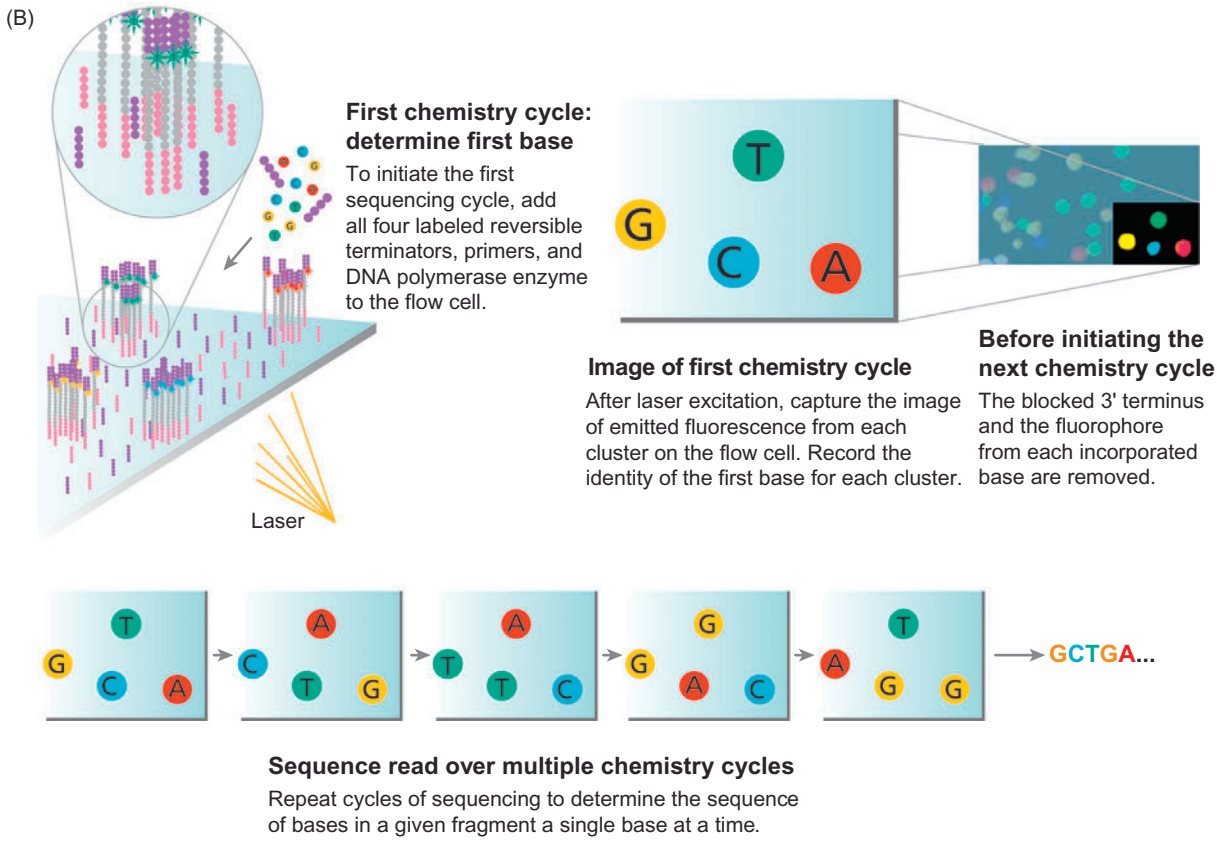

FIGURE 1.2 (Continued).

TABLE 1.2 General Steps for Preparation of Illumina Sequencing Libraries [22]

DNA fragmentation
End repair
3′ Adenylation (A-tailing)
Adapter ligation
Purification and enrichment
Library validation and quantification

Library Prep and Sequencing Chemistry

The workflow for Illumina sequencing begins with library preparation, in which the input DNA is processed to make it a suitable substrate for sequencing. The specific steps for library preparation will depend upon the application; protocols are readily available, usually paired with kits containing necessary reagents. An idealized workflow for general-purpose paired-end sequencing of genomic DNA based on a standard Illumina protocol (Table 1.2) [22] is presented here.

In clinical genomics applications, the input material will consist of patient samples such as peripheral blood, bone marrow aspirate, fresh tissue, or formalin-fixed paraffin-embedded (FFPE) tissue. Before library preparation can begin, DNA must be extracted by some method standardized for the laboratory and known to yield DNA of suitable quality for sequencing [23]. The DNA must be assayed for quality according to metrics and cutoffs established in the lab. DNA with $A_{260}/A_{280} = 1.8–2.0$ is generally considered suitable for library preparation. Gel electrophoresis may be performed to determine whether the input DNA is high molecular weight or has become degraded. However, given that library preparation begins with fragmentation, high molecular weight

is not mandatory. A smear seen on an agarose gel may indicate the presence of contaminants including detergents or proteins [22].

The input DNA is fragmented by sonication, nebulization, or endonuclease digestion to yield fragments of <800 bp. The fragment size can be controlled by adjusting parameters of the sonicator or other instrument. Fragment size is relevant in that it determines the eventual distance between mated reads in paired-end sequencing.

Fragmented DNA may have 5′ or 3′ overhangs of various lengths. To ensure uniform downstream reactivity, an end repair step is performed by simultaneous treatment with T4 DNA polymerase, Klenow fragment, and T4 polynucleotide kinase (PNK). The first two of these serve to remove 3′ overhangs (3′→5′ exonuclease activity) and fill in 5′ overhangs (5′→3′ polymerase activity). PNK adds a 5′ phosphate group.

The A-tailing (3′ adenylation) step starts from blunt-ended DNA and adds a 3′ A nucleotide using a recombinant "Exo−" Klenow fragment of *E. coli* DNA polymerase, which lacks 5′→3′ and 3′→5′ exonuclease activities. The A tail cannot serve as a template, which prevents the addition of multiple A nucleotides.

Illumina adapters allow hybridization to the flow cell and may also encompass an index sequence to allow multiplexing multiple samples on a single flow cell. Adapter ligation uses DNA ligase and a molar excess of indexed adapter oligonucleotides to place an adapter at both ends of each DNA fragment. The index is a 6-nucleotide sequence that serves as a barcode for each library. Libraries with different indexes can be multiplexed at the time of sequencing (i.e., run in a single lane) and informatically separated at a later time.

At this stage, libraries are size selected by gel electrophoresis to eliminate adapter dimers and other undesired sequence. It is at this stage that the final library insert size is selected, i.e., that the distance between paired ends is fixed (which is important for paired-end sequencing).

Each step of library preparation to this point has required a purification procedure, which in the aggregate results in marked reduction in the quantity of DNA present in the library. This is counteracted by a limited PCR amplification step. Although the library fragments had earlier acquired adapters carrying single-stranded DNA at each extremity, the PCR step renders the library fully double-stranded. The PCR step has several additional functions: it adds sequences necessary for hybridization to the flow cell; also, it ensures that the library carries adapters at both ends (otherwise, the primers would fail to anneal).

Library preparation is completed by a final round of agarose gel electrophoresis to purify the final product. Illumina recommends storing prepared library DNA at 10 nM concentration in 10 mM Tris−HCl buffer, pH 8.5, with 0.1% Tween-20 to prevent sample adsorption to the tube [22].

Choice of Platforms

Illumina sequencers in common use are the HiSeq and MiSeq systems. The HiSeq is a larger-scale, higher-throughput instrument, while the MiSeq is conceptualized as a benchtop or personal sequencer. The technical specifications of these platforms are likely to change rapidly over time. With that caveat, these platforms have the following features (Table 1.3).

In its current iteration, the HiSeq (model 2500) accepts one or two flow cells, each with eight lanes. In high output run mode, a single flow cell can generate up to 300 Gb of data at 2 × 100 bp paired-end read length. These runs require 11 days to automatically proceed from cluster generation to final sequence. In rapid run mode, one flow cell can generate 60 Gb of 2 × 100 bp reads or 90 Gb of 2 × 150 bp reads, requiring 27 or 40 h, respectively.

The MiSeq instrument accepts a single-lane flow cell. With current reagent kits, MiSeq is benchmarked to generate 5.1 Gb of 2 × 150 bp reads in about 24 h [27]. The typical use case for MiSeq is a scenario where the total amount of sequence desired is insufficient to justify use of an entire HiSeq flow cell, either because of a low number of samples to be batched together, small size of a target region to be sequenced, or low desired coverage (more common in constitutional, as opposed to oncology, testing). The FDA granted the MiSeqDx instrument marketing authorization for high-throughput genomic sequencing in December 2013, which has many implications for clinical NGS [28].

Phasing

In NGS, each sequence read originates from a single DNA fragment, so variants identified in close proximity to one another can be reliably assigned to discrete haplotypes. For example, tumors have occasionally been found to have point mutations in both *KRAS* codons 12 and 13. NGS data make it possible to determine whether nearby variants occurred on the same reads, indicating variants in *cis*, or on different reads, indicating variants in *trans*.

TABLE 1.3 Estimated Sequencing Yield for Illumina Instruments, as Reported by the Vendor [24–26]

	Run Mode/Kit	Read Length	Run Time	Number of Reads (Paired Ends)	Output
HiSeq 2000	High output	2 × 100 bp	11 d	6 billion	600 Gb
HiSeq 2500	Rapid run	2 × 150 bp	27 h	1.2 billion	120 Gb
	High output	2 × 100 bp	11 d	6 billion	600 Gb
MiSeq	Reagent kit v2	1 × 36 bp	4 h	24–30 million	540–610 Mb
		2 × 25 bp	5.5 h		750–850 Mb
		2 × 150 bp	24 h		4.5–5.1 Gb
		2 × 250 bp	39 h		7.5–8.5 Gb
	Reagent kit v3	2 × 75 bp	24 h	44–50 million	3.3–3.8 Gb
		2 × 300 bp	65 h		13.2–15 Gb
Genome Analyzer IIx	TruSeq SBS v5	1 × 35 bp	2 d	640 million	10–12 Gb
		2 × 50 bp	5 d		25–30 Gb
		2 × 75 bp	7 d		37.5–45 Gb
		2 × 100 bp	9.5 d		54–60 Gb
		2 × 150 bp	14 d		80–95 Gb

Sanger sequencing, in contrast, cannot determine phase, as the electropherogram represents a pooled sum of all molecules in the reaction.

When the variants to be phase-resolved do not lie on the same read or paired ends of the same read, there is no formal (guaranteed) means to resolve the phase, but it may be possible to do so by "walking" from one variant to the other. A tool for this purpose is included in the Genome Analysis Toolkit (GATK) [29].

SOLiD SEQUENCING

The Sequencing by Oligo Ligation Detection (SOLiD) platform was developed by Agencourt Bioscience (at the time, a subsidiary of Beckman Coulter, now a unit of Applied Biosystems, Inc., itself a brand of Life Technologies). The technique is distinctive in that instead of performing sequencing-by-synthesis one nucleotide at a time, it obtains sequence by determining the ability of oligonucleotides to anneal to a template molecule and become ligated to a primer.

To generate the substrate for sequencing, DNA fragments are fused to a "P1" adapter, bound to magnetic beads and amplified by emulsion PCR so that each bead is coated with a single clonal population. These beads are then fixed to the surface of a glass slide. More recent iterations of the platform use the "Wildfire" approach, omitting beads and allowing libraries to be prepared directly on a two-dimensional substrate by a template-walking process resembling colony PCR.

Five rounds of primer elongation are carried out, each of which is followed by a primer reset. In each round, a universal sequencing primer is hybridized to the template strands, positioning a 5′ end at the beginning of the region to be sequenced. This position is shifted by one nucleotide in each subsequent round of primer elongation. Multiple cycles of ligation, detection, and cleavage are then performed.

In each ligation step, a pool of 5′-labeled oligonucleotide 8-mers is introduced. The color of the fluorophore attached to each molecule corresponds to the sequence of the first two nucleotides (Figure 1.3). The remaining six nucleotides are degenerate (i.e., all possible combinations are present in the pool). An annealing and ligation reaction is carried out, so that the fluorophore attached to each bead indicates the dinucleotide sequence just downstream from the sequencing primer. The substrate is washed, and a digital image is acquired to document the color attached to each bead.

The fluorophores correspond to the 5′ dinucleotide of each 8-mer. Although it might appear that 16 colors would be needed to encode 16 possible dinucleotides, only four colors are used. Each nucleotide will be interrogated twice,

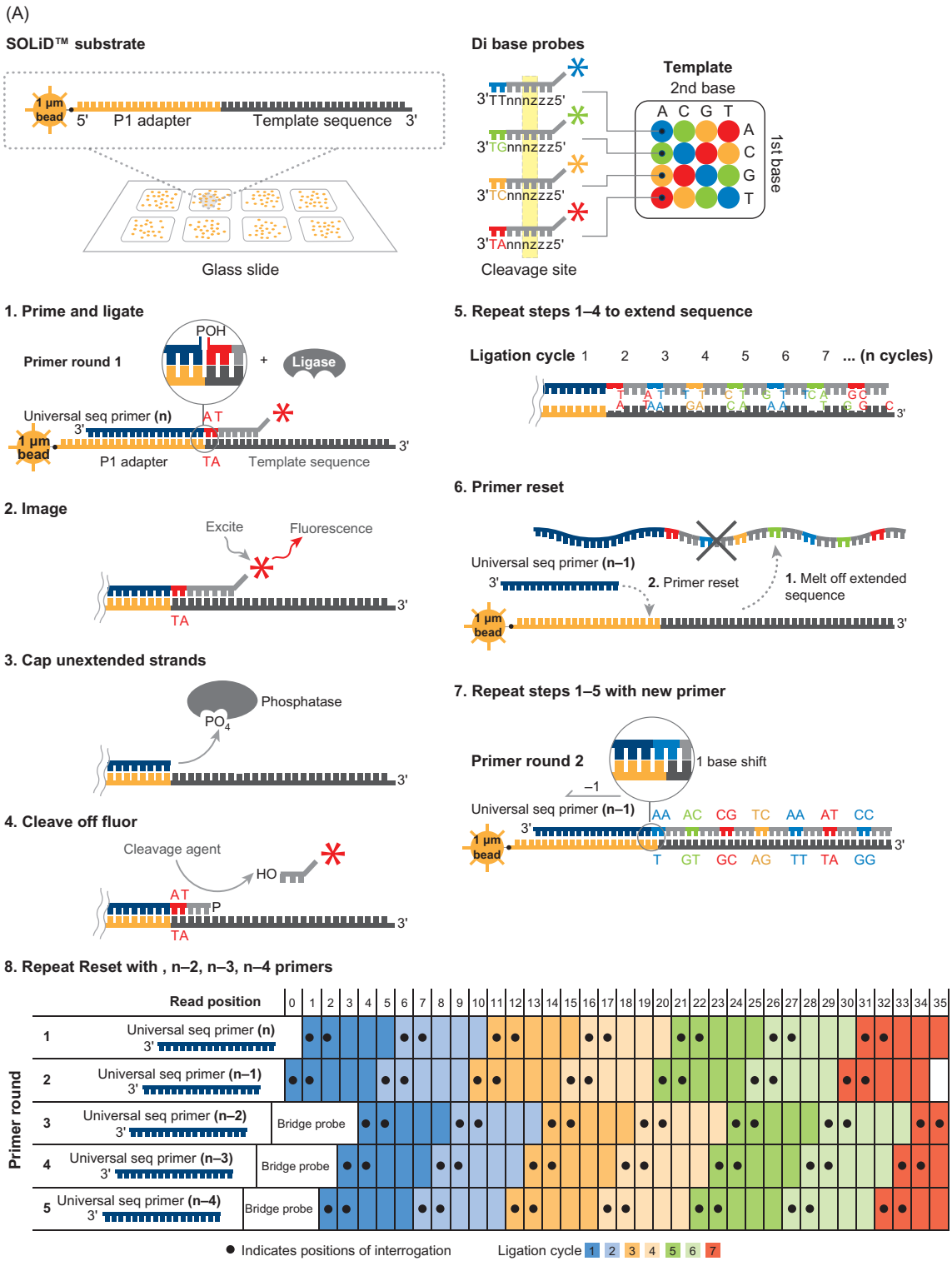

FIGURE 1.3 **Principle of SOLiD sequencing.** (A) Molecules to be sequenced are deposited on a flow cell slide, primers are annealed, and extension by is carried out by ligation of 8-mers carrying a label indicating their first two nucleotides. Several rounds of primer extension are performed. A "primer reset" is then carried out, applying a new primer of different length. (B) Data are collected in "color space" and are then processed to give the sequence in "base space." The sequencing is performed by querying each nucleotide twice, once as the first member of a dinucleotide and once as the second member. These queries occur in consecutive rounds of primer extension. To correctly interpret the meaning of the color encoding a dinucleotide, it is essential to know the identity of the preceding nucleotide. If nucleotide n − 1 is an A and the next dinucleotide is red, the next nucleotide must be T; but if nucleotide n − 1 is a C and the next dinucleotide is red, the next nucleotide is G. *Reprinted with permission from* Annual Review of Genomics and Human Genetics *by ANNUAL REVIEWS, copyright 2008.*

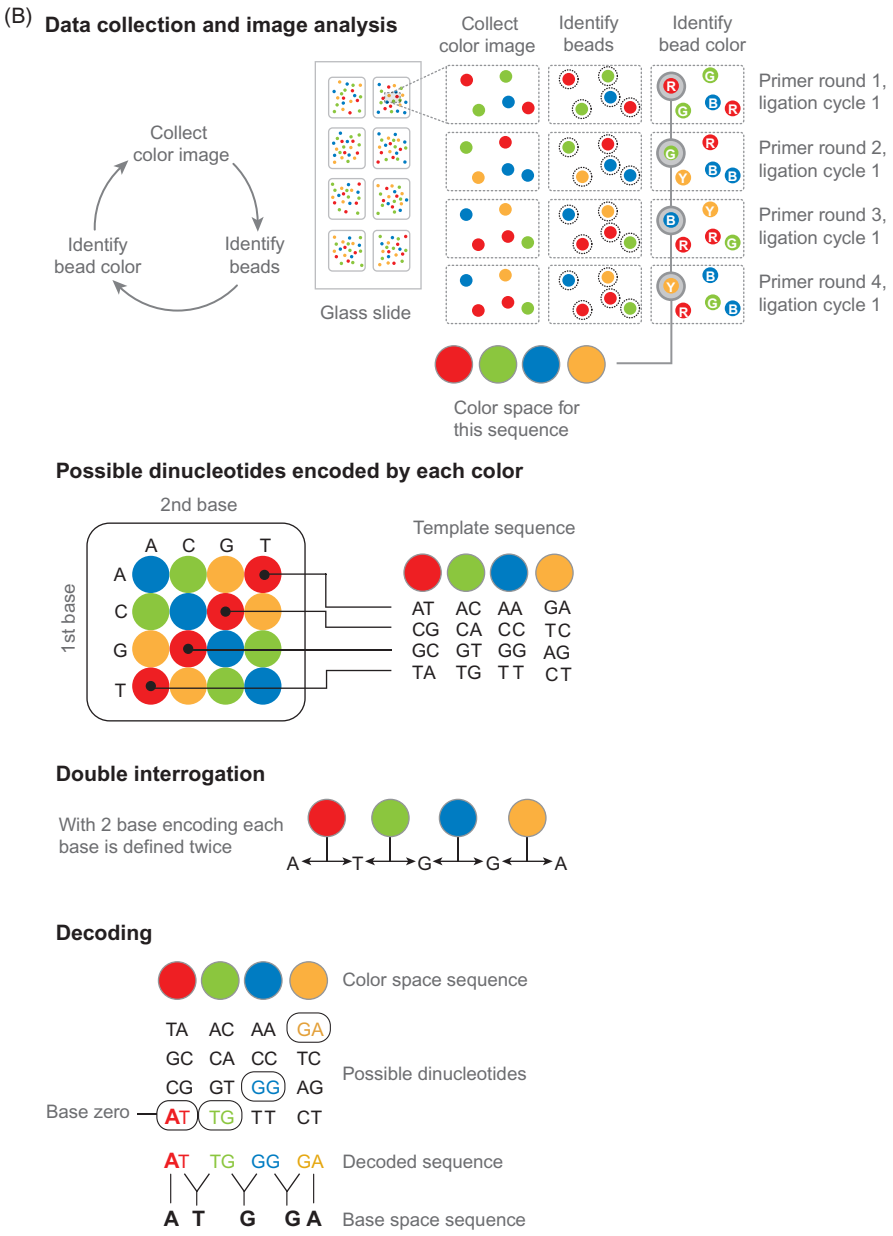

FIGURE 1.3 (Continued).

once in each of two successive primer elongation rounds. Because the reading frame shifts by one nucleotide in each round of primer elongation, the result of these two interrogations is a sequence of two colors: one indicating the nucleotide's identity as the first base of a dinucleotide, and one indicating its identity as the second base of a dinucleotide. There are 16 possible dinucleotides and 16 ordered pairs of colors. Thus the identity of each nucleotide is completely specified.

After each interrogation, a cleavage agent is introduced to cleave the 8-mers between the fifth and sixth nucleotides. A new ligation, detection, and cleavage cycle is then performed to query a new dinucleotide located five nucleotides downstream from the first. Up to nine of these cycles are performed.

A primer reset is then performed and a new round of primer elongation begins, in a new reading frame allowing a different set of dinucleotides to be queried. After five successive rounds of primer elongation, enough data have been gathered to reconstruct reads of 35–70 nucleotides, which are then used for downstream analysis.

SOLiD sequencing has several unique advantages. Annealing of the 8-mer oligonucleotides carries greater specificity than base pairing of a single nucleotide and helps to reduce errors. The primer extension methodology is

TABLE 1.4 Estimated Sequencing Yield for SOLiD Sequencing on the 5500 Series Genetic Analyzer Platform, October 2013, as Reported by the Vendor [30,31]

	Read Length	Reads/Run	Run Time	Output
SOLiD 4 System	2 × 35 bp	1.4 billion	8–9 days	40–56 Gb
	2 × 50 bp	1.4 billion	12–16 days	64–80 Gb
5500 W System (1 FlowChip)	1 × 50 bp		24 h	80 Gb
	1 × 75 bp			120 Gb
	2 × 50 bp			160 Gb
5500xl W System (2 FlowChips)	1 × 50 bp		24 h	160 Gb
	1 × 75 bp			240 Gb
	2 × 50 bp			320 Gb

also highly tolerant of repetitive sequence. The "color space" encoding of nucleotides allows a certain degree of protection against errors. Since each nucleotide is probed twice (once as the first member of a dinucleotide and once as the second member), true point mutations always manifest themselves as two color changes relative to the reference, in two successive rounds of primer elongation. A single color change will usually reflect a sequencing error.

Disadvantages of the SOLiD approach include its conceptual complexity. The short read lengths make sequence assembly difficult, particularly for *de novo* assembly.

Paired-end SOLiD sequencing is accomplished by preparing the library with a second ("P2") adapter distal to the magnetic bead. Oligonucleotide barcodes can also be incorporated, allowing multiplexing.

SOLiD sequencing can be carried out on the Applied Biosystems series 5500 Genetic Analyzers, which use one or two FlowChips. Each FlowChip has six lanes. In 1 × 75 bp fragment sequencing mode, each chip can generate up to 120 Gb of sequence data; in 2 × 50 bp mode, each chip generates up to 160 Gb [30]. The other available platform is the older Applied Biosystems SOLiD 4 System, which has lower throughput (Table 1.4).

ION TORRENT SEQUENCING

All of the approaches described in this chapter involve translating DNA sequence into a detectable physical event. Whereas Sanger sequencing, Illumina sequencing, and SOLiD are based on detecting fluorescence, the Ion Torrent approach (Life Technologies) is to convert sequence into very small changes in pH that occur as a result of H^+ release when nucleotides are incorporated into an elongating sequence.

Library preparation for Ion Torrent sequencing involves fragmentation, end repair, adapter ligation, and size selection. The library molecules are bound to the surface of beads and amplified by emulsion PCR so that each bead is coated with a homogeneous population of molecules. Each bead is then dispersed into a discrete well on the surface of a semiconductor sensor array chip constructed by complementary metal-oxide semiconductor (CMOS) technology [32]. This is the same technology that underlies most microchips, including microprocessors, so the significant advances that have been made in CMOS semiconductor technology are accrued to the Ion Torrent as a matter of course.

The substrate in Ion Torrent sequencing has the property that each well, containing an embedded ion-sensitive field effect transistor, functions as an extremely sensitive pH meter (Figure 1.4). Sequencing is accomplished by flooding the plate with one deoxynucleotide species (dA, dC, dG, dT) at a time. In each well, incorporation of a nucleotide causes release of pyrophosphate and a H^+ ion, resulting in a pH change. The change in H^+ concentration supplies a voltage change at the gate of the transistor, allowing current to flow across the transistor and resulting in a signal [32]. Homopolymers cause incorporation of a greater number of nucleotides, with a correspondingly larger pH change and larger signal. The sequencing reaction is highly analogous to pyrosequencing, with the difference that pH is detected instead of light.

The throughput of the method depends on the number of wells per chip, which in turn is related to semiconductor fabrication constraints. The current Proton series of chips carries 154 million wells, each 1.25 μm in diameter and spaced 1.68 μm apart [32]. Additional improvements are expected to reflect Moore's law (doubling of the number of wells every 18 months), given that the substrate is a semiconductor microchip.

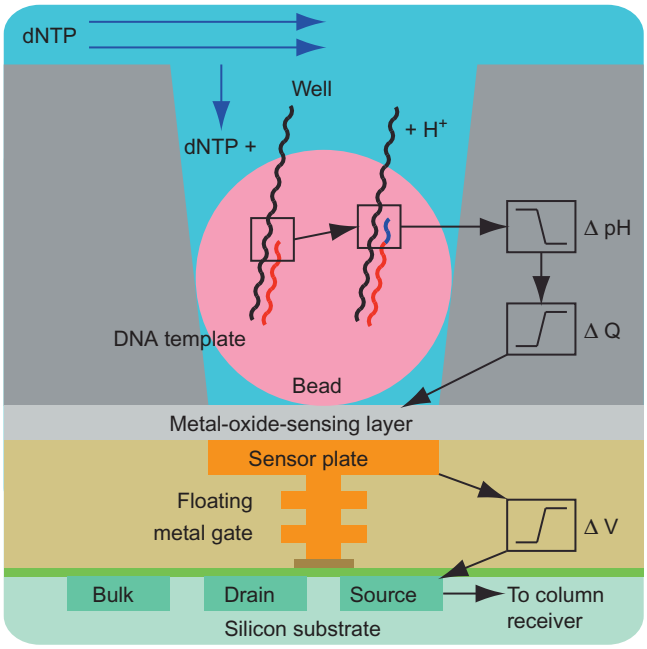

FIGURE 1.4 **Schematic diagram of Ion Torrent sensor.** The diagram shows a single well, within which is lodged a bead containing DNA template, along with the underlying sensor. When nucleotides are incorporated, H^+ ions are released and the resulting pH change is detected as a voltage change. *Reprinted by permission from Macmillan Publishers Ltd:* Nature *475 (7356) 348, copyright 2011.*

Ion Torrent chemistry is unique in using natural deoxynucleotides rather than a derivative thereof, which can reduce sequencing biases related to incorporation of unnatural nucleotides. Compared with other platforms, reads are relatively long, and reaction times are very short (3 h for a 300 base run). Several paired-end modes are available but require off-instrument repriming [32]. The software associated with Ion Torrent sequencing, including an aligner, variant caller, and other plug-ins, is distributed as an open-source product, contributing to its relatively wide adoption.

An often cited disadvantage of Ion Torrent methodology is its relatively high error rate compared with Illumina-based platforms, particularly in homopolymeric regions [33]. Some fraction of reads is rejected due to insufficient quality metrics including signal to noise issues, bead clumping, nonclonal beads, or failure to incorporate a bead at a given sensor. Once these 20–25% of failed reads are excluded, an accuracy of 99.5% (fewer than 0.5% base error rate) for 250 bp reads has been reported [32]. However, it has recently been demonstrated that a significant number of errors are introduced by false priming events during the multiplex amplification step of library preparation, errors which have been unrecognized in prior studies of accuracy [34].

The most problematic source of error is related to homopolymers, whose length must be deduced from the peak height of a multiple incorporation event. Even a quoted 99.3% base accuracy within homopolymers means that a 5-mer will be identified correctly only 96.5% of the time [32], which is significant given that runs of this length or longer are not uncommon in the genome. Difficulties sequencing across homopolymer runs cause uncertainty in the sequence of the run itself, but also cause "dephasing" of subsequent nucleotides due to incomplete extension, although the order in which nucleotides are flowed across the chip can be changed to mitigate dephasing issues. The simplest flow order, A–C–G–T, returns to each nucleotide only on every fourth cycle, and wells with incomplete extension must wait three cycles to catch up. So-called minimally dephasing flow orders repeat each nucleotide at a shorter interval (e.g., A–C–A), allow lagging wells to catch up and reduce sequencing errors in polymeric regions.

Ion Torrent sequencing is currently available on two instruments: Ion PGM (Personal Genome Machine) and Ion Proton. The Ion PGM is a benchtop instrument that uses Ion 314, 316, or 318 chips yielding 200 or 400 base reads. At the high end of the performance range for Ion PGM, current benchmarks are 4–5.5 million reads of 400 bp (1.2–2 Gb of sequence) with a run time of 7.3 h [35] (Table 1.5).

TABLE 1.5 Estimated Sequencing Yield for Ion PGM and Ion Proton Instruments, as Reported by the Vendor [36,37]

Platform	Chip	Sequencing Run Time		Output		Expected Reads
		200-Base Reads	400-Base Reads	200-Base Reads	400-Base Reads	
Ion PGM	Ion 314™ v2	2.3 h	3.7 h	30–50 Mb	60–100 Mb	400–500 thousand
	Ion 316™ v2	3.0 h	4.9 h	300–600 Mb	600 Mb–1 Gb	2–3 million
	Ion 318™ v2	4.4 h	7.3 h	600 Mb–1 Gb	1.2–2 Gb	4–5.5 million
Ion Proton	Ion PI™	2–4 h		10 Gb		60–80 million

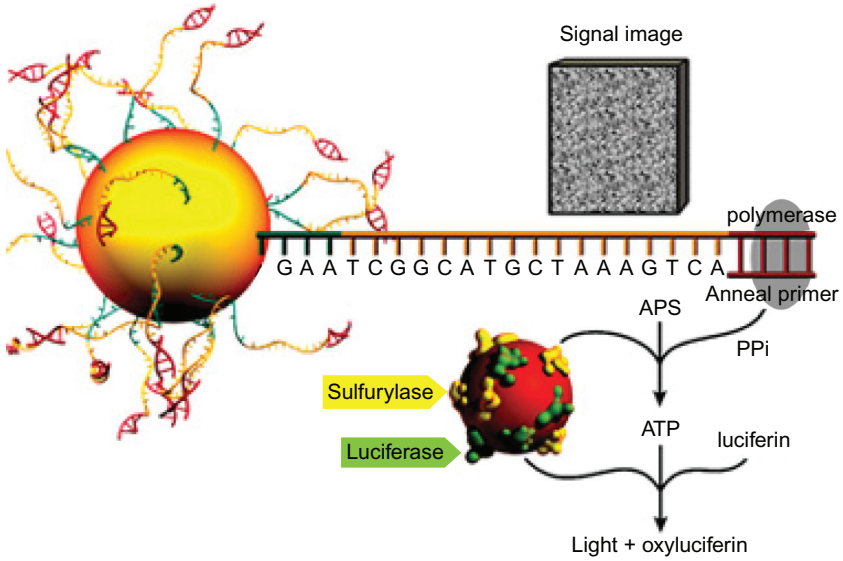

FIGURE 1.5 **Pyrosequencing principle as implemented in the Roche 454 system.** A population of identical template molecules is immobilized on the surface of a bead. Incorporation of a nucleotide phosphosulfate such as adenosine phosphosulfate (APS) is enzymatically converted into light emission. *Reprinted from Ref. [37], copyright 2008, with permission from Elsevier.*

AmpliSeq Library Preparation

Library preparation for targeted Ion Torrent sequencing requires either hybridization-based capture or PCR-based amplification. Life Technologies has developed a platform, AmpliSeq, that allows single-tube PCR amplification of up to 4000 amplicons, as an input for sequencing. The quoted input DNA requirement is 10 ng. Several off-the-shelf panels are available, including a Cancer Mutation Hotspot Panel (46 genes), a Comprehensive Cancer Panel (406 genes), and an Inherited Disease Panel (325 genes). The company provides a tool for designing custom panels.

ROCHE 454 GENOME SEQUENCERS

The Roche 454 platforms are based on pyrosequencing. In non-high-throughput (conventional, non-array-based) configuration, pyrosequencing is a highly sensitive method for sequencing by synthesis. A template DNA strand is immobilized and exposed in turn to individual deoxynucleotides in the presence of DNA polymerase, luciferase, and ATP sulfurylase (Figure 1.5). Luciferin is present as a luciferase substrate, and adenosine 5′ phosphosulfate as an ATP sulfurylase substrate. Each incorporated nucleotide causes release of a pyrophosphate (PP_i) moiety, and this PP_i is combined with adenosine 5′ phosphosulfate to form ATP in a reaction catalyzed by ATP sulfurylase. In the presence of ATP, luciferase converts luciferin to oxyluciferin and light is emitted, resulting in a signal. To prevent consumption of dATP by the luciferase, with resulting emission of aberrant signals, dATPαS is used as the source of

deoxyadenosine nucleotides. Apyrase is added to degrade unincorporated nucleotide triphosphates to nucleotide monophosphates plus inorganic phosphate, P_i, and another cycle begins.

Pyrosequencing was adapted to high-throughput sequencing by the 454 Corporation with the development of a parallel, highly multiplexed platform. Library DNA fragments are captured on the surface of small beads, at a limiting dilution that ensures that each bead receives a single molecule (Figure 1.6). These DNA molecules are

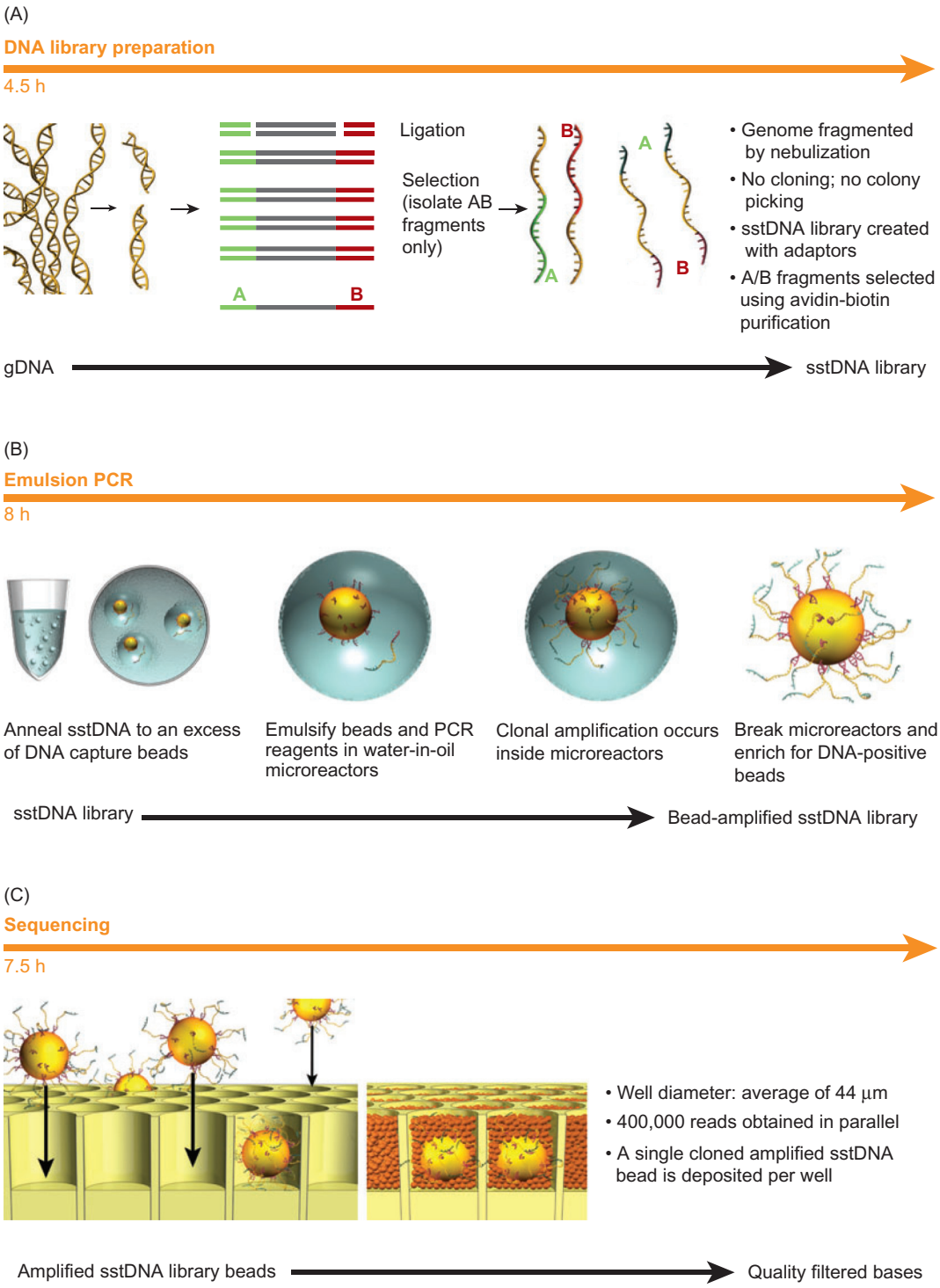

FIGURE 1.6 **Workflow in Roche 454 sequencing.** (A) Library preparation, (B) emulsion PCR, and (C) chip loading and sequencing. *Reprinted with permission from* Annual Review of Genomics and Human Genetics *by ANNUAL REVIEWS, copyright 2008.*

TABLE 1.6 Estimated Sequencing Yield for Roche 454 Genome Sequencers, as Reported by the Vendor [39,40]

	Read Length	Reads/Run	Sequence/Run	Sequencing Run Time	Consensus Accuracy
GS FLX Titanium XL+	Up to 1000 bp	10^6 shotgun	700 Mb	23 h	99.997%
GS FLX Titanium XLR70	Up to 600 bp	10^6 shotgun, 7×10^5 amplicon	450 Mb	10 h	99.995%
GS Junior	400 bp	10^5 shotgun, 7×10^4 amplicon	35 Mb	10 h	NR

Accuracy is reported as concordance between reference and consensus sequence at 15× coverage. NR, not reported.

amplified in an oil–water emulsion (emulsion PCR) on the surface of the beads, then immobilized on the surface of a PicoTiter plate within a 29 μm well. Fiber optics allows light emission in each well to be discretely detected.

The 454 GS20 Genome Sequencer, released in 2005, was the first NGS platform to become commercially available. Current iterations of the platform include the GS FLX+ system and a smaller benchtop instrument, the GS Junior. These instruments incorporate the fluidics and optics needed to perform the sequencing reaction and capture the resulting data. GS FLX+ in combination with GS FLX Titanium chemistry is cited as giving read lengths up to 1000 bp with throughput of 700 Mb per 23 h run, with consensus accuracy of 99.997% at 15× coverage. The GS Junior, with GS Junior Titanium chemistry, gives quoted read lengths of 400 bp, throughput of 35 Mb per 10 h run, and consensus accuracy of 99%.

Advantages of the 454 platform include long read length and short turnaround time. Cost per base sequenced, however, is high, on the order of $10 per million bases [38]. As with Ion Torrent sequencing, in pyrosequencing, homopolymers are detected as incorporation events of larger magnitude than expected, as multiple nucleotides are incorporated. The relation between run length and signal intensity (or area under the curve) is not strictly stoichiometric, resulting in difficulties sequencing across homopolymers greater than 5 bases long (Table 1.6).

THIRD-GENERATION SEQUENCING PLATFORMS

The NGS approaches discussed above have been described as second-generation approaches, in anticipation of the coming availability of third-generation sequencing. Second-generation methods require library preparation and an enrichment or amplification step. These steps are time-consuming, introduce biases related to preferential capture or amplification of certain regions, and also can result in PCR errors which are propagated into the eventual sequence data. Third-generation methods circumvent these problems by sequencing individual molecules, without the need for an enrichment or amplification step. The major disadvantage of single-molecule methods is that they require the ability to detect fantastically small signals, without compromising accuracy. These methods are further discussed in Chapter 5.

References

[1] Leonard DGB. Diagnostic molecular pathology. 1st ed. Philadelphia, PA: W.B. Saunders; 2003.
[2] Pfeifer JD. Molecular genetic testing in surgical pathology. Philadelphia, PA: Lippincott, Williams & Wilkins; 2006.
[3] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977;74(12):5463–7.
[4] Slatko BE, Albright LM, Tabor S, Ju J. DNA sequencing by the dideoxy method. Curr Protoc Mol Biol 2001; Chapter 7: Unit7 4A.
[5] Klenow H, Henningsen I. Selective elimination of the exonuclease activity of the deoxyribonucleic acid polymerase from *Escherichia coli* B by limited proteolysis. Proc Natl Acad Sci USA 1970;65(1):168–75.
[6] Tabor S, Richardson CC. Effect of manganese ions on the incorporation of dideoxynucleotides by bacteriophage T7 DNA polymerase and *Escherichia coli* DNA polymerase I. Proc Natl Acad Sci USA 1989;86(11):4076–80.
[7] Tabor S, Richardson CC. Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by in vitro mutagenesis. J Biol Chem 1989;264(11):6447–58.
[8] Sears LE, Moran LS, Kissinger C, Creasey T, Perry-O'Keefe H, Roskey M, et al. CircumVent thermal cycle sequencing and alternative manual and automated DNA sequencing protocols using the highly thermostable VentR (exo-) DNA polymerase. BioTechniques 1992;13(4):626–33.
[9] Shendure JA, Porreca GJ, Church GM, Gardner AF, Hendrickson CL, Kieleczawa J, et al. Overview of DNA sequencing strategies. Curr Protoc Mol Biol 2011;96:7.1.–7.1.23.
[10] Martin C, Bresnick L, Juo RR, Voyta JC, Bronstein I. Improved chemiluminescent DNA sequencing. BioTechniques 1991;11(1):110–3.
[11] Applied Biosystems. Product Bulletin: BigDye Terminator v3.1 and v1.1 Cycle Sequencing Kits. 2002. Available from: <http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_040026.pdf>.

[12] Slatko BE, Kieleczawa J, Ju J, Gardner AF, Hendrickson CL, Ausubel FM. "First generation" automated DNA sequencing technology. Curr Protoc Mol Biol 2011; Chapter 7: Unit 7 2.
[13] Applied Biosystems. Applied Biosystems Genetic Analyzers at a glance. 2006. Available from:<http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_040402.pdf>.
[14] Dong C, Yu B. Mutation surveyor: an in silico tool for sequencing analysis. Methods Mol Biol 2011;760:223–37.
[15] Sikkema-Raddatz B, Johansson LF, de Boer EN, Almomani R, Boven LG, van den Berg MP, et al. Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. Hum Mutat 2013;34(7):1035–42.
[16] Sivakumaran TA, Husami A, Kissell D, Zhang W, Keddache M, Black AP, et al. Performance evaluation of the next-generation sequencing approach for molecular diagnosis of hereditary hearing loss. Otolaryngol Head Neck Surg 2013;148(6):1007–16.
[17] International HapMap 3 Consortium, Integrating common and rare genetic variation in diverse human populations. Nature 2010;467(7311):52–8.
[18] 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. Nature 2010;467(7319):1061–73.
[19] Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem 2013;6:287–303.
[20] Junemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, et al. Updating benchtop sequencing performance comparison. Nat Biotechnol 2013;31(4):294–6.
[21] Illumina, Inc. Technology Spotlight: Illumina Sequencing. HiSeq System 2010 [6/19/14]. Available from:<http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf>.
[22] Illumina, Inc. Paired-End Sample Preparation Guide, Rev. E. 2011 [6/19/14]. Available from: <http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf>.
[23] Liu X, Harada S. DNA isolation from mammalian samples. Curr Protoc Mol Biol 2013; 102:2.14.1–2.14.13. [PubMed PMID: 23547013].
[24] Illumina, Inc. HiSeq Systems Comparison 2013 [10/25/13]. Available from: <http://www.illumina.com/systems/hiseq_comparison.ilmn>.
[25] Illumina, Inc. MiSeq Performance Specifications 2013 [10/25/13]. Available from: <http://www.illumina.com/systems/miseq/performance_specifications.ilmn>.
[26] Illumina, Inc. Genome Analyzer IIx Performance and Specifications 2013 [10/25/13]. Available from: <http://www.illumina.com/systems/genome_analyzer_iix/performance_specifications.ilmn>.
[27] Illumina, Inc. Specification Sheet: Illumina Sequencing. MiSeq System 2013 [6/19/14]. Available from: <http://res.illumina.com/documents/products/datasheets/datasheet_miseq.pdf>.
[28] Collins FS, Hamburg MA. First FDA authorization for next-generation sequencer. N Engl J Med 2013;369(25):2369–71.
[29] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20(9):1297–303.
[30] Applied Biosystems. 5500W Series Genetic Analyzers v2.0 Specification Sheet. 2012 [6/19/14]. Available from:<http://tools.lifetechnologies.com/content/sfs/brochures/5500-w-series-spec-sheet.pdf>.
[31] Applied Biosystems by Life Technologies. Specification Sheet: Applied Biosystems SOLiD 4 System 2010 [6/19/14]. Available from: <http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_078637.pdf>.
[32] Merriman B, Ion Torrent R, Team D, Rothberg JM. Progress in Ion Torrent semiconductor chip based sequencing. Electrophoresis 2012;33(23):3397–417.
[33] Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 2012;30(5):434–9.
[34] Eshleman JR, et al. False positives in multiplex PCR-based next-generation sequencing have unique signatures. J Mol Diagn in press.
[35] Life Technologies. Ion PGM System Specifications 2013 [10/25/13]. Available from: <http://www.lifetechnologies.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-pgm-system-for-next-generation-sequencing/ion-pgm-system-specifications.html>.
[36] Life Technologies. Ion Proton System Specifications 2013 [10/25/13]. Available from: <http://www.lifetechnologies.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-proton-system-for-next-generation-sequencing/ion-proton-system-specifications.html>.
[37] Droege M, Hill B. The genome sequencer FLX™ system. J Biotechnol 2008;136:3–10.
[38] Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012;2012:251364.
[39] Roche Diagnostics. GS FLX+ System: Performance 2013 [10/25/13]. Available from: <http://454.com/products/gs-flx-system/>.
[40] Roche Diagnostics. GS Junior: Performance 2013 [10/25/13]. Available from: <http://454.com/products/gs-junior-system/index.asp>.