Published in final edited form as:

Cancer Lett. 2013 November 01; 340(2): 284–295. doi:10.1016/j.canlet.2012.11.025.

Next-generation sequencing in the clinic: Promises and challenges

Jiekun Xuan[a,b], **Ying Yu**[a], **Tao Qing**[a], **Lei Guo**[b,*], and **Leming Shi**[a,b,*]

[a]School of Pharmacy, Fudan University, 826 Zhangheng Road, Shanghai 201203, China

[b]National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA

Abstract

The advent of next generation sequencing (NGS) technologies has revolutionized the field of genomics, enabling fast and cost-effective generation of genome-scale sequence data with exquisite resolution and accuracy. Over the past years, rapid technological advances led by academic institutions and companies have continued to broaden NGS applications from research to the clinic. A recent crop of discoveries have highlighted the medical impact of NGS technologies on Mendelian and complex diseases, particularly cancer. However, the ever-increasing pace of NGS adoption presents enormous challenges in terms of data processing, storage, management and interpretation as well as sequencing quality control, which hinder the translation from sequence data into clinical practice. In this review, we first summarize the technical characteristics and performance of current NGS platforms. We further highlight advances in the applications of NGS technologies towards the development of clinical diagnostics and therapeutics. Common issues in NGS workflows are also discussed to guide the selection of NGS platforms and pipelines for specific research purposes.

Keywords

Whole-genome sequencing; Exome sequencing; RNA-Seq; Bioinformatics; FFPE; Tumor heterogeneity; Clinical applications

1. Introduction

Increased awareness that decoding the human genome provides critical clues to the genetics of diseases as well as the development of more specific preventive, diagnostic and therapeutic strategies has driven extensive sequencing and mapping efforts in the past decades. After the completion of the first human genome sequence in 2004 [1], the growing need to sequence a large number of individual genomes in a fast, low-cost and accurate way has directed a shift from traditional Sanger sequencing methods towards new high-throughput genomic technologies. In 2005, the first massively parallel DNA sequencing platforms emerged, ushering in a new era of next-generation sequencing (NGS) [2,3]. To

*Corresponding authors. Address: School of Pharmacy, Fudan University, 826 Zhangheng Road, Shanghai 201203, China. lei.guo@fda.hhs.gov (L. Guo), leming.shi@gmail.com (L. Shi).

date, the development of NGS technologies has vastly accelerated the pace of data generation—on the order of hundreds of gigabases of nucleotide sequence per instrument run, while reducing sequencing cost by over five orders of magnitude. Owing to these advantages, NGS technologies have been widely used for many applications, such as rare variant discovery by whole genome resequencing or targeted sequencing, transcriptome profiling of cells, tissues and organisms, and identification of epigenetic markers for disease diagnosis.

Here, we first provide a brief overview of the characteristics, strengths and limitations of current NGS platforms (Table 1). We then discuss the major applications of NGS technologies, with a focus on cancer diagnosis and prognosis. Finally, we discuss the bioinformatics tools and challenges in NGS data analysis.

2. Overview of NGS technologies

The technical details of the three major NGS platforms have been well described elsewhere [4]. The following section primarily focuses on the performance of each sequencing system.

2.1. Roche/454

The 454 sequencing system is based on the combination of emulsion PCR and pyrosequencing technology [2]. In emulsion PCR, single-stranded template carrying beads are confined to individual emulsion droplets in which millions of copies of each template are produced by PCR amplification. Amplicon-bearing beads are subsequently enriched and deposited into individual wells of a picotiter plate where solid-phase pyrosequencing is carried out. In this sequencing-by-synthesis method, luminescence emission from the release of pyrophosphate upon template-directed nucleotide incorporation is monitored in real time. The strength of the 454 system lies in its ability to sequence long reads. The latest 454 GS FLX platform with Titanium chemistry can produce approximately one million reads with lengths of up to 1000 bp per instrument run. Owing to this advantage, despite the higher costs compared with other NGS platforms, the 454 platform is best suitable for several applications, including *de novo* assembly [5] and metagenomics [6]. However, the 454 technology has an inherent problem in the detection of homopolymers, stretches of the same nucleotide. Due to the lack of a terminating moiety, multiple incorporations of identical nucleotides can occur in homopolymeric regions during a single sequencing cycle. This can lead to nonlinearity between the signal intensity and the length of homopolymer stretches when more than three or four nucleotides are consecutively incorporated. Consequently, the 454 system has a relatively high error rate for calling insertions and deletions (indels) in homopolymers [7].

2.2. Illumina/Solexa

The Illumina sequencing system employs an array-based DNA sequencing-by-synthesis technology with reversible terminator chemistry [8]. In this approach, template DNA fragments are hybridized to a reaction chamber on an optically transparent solid surface (i.e., flow cell). Reversible terminators [9], a series of four modified nucleotides each labeled with a different removable fluorescent dye at the 3′-hydroxyl terminus, are used for step-by-step

DNA synthesis. Millions of clonal clusters can be generated in each lane of the flow cell, which contains eight independent lanes for multiple libraries to be sequenced in parallel. The Genome Analyzer (GA), the first Illumina sequencing platform, originally produced 35-bp reads and generate more than 1 gigabase (Gb) of high-quality sequence per run in 2–3 days. The upgraded platforms, such as GA IIx and HiSeq 2000, yield much higher sequence output with increased read lengths. Despite its ultra-high-throughput and cost-effective advantages, the utility of the Illumina systems is limited to short-read sequencing. The limitation in read lengths is primarily due to dephasing effects [4]. Decreased or increased efficacy of nucleotide incorporation and failures in removing or adding terminating moieties in any given cycle can cause incomplete extension or overextension of the growing strand along the template, resulting lagging-strand or leading-strand dephasing. Moreover, signal dephasing can be caused by decay in fluorescent signal, incorporation of nucleotides without a fluorescent label (dark nucleotides) or incomplete removal of fluorescent labels, leading to base-calling errors. Consequently, base substitution error rates increase with read length. In addition, uneven read coverage has been revealed across AT-rich and GC-rich regions, with a bias towards the latter.

2.3. Life Technologies/SOLiD

The SOLiD system uses a ligation-based sequencing technology originated from previous work [3]. The sequencing library is prepared by emulsion PCR as in the 454 protocol. Sequencing is performed through successive cycles of ligation, in which each sequencing primer is ligated to a specific fluorescence-labeled octamer probe according to the complementarity between the di-bases of the probe and the template. Since each four di-bases (e.g., AG, GA, TC, CT) are tagged with one of four fluorescent dyes, the di-nucleotides at the same positions of each template are associated with a unique fluorescent color. Across ligation cycles, di-nucleotides are read at intervals of five bases, that is, di-nucleotides at position 4– 5, 9–10, 14–15, 19–20 and so forth. After five ligation rounds, each nucleotide in the template is read twice by two fluorescent signals, greatly improving base-calling accuracy. Among the current NGS platforms, the SOLiD system presents the lowest error rate. Its most common error type is substitution. In addition, an underrepresentation of AT-rich regions has also been shown in the SOLiD data [10].

2.4. Emerging technologies

The emergence of single-molecule sequencing has provided a technological leap forward in the evolution of next generation sequencing. The promise of this technology lies in its ability to directly sequence single DNA or RNA molecules in biological samples without amplification. The single-molecule sequencing strategy promises significant advantages over current NGS technologies in that it minimizes sample handling, reduces sample input requirements, avoids amplification-induced bias and errors, increases read length flexibility and enables accurate quantitation of nucleic acid molecules. The simplicity, sensitivity and quantitative capabilities of single-molecule sequencing make it highly promising for molecular diagnostics [11].

The Helicos Genetic Analysis System is the first commercially available single-molecule sequencing platform [12]. In this system, poly(A)-tailed single-stranded DNA templates are

captured by poly(T) oligonucleotide primers tethered to the surface of a flow cell. Sequencing is performed through iterative cycles of DNA polymerase- mediated single-base primer extension using a series of four fluorescent Virtual Terminator nucleotides, each of which represents a 3′-unblocked reversible terminator with a fluorophore-labeled inhibitory moiety [13]. In a standard run, the sequencer with two 25-channel flow cells is capable of capturing billions of single DNA molecules and generating over 21–35 Gb of sequence data with an average read length of 35 bp. Although the sequencing process is asynchronous in the Helicos system, dephasing effects that commonly exist in amplification-based sequencing platforms are not present. Moreover, there is no GC-content bias in read coverage. However, the current error rate in Helicos reads is relatively high (~3–5%), and the dominant error type is deletion, which presumably results from incorporation of unlabeled nucleotides and/or detection errors. The use of Virtual Terminator chemistry solves the homopolymer sequencing problem, and the base-by-base incorporation manner results in very low substitution error rates (typically 0.2%). When a two-pass strategy is applied, in which individual template molecules are sequenced twice, the error rates can be further reduced.

Other single-molecule sequencing technologies with longer read lengths, higher sequencing speed or lower overall cost are also emerging. One example is the PacBio *RS*, a single-molecule real-time sequencing system developed by Pacific Biosciences [14]. In this system, a single template-bound DNA polymerase molecule is immobilized to the bottom of a zero-mode waveguide, which functions as a nanophotonic visualization chamber for monitoring the polymerization reaction in a detection volume on the order of zeptoliters (10^{-21} l). During sequencing, template-directed incorporation of four fluorescent phospholinked nucleotides into the growing complementary strand is optically recorded in real time. The fluorescent dye attached to the terminal phosphate moiety of each phospholinked nucleotide is naturally removed by enzymatic cleavage upon incorporation. This allows rapid and processive DNA synthesis by the polymerase, yielding sequence reads of thousands of bases. Nonetheless, the PacBio system presently offers a throughput of approximately 50–100 Mb per run, which is much lower than current NGS platforms. Moreover, the single-read error rate is typically 15%, exceeding the error tolerance of many applications. A second example is nanopore sequencing technologies, in which single-stranded nucleic acid molecules are electrophoretically driven through a nanometer-sized pore and detected by their effect on an ionic current or optical signal [15]. Nanopore sequencing potentially offers long read lengths of up to tens of kilobases, minimal requirements of reagent and sample preparation, and high sequencing pace at low cost. However, several problems remain to be solved before the application of nanopore sequencing. The high speed of DNA translocation through nanopores makes it challenging to distinguish base signals from background noises by an electronic sensor. The random motion of molecules during translocation also adds to the difficulty in reaching single-base resolution. As a solution, IBM is developing a DNA transistor technology that incorporates alternating layers of metal and dielectric material within a nanopore to control the rate of DNA translocation [16]. Oxford Nanopore Technologies is also commercializing electronic nanopore sensing systems based on exonuclease and strand sequencing techniques. Exonuclease sequencing [17] employs a modified α-hemolysin nanopore with a bound exonuclease that cleaves off single

nucleotides, allowing successive passage and detection of each nucleotide in a polymer, while strand sequencing uses a polymerase to pass single-stranded DNA polymers through the nanopore.

Other new sequencing technologies are also under development, such as fluorescence resonance energy transfer (FRET)-based single-molecule sequencing technology from VisiGen Biotechnologies, Ion semiconductor sequencing technology from Ion Torrent, now part of Life Technologies, and DNA nanoball sequencing technology from Complete Genomics. Despite remarkable advantages of these new technologies, there remains much room for improvement before introducing them into clinical practice.

The LifeTech Ion Torrent and Proton platforms are strikingly different from other NGS platforms, because they measure pH changes rather than e.g. fluorescence during sequencing. While the data output is still relatively low per chip, the fast turnaround time per chip makes the Ion Torrent and Proton very suitable to smaller, focused sequencing projects, 16S sequencing projects, SNP detection and validation, as well as sequencing of small genomes. Hence, they may gain more relevance for the clinics with increase in read length in the coming years and decrease in costs. Similarly, smaller and more affordable sequencing instruments, like the Illumina MiSeq, may become an instrument of choice for diagnostic and prognostic clinical centers and smaller laboratories.

Cost per platform, especially running cost, is important to the end user but difficult to estimate. Based on our own recent experience in a large sequencing project, the Sequencing Quality Control (SEQC) project, the cost per Gb data for the three major platforms is $46.8 for HiSeq 2000, $77.2 for SOLiD4, and $12,210 for Roche 454. That is, the ratio of per Gb cost is 1:1.65:261 for Illumina:LifeTech:454.

Although the Illumina HiSeq 2000 or 2500 platform is the most cost-effective tool for whole-genome sequencing, but it still takes days to have one genome sequenced. The expected rate of dropping in sequencing cost and increase in sequencing throughput has not been maintained over the past one and half years because of the lack of competition – Illumina has been dominating the sequencing market. Newer sequencing platforms, like nanopore-based technologies, have yet to demonstrate their viability in terms sequencing throughput and accuracy. In the next two to three years, Illumina will likely continue to dominate the sequencing market and the cost per genome at 30× coverage is almost impossible to drop to below $1000. Note that the cost for storing, analysis, and interpreting whole-genome sequencing data is even higher than the cost of generating the data. Therefore, sequencing cost will remain a concern for most people including patients and insurers.

3. Applications of NGS

3.1. Genome-wide discovery of causal variants

Most common diseases and quantitative traits in human populations have a complex genetic basis. The identification of genetic variants that underlie susceptibility to common diseases and traits has been challenging. Genome-wide association studies, which have thus far

focused on common variants (minor allele frequency, MAF > 5%), have achieved only modest success in explaining the heritability of most complex traits. Although over 7000 strong SNP-trait associations ($p < 1.0 \times 10^{-5}$) have been identified to date, as listed in the National Human Genome Research Institute (NHGRI)'s Catalog of Published Genome-Wide Association Studies, the majority of these SNPs only have small effect sizes and very few are causal variants. The 'common disease – common variant' hypothesis [18], which posits that common traits are most likely attributed to genetic variants with high frequencies has been refuted. The missing heritability observed in GWA studies has been attributed to several causes, including rare variants of large effects undetected by available genotyping arrays, structural variants poorly captured by current technologies, a large number of small-effect common variants uncovered by existing arrays, insufficient power to identify gene–gene interactions, and underestimate of environmental and epigenetic effects. An important role of rare variants in conferring susceptibility to common diseases and traits has been recognized.

With the advent of NGS technologies, it has become feasible to sequence whole genomes in relatively large cohorts of individuals to identify rare variants. The 1000 Genomes Project, as an example, is developing a more detailed catalog of genetic variants with frequencies down to 1% in multiple human populations [19,20]. The UK10K project is a more ambitious sequencing effort underway, which will sequence 10,000 human genomes from two well-phenotyped UK population-based cohorts to identify rare variants in several types of diseases. Large-scale genome sequencing of other species has also been launched [21]. These advances will largely facilitate the discovery of disease-causing variants. Of note, recent whole genome sequencing studies have discovered a number of novel abnormalities in cancer genomes [22–28] (Table 2). Furthermore, the characterization of unique mutational patterns of individual cancer genomes has shown great promise in personalized cancer therapy [29,30].

3.2. Targeted sequencing

Although whole-genome sequencing (WGS) is the most straightforward and comprehensive strategy for genome analysis, large-scale WGS studies are still unaffordable for many research laboratories and clinical settings. In comparison, targeted sequencing can yield much higher coverage of genomic regions of interest while reducing the sequencing cost and time. The most commonly used target-enrichment techniques include PCR and array-based or solution-based hybridization [31].

PCR-based enrichment methods have the advantage of even coverage and high specificity. Generally, primer sets specifically targeting genomic sequences of interest are used to generate multiple overlapping amplicons, which can be pooled and converted into a library for sequencing. The utilization of PCR-based enrichment methods for massively parallel targeted sequencing faces several challenges. First, PCR specificity largely depends on primer design and reaction optimization, which can require extensive computational analysis to avoid non-specific priming, primer cross-reactivity and dimer formation. PCR primers are usually derived from reference gene sequences. The existence of variants in primer annealing sites may decrease priming efficiency and cause allelic bias or dropout [32].

Moreover, large rearrangements (e.g., insertions, deletions, translocations) in cancer genomes may be undetectable unless primer pairs in flanking regions are defined. Second, multiplex amplification of target sequences using multiple primer sets in a single tube can result in an excessive amount of non-specific products. Several technologies have been developed to get around this problem, including microfluidic PCR [33,34], emulsion PCR [35,36], and microdroplet PCR [37]. These strategies enable many singleplex PCRs to be performed independently and simultaneously in a single reaction, where primer-pair cross-interactions and product-product hybridization are prevented. Third, the size of PCR amplicons is limited (<10 kb) due to the potential loss of fidelity and efficiency in longer PCRs. Thus, PCR amplification of large regions can involve thousands of parallel reactions that need tedious optimization, resulting in dramatic increases in cost, labor intensity and input DNA amount. Consequently, the scale of PCR-based target enrichment is limited to several megabases. Despite these drawbacks, PCR-based strategies may find wide applications in clinical practice, considering their high specificity that can ensure clinical accuracy. Two major commercial products for PCR-based target enrichment are RainDance's RainStorm (microdroplet-based) and Fluidigm's Access Array (microfluidic-based). Their applications in the context of NGS have successfully identified disease-causing mutations for diagnostic testing [38– 40].

Hybridization-based enrichment methods can directly capture targets of interest from a NGS library using complementary oligonucleotides either in solution or on array. The major advantage of the methods is their capability to cover large genomic regions, which have been scaled to the entire human exome (~30 Mb) [41–43]. In array-based hybridization methods, target-specific capture probes are synthesized on high-density DNA microarrays. A number of studies have shown the flexibility of on-array hybridization in capturing both short-non-continuous regions and long-continuous regions [41,42,44–47]. Nevertheless, these methods are difficult to scale to large numbers of samples due to the high cost of microarrays. Moreover, a vast excess of input DNA is required to ensure sufficient hybridization of each probe with its target. Solution- based hybridization methods have been developed to overcome these disadvantages. In the methods, an excess of biotinylated single-stranded RNA [48,49] or DNA oligos [43] is presented as capture probes in solution. A relatively small amount of input DNA-fragment library is required for hybridization reactions in aqueous phase. Consequently, both sample DNA quantity and cost per assay for target enrichment are substantially reduced. Currently, three major commercial products including Agilent's Sure- Select (array-based and solution-based), Nimblegen's SeqCap (array-based and solution-based) and Illumina's TruSeq (solution-based) in conjunction with NGS platforms have been proven highly effective in exome sequencing [50,51]. Thus far, a number of studies have reported the identification of causal mutations by exome sequencing for many genetic diseases [52,53] and cancers [54–63] (Table 2).

3.3. RNA-Seq

RNA-Seq applies NGS technologies to qualitatively and quantitatively profile the full set of transcripts (i.e., transcriptome), including mRNAs, small RNAs and other non-coding RNAs. Transcriptome profiling provides a snapshot of gene expression patterns and regulatory elements in a cell, tissue or organism under different physiological states, which

is important to the understanding of biological processes in development and disease. Though a transcriptome only represents a small fraction of the human genome (<5%) [64], it is very complex in that transcripts derived from alternative splicing [65], gene fusion [66], antisense transcription [67] and RNA editing [68] largely increase the diversity of transcriptome. Initial high-throughput analysis of transcriptomes mainly relied on microarray technologies, but their abilities are limited due to the dependence on prior knowledge about the genome, the limited dynamic range of detection and cross-hybridization issues [69]. As an alternative approach, RNASeq offers a direct sequencing strategy that is able to define at single base resolution the complete repertoire of RNA transcripts across a broad range of expression levels. Recently, RNA-Seq has been increasingly applied to study complex diseases, particularly cancer, taking advantage of its superior sensitivity and efficiency in detecting allele-specific expression, fusion transcripts and non-coding RNAs.

Gene fusions represent a common feature of cancer [70]. They have mostly been found in hematological malignancies and bone and soft tissue sarcomas but less frequently in epithelial carcinomas [71]. The identification of gene fusions in common solid tumors has been largely hindered by the limitations of cytogenetic techniques (FISH, metaphase karyotyping and array-CGH) and clonal heterogeneity. RNA-Seq has greatly facilitated the discovery of novel gene fusions in various cancer types, including prostate cancer [72–77], breast cancer [78,79], lymphoma [80,81], sarcoma [82] and melanoma [83]. Its superiority has been shown for unveiling not only recurrent gene fusions arising from chromosomal rearrangements that are not detectable at the genomic level, but also recurrent chimeric read-through transcripts (e.g., *SLC45A3-ELK4, CDK2-RAB5B*) in the absence of DNA aberrations. The occurrence of some fusion events has been linked to the mechanisms of carcinogenesis in specific tissues or organs, which could be used to develop diagnostic markers. For example, the MHC class II transactivator CIITA has been found to present as a partner of various gene fusions in B-cell lymphomas, suggesting that CIITA rearrangements may represent a novel oncogenic mechanism in lymphoid cancers [80]. Additionally, a new subtype of bone sarcoma caused by a *BCOR-CCNB3* gene fusion mechanism has been defined [82]. On the contrary, some gene fusions are present across different cancer types. For example, recurrent fusions involving RAF pathway genes have been identified in prostate cancer, gastric cancer and melanoma, providing therapeutic targets for all three cancers [74].

Non-coding RNAs (ncRNAs) are emerging as functional elements in a wide range of biological processes, including proliferation, differentiation, development and apoptosis. Aberrant functions of ncRNAs have been implicated in the pathogenesis of many human diseases, particularly cancer [84]. Long non-coding RNAs (lncRNAs) are a group of ncRNAs more than 200 nucleotides in length that take part in a broad spectrum of cellular functions, including epigenetic modifications, transcriptional regulation and post-transcriptional processing of mRNAs [85]. The involvement of lncRNAs in cancer biology has been proven by the characterization of dozens of lncRNAs (e.g., *DD3/PCA3* [86,87], *MALAT-1* [88,89], *HULC* [90,91], *ANRIL* [92,93], *HOTAIR* [94–97]) that are differentially expressed in cancers [98]. However, though it has been recognized that the human genome produces a vast repertoire of lncRNAs, including intergenic, antisense, and intronic

transcripts, their functional catalog is still far from fully defined [99,100]. A recent study reported the discovery of a novel lncRNA *PCAT-1* as an important actor in prostate cancer progression through transcriptome analysis of 102 prostate cancer tissues and cell lines by RNA-Seq, highlighting the potential of NGS technologies to comprehensively identify unannotated ncRNAs that can be clinically valuable markers of disease states [101].

MicroRNAs (miRNAs) are small ncRNAs of ~22 nucleotides that function as key regulators of gene expression by binding to complementary sequences in target sites and inducing mRNA degradation and/or translational repression [102]. Dysregulated expression of miRNAs is a hallmark of many cancers, in which miRNAs can act as oncogenes or tumor suppressors. Previous studies have shown distinct miRNA expression patterns in paired normal and tumor tissues and also in different tumors, suggesting that miRNA profiles can be accurate indicators of tumor type and malignant progression [103–106]. miRNA profiling by next-generation sequencing has been proven a powerful tool in the identification of novel cancer signatures for diagnosis and prognosis in recent studies [107– 114]. Notably, the discovery of tumor-specific circulating miRNAs in body fluids opens up promising avenues for developing non-invasive diagnostic and prognostic markers in cancer [115]. Other ncRNAs, such as PIWI-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs), transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), have also been linked to cancer development [84,116]. A recent NGS analysis of small ncRNA species in organ-confined and metastatic prostate cancer successfully revealed differentially expressed snoRNAs and tRNAs as well as miRNA signatures that are associated with disease promotion [114].

The reliable quantification of gene expression abundance with RNA-Seq depends on the depth of sequence data collected on each sample and the inherent expression level of the gene of interest. For genes expressed at lower abundance in a sample, many more sequence reads are required to achieve accurate quantification, whereas for highly expressed genes, much fewer reads are required for their accurate quantification. Roughly, 50–100 million reads per sample are required for an RNA-Seq profile to reach similar performance of an Affymetrix gene expression microarray. One emerging approach for large-scale clinical genomic studies is to first use deep RNA-Seq as a discovery tool to identify transcriptomic features relevant to the disease process or treatment outcome and then use custom-designed arrays to reliably screen a large number of patient samples as a validation phase or in routine clinical applications [117].

3.4. Epigentic profiling

Epigenetic modification of DNA is involved in many aspects of cancer progression. Heyn and Esteller [118] have presented a general overview of the importance, applications, and challenges of DNA methylation monitoring in the clinic, and it is expected that DNA methylation status will be valuable for future diagnosis, prognosis and prediction of response to therapies. Carvalho et al. [119] conducted genome-wide DNA methylation profiling on non-small cell lung carcinomas and found that while hypomethylated differentially methylated regions (DMRs) did not correlate to any particular functional category of genes, the hypermethylated DMRs were strongly associated with genes encoding transcriptional regulators.

4. Bioinformatics challenges and solutions

The dramatic reduction in sequencing cost and time allows current NGS platforms to generate an unprecedented volume of sequence data, as many as millions or billions of short reads (~50– 150 bp) per run, posing great bioinformatics challenges in terms of NGS data storage, quality control, alignment, assembly and annotation. Thus far, a number of computational tools have been developed for analyzing NGS data in the following scopes: (i) base calling; (ii) alignment of sequence reads to a reference; (iii) *de novo* assembly; and (iv) variant detection and genome annotation (Table 3).

Base calling is a process of identifying nucleotide sequences of DNA templates from fluorescence intensity signals produced by sequencers. This procedure is critical to the interpretation of NGS data, since any introduced sequence errors would have an influence on downstream analyses, including alignment, SNP and genotype calling. The technological differences among NGS platforms combined with the use of different base-calling algorithms lead to platform-dependent error characteristics. The 454 platform tends to have insertion and deletion (indel) errors caused by broaden signal distributions in homopolymers [167]. In the Illumina platform, the average raw error rate is typically 1% [168]. Error rates increase towards the end of reads due to the accumulation of asynchrony in the synthesis process. Substitution errors are more frequent than indel errors, and certain substitution (A > C transversion) errors are more prevalent [169]. In the SOLiD platform, the raw error rate increases from ~2% to ~8% towards the 3′-end [170]. The bias in the distributions of fluorescence intensities appears in later sequencing cycles, which can be alleviated by an intensity normalization [171]. A number of improved base callers have been developed to reduce the error rate for each platform, including Rsolid [171] for the SOLiD platform, Pyrobayes [172] for the 454 platform, and BayesCall [173], Ibis [174], Seraphim [175], and AYB [176] for the Illumina platform. Base-calling algorithms use quality scores to estimate error probabilities for each base call, most of which can be transformed to Phred quality score (Q) [177]. An estimated base-calling error rate is $10^{-Q/10}$. That is, if a Phred quality score of 20 is assigned, the probability to erroneously call a base is 1% (10^{-2}). Improvement of base-calling performance remains a computational challenge. Theoretically, faster and more accurate base callers should be developed for various NGS platforms. Unfortunately, in sequencing practice, the fluorescent intensity signal data files are usually no longer provided to the end user because of their huge storage requirements. Thus, the end user has to live with whatever the platform provider has implemented in their base calling algorithms.

Incorrect mapping of reads can readily lead to erroneous identification of sequence variants, highlighting the importance of alignment accuracy. The most common alignment problem arises from reads that map to multiple locations on the reference sequence, so-called multi-reads. Thus far, it remains challenging to correctly assign multi-reads to their original sites. Three strategies have been used to cope with multi-reads. The first strategy is to discard all multi-reads and only utilize reads that map uniquely to the reference genome. However, this can cause the omission of up to 30% of mappable reads, rendering those variants in repetitive elements and gene families that may be of functional significance undetectable. In contrast, the other two strategies will not limit alignments to repetitive sequences. One is the best-match approach, in which each read is assigned to the location(s) with the fewest

mismatches. When more than one best-match location is found, the alignment program will report all locations or one of them by random selection. The third strategy is to report all alignments until the predefined maximum number is reached. However, these two strategies can introduce incorrect alignments, especially in regions of high diversity between the sequenced genome and the reference genome. In this case, use of longer reads and paired-end reads can efficiently reduce the ambiguity and improve alignment accuracy.

Calling SNPs from DNA sequence data remains a major challenge. For a given sample, the number and identities (locations) of SNPs called by different software packages can be quite divergent. Consequently, SNPs identified to be associated with the phenotype (disease status or treatment response) can be largely different, making validation of the findings very difficult. Without a “gold standard” SNP calling algorithm, one may focus on those SNPs that are called by two or more SNP calling algorithms to ensure a better chance of validation.

Reconstruction of a genome in the absence of a reference sequence requires *de novo* assembly of reads into longer contiguous sequences (i.e., contigs) followed by correctly ordering contigs into scaffolds. The short length of NGS reads greatly adds to the difficulty of genome assembly. A compromise solution is to increase the depth of coverage, but it’s still unable to counteract the problems with assembly of repetitive regions. It is a substantial challenge for an assembler to distinguish genomic regions that share repeats, especially when the repeats are longer than the reads used for assembly. Such repeats not only create gaps, their flanking regions can also be incorrectly connected, thus generating misassembled chimeras by linking distant regions together. Two approaches have been developed to solve the problem: overlap graph and de Bruijn graph [178]. The latter is superior in short-read assembly but requires information from paired-end reads to resolve repeats. Hence, efficient generation and algorithmic analysis of read pairs would be the key to assembling large genomes with short reads. Although there is rarely a need for *de novo* assembly of the human genome in clinical applications, the relevance of metagenomes [179,180] to human health makes *de novo* assembly a necessity.

Transcriptome analysis with RNA-Seq data brings additional computational challenges. One challenge is to map reads that span splice or fusion junctions. Conventional mapping programs such as ELAND, Bowtie [121] and MAQ [123] that need to allocate reads to contiguous sequences are inappropriate for spliced alignment. New algorithms have been developed to map splice-crossing reads, some of which utilize previously known splice events (e.g., ERANGE [129]), while others (e.g., GSNAP [130], MapSplice [131], RUM [132], SpliceMap [133], TopHat [134]) do not rely upon prior knowledge. In particular, some algorithms have been specifically designed for the identification of gene fusions, including deFuse [157], FusionSeq [158], ShortFuse [159] and TopHat-Fusion [160]. Despite high sensitivity of these aligners in detecting junctions, misalignment of multi-reads can readily occur and lead to a high false positive rate of identification. Most algorithms handle this problem by discarding multi-reads. However, this may result in inaccurate estimation of the expression levels of genes located in repetitive regions.

Gene expression measurement is another challenge in RNA-Seq data analysis. The standard approach to estimate the expression level of a gene is to calculate the count of reads mapped to that gene. However, the read count is a function of the length and molar concentration of the transcript. One common solution is to normalize the read count by the transcript length and the number of million mappable reads or fragments (read pairs) to obtain the measurement, which is expressed as reads per kilobase per million (RPKM) or fragments per kilobase per million (FPKM). Moreover, multireads that originate from multiple isoforms of the same gene and homologs of gene families can lead to incorrect estimation of gene expression. Given the disadvantages of discarding multi-reads, an alternative strategy has been developed to rescue multi-reads by allocating them in proportion to the number of reads uniquely mapped at the same loci [181]. Several methods have also been reported to estimate expression levels of isoforms and homologous genes in the presence of multi-reads with a probabilistic generative model optimized by an Expectation–Maximization (EM) algorithm [182,183].

NGS technology is coupled with immense data storage requirements, which needs to be considered prior to the decision of employing such platforms in clinical practices, at least for now. This is because the huge amount of raw sequence data for each patient, usually hundreds of Gbs, needs to be stored for potential future analysis and interpretation when analytical algorithms improve. The transfer of NGS data from the sequencing facility to the data analysis center presents another challenge. In our experience, shipping the data on hard drives with 2 Tbs or 3 Tbs storage space is a routine mechanism. Making the Tbs of sequence data readily accessible to the computing power could be the bottleneck. However, with the improvement and standardization of sequencing platforms and data analysis algorithms, within a few years down the road it may become unnecessary to store the raw sequence data anymore. What need to be stored and transferred may be the variant calling results that are much smaller in size compared to the raw sequence data.

5. Cancer-specific concerns

5.1. Issues with tumor samples

Fresh or fresh-frozen tumor tissues can provide good-quality samples for massively parallel sequencing of cancer genomes, but they are limited in supply in most hospitals. Clinical tissue samples are routinely formalin-fixed and paraffin-embedded (FFPE) for histopathological examination and long-term storage. Undoubtedly, the use of FFPE material could provide a rich sample source for molecular studies of cancer. However, the preparation procedure of FFPE samples can lead to significant degradation and chemical modification of nucleic acids. It is known that formalin fixation adds hydroxymethyl groups to nucleic acid bases and induces cross-linking with proteins. As a consequence, artificial sequence alterations can occur in DNA extracted from FFPE samples during PCR. It has been assumed that such artifacts are caused by formalin cross-linking of cytosine nucleotides, which misleads DNA polymerase to incorporate an adenine instead of a guanosine, resulting in a C > T or G > A transition [184]. On the other hand, the poly(A) tail of mRNA isolated from FFPE samples is likely to be heavily modified, blocking the annealing of oligo (dT) primers to the poly(A) tail in the reverse transcription reaction [185].

Given these concerns, it is necessary to assess fixation-induced nucleic acid damage and minimize error rates when performing massively parallel sequencing of FFPE samples. One strategy for removing damage-derived artifacts is to use more stringent alignment criteria. However, this can lead to a reduction in coverage depth and may also remove genuine mutations. As a solution, a recent study reported a novel post-alignment filtering method that integrates global nucleotide mismatch rates and local mismatch rates to remove false positive calls caused by formalin fixation [186]. Another strategy is to increase sequencing depth. In a whole-exome sequencing study, a high rate of discordant loci as false positives in FFPE tissues compared to paired snap-frozen tissues was detected at 20× coverage [187]. While false positives were reduced but still present at 40× coverage, no discordance was observed at 80× coverage. Hence, it is seen that accurate detection of somatic mutations in FFPE tumor samples can be achieved at high coverages, especially using targeted sequencing approaches [187,188].

5.2. Tumor heterogeneity

Phenotypic and functional heterogeneity across different tumors in the same individual (inter-tumor heterogeneity) as well as cell subpopulations within a single tumor (intra-tumor heterogeneity) has been recognized as a hallmark of cancer. The heterogeneous nature of tumors can lead to inaccurate diagnosis and therapeutic resistance. Several mechanisms have been postulated to account for tumor cell heterogeneity but their relative contribution remains controversial [189]. The most widely accepted mechanism involves clonal evolution of tumor cell populations driven by genomic instability, which has been supported by morphological and cytogenetic findings. However, thus far, tumor heterogeneity has not been well defined at the molecular level due to limitations in experimental and analytical tools. Recently, taking advantage of NGS technologies, high-resolution genome-wide studies of genetic diversity among tumor cells are emerging, leading to a deeper understanding of tumor clonal architecture and evolution. In a pilot study, Gerlinger et al. [190] revealed branched mutation profiles in multiple spatially separated specimens taken from a single tumor by whole-exome sequencing. A more sophisticated view of intra-tumor heterogeneity could be obtained with the development of single-cell sequencing technology. Several studies have made breakthroughs in reconstructing mutational pathways and tumor evolution history based on single-cell genetic architecture [191]. In addition to genomic heterogeneity, transcriptome diversity in individual tumors has also been revealed at the single-cell level [192]. Single-cell RNA-Seq technology has been established for identifying gene expression signatures as well as tumor-associated mutations with small amounts of sample RNA [193]. Despite these advances, single-cell sequencing is still not applicable in clinical settings until sequencing cost and time are reduced to a reasonable level to make the analysis of hundreds of single cells affordable. However, it holds great promise to translate the knowledge of tumor heterogeneity into clinical practice, enabling precise diagnosis, targeted therapy and personalized medicine to improve the clinical management of cancer patients.

6. Concluding remarks

In the past few years, NGS technologies have made great strides in both basic and clinical research, providing deeper insights into the complex genomic landscapes of many diseases.

However, implementing NGS into clinical settings still faces some hurdles. First, the rapid generation of enormous amounts of sequence data presents a huge challenge for data integration. For instance, whole genome sequencing can easily discover numerous genetic variations between patients and healthy volunteers, but it will be difficult to extract clinically useful and actionable information and validate significant genotype-phenotype associations. Notably, a person's genetic make-up is not the only determinant of disease risk and drug response. A variety of demographic and clinical factors (such as age, gender, ethnicity, pathological stage, and medical history) could potentially complicate clinical decision-making. Therefore, it is important to control for confounding factors in the development of reliable and reproducible molecular markers. Second, there lacks a standard for quality control of sequence data. The problem of sequencing errors remains significant, since such errors are not distinguishable from genetic variants and could be misidentified as phenotype-associated mutations. It is known that all commercially available NGS platforms have different error types and error rates. Each error type needs to be carefully assessed and corrected in order to minimize the potential impact on downstream data analysis. Additionally, applying different bioinformatics strategies could significantly affect the output of NGS data analysis. Thus, it is necessary to understand the principles, advantages and limitations of bioinformatics tools so as to establish appropriate pipelines capable of generating reliable analytical results. These issues are being addressed by the Sequencing Quality Control (SEQC) project, a community-wide collaborative effort led by the US Food and Drug Administration (www.fda.gov/Micro-ArrayQC/) as a follow-up of its MicroArray Quality Control (MAQC) project [194,195]. How to implement regulatory quality control standards in translational research and clinical testing requires more attention in the future. There is a need to establish reference DNA/RNA samples and reference data sets. Third, the massive accumulation of genomic data also raises ethical issues. Whether and how to return sequence results to patients remains questionable. The reality with NGS data is that clinically and biologically important information is frequently buried in huge amounts of noise or false positives and the cost of false-positive diagnosis can be tremendous. Protection of the privacy and confidentiality of individual genomic information is also of concern. Only when patients are willing to share their medical information and their sequence data can new biomedically important findings be discovered and utilized for the benefits of large populations. Finally, it is important to note that demonstrated technical performance of NGS or any other genomics technologies does not automatically translate to diagnostic assays of clinical utilities, because markers reported by most studies are not sufficiently predictive for taking clinical actions. Despite these challenges, NGS has provided unprecedented opportunities for clinical diagnostics and personalized medicine.

References

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004; 431:931–945. [PubMed: 15496913]
2. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–380. [PubMed: 16056220]
3. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005; 309:1728–1732. [PubMed: 16081699]

4. Metzker ML. Sequencing technologies – the next generation. Nat Rev Genet. 2010; 11:31–46. [PubMed: 19997069]
5. Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin JF. Accuracy and quality assessment of 454 GS-FLX titanium pyrosequencing. BMC Genomics. 2011; 12:245. [PubMed: 21592414]
6. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012; 30:434–439. [PubMed: 22522955]
7. Wommack KE, Bhavsar J, Ravel J. Metagenomics: read length matters. Appl Environ Microbiol. 2008; 74:1453–1463. [PubMed: 18192407]
8. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456:53–59. [PubMed: 18987734]
9. Turcatti G, Romieu A, Fedurco M, Tairi AP. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Res. 2008; 36:e25. [PubMed: 18263613]
10. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009; 10:R32. [PubMed: 19327155]
11. Milos PM. Emergence of single-molecule sequencing and potential for molecular diagnostic applications. Expert Rev Mol Diagn. 2009; 9:659–666. [PubMed: 19817551]
12. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-molecule DNA sequencing of a viral genome. Science. 2008; 320:106–109. [PubMed: 18388294]
13. Bowers J, Mitchell J, Beer E, Buzby PR, Causey M, Efcavitch JW, et al. Virtual terminator nucleotides for next-generation DNA sequencing. Nat Methods. 2009; 6:593–595. [PubMed: 19620973]
14. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009; 323:133–138. [PubMed: 19023044]
15. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. Nat Biotechnol. 2008; 26:1146–1153. [PubMed: 18846088]
16. Polonsky S, Rossnagel S, Stolovitzky G. Nanopore in metal-dielectric sandwich for DNA position control. Appl Phys Lett. 2007; 91:153103.
17. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009; 4:265–270. [PubMed: 19350039]
18. Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet. 2001; 17:502–510. [PubMed: 11525833]
19. Kaiser J. DNA sequencing. A plan to capture human diversity in 1000 genomes. Science. 2008; 319:395. [PubMed: 18218868]
20. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
21. Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered. 2009; 100:659–674. [PubMed: 19892720]
22. Roberts KG, Morin RD, Zhang J, Hirst M, Zhao Y, Su X, et al. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. Cancer Cell. 2012; 22:153–166. [PubMed: 22897847]
23. Stacey SN, Sulem P, Jonasdottir A, Masson G, Gudmundsson J, Gudbjartsson DF, et al. A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. Nat Genet. 2011; 43:1098–1103. [PubMed: 21946351]
24. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat Genet. 2011; 43:964–968. [PubMed: 21892161]
25. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat Genet. 2012; 44:760–764. [PubMed: 22634756]

26. Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. Nature. 2012; 485:502–506. [PubMed: 22622578]
27. Cheung NK, Zhang J, Lu C, Parker M, Bahrami A, Tickoo SK, et al. St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project. Association of age at diagnosis and genetic mutations in patients with neuroblastoma. JAMA. 2012; 307:1062–1071. [PubMed: 22416102]
28. Rafnar T, Gudbjartsson DF, Sulem P, Jonasdottir A, Sigurdsson A, Jonasdottir A, et al. Mutations in BRIP1 confer high risk of ovarian cancer. Nat Genet. 2011; 43:1104–1107. [PubMed: 21964575]
29. 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; 3:111ra121.
30. Guan YF, Li GR, Wang RJ, Yi YT, Yang L, Jiang D, et al. Application of next-generation sequencing in clinical oncology to advance personalized treatment of cancer. Chin J Cancer. 2012; 31:463–470. [PubMed: 22980418]
31. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010; 7:111–1118. [PubMed: 20111037]
32. Ikegawa S, Mabuchi A, Ogawa M, Ikeda T. Allele-specific PCR amplification due to sequence identity between a PCR primer and an amplicon: is direct sequencing so reliable? Hum Genet. 2002; 110:606–608. [PubMed: 12107448]
33. Kirkness EF. Targeted sequencing with microfluidics. Nat Biotechnol. 2009; 27:998–999. [PubMed: 19898452]
34. Zhang Y, Ozdemir P. Microfluidic DNA amplification – a review. Anal Chim Acta. 2009; 638:115–125. [PubMed: 19327449]
35. Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, Griffiths AD. Amplification of complex gene libraries by emulsion PCR. Nat Methods. 2006; 3:45–50.
36. Xu MY, Aragon AD, Mascarenas MR, Torrez-Martinez N, Edwards JS. Dual primer emulsion PCR for next-generation DNA sequencing. Biotechniques. 2010; 48:409–412. [PubMed: 20569215]
37. 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; 27:1025–1031. [PubMed: 19881494]
38. Hopp K, Heyer CM, Hommerding CJ, Henke SA, Sundsbak JL, Patel S, et al. B9D1 is revealed as a novel Meckel syndrome (MKS) gene by targeted exon-enriched next-generation sequencing and deletion analysis. Hum Mol Genet. 2011; 20:2524–2534. [PubMed: 21493627]
39. Jones MA, Bhide S, Chin E, Ng BG, Rhodenizer D, Zhang VW, et al. Targeted polymerase chain reaction-based enrichment and next generation sequencing for diagnostic testing of congenital disorders of glycosylation. Genet Med. 2011; 13:921–932. [PubMed: 21811164]
40. Hollants S, Redeker EJ, Matthijs G. Microfluidic amplification as a tool for massive parallel sequencing of the familial hypercholesterolemia genes. Clin Chem. 2012; 58:717–724. [PubMed: 22294733]
41. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 2007; 39:522–527.
42. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461:272–276. [PubMed: 19684571]
43. Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ, D'Ascenzo M, et al. Whole exome capture in solution with 3 Gbp of data. Genome Biol. 2010; 11:R62. [PubMed: 20565776]
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; 4:903–905. [PubMed: 17934467]
45. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. Microarray-based genomic selection for high-throughput resequencing. Nat Methods. 2007; 4:907–909. [PubMed: 17934469]

46. Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin Gordon D, Brizuela L, et al. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat Protoc. 2009; 4:960–974. [PubMed: 19478811]
47. Mokry M, Feitsma H, Nijman IJ, de Bruijn E, van der Zaag PJ, Guryev V, Cuppen E. Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res. 2010; 38:e116. [PubMed: 20164091]
48. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009; 27:182–189. [PubMed: 19182786]
49. Tewhey R, Nakano M, Wang X, Pabón-Peña C, Novak B, Giuffre A, et al. Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biol. 2009; 10:R116. [PubMed: 19835619]
50. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol. 2011; 29:908–914. [PubMed: 21947028]
51. Casci T. DNA sequencing: exome sequencing technologies compared. Nat Rev Genet. 2011; 12:741. [PubMed: 22005978]
52. Wu CH, Fallini C, Ticozzi N, Keagle PJ, Sapp PC, Piotrowska K, et al. Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. Nature. 2012; 488:499–503. [PubMed: 22801503]
53. Veltman JA, Brunner HG. De novo mutations in human genetic disease. Nat Rev Genet. 2012; 13:565–575. [PubMed: 22805709]
54. Yan XJ, Xu J, Gu ZH, Pan CM, Lu G, Shen Y, et al. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat Genet. 2011; 43:309–315. [PubMed: 21399634]
55. Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature. 2012; 486:405–409. [PubMed: 22722202]
56. Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science. 2011; 333:1154–1157. [PubMed: 21798897]
57. Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat Genet. 2011; 44:47–52. [PubMed: 22158541]
58. Xiong D, Li G, Li K, Xu Q, Pan Z, Ding F, et al. Exome sequencing identifies MXRA5 as a novel cancer gene frequently mutated in non-small cell lung carcinoma from Chinese patients. Carcinogenesis. 2012; 33:1797–1805. [PubMed: 22696596]
59. Liu P, Morrison C, Wang L, Xiong D, Vedell P, Cui P, et al. Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. Carcinogenesis. 2012; 33:1270–1276. [PubMed: 22510280]
60. Krauthammer, M., Kong, Y., Ha, BH., Evans, P., Bacchiocchi, A., McCusker, JP., et al. Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. Nat Genet. 2012. http://dx.doi.org/10.1038/ng.2359
61. Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, Davis S, Stemke-Hale K, Davies MA, Gershenwald JE, Robinson W, Robinson S, Rosenberg SA, Samuels Y. NISC Comparative Sequencing Program. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nat Genet. 2011; 43:442–446. [PubMed: 21499247]
62. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012; 44:685–689. [PubMed: 22610119]
63. Zang ZJ, Cutcutache I, Poon SL, Zhang SL, McPherson JR, Tao J, et al. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. Nat Genet. 2012; 44:570–574. [PubMed: 22484628]

64. Ponting CP. The functional repertoires of metazoan genomes. Nat Rev Genet. 2008; 9:689–698. [PubMed: 18663365]

65. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet. 2010; 11:345–355. [PubMed: 20376054]

66. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, et al. Transcription-mediated gene fusion in the human genome. Genome Res. 2006; 16:30–36. [PubMed: 16344562]

67. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, et al. RIKEN Genome Exploration Research Group; Genome Science Group (Genome Network Project Core Group); FANTOM Consortium. Antisense transcription in the mammalian transcriptome. Science. 2005; 309:1564–1566. [PubMed: 16141073]

68. Gott JM, Emeson RB. Functions and mechanisms of RNA editing. Annu Rev Genet. 2000; 34:499–531. [PubMed: 11092837]

69. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10:57–63. [PubMed: 19015660]

70. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458(2009):719–724. [PubMed: 19360079]

71. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer. 2007; 7:233–245. [PubMed: 17361217]

72. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, et al. Transcriptome sequencing to detect gene fusions in cancer. Nature. 2009; 458:97–101. [PubMed: 19136943]

73. 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 USA. 2009; 106:12353–12358. [PubMed: 19592507]

74. Palanisamy N, Ateeq B, Kalyana-Sundaram S, Pflueger D, Ramnarayanan K, Shankar S, et al. Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat Med. 2010; 16:793–798. [PubMed: 20526349]

75. 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. [PubMed: 21036922]

76. Nacu S, Yuan W, Kan Z, Bhatt D, Rivers CS, Stinson J, et al. Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. BMC Med Genomics. 2011; 4:11. [PubMed: 21261984]

77. Ren S, Peng Z, Mao JH, Yu Y, Yin C, Gao X, et al. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Res. 2012; 22:806–821. [PubMed: 22349460]

78. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol. 2011; 12:R6. [PubMed: 21247443]

79. Ha KC, Lalonde E, Li L, Cavallone L, Natrajan R, Lambros MB, et al. Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines. BMC Med Genomics. 2011; 4:75. [PubMed: 22032724]

80. Steidl C, Shah SP, Woolcock BW, Rui L, Kawahara M, Farinha P, et al. MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. Nature. 2011; 471:377–381. [PubMed: 21368758]

81. Scott DW, Mungall KL, Ben-Neriah S, Rogic S, Morin RD, Slack GW, et al. TBL1XR1/TP63: a novel recurrent gene fusion in B-cell non-Hodgkin lymphoma. Blood. 2012; 119:4949–4952. [PubMed: 22496164]

82. Pierron G, Tirode F, Lucchesi C, Reynaud S, Ballet S, Cohen-Gogo S, et al. A new subtype of bone sarcoma defined by BCOR-CCNB3 gene fusion. Nat Genet. 2012; 44:461–466. [PubMed: 22387997]

83. Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, et al. Integrative analysis of the melanoma transcriptome. Genome Res. 2010; 20:413–427. [PubMed: 20179022]

84. Esteller M. Non-coding RNAs in human disease. Nat Rev Genet. 2011; 12:861–874. [PubMed: 22094949]

85. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009; 10:155–159. [PubMed: 19188922]

86. Bussemakers MJ, van Bokhoven A, Verhaegh GW, Smit FP, Karthaus HF, Schalken JA, et al. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. Cancer Res. 1999; 59:5975–5979. [PubMed: 10606244]

87. de Kok JB, Verhaegh GW, Roelofs RW, Hessels D, Kiemeney LA, Aalders TW, et al. DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. Cancer Res. 2002; 62:2695–2698. [PubMed: 11980670]

88. Ji P, Diederichs S, Wang W, Böing S, Metzger R, Schneider PM, et al. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene. 2003; 22:8031–8041. [PubMed: 12970751]

89. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell. 2010; 39:925–938. [PubMed: 20797886]

90. Panzitt K, Tschernatsch MM, Guelly C, Moustafa T, Stradner M, Strohmaier HM, et al. Characterization of HULC, a novel gene with striking upregulation in hepatocellular carcinoma, as non-coding RNA. Gastroenterology. 2007; 132:330–342. [PubMed: 17241883]

91. Wang J, Liu X, Wu H, Ni P, Gu Z, Qiao Y, et al. CREB up-regulates long noncoding RNA, HULC expression through interaction with microRNA-372 in liver cancer. Nucleic Acids Res. 2010; 38:5366–5383. [PubMed: 20423907]

92. Pasmant E, Laurendeau I, Héron D, Vidaud M, Vidaud D, Bièche I. Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. Cancer Res. 2007; 67:3963–3969. [PubMed: 17440112]

93. Yap KL, Li S, Muñoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, et al. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol Cell. 2010; 38:662–674. [PubMed: 20541999]

94. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 2007; 129:1311–1323. [PubMed: 17604720]

95. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010; 464:1071–1076. [PubMed: 20393566]

96. Kogo R, Shimamura T, Mimori K, Kawahara K, Imoto S, Sudo T, et al. Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. Cancer Res. 2011; 71:6320–6326. [PubMed: 21862635]

97. Kim, K., Jutooru, I., Chadalapaka, G., Johnson, G., Frank, J., Burghardt, R., et al. HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. Oncogene. 2012. http://dx.doi.org/10.1038/onc.2012.193

98. Spizzo, R., Almeida, MI., Colombatti, A., Calin, GA. Long non-coding RNAs and cancer: a new frontier of translational research?. Oncogene. 2012. http://dxdoi.org/10.1038/onc.2011.621

99. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009; 458:223–227. [PubMed: 19182780]

100. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell. 2009; 136:629–641. [PubMed: 19239885]

101. 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; 29:742–749. [PubMed: 21804560]

102. Pasquinelli AE. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nat Rev Genet. 2012; 13:271–282. [PubMed: 22411466]

103. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature. 2005; 435:834–838. [PubMed: 15944708]

104. Calin GA, Ferracin M, Cimmino A, Di Leva G, Shimizu M, Wojcik SE, et al. A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N Engl J Med. 2005; 353:1793–1801. [PubMed: 16251535]

105. Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, et al. A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Natl Acad Sci USA. 2006; 103:2257–2261. [PubMed: 16461460]

106. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, et al. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell. 2006; 9:189–198. [PubMed: 16530703]

107. Chen X, Ba Y, Ma L, Cai X, Yin Y, Wang K, et al. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell Res. 2008; 18:997–1006. [PubMed: 18766170]

108. Vaz C, Ahmad HM, Sharma P, Gupta R, Kumar L, Kulshreshtha R, Bhattacharya A. Analysis of microRNA transcriptome by deep sequencing of small RNA libraries of peripheral blood. BMC Genomics. 2010; 11:288. [PubMed: 20459673]

109. Schulte JH, Marschall T, Martin M, Rosenstiel P, Mestdagh P, Schlierf S, et al. Deep sequencing reveals differential expression of microRNAs in favorable versus unfavorable neuroblastoma. Nucleic Acids Res. 2010; 38:5919–5928. [PubMed: 20466808]

110. Persson H, Kvist A, Rego N, Staaf J, Vallon-Christersson J, Luts L, et al. Identification of new microRNAs in paired normal and tumor breast tissue suggests a dual role for the ERBB2/Her2 gene. Cancer Res. 2011; 71:78–86. [PubMed: 21199797]

111. Han Y, Chen J, Zhao X, Liang C, Wang Y, Sun L, et al. MicroRNA expression signatures of bladder cancer revealed by deep sequencing. PLoS ONE. 2011; 6:e18286. [PubMed: 21464941]

112. Ugras S, Brill E, Jacobsen A, Hafner M, Socci ND, Decarolis PL, et al. Small RNA sequencing and functional characterization reveals MicroRNA-143 tumor suppressor activity in liposarcoma. Cancer Res. 2011; 71:5659–5669. [PubMed: 21693658]

113. Volinia S, Galasso M, Sana ME, Wise TF, Palatini J, Huebner K, Croce CM. Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. Proc Natl Acad Sci USA. 2012; 109:3024–3029. [PubMed: 22315424]

114. Martens-Uzunova ES, Jalava SE, Dits NF, van Leenders GJ, Møller S, Trapman J, et al. Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. Oncogene. 2012; 31:978–991. [PubMed: 21765474]

115. Cortez MA, Bueso-Ramos C, Ferdin J, Lopez-Berestein G, Sood AK, Calin GA. MicroRNAs in body fluids – the mix of hormones and biomarkers. Nat Rev Clin Oncol. 2011; 8:467–477. [PubMed: 21647195]

116. White RJ. RNA polymerases I and III, non-coding RNAs and cancer. Trends Genet. 2008; 24:622–629. [PubMed: 18980784]

117. Xu W, Seok J, Mindrinos MN, Schweitzer AC, Jiang H, Wilhelmy J, et al. Human transcriptome array for high-throughput clinical studies. Proc Natl Acad Sci USA. 2011; 108:3707–3712. [PubMed: 21317363]

118. Heyn H, Esteller M. DNA methylation profiling in the clinic: applications and challenges. Nat Rev Genet. 2012; 13:679–692. [PubMed: 22945394]

119. Carvalho RH, Haberle V, Hou J, van Gent T, Thongjuea S, van Ijcken W, et al. Genome-wide DNA methylation profiling of non-small cell lung carcinomas. Epigenetics Chromatin. 2012; 5:9. [PubMed: 22726460]

120. Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. PLoS ONE. 2009; 4:e7767. [PubMed: 19907642]

121. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009:R25. [PubMed: 19261174]

122. Li H, Durbin R. Fast and accurate short read alignment with Burrows– Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

123. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008; 18:1851–1858. [PubMed: 18714091]

124. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009; 41:1061–1067. [PubMed: 19718026]
125. Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics. 2008; 9:128. [PubMed: 18307793]
126. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol. 2009:e1000386. [PubMed: 19461883]
127. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009; 25:1966–1967. [PubMed: 19497933]
128. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome Res. 2001; 11:1725–1729. [PubMed: 11591649]
129. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5:621–628. [PubMed: 18516045]
130. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010; 26:873–881. [PubMed: 20147302]
131. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice. accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010; 38:e178. [PubMed: 20802226]
132. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics. 2011; 27:2518–2528. [PubMed: 21775302]
133. Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res. 2010; 38:4570–4578. [PubMed: 20371516]
134. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
135. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009; 19:1117–1123. [PubMed: 19251739]
136. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 2008; 18:810–820. [PubMed: 18340039]
137. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, et al. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics. 2008; 24:2818–2824. [PubMed: 18952627]
138. Bryant DW Jr, Wong WK, Mockler TC. QSRA: a quality-value guided de novo short read assembler. BMC Bioinformatics. 2009; 10:69. [PubMed: 19239711]
139. Li RQ, Zhu HM, Ruan J, Qian W, Fang XD, Shi ZB, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010; 20:265–272. [PubMed: 20019144]
140. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. [PubMed: 18349386]
141. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28:511–515. [PubMed: 20436464]
142. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012; 28:1086–1092. [PubMed: 22368243]
143. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010; 28:503–510. [PubMed: 20436462]
144. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010; 7:909–912. [PubMed: 20935650]
145. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29:644–652. [PubMed: 21572440]

146. 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; 20:1297–1303. [PubMed: 20644199]

147. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. Genome project data processing subgroup. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]

148. Simola DF, Kim J. Sniper: improved SNP discovery by multiply mapping deep sequenced reads. Genome Biol. 2011; 12:R55. [PubMed: 21689413]

149. Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, et al. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics. 2010; 26:730–736. [PubMed: 20130035]

150. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. SNP detection for massively parallel whole-genome resequencing. Genome Res. 2009; 19:1124–1132. [PubMed: 19420381]

151. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009; 6:677–681. [PubMed: 19668202]

152. Sindi SS, Onal S, Peng LC, Wu HT, Raphael BJ. An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biol. 2012; 13:R22. [PubMed: 22452995]

153. Lee S, Hormozdiari F, Alkan C, Brudno M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. Nat Methods. 2009; 6:473–474. [PubMed: 19483690]

154. Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol. 2009; 10:R23. [PubMed: 19236709]

155. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25:2865–2871. [PubMed: 19561018]

156. Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, et al. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics. 2010; 26:i350–357. [PubMed: 20529927]

157. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, et al. DeFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol. 2011; 7:e1001138. [PubMed: 21625565]

158. Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, Rozowsky JS, et al. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol. 2010; 11:R104. [PubMed: 20964841]

159. Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V. Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. Bioinformatics. 2011; 27:1068–1075. [PubMed: 21330288]

160. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. 2011; 12:R72. [PubMed: 21835007]

161. Zhang Q, Ding L, Larson DE, Koboldt DC, McLellan MD, Chen K, et al. CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. Bioinformatics. 2010; 26:464–469. [PubMed: 20031968]

162. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011; 21:974–984. [PubMed: 21324876]

163. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012; 28:311–317. [PubMed: 22155872]

164. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009; 25:2283–2285. [PubMed: 19542151]

165. Browning BL, Yu Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet. 2009; 85:847–861. [PubMed: 19931040]

166. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009; 5:e1000529. [PubMed: 19543373]

167. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 2007; 8:R143. [PubMed: 17659080]

168. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci USA. 2011; 108:9530–9535. [PubMed: 21586637]

169. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. 2011; 12:R112. [PubMed: 22067484]

170. Sasson A, Michael TP. Filtering error from SOLiD output. Bioinformatics. 2010; 26:849–850. [PubMed: 20207696]

171. Wu H, Irizarry RA, Bravo HC. Intensity normalization improves color calling in SOLiD sequencing. Nat Methods. 2010; 7:336–337. [PubMed: 20431543]

172. Quinlan AR, Stewart DA, Strömberg MP, Marth GT. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. Nat Methods. 2008; 5:179–181. [PubMed: 18193056]

173. Kao WC, Stevens K, Song YS. BayesCall: a model-based base-calling algorithm for high-throughput short-read sequencing. Genome Res. 2009; 19:1884–1895. [PubMed: 19661376]

174. Kircher M, Stenzel U, Kelso J. Improved base calling for the Illumina genome analyzer using machine learning strategies. Genome Biol. 2009; 10:R83. [PubMed: 19682367]

175. Bravo HC, Irizarry RA. Model-based quality assessment and base-calling for second-generation sequencing data. Biometrics. 2010; 66:665–674. [PubMed: 19912177]

176. Massingham T, Goldman N. All your base: a fast and accurate probabilistic approach to base calling. Genome Biol. 2012; 13:R13. [PubMed: 22377270]

177. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998; 8:186–194. [PubMed: 9521922]

178. Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat Methods. 2009; 6:S6–S12. [PubMed: 19844229]

179. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012; 486:207–214. [PubMed: 22699609]

180. Human Microbiome Project Consortium. A framework for human microbiome research. Nature. 2012; 486:215–221. [PubMed: 22699610]

181. Faulkner GJ, Forrest AR, Chalk AM, Schroder K, Hayashizaki Y, Carninci P, Hume DA, Grimmond SM. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics. 2008; 91:281–288. [PubMed: 18178374]

182. Paşaniuc B, Zaitlen N, Halperin E. Accurate estimation of expression levels of homologous genes in RNA-seq experiments. J Comput Biol. 2011; 18:459–468. [PubMed: 21385047]

183. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323. [PubMed: 21816040]

184. Williams C, Pontén F, Moberg C, Söderkvist P, Uhlén M, Pontén J, et al. A high frequency of sequence alterations is due to formalin fixation of archival specimens. Am J Pathol. 1999; 155:1467–1471. [PubMed: 10550302]

185. Srinivasan M, Sedmak D, Jewell S. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. Am J Pathol. 2002; 161:1961–1971. [PubMed: 12466110]

186. Yost SE, Smith EN, Schwab RB, Bao L, Jung H, Wang X, et al. Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. Nucleic Acids Res. 2012; 40:e107. [PubMed: 22492626]

187. Kerick M, Isau M, Timmermann B, Sültmann H, Herwig R, Krobitsch S, et al. Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE)

tumor tissues, input amount and tumor heterogeneity. BMC Med Genomics. 2011; 4:68. [PubMed: 21958464]

188. Wagle N, Berger MF, Davis MJ, Blumenstiel B, Defelice M, Pochanard P, et al. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer Discov. 2012; 2:82–93. [PubMed: 22585170]
189. Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature. 2009; 462:1005–1010. [PubMed: 20033038]
190. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012; 366:883–892. [PubMed: 22397650]
191. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011; 472:90–94. [PubMed: 21399628]
192. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell. 2012; 148:873–885. [PubMed: 22385957]
193. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell. 2012; 148:886–895. [PubMed: 22385958]
194. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006; 24:1151–1161. [PubMed: 16964229]
195. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, et al. The MicroArray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol. 2010; 28:827–838. [PubMed: 20676074]

Table 1

Overview of major next-generation sequencing platforms.

Company	Platform	Amplification	Sequencing	Read length	Throughput/ time per run	Dominant error type	Overall error rate
Roche/454 Life Sciences	GS FLX Titanium XL+	Emulsion PCR	Pyrosequencing	Up to 1 kb	700 Mb/23 h	Indel	0.5%
	GS FLX Titanium XLR70			Up to 600 bp	450 Mb/10 h		
	GS Junior			~400 bp	35 Mb/10 h		
	HiSeq 2000			36–100 bp	105–600 Gb/2– 11 days		
Illumina	Genome Analyzer IIx	Bridge PCR	Sequencing-by- synthesis with reversible terminator	35–150 bp	10–95 Gb/2– 14 days	Substitution	0.2%
	MiSeq			36–250 bp	540 Mb–8.5 Gb/4–39 h		
Life Technologies/ Applied Biosystems	5500xl SOLiD™ system	Emulsion PCR	Sequencing by ligation	35–75 bp	10–15 Gb/day	Substitution	0.1%
	SOLiD™ 4 system			25–50 bp	25–100 Gb/3.5– 16 days		
Life Technologies/Ion Torrent	Ion Proton™ sequencer (Proton I chip)	Emulsion PCR	Ion semiconductor sequencing	Up to 200 bp	Up to 10 Gb/2–4 h	Indel	1%
	Ion PGM™ sequencer (318 chip)			35–200 bp	300 Mb–1 Gb/0.9–4.5 h		
Helicos BioSciences	HeliScope™ single molecule sequencer	NONE	Single molecule sequencing	25–55 bp	21–35 Gb/8 days	Deletion	5%
Pacific Biosciences	PacBio RS	NONE	Single molecule sequencing	250 bp– 10 kb	NA	Indel	15%

Table 2

Cancer driver mutations discovered by large-scale next generation sequencing.

Gene	Aberration type	Tumor type	Biological function	Tumor effect	Sequencing method	Number of samples	Sample type	Reference
EBF1-PDGFRB, BCR-JAK2, NUP214-ABL1	Fusion	ALL	Kinase signaling	Activating	Whole- genome	15	Acute lymphoblastic leukemia	22
IL7R, SH2B3	Mutation	ALL	Cytokine signaling	Activating	Whole- genome	15	Acute lymphoblastic leukemia	22
TP53	Mutation	Cell Carcinoma	Cell cycle regulation	Inactivating	Whole- genome	457	Peripheral blood	23
VTI1A-TCF7L2	Fusion	Colon	Transcription factor	Activating	Whole- genome	9	Colorectal adenocarcinomas	24
ARID1A, ARID1B, ARID2, MLL, MLL3	Mutation	Liver	Chromatin regulation	Inactivating	Whole- genome	27	Hepatocellular carcinoma	25
PREX2	Mutation	Melanoma	Rac exchange factor	Inactivating	Whole- genome	25	Melanomas	26
ATRX	Mutation	Neuroblastoma	Telomere maintenance	Inactivating	Whole- genome	40	Neuroblastomas	27
BRIP1	Mutation	Ovary	DNA repair	Inactivating	Whole- genome	457	Peripheral blood	28
DNMT3A	Mutation	AML	DNA methylation	Inactivating	Exome	112	Acute monocytic leukemias	54
CBFB	Mutation	Breast	Transcription factor	Inactivating	Exome	103	Breast cancers	55
MAGI3-AKT3	Fusion	Breast	Cell signaling	Activating	Exome	103	Breast cancers	55
N0TCH1	Mutation	Cell carcinoma	Cell signaling	Inactivating	Exome	32	Head and neck squamous cell carcinomas	56
SF3B1	Mutation	CML	mRNA splicing	Inactivating	Exome	105	Chronic lymphocytic leukemias	57
MXRA5	Mutation	Lung	Matrix remodeling	Activating	Exome	14	Non-small cell lung carcinomas	58
CSMD3	Mutation	Lung	Unknown	Inactivating	Exome	31	Non-small cell lung carcinomas	59
RAC1	Mutation	Melanoma	Cell signaling	Activating	Exome	147	Melanomas	60
GRIN2A	Mutation	Melanoma	Glutamate receptor	Unknown	Exome	14	Melanomas	61
SPOP, FOXA1, MED 12	Mutation	Prostate	Transcription regulation	Unknown	Exome	112	Prostate tumors	62
FAT4	Mutation	Stomach	Cell adhesion	Inactivating	Exome	15	Gastric adenocarcinomas	63
ARID1A	Mutation	Stomach	Chromatin remodeling	Inactivating	Exome	15	Gastric adenocarcinomas	63

Table 3

Bioinformatics tools for next-generation sequencing analysis.

Primary category	Program	Author(s)	URL
Unspliced alignment	BFAST	Homer et al. [120]	http://sourceforge.net/apps/mediawiki/bfast/
	Bowtie	Langmead et al. [121]	http://bowtie-bio.sourceforge.net/
	BWA	Li et al. [122]	http://bio-bwa.sourceforge.net/
	Cross_match	Phil Green and co-workers	http://www.phrap.org/phredphrapconsed.html
	ELAND	Anthony J. Cox	http://www.illumina.com
	MAQ	Li et al. [123]	http://maq.sourceforge.net/
	Mosaik	Michael Strömberg	http://bioinformatics.bc.edu/marthlab/Mosaik
	mrFAST	Alkan et al. [124]	http://mrfast.sourceforge.net/
	RMAP	Smith et al. [125]	http://rulai.cshl.edu/rmap/
	SHRiMP	Rumble et al. [126]	http://compbio.cs.toronto.edu/shrimp/
	SOAP2	Li et al. [127]	http://soap.genomics.org.cn/soapaligner.html
	SSAHA2	Ning et al. [128]	http://www.sanger.ac.uk/resources/software/ssaha2/
Spliced alignment	ERANGE	Mortazavi et al. [129]	http://woldlab.caltech.edu/rnaseq
	GSNAP	Wu et al. [130]	http://research-pub.gene.com/gmap/
	MapSplice	Wang et al. [131]	http://www.netlab.uky.edu/p/bioinfo/MapSplice
	RUM	Grant et al. [132]	http://cbil.upenn.edu/RUM/
	SpliceMap	Au et al. [133]	http://www.stanford.edu/group/wonglab/SpliceMap/
	TopHat	Trapnell et al. [134]	http://tophat.cbcb.umd.edu/
De *novo* genome assembly	ABySS	Simpson et al. [135]	http://www.bcgsc.ca/platform/bioinfo/software/abyss/
	ALLPATHS-LG	Butler et al. [136]	http://www.broadinstitute.org/software/allpaths-lg/blog/
	CABOG	Miller et al. [137]	http://sourceforge.net/apps/mediawiki/wgs-assembler/
	Newbler	Margulies et al. [2]	http://454.com/
	QSRA	Bryant et al. [138]	http://qsra.cgrb.oregonstate.edu/
	SOAPdenovo	Li et al. [139]	http://soap.genomics.org.cn/soapdenovo.html
	Velvet	Zerbino et al. [140]	http://www.ebi.ac.uk/zerbino/velvet/
Transcriptome assembly	Cufflinks	Trapnell et al. [141]	http://cufflinks.cbcb.umd.edu/
	Oases	Schulz et al. [142]	http://www.ebi.ac.uk/zerbino/oases/
	Scripture	Guttman et al. [143]	http://www.broadinstitute.org/software/scripture/
	Trans-ABySS	Robertson et al. [144]	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss
	Trinity	Grabherr et al. [145]	http://trinityrnaseq.sourceforge.net/
SNP detection	GATK	McKenna et al. [146]	http://www.broadinstitute.org/gatk/
	SAMtools	Li et al. [147]	http://samtools.sourceforge.net/
	Sniper	Simola et al. [148]	http://kim.bio.upenn.edu/software/sniper.shtml
	SNVMix	Goya et al. [149]	http://www.bcgsc.ca/platform/bioinfo/software/SNVMix
	SOAPsnp	Li et al. [150]	http://soap.genomics.org.cn/soapsnp.html
Structural variation detection	BreakDancer	Chen et al. [151]	http://gmt.genome.wustl.edu/breakdancer/current/
	GASVPro	Sindi et al. [152]	http://compbio.cs.brown.edu/software.html
	MoDIL	Lee et al. [153]	http://compbio.cs.toronto.edu/modi1/
	PEMer	Korbel et al. [154]	http://sv.gersteinlab.org/pemer/

Primary category	Program	Author(s)	URL
	Pindel	Ye et al. [155]	https://trac.nbic.nl/pindel/
	VariationHunter	Hormozdiari et al. [156]	http://compbio.cs.sfu.ca/strvar.htm
Fusion detection	deFuse	McPherson et al. [157]	http://defuse.sourceforge.net/
	FusionSeq	Sboner et al. [158]	http://archive.gersteinlab.org/proj/rnaseq/fusionseq/
	ShortFuse	Kinsella et al. [159]	http://exon.ucsd.edu/ShortFuse
	TopHat-Fusion	Kim et al. [160]	http://tophat.cbcb.umd.edu/fusion_index.html
CNV detection	CMDS	Zhang et al. [161]	https://dsgweb.wustl.edu/qunyuan/software/cmds/
	CNVnator	Abyzov et al. [162]	http://sv.gersteinlab.org/cnvnator/
Somatic variant detection	SomaticSniper	Larson et al. [163]	http://gmt.genome.wustl.edu/somatic-sniper/current/
Somatic and germline variant detection	VarScan	Koboldt et al. [164]	http://genome.wustl.edu/software/varscan
Genotype Calling	BEAGLE	Browning et al. [165]	http://faculty.washington.edu/browning/beagle/beagle.html
	IMPUTE2	Howie et al. [166]	http://mathgen.stats.ox.ac.uk/impute/impute_v2.html