Review

Cancer Prevention Research

Key Principles and Clinical Applications of "Next-Generation" DNA Sequencing

Jason M. Rizzo and Michael J. Buck

Abstract

Demand for fast, inexpensive, and accurate DNA sequencing data has led to the birth and dominance of a new generation of sequencing technologies. So-called "*next-generation*" sequencing technologies enable rapid generation of data by sequencing massive amounts of DNA in parallel using diverse methodologies which overcome the limitations of Sanger sequencing methods used to sequence the first human genome. Despite opening new frontiers of genomics research, the fundamental shift away from the Sanger sequencing that next-generation technologies has created has also left many unaware of the capabilities and applications of these new technologies, especially those in the clinical realm. Moreover, the brisk evolution of sequencing technologies has flooded the market with commercially available sequencing platforms, whose unique chemistries and diverse applications stand as another obstacle restricting the potential of next-generation sequencing. This review serves to provide a primer on next-generation sequencing technologies for clinical researchers and physician scientists. We provide an overview of the capabilities and clinical applications of DNA sequencing technologies to raise awareness among researchers about the power of these novel genomic tools. In addition, we discuss that key sequencing principles provide a comparison between existing and near-term technologies and outline key advantages and disadvantages between different sequencing platforms to help researchers choose an appropriate platform for their research interests. *Cancer Prev Res; 5(7); 887–900. ©2012 AACR.*

Introduction

Initial sequencing of the human genome took more than a decade and cost an estimated $70 million dollars (1). Sequencing for the Human Genome Project (HGP) relied primarily on automation of sequencing methods first introduced by Sanger in 1977 (2). Despite the successful use this technology to generate early maps of the human genome (3–5), the limitations of Sanger sequencing created a high demand for more robust sequencing technologies capable of generating large amounts of data, quickly, and at lower costs.

Recognizing this need, the National Human Genome Research Institute (NHGRI) initiated a funding program in 2004 aimed at catalyzing sequencing technology development and with a goal of reducing the cost of genome sequencing to ~$100,000 in 5 years and, ultimately, $1,000 in 10 years (6–8). The initiative has been widely successful to date, and a bevy of new technologies has emerged in the sequencing marketplace over the past 5 years. New technologies offer radically different approaches that enable the sequencing of large amounts of DNA in parallel and at substantially lower costs than conventional methods. The terms "*next-generation sequencing*" and "*massive-parallel sequencing*" have been used loosely to collectively refer to these new high-throughput technologies.

First-Generation Sequencing

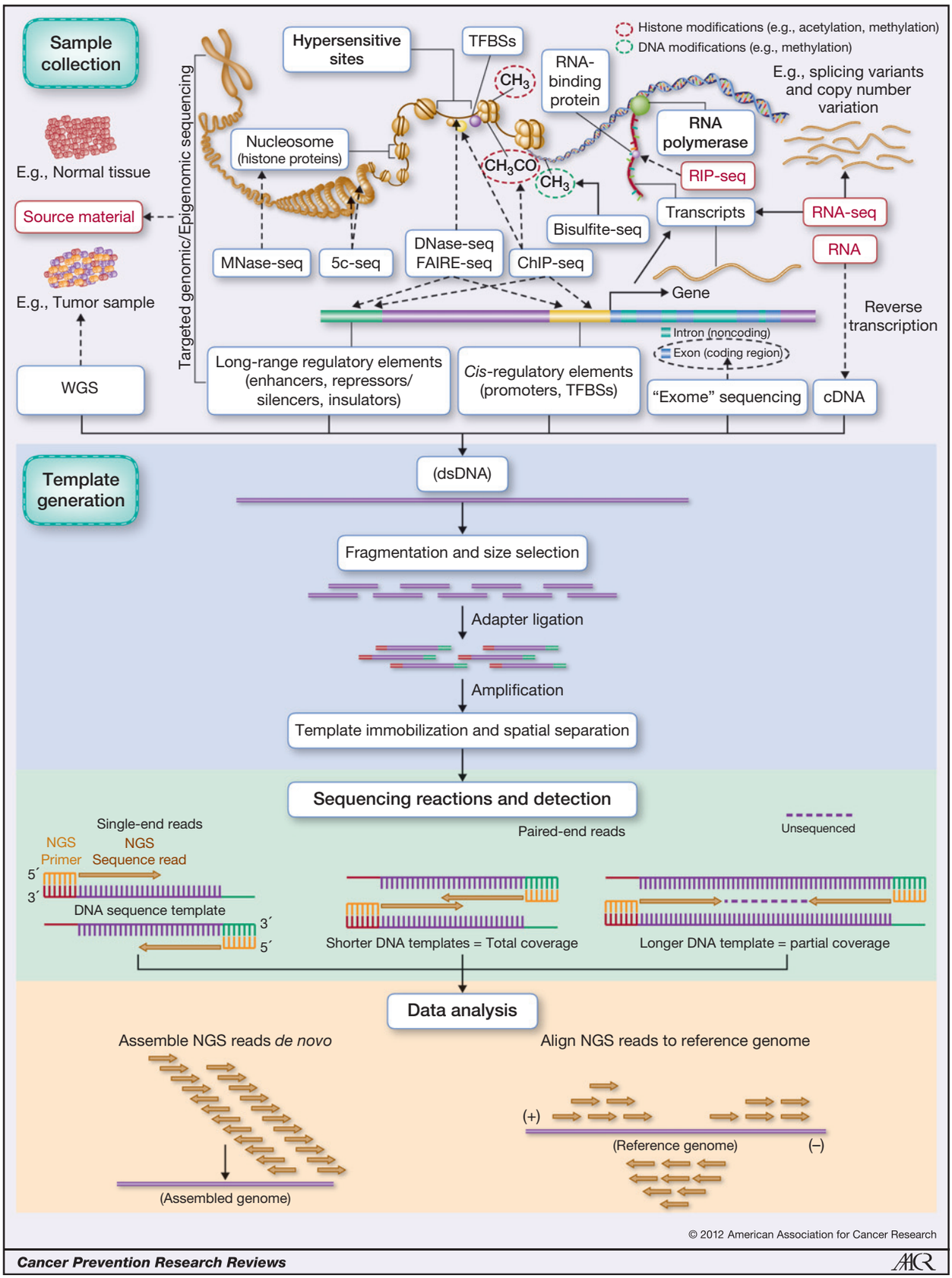
Automated Sanger sequencing is now considered the "*first-generation*" of DNA sequencing technologies. Technically, standard Sanger sequencing identifies linear sequences of nucleotides by electrophoretic separation of randomly terminated extension products (2). Automated methods use fluorescently labeled terminators, capillary electrophoresis separation, and automated laser signal detection for improved nucleotide sequence detection [ref. 9; for reviews, see the studies of Hutchinson (ref. 10) and Metzker (ref. 11)]. As a key strength, Sanger sequencing remains the most available technology today and its well-defined chemistry makes it is the most accurate method for sequencing available now. Sanger sequencing reactions can read DNA fragments of 500 bp to 1 kb in length, and this method is still used routinely for sequencing small amounts of DNA fragments and is the gold-standard for clinical cytogenetics (12).

Despite strong availability and accuracy, however, Sanger sequencing has restricted applications because of technical limitations of its workflow. The main limitation of Sanger

Authors' Affiliation: Department of Biochemistry and Center of Excellence in Bioinformatics and Life Sciences, State University of New York at Buffalo, Buffalo, New York

Corresponding Author: Michael J. Buck, State University of New York at Buffalo, 701 Elicott St., Buffalo, NY 14203. Phone: 716-881-7569; Fax: 716-849-6655; E-mail: mjbuck@buffalo.edu

doi: 10.1158/1940-6207.CAPR-11-0432



sequencing is one of throughput, that is, the amount of DNA sequence that can be read with each sequencing reaction. Throughput is a function of sequencing reaction time, the number of sequencing reactions that can be run in parallel, and lengths of sequences read by each reaction. The requirement for electrophoretic separation of DNA fragments for reading DNA sequence content in Sanger-based sequencing is the primary bottleneck for throughput with this method, increasing time and limiting the number of reactions that can be run in parallel (13). Despite efficient automation, each Sanger instrument can only read 96 reactions in parallel, and this restricts the technology's throughput to approximately 115 kb/day (1,000 bp; ref. 14). Current estimates suggest a cost of approximately $5 to 30 million USD to sequence an entire human genome using Sanger-based methods, and on one machine, it would take around 60 years to accomplish this task (8, 13). Together, these cost and time constraints limit access to and application of genome sequencing efforts on this platform.

Next-Generation Sequencing

Overview

"*Next-generation*" and "*massive-parallel*" DNA sequencing are blanket terms used to refer collectively to the high-throughput DNA sequencing technologies available which are capable of sequencing large numbers of different DNA sequences in a single reaction (i.e., in parallel). All next-generation sequencing (NGS) technologies monitor the sequential addition of nucleotides to immobilized and spatially arrayed DNA templates but differ substantially in how these templates are generated and how they are interrogated to reveal their sequences (15). A basic workflow for NGS sequencing technologies is presented in Fig. 1.

Template generation

In general, the starting material for all NGS experiments is double-stranded DNA; however, the source of this material may vary (e.g., genomic DNA, reverse-transcribed RNA or cDNA, immunoprecipitated DNA). All starting material must be converted into a library of sequencing reaction templates (sequencing library), which require common steps of fragmentation, size selection, and adapter ligation (15). Fragmentation and size selection steps serve to break the DNA templates into smaller sequence-able fragments, the size of which depend on each sequencing platform's specifications. Adapter ligation adds platform-specific synthetic DNAs to the ends of library fragments, which serve as primers for downstream amplification and/or sequencing reactions. Ideally, the above steps create an unbiased sequencing library that accurately represents the sample's DNA population. Depending on the NGS technology used, a library is either sequenced directly (single-molecule templates) or is amplified then sequenced (clonally amplified templates). Template generation also serves to spatially separate and immobilize DNA fragment populations for sequencing, typically by attachment to solid surfaces or beads. This allows the downstream sequencing reaction to operate as millions of microreactions carried out in parallel on each spatially distinct template (16).

Clonally amplified versus single-molecule templates

Most sequencing platforms cannot monitor single-molecule reactions and template amplification is therefore required to produce sufficient signal for detection of nucleotide addition by the instrument's system (1, 14). Amplification strategies vary between platforms and commonly include emulsion PCR or bridging amplification strategies (Table 1). Importantly, all amplification steps can introduce sequencing errors into experiments as reagents (DNA polymerases) are not 100% accurate and can introduce mutations into the clonally amplified template populations, which subsequently masquerade as sequence variants in downstream analysis (1). Amplification steps also reduce the dynamic range of sequence detection and potentially remove low-abundance variants from sequenced populations.

Single-molecule template sequencing bypasses the need for amplification steps and requires far less starting material for sequence detection (17). The ability to sequence extremely small quantities of DNA without manipulation gives single-molecule sequencing approaches a greater (potentially unlimited) dynamic range of sequence detection, including the possibility of sequencing DNA from only a single cell (18, 19). Currently, single-molecule sequencers are just beginning to enter the market; however, despite promises to improve signal quality and expand the types of data produced, the availability of this platform is limited, and downstream analysis pipelines are very immature compared with those of clonally amplified signals. In addition, advantages over amplification-based platforms have yet to be realized and remain equally uncertain. Amplification-based and single-molecule sequencing technologies have been referred to as "*second-generation*" and "*third-generation*"

Figure 1. Basic workflow for NGS experiments. NGS experiments consist of 4 phases: sample collection (purple), template generation (blue), sequencing reactions and detection (green), and data analysis (orange). Experiments can have broad applications, depending on the source and nature of input DNA used for sequencing. Source materials include normal and diseased tissues, from which NGS experiments can sequence the whole genome (WGS) or targeted genomic/epigenetic elements. Table 2 lists experimental approaches, descriptions, and key references for the sample collection strategies illustrated. Illustration of sample collection is modified from the study of Myers and colleagues (82). Illustration of sequencing reactions and detection is a generalized schematic and substantially differs based on the platform used (Table 1). NGS experiments can be grouped broadly into 2 general categories: *de novo* assembly and resequencing. Assembled genomes are built from scratch, without the use of an existing scaffold, whereas resequencing experiments align sequence reads back to a reference genome (orange). Since the HGP, all human genome sequencing efforts have been resequencing, as it is not cost-effective, extremely difficult, and of limited (immediate) value to reassemble a human genome (95, 96). Conversely, smaller genomes such as those of novel bacteria are routinely assembled *de novo* (97). dsDNA, double-stranded DNA; TFBS, transcription factor–binding sites.

sequencing technologies, respectively, in the literature (ref. 16; Table 1).

Sequencing reactions

Each sequencing platform uses a series of repeating chemical reactions that are carried out and detected automatically. Reactions typically use a flow cell that houses the immobilized templates and enables standardized addition and detection of nucleotides, washing/removal of reagents, and repetition of this cyclical process on a nucleotide-by-nucleotide basis to sequence all DNA templates (i.e., sequencing library) in parallel. While all sequencing platforms differ in their proprietary chemistries (Table 1), the use of DNA polymerase or DNA ligase enzyme is a common feature, and these methods have been referred to collectively as "*sequencing by synthesis*" (SBS) strategies in the literature (20).

Overall, the reading of sequence content on a nucleotide-by-nucleotide stepwise fashion used by NGS overcomes the limitations of discrete separation and detection requirements of first-generation methods and has radically improved the throughput of sequencing reactions by several orders of magnitude. Such improvements have allowed the per-base cost of sequencing to decrease by more than 100,000-fold in the past 5 years with further reductions expected (21). Cost reductions in sequencing technologies have enabled widespread use and diverse applications of these methods (see Clinical Applications of NGS and Preclinical Applications of NGS)

Paired-end and mate-paired sequencing

Typically, NGS methods sequence only a single end of the DNA templates in their libraries, with all DNA fragments afforded an equal probability of occurring in forward or reverse direction reads. Depending on the instrument and library construction protocol used, however, forward and reverse reads can be paired to map both ends of linear DNA fragments during sequencing ("paired-end sequencing") or both ends of previously circularized DNA fragments ("mate-pair sequencing"). The choice of pair-end or mate-pair depends on the clinical application and is discussed later. It is important to note that both paired-end and mate-pair approaches still only sequence the ends of DNA fragments included in sequencing libraries and therefore do not provide sequence information for the internal portion of longer templates (Fig. 1).

Limitations

The increased throughput of NGS reactions comes at the cost of read length, as the most readily available sequencing platforms (Illumina, Roche, SoLiD) offer shorter average read lengths (30–400 bp) than conventional Sanger-based methods (500–1 kb; ref. 13). Several third-generation technologies hold the promise of longer read lengths; however, these are not widely available and, as mentioned, are exceedingly immature technology platforms (ref. 22; Table 1).

Shorter read lengths place restrictions on the types of experiments NGS methods can conduct. For instance, it is difficult to assemble a genome *de novo* using such short fragment lengths (23); therefore, most application of these technologies focus on comparing the density and sequence content of shorter reads to that of an existing reference genome (known as genome "re-sequencing"; Fig. 1). In addition, shorter read lengths may not align or "map" back to a reference genome uniquely, often leaving repetitive regions of the genome unmappable to these types of experiments. Sequence alignment is also challenging for regions with higher levels of diversity between the reference genome and the sequenced genome, such as structural variants (e.g., insertions, deletions, translocations; ref. 24). These issues are combated through the use of longer read lengths or paired-end/mate-pair approaches (Fig. 2A and B). Given the relative immaturity of third-generation NGS platforms, nearly all human genome resequencing conducted today relies on the paired-end or mate-paired approaches of second-generation platforms. Paired-end sequencing is much easier than mate-paired sequencing and requires less DNA, making it the standard means by which human genomes are resequenced (14). Although more expensive and technically challenging, mate-paired libraries can sample DNA sequence over a larger distance (1.5–20 kb) than paired-end approaches (300–500 bp) and are therefore better suited for mapping very large structural changes (14).

Data analysis

Ironically, one of the key limitations of NGS also serves as its greatest strength, the high volume of data generation. NGS reactions generate huge sequence data sets in the range of megabases (millions) to gigabases (billions), the interpretation of which is no trivial task (16). Moreover, the scale and nature of data produced by all NGS platforms place substantial demands on information technology at all stages of sequencing, including data tracking, storage, and quality control (25). Together, these extensive data gathering capabilities now double as constraints, shifting the bottlenecks in genomics research from data acquisition to those of data analysis and interpretation (26). NGS machines are generating data at such a rapid pace that supply cannot keep up with demand for new analytic approaches capable of mining NGS data sets (see Future Directions and Challenges). Data analysis is a critical feature of any NGS project and will depend on the goal and type of project. The initial analysis or base calling is typically conducted by proprietary software on the sequencing platform. After base calling, the sequencing data are aligned to a reference genome if available or a *de novo* assembly is conducted. Sequence alignment and assembly is an active area of computational research with new methods being developed (see review by Flicek and Birney; ref. 27). Once the sequence is aligned to a reference genome, the data need to be analyzed in an experiment-specific fashion [for reviews, see the studies of Martin and Wang for RNA-seq (ref. 28), Park for ChIP-seq (ref. 29), Bamshad and

Table 1. Comparison of NGS platforms

Sequencing platform	Library/template preparation	Sequencing reaction chemistry	Maximum read length, bp	Run time, d	Maximum throughput per run (total bp sequenced)	Strengths	Limitations
Illumina HiSeq 2000	Bridging amplification	Reverse terminator	100	2[a], 11[b]	95–600 Gb[d]	Most widely used platform; large throughput	All samples on flow cell sequenced at same read length
Illumina MiSeq	Bridging amplification	Reverse terminator	250	0.17[a], 1.1[b]	440 Mb–7 Gb	Short run times	Low number of total reads (~15 million)
Roche Genome Sequencer FLX	Emulsion PCR	Pyrosequencing	400	0.4	0.5–0.6 Gb	Longer reads; fast run times	High reagent cost; lowest number of total reads (~1 million)
SoLiD/ABI 5500	Emulsion PCR	Ligation sequencing	75	2[a], 7[b]	90–300 Gb	Independent flow cell lanes; high capacity for multiplexing	Short read lengths
Ion Personal Genome Machine	Emulsion PCR	Ion sequencing	200	0.1	1 Gb	Short run times; low-cost scalable machine	Low number of total reads (~11 million)
Complete Genomics	PCR on DNA nanoballs	Ligation sequencing	70	12	20–60 Gb	Complete service for human sequencing	High cost per sample; only available for human resequencing
Helicos	Single molecule	Reverse terminator	55	8	21–35 Gb	No amplification bias	Machine not widely used; sequencing service available through company
PacBio RS	Single molecule	Real-time	1,000	<0.1[c]	N/A	Potential for longest read lengths and shorter run times	Highest error rates

NOTE: Amplification-based and single-molecule sequencing technologies have been referred to as second- and third-generation sequencing technologies, respectively, in the literature (16). The term third-generation sequencing has also been used to refer to near-term nanopore sequencing technologies. Nanopore sequencing is not covered in this review, and readers are directed to an article by Branton and colleagues (90).

Abbreviation: N/A, not applicable.

[a]Single-end sequencing.

[b]Pair-end sequencing.

[c]Company estimate.

[d]Two flow cells.

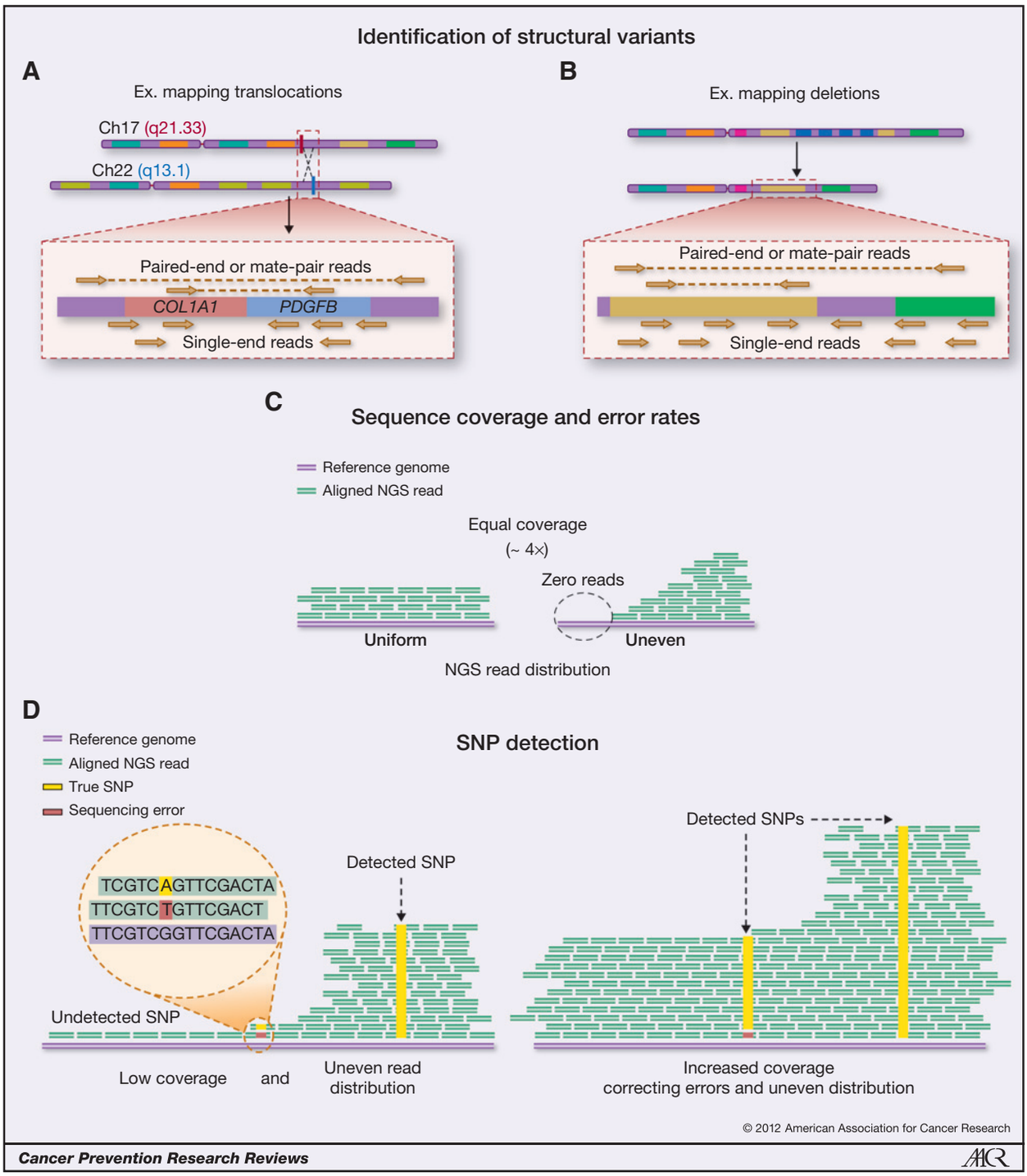


Figure 2. Key NGS principles. A and B, identification of structural variants: Longer (paired-end or mate-pair) sequencing reads are more adept at mapping large structural variations (e.g., translocations and deletions) because they provide added information concerning which sequences co-occur on the same template. A, illustration of well-documented translocation between chromosomes 17 and 22 which places the platelet-derived growth factor-β (*PDGFB*) gene under control of the highly active collagen type 1A1 promoter (*COL1A1*). This translocation is implicated in the pathogenesis of the rare cutaneous malignancy dermatofibrosarcoma protuberans (DFSP; ref. 98). Notice how alignment of short reads does not distinguish the mutated sequence (translocation) from the normal reference genome. B, illustration of a hypothetical chromosomal deletion mutant. Again, notice how alignment of short reads does not distinguish the mutated sequence (deletion) from the normal reference genome. C, sequence coverage and error rates. Illustration of uniform and uneven NGS read distributions for resequencing experiments. An uneven read distribution can leave regions of the genome uncovered (black circle). D, SNP detection. Left, illustration of how low coverage, uneven read distributions, and high error rates can interact to confound genotype detection, including SNP calling (as illustrated). Right, illustration of how uneven read distributions and errors can be overcome by higher coverage rates.

colleagues for exome sequencing (ref. 30), Medvedev and colleagues for whole-genome resequencing (ref. 31), and Wooley and colleagues for metagenomics (ref. 32)].

Choosing a Sequencing Technology

Genomics experiments are largely descriptive and afford researchers the opportunity to explore a biologic question in a comprehensive manner. Experimental design is paramount for the success of all genomics experiments, and choice of sequencing strategy should be informed by the goal(s) of the project.

Experimental design and biases

It is important to recognize that bias can be introduced at all steps in an experimental NGS protocol. This principle is best illustrated by the example of template amplification steps, which can introduce mutations into clonally amplified DNA templates that subsequently masquerade as sequence variants. Amplification steps also reduce the dynamic range of sequence detection and potentially remove low-abundance variants from sequenced populations. Much like amplification steps, *any* sample manipulation can cause quantitative and qualitative artifacts in downstream analysis (22, 33). Therefore, it is important to design your experiment in a way that maximizes the collection of sequence information you covet and minimizes the biases against it. For example, if your sample is being extensively manipulated before sequencing, a reference DNA sample with known sequence content and similar size/quantity should also be carried through your sequencing protocol and analyzed in parallel as a control.

Another important consideration is the quantity and quality of the DNA you choose to sequence. Most NGS platforms have proprietary library preparations that are optimized for a specific DNA quantity and quality. These input metrics are typically easy to achieve using unmodified fresh or fresh-frozen samples, however, this may be more challenging with clinical specimens, especially archived formalin-fixed, paraffin-embedded (FFPE) samples. Thankfully, the use of clinical specimens with limited DNA quality/quantity for NGS is an area of active research, and work has shown that NGS platforms can handle this input material (34, 35). Despite this evidence, it is highly likely that any deviation in sample quality/quantity from an NGS platform's optimized protocol will still require extensive troubleshooting by the user. An appropriate control under such circumstances would again be the sequencing of reference DNA treated with the same conditions (e.g., FFPE) before sequencing.

Sequencing coverage and error rates

Both the quality and quantity of sequence data derived from NGS experiments will ultimately determine how comprehensive and accurate downstream analyses can be (36). Qualitatively, individual base calling error rates vary between NGS platforms (Table 1). All NGS platforms provide confidence scores for each individual base call, enabling researchers to use different quality filters when mining their sequence data. More generally, the chemistries of most NGS reactions are such that the initial portion of each sequencing read is typically more accurate than the latter (due to signal decay).

Quantitatively, the amount of sequence data can be assessed by the metric of sequencing "coverage." Generally speaking, sequence coverage (also called "depth") refers to the average number of times a base pair is sequenced in a given experiment. More specifically, this coverage metric is best viewed in the context of the physical locations (distribution) of these reads, as NGS reactions may not represent all genomic locations uniformly (due to handling, platform biases, run-to-run variation; ref. 36). For example, local sequence content has been shown to exert a bias on the coverage of short read NGS platforms, whereby higher read densities are more likely to be found at genomic regions with elevated GC content (37). Such coverage biases can interfere with quantitative applications of NGS, including gene expression profiling (RNA-seq) or copy number variation analysis. Several methods have been developed to account for the nonuniformity of coverage and adjust signals for GC bias to improve the accuracy of quantitative analysis (37–39). Qualitatively, uneven sequence coverage can also interfere with the analysis of sequence variants. For example, a deeply sequenced sample with nonuniform read distribution can still leave a substantial portion of the genome unsequenced or undersequenced, and analysis of these regions will not be able to identify sequence variations such as single-nucleotide polymorphisms (SNP), point mutations, or structural variants, because these locations will either be unsequenced or confounded by sequencing errors (Fig. 2C and D).

Ultimately, coverage depth, distribution, and sequence quality determine what information can be retrieved from each sequencing experiment. In theory, an experiment with 100% accuracy and uniform coverage distribution would provide all sequence content information (including identification of SNPs and complex structural variants) with just 1× coverage depth. In reality, however, accuracy is never 100% and coverage is not uniform; therefore, deeper sequence coverage is needed to enable correction of sequencing errors and to compensate for uneven coverage (Fig. 2D). For discovery of structural variants (e.g., insertions, deletions, translocations), accurate identification of a complete human genome sequence with current (second-generation) platforms, requires approximately 20× to 30× sequence coverage to overcome the uneven read distributions and sequencing errors (22, 24). Higher coverage levels are required to make accurate SNP calls from an individual genome sequence, as these experiments are powered differently (36). Standards are evolving and current recommendations range from 30× to 100× coverage, depending on both platform error rate and the analytic sensitivity and specificity desired (12, 24, 36, 40). These higher coverage requirements are why the cost of whole-genome sequencing (WGS) still remains above $1,000 for many sequencing applications (16). Newer single-molecule sequencers

promise more evenly distributed and longer reads, potentially providing a complete genome sequence at lower costs; however, this ability often comes at the cost of higher error rates (1, 22). To save costs, population-scale sequencing projects, such as the 1,000 genomes project, have used low-coverage pooled data sets and are able to detect SNP variants with frequencies as low as 1% in DNA populations with 4× coverage and error rates common to second-generation platforms (41). With this approach, an investigator probing for high-frequency variants (>1%) in a select population could also succeed with lower coverage sequencing experiments to reduce costs.

Clinical Applications of NGS

Investment in the development of NGS technologies by the NHGRI was made with the goal of expediting the use of genome sequencing data in the clinical practice of medicine. As the cost of DNA sequencing continues to drop, the actual translation of base pair reads to bedside clinical applications has finally begun. Several small-scale studies have laid the foundation for personalized genome-based medicine, showing the value of both whole-genome and targeted sequencing approaches in the diagnosis and treatment of diseases. These findings are the first of many to follow, and this progression, coupled to the expanding definition of genetic influences on clinical phenotypes identified by preclinical studies, will place NGS machines among the most valuable clinical tools available to modern medicine.

Exome and targeted sequencing

At present, a small percentage of the human genome's sequence is characterized (<10%), and limited clinically valuable information can be immediately gained from having a patient's complete genome sequence at this time. Therefore, it is often more cost-effective for clinical researchers to sequence only the exome (the 2% of the genome represented by protein-coding regions or exons), the "Mendelianome" (coding regions of 2,993 known disease genes), or targeted disease gene panels to screen for relevant mutations in the diagnosis and treatment of disease (30, 42). In targeted NGS reactions, sequencing reads are intentionally distributed to specific genomic locations, which allows for higher sequencing coverage and therefore ensures accurate detection of sequence variants at these loci, regardless of platform error rates. To target sequencing reads to specific genomic locations, DNA regions of interest are enriched before sequencing reactions using capture strategies. Enriched regions are then loaded as input onto the sequencer in place of using whole-genome DNA. Commonly used enrichment strategies include hybrid capture, microdroplet PCR, or array capture techniques [refs. 42–44; for performance review, see the study of Mamanova and colleagues (ref. 45)].

Targeted whole-exome sequencing has already proven valuable in the clinic, helping physicians make difficult diagnoses by providing a comprehensive view of the genetic makeup of their patients. Choi and colleagues first showed the value of whole-exome sequencing in the clinic by making genetic diagnoses of congenital chloride diarrhea in a cohort of patients referred with suspected cases of Bartter syndrome, a renal salt-wasting disease (46). In this study, exome sequences were captured using array hybridization and then sequenced using the paired-end approach (Illumina platform). Exome sequencing was conducted on 6 patients who lacked mutations in known genes for Bartter syndrome on standard clinical (site-specific) sequencing. Results revealed homozygous deleterious mutations in the *SLC26A3* locus for all 6 patients, which enabled a molecular diagnosis of congenital chloride diarrhea that was subsequently confirmed on clinical evaluation. This result was the first to show the value of exome sequencing in making a clinical diagnosis and several similar studies have followed. For example, both Bolze and colleagues and Worthey and colleagues used exome-sequencing to aid molecular diagnoses in patients with novel diseases when standard diagnostic approaches were exhausted for cases of autoimmune lymphoproliferative syndrome (ALPS) and childhood inflammatory bowel disease, respectively (47, 48).

Other targeted sequencing strategies have also been used successfully in the clinic. For example, NGS methods have been used to extend preconception screening tests to include 448 severe recessive childhood diseases that were previously impractical under single-gene testing models (49). Similarly, targeted NGS efforts have been used to screen panels of disease-relevant genes, including the successful development and implementation rapid diagnostic cancer gene panels in the genetics departments at Washington University Medical School (St. Louis, MO) and Baylor College of Medicine (Houston, TX; refs. 42, 50–52).

Overall, this ability to target multiple candidate genes and retrieve fast, accurate, and cheap sequencing data is revolutionizing the field of cytogenetics. Previously, high costs excluded the use of genetic testing in diagnosing diseases featuring complex genetics (e.g., multiple genes contributing) or those without distinct clinical features (i.e., uncertain which gene to test; ref. 12). At present, NGS research findings are currently used to guide diagnostic Sanger sequencing for confirmation of clinical diagnosis, as clinical standards have not been established for the accuracy and interpretation of research-grade NGS data (12, 49) Ultimately, improvements in sequencing accuracy, costs, and analysis, including development of testing standards will make NGS testing the gold-standard for clinical cytogenetics.

Preclinically, targeted sequencing is also helping to expand the characterization of genetic contributions to different diseases. For example, Jiao and colleagues used exome sequencing to explore the genetic basis of 10 nonfamilial pancreatic neuroendocrine tumors and then screened the most commonly mutated genes in 58 additional samples using standard Sanger sequencing (53). Results identified frequent mutations in *MEN1* (44% tumors) *DAXX*/*ATRX* (43%), and mTOR pathway genes (14%). Interestingly, mutations *MEN1* and *DAXX*/*ATRX* genes showed clinical potential, as their presence was associated with an improved

prognosis. In addition, implication of mutations in the mTOR pathway was also of clinical interest, given the availability of mTOR inhibiting therapeutics.

Whole-genome sequencing

Thanks to the significant decrease in sequencing costs afforded by NGS technologies, it is becoming more cost-effective to resequence entire human genomes from clinical samples and this will soon be routine in the clinical practice of medicine (12, 54). Welch and colleagues provide an early example of how WGS with NGS has been successfully used in a clinically relevant time frame to alter the treatment plan of a patient with cancer (55). Results from this study successfully identified a *PML-RARA* fusion event using NGS-WGS on a difficult diagnostic case of acute promyelocytic leukemia. Importantly, the fusion event detected by WGS-NGS was identified and validated in just 7 weeks from biopsy and allowed for a change in treatment plan of this patient to be realized clinically. More importantly, however, this genetic rearrangement was not detectable using standard cytogenetic techniques, and the diagnosis and tailored therapy would not have been possible without using WGS-NGS methods.

In a similar fashion, Roychowdhury and colleagues implemented a pilot study to explore the practical challenges of applying NGS to clinical oncology, including assessing the ability of NGS methods to identify informative mutations in a clinically relevant time frame (56). This study used an integrative NGS approach that included WGS and targeted exome sequencing and was able to successfully develop and implement a clinical protocol that identified individualized mutational landscapes for tumors from patients with metastatic colorectal cancer and malignant melanoma. Importantly, mutations in these tumors were identified within 24 days of biopsy and ultimately enabled enrollment in biomarker-driven clinical trials in oncology.

Preclinically, WGS experiments offer enormous potential for identifying novel disease-relevant genetic abnormalities, especially for complex diseases such as cancer. Thus far, a variety of cancer genomes have been successfully sequenced using NGS methods, and results have yielded unprecedented insights into the mutational processes and gene regulatory networks implicated in disease progression. For example, Pleasance and colleagues used paired-end sequencing to generate a comprehensive catalog of somatic mutations in a malignant melanoma genome (57). Interestingly, results revealed a dominant mutational signature relating prior UV exposure in addition to identifying 470 novel somatic substitutions and 42 previously cataloged mutations. Ding and colleagues also used NGS to sequence cancer genomes from metastasis and primary breast cancer samples to investigate the mutational evolution of cancer cells, identifying two *de novo* mutations and a large deletion acquired by metastatic cells during disease progression and 20 mutations shared by primary and metastatic cell populations (58).

Ultimately, more WGS, rather than targeted efforts, will be needed to assign functionality to the remainder (and majority) of the genome's sequence and its role(s) in diseases such as cancer. As sequencing costs continue to drop, the economies of targeted genome sequencing will no longer exceed the value of having a complete genome sequence. Results from present association studies underscore the use of studying a complete genome sequence, finding that the vast majority (>80%) of disease-associated sequence variants fall outside of coding regions (59, 60). This finding is not incredibly surprising given that the majority of the genome consists of noncoding regions, which include gene promoters, enhancers, and intronic regions (61).

Other clinical applications

Additional clinical applications of NGS include the sequencing of cell-free DNA fragments circulating in a patient's bloodstream. For example, Snyder and colleagues used NGS to show that increased levels of cell-free DNA from a heart transplant donor's genome can be found in a recipients bloodstream when a transplant recipient is undergoing an acute cellular rejection, as validated by endomyocardial biopsy (62). Similar to detection of a low-frequency variant, this experiment relied on deep sequencing (very high coverage) of cell-free DNA to detect changes in a small fraction of that DNA population which belonged to the organ donor. This result shows the potential of using NGS as a noninvasive method for detecting solid organ transplant rejection. Similarly, Palomaki and colleagues showed the potential of NGS as a noninvasive method for detecting Down syndrome and other fetal aneuploidies (trisomy 13 and 18) by sequencing the subpopulation of cell-free DNA in a pregnant mother's bloodstream belonging to her fetus (63, 64). Results from this study showed the promise of an NGS plasma-based DNA test that can detect Down syndrome and other aneuploidies with high sensitivity and specificity. Together, sequencing of cell-free DNA by NGS in both of these examples offers enormous potential to reduce invasive medical procedures and associated morbidity/mortality.

Preclinical Applications of NGS

Conceptually, an NGS instrument can be thought of as a digital measuring stick for genome cartography efforts, providing information on the nucleotide content, relative abundance, and genomic locations of the DNA templates and its sequences. While the actual sequencing conducted by NGS machines has been automated, the source of DNA loaded onto each machine can vary, which radically expands the genomic landscapes that NGS technologies can map (Fig. 1). Preclinical applications of NGS include all experiments that characterize the genetic and/or epigenetic profiles of disease states to enhance our understanding of the molecular basis for disease pathogenesis. Results from preclinical studies facilitate the identification of novel mutations and biomarkers for future diagnostic, prognostics, or therapeutic interventions; however, these have yet to be directly applied to care in a clinical setting.

Transcriptome sequencing (RNA-seq)

In addition to sequencing DNA to catalog the genetic alterations in normal and disease samples, NGS technologies can also be used to sequence RNA populations to identify all of the genes that are transcribed from that DNA (termed the "transcriptome"). Transcribed sequences also include untranslated (non–protein-coding) RNA species such as microRNAs. For sequencing with NGS, RNA sequences must first be converted to cDNA by reverse transcription, as NGS sequencing reactions require DNA substrates.

Results from RNA-sequencing provide information on how DNA sequences have been rearranged before their transcription and also provide quantitative information on relative abundance (expression level) of those RNA sequences. Because these data only provide information on transcribed ("expressed") sequences, it potentially enriches for functionally relevant mutations. For example, Shah and colleagues analyzed the transcriptomes of 4 adult granulose cell tumors (GCT) using paired-end RNA-sequencing and compared their sequencing results with the transcriptomes of 11 epithelial ovarian tumors and published sequences of the human genome (65). Results from this transcriptomic analysis were able to identify a single recurrent somatic in the *FOX2* gene that was specific to 3 of the GCT tumors and was also present in 97% of 89 additional adult GCTs they tested on follow-up Sanger sequencing. In addition, the *FOXL2* gene is known to encode a transcription factor known to be involved in granulose cell development (65). Together these results provide evidence that mutation in *FOXL2* is a potential driver in the pathogenesis of adult GCTs. More importantly, the sensitivity, specificity, and reproducibility of these results testify to the power of NGS approaches at identifying functionally relevant somatic DNA mutations by transcriptomic mapping. Similar RNA-seq studied using NGS have continued to identify and implicate key somatic mutations in oncogenesis, including frequent disruption of the *ARID1A* tumor suppressor gene identified by Wiegand and colleagues in clear cell and endometroid carcinomas (66) and mutations in the *DICER1* genes of nonepithelial ovarian cancers identified by Heravi-Moussavi and colleagues (67). Interestingly, oncogenic mutations in *ARID1A* and *DICER1* genes were again shown to be reproducibly identified in other samples with high specificity by alternative means. More importantly, these mutations were also shown to alter the gene function *in vivo* in a fashion that correlated with the tumor's clinical behavior.

RNA-seq experiments can also detect important gene rearrangements that lead to fusion genes. Maher and colleagues identified novel fusion genes in commonly used cancer cell lines and successfully applied the same technique to identify novel *ETS* gene fusions in 2 prostate tumor samples (68). Pflueger and colleagues expanded on these findings to identify 7 novel cancer-specific gene fusions by conducting RNA-seq on 25 prostate cancer samples. Interestingly, findings from this study again identified common gene fusions involving *ETS* genes along with non-*ETS* gene fusions of lower frequencies, suggesting that *ETS* rearrangements in prostate cancer may predispose tumors to additional gene rearrangements (69). In a similar fashion, Prensner and colleagues also sequenced RNA populations from prostate tissues and cell lines but instead focused on the identification of RNA sequences without associated protein products (noncoding RNAs; ncRNAs). Results from this study identified 121 novel prostate cancer–associated noncoding RNA transcripts (*PCAT*), whose expression patterns stratified patient tissues into molecular subtypes and therefore could potentially serve as biomarkers for disease (70).

Other preclinical applications

Additional preclinical applications of NGS include targeted sequencing of genomic loci harboring specific features or modifications relevant to genome biology. For example, methylation of DNA is a heritable epigenetic modification whose localization can be mapped across a genome using NGS following specific DNA treatment/enrichment protocols (bisulfite-seq; refs. 71, 72). Zhang and colleagues recently used this technique to show how expression of cancer-related genes is affected by the epigenetic marks in retinoblastoma tumors (73).

Other preclinical applications of NGS include targeted sequencing of transcription factor–binding sites (ChIP-Seq; ref. 29), regulatory regions (FAIRE-Seq and DNase-Seq; refs. 74–76), and chromatin structure (MNase-Seq; refs. 77, 78; Fig. 1). Moreover, NGS technologies can also be used to characterize microbial populations in a culture-independent fashion, by probing for the presence and abundance of microbe-specific DNA sequences in both tumor and normal human cell environments (metagenomics; refs. 32, 79, 80). A more comprehensive list of NGS applications and key references is provided in Table 2.

Large-Scale Genome Sequencing Projects

Several large-scale preclinical genome sequencing efforts are also underway to help expedite the characterization of both normal and tumor genomes. These efforts involve multiple institutions and are funded by the NIH (Bethesda, MD) with a mandate of making all of their sequencing data and protocols publicly available, allowing researchers to use, integrate, and expand upon this work.

For example, a population-scale sequencing effort known as the 1,000 Genomes project is collecting WGS data from a diverse sampling of individuals to map patterns of inheritance, typically focusing on the most common form of genetic variation, the SNP (41, 59). These so-called genome-wide association studies (GWAS) are powered by their scale (identifying millions of SNPs in thousands of diverse individuals) and have already provided valuable insights on the genetics of complex diseases (60).

Another large-scale project is the Cancer Genome Atlas (TCGA), which is funding a national network of research and technology teams working to comprehensively identify and catalog all genetic alterations found in all cancer types

Table 2. Applications of NGS technologies

Experiment	Source DNA (input)	Description	References
WGS	gDNA	Identifies an individual's complete genome sequence (coding and noncoding regions); including copy number variation (e.g., repeats, indels) and structural rearragements (e.g., translocations)	(1, 16)
Targeted "exome" sequencing	Protein-encoding gDNA (i.e., exons)	Identifies the sequence for all coding regions (exons), including copy number variation (e.g., repeats, indels) and structural rearrangements (e.g. translocations)	(16, 45)
RNA-seq	cDNA made from various sources of RNA	Can identify all transcribed sequences (transcriptome) or just coding RNA sequences; can also provide information on sequence content (e.g., splicing variants) and copy number/abundance (e.g., gene expression profiling)	(28)
Bisulfite-seq	Bisulfite-treated DNA	Identifies sites of DNA methylation (e.g., genetic imprinting)	(71, 72)
ChIP-seq	Immunoprecipitated DNA	Identifies sites of protein–DNA interactions such as transcription factor–binding sites	(29)
RIP-seq	cDNA made from immunoprecipitated RNA	Identifies sites of protein–RNA interactions; a ChIP-seq for RNA-binding proteins	(91)
DNase-seq	DNase-digested chromatin DNA	Identifies genomic regions susceptible to enzymatic cleavage by DNase, i.e., hypersensitive sites and potential regulatory regions	(75, 76)
FAIRE-seq	Open/accessible chromatin DNA	Identifies open/accessible chromatin regions, i.e., hypersensitive sites and potential regulatory regions	(74, 75)
MNase-seq	Nucleosome-associated DNA	Identifies nucleosome positions on genomic DNA (i.e., primary chromatin structure); also provides information on histone/ nucleosome density at each location	(77, 92)
Hi-C/5C-seq	Captured chromosome conformations	Identifies intra- and interchromosomal interactions; determines the spatial organization of chromosomes at high resolution	(93, 94)
Metagenomics	Microbial DNA populations	Genomic analysis of microbial communities; identifies bacterial/viral populations present in specific environments (e.g., human gut and tumor samples)	(32, 79, 80)

NOTE: Immunoprecipitated (IP) DNA and RNA can be collected for any protein that has an antibody or using an epitope-tagged protein. IP DNA sources can include histone proteins (e.g., histone H3 or H4), as a paired or alternative approach to MNase-seq experiments (77). IP DNA sources can also include covalently modified histone proteins (i.e., specific histone acetylations/methylations; e.g., H3K36Me3)–to map "histone code."

Abbreviations: cDNA, reverse-transcribed RNA or "complementary DNA" (i.e., introns removed during RNA splicing); Chromatin, the collection of DNA and proteins in the nucleus; openness/accessibility of chromatin, regions of looser DNA packaging, susceptible to enzymatic cleavage and protein binding/gene regulation (see Fig. 1: "Hypersensitive Sites"); indels, insertions/deletions; transcriptome, all transcribed DNA sequences, includes small noncoding RNAs, miRNAs, and coding RNAs (i.e. genes).

using NGS methods (81). Similarly, projects such as the Epigenomics Roadmap and the Encyclopedia of DNA Elements (ENCODE) are funding consortiums of researchers to leverage NGS experimental pipelines to generate comprehensive maps of genomic and epigenomic elements in normal human and cancer cell lines (ENCODE; ref. 82) and stem cells and primary *ex vivo* tissues (Epigenomics Roadmap; ref. 83).

Future Directions and Challenges

Despite opening broad new vistas in the exploration of biomedical questions on genome-wide scales, the rapid ascent and evolution of NGS technologies has outpaced the development of other vital resources needed for these technologies to achieve their full potential.

Data analysis and computational infrastructure

The most pressing challenge to NGS technologies is the need to expedite the use of NGS data in the clinical practice of medicine, translating base pair reads to bedside applications (54). Certainly, faster sequencing machines, such as real-time sequencers, can dramatically reduce the time it takes to retrieve raw sequence data (84). However, despite the availability of faster data acquisition techniques, the primary bottleneck in realizing the clinical potential of NGS technologies is one of data analysis (26). Analysis of large sequencing data sets mandates an advanced computational infrastructure dedicated to preserving, processing, and analyzing NGS data. NGS platforms such as the Illumina MiSeq and the Ion Torrent (Table 1) have been developed with the aim of providing a more user-friendly sequencing experience, including automated data analysis and storage pipelines; however, these machines come at the cost of lower throughput which restricts the types of analyses they can conduct. Accordingly, there is a dire need for the development of novel data analysis techniques to interpret NGS data sets from human samples and for more efficient and user-friendly bioinformatics pipelines for all platforms to incorporate these methods and expand their application in the scientific and medical communities (85).

Integration of NGS data sets also presents added challenges, as genomics experiments now span multiple technology platforms, different from the single Sanger sequencing platform previously used for genetic analyses. Moreover, early studies suggest that NGS data may be most valuable clinically when used in an integrative approach that gathers multiple genomic data sets (e.g., RNA-seq, WGS, and exome-seq) to facilitate clinical action (56). Ultimately, one can envision a scenario where various NGS data sets are collected for a patient over their lifetime to help guide clinicians in tailoring both preventative medicine and medical diagnosis/treatments.

Medical education

Appropriate and effective use of NGS will only be realized once practicing physicians have been trained and educated to its uses. Training of both physicians and genetic counselors will have to adapt to increase focus on NGS technology and whole-genome analyses in addition to the single disease gene focus of current classical Mendelian genetics. In addition, institutional changes will also be necessary, including the addition of genetic counselors to all clinical departments to facilitate the application of genome-based medicine. Success stories from early clinical pilot studies using NGS underscore the value of using a mutidisciplinary team dedicated to the collection and interpretation of NGS data in a clinically relevant time frame (55, 56).

Political and social issues

Finally, numerous political and societal challenges are presented by the burgeoning era of personal genomics and stand as major obstacles limiting the potential of all NGS technologies (86, 87). Issues surrounding the appropriate disclosure and use of NGS data, necessary adaptations of patent law to facilitate the clinical use of NGS technologies, and establishment of insurance reimbursement protocols for NGS testing must also be addressed in the coming years to enable society to embrace the transition to a post-genomic era of medicine (87–89).

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Grant Support

This work was funded by an NSF grant IIS1016929 to M.J. Buck and a PhRMA predoctoral fellowship in Informatics to J.M. Rizzo.

Received September 19, 2011; revised March 16, 2012; accepted April 18, 2012; published OnlineFirst May 22, 2012.

References

1. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet 2010;11:31–46.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977;74:5463–7.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
4. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291:1304–51.
5. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931–45.
6. Mardis ER. Anticipating the 1,000 dollar genome. Genome Biol 2006;7:112.
7. Schloss JA. How to get genomes at one ten-thousandth the cost. Nat Biotechnol 2008;26:1113–5.
8. Bennett ST, Barnes C, Cox A, Davies L, Brown C. Toward the 1,000 dollars human genome. Pharmacogenomics 2005;6:373–82.
9. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 1987;238:336–41.
10. Hutchison CA III. DNA sequencing: bench to bedside and beyond. Nucleic Acids Res 2007;35:6227–37.
11. Metzker ML. Emerging technologies in DNA sequencing. Genome Res 2005;15:1767–76.
12. Kingsmore SF, Saunders CJ. Deep sequencing of patient genomes for disease diagnosis: when will it become routine? Sci Transl Med 2011;3:87ps23.
13. Hert DG, Fredlake CP, Barron AE. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. Electrophoresis 2008;29:4618–26.
14. Mardis ER. A decade's perspective on DNA sequencing technology. Nature 2011;470:198–203.

15. Linnarsson S. Recent advances in DNA sequencing methods - general principles of sample preparation. Exp Cell Res 2010;316:1339–43.
16. Natrajan R, Reis-Filho JS. Next-generation sequencing applied to molecular diagnostics. Expert Rev Mol Diagn 2011;11:425–44.
17. Hart C, Lipson D, Ozsolak F, Raz T, Steinmann K, Thompson J, et al. Single-molecule sequencing: sequence methods to enable accurate quantitation. Methods Enzymol 2010;472:407–30.
18. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 2009;6:377–82.
19. Navin N, Hicks J. Future medical applications of single-cell sequencing in cancer. Genome Med 2011;3:31.
20. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, et al. The challenges of sequencing by synthesis. Nat Biotechnol 2009;27:1013–23.
21. Lander ES. Initial impact of the sequencing of the human genome. Nature 2011;470:187–97.
22. Thompson JF, Milos PM. The properties and applications of single-molecule DNA sequencing. Genome Biol 2011;12:217.
23. Nagarajan N, Pop M. Sequencing and genome assembly using next-generation technologies. Methods Mol Biol 2010;673:1–17.
24. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 2011;12: 443–51.
25. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet 2008;24:142–9.
26. McPherson JD. Next-generation gap. Nat Methods 2009;6:S2–5.
27. Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat Methods 2009;6:S6–12.
28. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet 2011;12:671–82.
29. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 2009;10:669–80.
30. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 2011;12:745–55.
31. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 2009;6:S13–20.
32. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol 2010;6:e1000667.
33. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 2008;36:e105.
34. Wood HM, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic Acids Res 2010;38:e151.
35. Schweiger MR, Kerick M, Timmermann B, Albrecht MW, Borodina T, Parkhomchuk D, et al. Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis. PLoS One 2009;4:e5548.
36. Ajay SS, Parker SC, Ozel Abaan H, Fuentes Fajardo KV, Margulies EH. Accurate and comprehensive sequencing of personal genomes. Genome Res 2011;29:1498–505.
37. Taub MA, Corrada Bravo H, Irizarry RA. Overcoming bias and systematic errors in next generation sequencing data. Genome Med 2010;2:87.
38. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 2010;465:473–7.
39. Li J, Jiang H, Wong WH. Modeling non-uniformity in short-read rates in RNA-Seq data. Genome Biol 2010;11:R50.
40. Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, et al. A highly annotated whole-genome sequence of a Korean individual. Nature 2009;460: 1011–5.
41. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061–73.
42. Shen P, Wang W, Krishnakumar S, Palm C, Chi AK, Enns GM, et al. High-quality DNA sequence capture of 524 disease candidate genes. Proc Natl Acad Sci U S A 2011;108:6549–54.
43. Tewhey R, Warner JB, Nakano M, Libby B, Medkova M, David PH, et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol 2009;27:1025–31.
44. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007;4:903–5.
45. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods 2010;7:111–8.
46. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A 2009;106:19096–101.
47. Bolze A, Byun M, McDonald D, Morgan NV, Abhyankar A, Premkumar L, et al. Whole-exome-sequencing-based discovery of human FADD deficiency. Am J Hum Genet 2010;87:873–81.
48. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med 2011;13:255–62.
49. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 2011;3:65ra4.
50. Holbrook JD, Parker JS, Gallagher KT, Halsey WS, Hughes AM, Weigman VJ, et al. Deep sequencing of gastric carcinoma reveals somatic mutations relevant to personalized medicine. J Transl Med 2011;9:119.
51. Heger M. Wash U Med School Offers 28-Gene Cancer Dx Panel on HiSeq through CLIA Lab. Clinical Sequencing News 2011 [cited 2011 Nov 30]. Available from: http://www.genomeweb.com/sequencing/wash-u-med-school-offers-28-gene-cancer-dx-panel-hiseq-through-clia-lab.
52. Karow J. Baylor's Cancer Genetics Lab to Offer Ion AmpliSeq Cancer Panel on PGM. Clinical Sequencing News 2011 [cited 2011 Nov 30]. Available from: http://www.genomeweb.com/sequencing/baylors-cancer-genetics-lab-offer-ion-ampliseq-cancer-panel-pgm
53. Jiao Y, Shi C, Edil BH, de Wilde RF, Klimstra DS, Maitra A, et al. DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. Science 2011;331:1199–203.
54. Green ED, Guyer MS. Charting a course for genomic medicine from base pairs to bedside. Nature 2011;470:204–13.
55. Welch JS, Westervelt P, Ding L, Larson DE, Klco JM, Kulkarni S, et al. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. JAMA 2011;305:1577–84.
56. Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM, Cao X, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci Transl Med 2011;3:111ra21.
57. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 2010;463:191–6.
58. Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature 2010;464:999–1005.
59. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747–53.
60. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med 2010;363:166–76.
61. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. Annotating non-coding regions of the genome. Nat Rev Genet 2010;11:559–71.
62. Snyder TM, Khush KK, Valantine HA, Quake SR. Universal noninvasive detection of solid organ transplant rejection. Proc Natl Acad Sci U S A 2011;108:6229–34.
63. Palomaki GE, Kloza EM, Lambert-Messerlian GM, Haddow JE, Neveux LM, Ehrich M, et al. DNA sequencing of maternal plasma to detect Down syndrome: an international clinical validation study. Genet Med 2011;13:913–20.

64. Palomaki GE, Deciu C, Kloza EM, Lambert-Messerlian GM, Haddow JE, Neveux LM, et al. DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. Genet Med 2012;14:296–305.
65. Shah SP, Kobel M, Senz J, Morin RD, Clarke BA, Wiegand KC, et al. Mutation of FOXL2 in granulosa-cell tumors of the ovary. N Engl J Med 2009;360:2719–29.
66. Wiegand KC, Shah SP, Al-Agha OM, Zhao Y, Tse K, Zeng T, et al. ARID1A mutations in endometriosis-associated ovarian carcinomas. N Engl J Med 2010;363:1532–43.
67. Heravi-Moussavi A, Anglesio MS, Cheng SW, Senz J, Yang W, Prentice L, et al. Recurrent somatic DICER1 mutations in nonepithelial ovarian cancers. N Engl J Med 2012;366:234–42.
68. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A 2009;106:12353–8.
69. Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin PC, et al. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res 2011;21:56–67.
70. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol 2011;29:742–9.
71. Hajkova P, el-Maarri O, Engemann S, Oswald J, Olek A, Walter J. DNA-methylation analysis by the bisulfite-assisted genomic sequencing method. Methods Mol Biol 2002;200:143–54.
72. Tost J, Gut IG. DNA methylation analysis by pyrosequencing. Nat Protoc 2007;2:2265–75.
73. Zhang J, Benavente CA, McEvoy J, Flores-Otero J, Ding L, Chen X, et al. A novel retinoblastoma therapy from genomic and epigenetic analyses. Nature 2012;481:329–34.
74. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 2007;17:877–85.
75. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 2011;21:1757–67.
76. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc 2010;2010:pdb.prot5384.
77. Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 2009;10:161–72.
78. Zhang Z, Pugh BF. High-resolution genome-wide mapping of the primary structure of chromatin. Cell 2011;144:175–86.
79. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. Annu Rev Genet 2004;38:525–52.
80. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet 2005;6:805–14.
81. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature 2011;474:609–15.
82. Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, et al. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 2011;9:e1001046.
83. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 2010;28:1045–8.
84. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science 2009;323:133–8.
85. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. Bioinformatics challenges for personalized medicine. Bioinformatics 2011;27:1741–8.
86. This time it's personal. Nature 2008;453:697.
87. Hudson KL. Genomics, health care, and society. N Engl J Med 2011;365:1033–41.
88. Cook-Deegan R, Heaney C. Patents in genomics and human genetics. Annu Rev Genomics Hum Genet 2010;11:383–425.
89. Fujiwara Y. Genomics, health care, and society. N Engl J Med 2011;365:2339.
90. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. Nat Biotechnol 2008;26:1146–53.
91. Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell 2010;40:939–53.
92. Rizzo JM, Mieczkowski PA, Buck MJ. Tup1 stabilizes promoter nucleosome positioning and occupancy at transcriptionally plastic genes. Nucleic Acids Res 2011;39:8803–19.
93. Dostie J, Dekker J. Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc 2007;2:988–1002.
94. van Steensel B, Dekker J. Genomics tools for unraveling chromosome architecture. Nat Biotechnol 2010;28:1089–95.
95. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods 2011;8:61–5.
96. Scheibye-Alsing K, Hoffmann S, Frankel A, Jensen P, Stadler PF, Mang Y, et al. Sequence assembly. Comput Biol Chem 2009;33:121–36.
97. Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. Genome Res 2008;18:324–30.
98. Simon MP, Pedeutour F, Sirvent N, Grosgeorge J, Minoletti F, Coindre JM, et al. Deregulation of the platelet-derived growth factor B-chain gene via fusion with collagen gene COL1A1 in dermatofibrosarcoma protuberans and giant-cell fibroblastoma. Nat Genet 1997;15:95–8.