Chapter 6
Advancements in Next-Generation Sequencing for Detecting Minimal Residual Disease

Erin L. Crowgey and Nitin Mahajan

Introduction

The sequencing of DNA molecules enables the precise identification and order of nucleotides. These techniques include any method or technology that is used to determine the order of the four bases—adenine (A), guanine (G), cytosine (C), and thymine (T)—in a strand of DNA. Massively parallel sequencing, also known as next-generation sequencing (NGS),

DNA is a giant resource that will change mankind, like the printing press.
– James D. Watson [1]

E. L. Crowgey (✉)
Nemours Alfred I. duPont Hospital for Children, Biomedical Research Department, Wilmington, DE, USA
e-mail: erin.crowgey@nemours.org

N. Mahajan
Washington University in St. Louis, Pediatric Hematology and Oncology, St Louis, MO, USA

T. E. Druley (ed.), *Minimal Residual Disease Testing*,
https://doi.org/10.1007/978-3-319-94827-0_6

is a method of simultaneously sequencing millions of fragments of DNA (or complementary DNA) simultaneously. These advancements have revolutionized the field of molecular biology [2–4] and are routinely used in a wide variety of research and clinical settings. Indispensable knowledge attained with modern DNA sequencing technology has been instrumental to unveil the plethora of previously hidden facts not only in the medical field but also in plant biology, forensic science, and evolution [3, 5].

Since the inexplicable relationship between genetic instability and tumorigenesis was proposed by Nowell in 1976 [6], progress in cancer genomics has strengthened and provided strong evidence to support this fundamental hypothesis. Most genetic disorders have been associated with modifications within regions that affect the coding of proteins, and are often divided into three types of categories: (1) single-gene disorder, (2) chromosomal disorders, and (3) complex disorders. There are many challenges associated with characterizing a genetic disorder, creating a unique niche for developing appropriate bioinformatics methodologies.

New developments in molecular biology techniques, which are often used to gain insight into genetic disorders, are generating massive amounts of data that need processed and refined prior to being incorporated into electronic health records (EHR) or clinical decision support applications [7]. Although these new techniques, such as NGS, are gaining momentum in the clinical field, there are no gold standards for clinical data analysis, interpretation, and integration that can be broadly applied to all disorders. Technical advancements in NGS led to a dramatic decrease in sequencing costs, mostly due to increase in the volume of data generated over the same period of time, to the point where an entire human genome in 2018 can be sequenced for less than $1000 [8], which ultimately increases its accessibility to researchers. Unfortunately, this expansion in data production has not been accompanied by an equivalent improvement in sequencing fidelity, as the chemistry needed for speed and volume currently comes at the price of precision, which is of little consequence when looking for sequence changes that are heterozygous or homozygous. To be clear, no amount of "deep sequencing" will be able to recover true mutations occurring at frequencies below the error rate of the sequencing platform itself as stochastic errors are continually generated at a constant rate [9]. Yet, the promise of genome-driven information in medical science is undoubtedly inspiring as witnessed by the targeted therapies based on the detection of oncogenic drivers [10, 11].

The corollary to the volumes of sequence data generated is the necessity for subsequent computational strategies to handle the previously incomprehensible volumes of data. Ultimately, there is a unique niche in the informatics field, especially in bioinformatics, to develop and deliver robust methodologies capable of analyzing the massive amount of data being generated by new technologies. Bioinformatics, an interdisciplinary field, requires the intersection of computer science, statistics, mathematics, biology, and engineering, with the ultimate goal of being applied in the clinical field for translating large biological datasets into diagnostic or predictive knowledge, clearly requiring a team science approach.

In Table 6.1, we list a brief history of the advances and applications of DNA (cDNA) sequencing techniques, as well as bioinformatic analyses, along with their applicability and challenges in detecting the minimal residual disease (MRD) in leukemia patients, which is highlighted in Table 6.2.

Cancer Genetics

The French-American-British (FAB) classification is a morphology-based system that was introduced several decades ago to help classify specific leukemias into subgroups [12]. Unfortunately, throughout the last several years, it has become apparent that a general classification system, such as FAB, does not apply broadly and appropriately apply to all cancer types and age ranges, as the genetic landscape between adult and pediatric cancers can be much different [13].

TABLE 6.1 History Ssequencing

Year	Lead researcher/ association	Highlight
1865	Gregor Mendel	Uses peas to figure out the fundamental of principles of heredity
1871	Friedrich Miescher	Identified the presence of "nuclein" (now known as DNA) and associated proteins, in the cell nucleus
1904	Walter Sutton and Theodor Boveri	Proposed the chromosome theory of heredity after finding that chromosomes occur in matched pairs, one inherited from the mother and one from the father
1910	Albrecht Kossel	Discovered the five nucleotide bases, adenine, cytosine, guanine, thymine, and uracil
1950	Erwin Chargaff	Suggested pairing pattern of the bases A, C, G, and T
1952	Alfred Hershey and Martha Chase	Demonstrated DNA, rather than protein, carries genetic information
1953	James Watson and Francis Crick	Published the double helix structure of DNA
1961	Marshall Nirenberg, Har Gobind Khorana, and colleagues	Identified how to read the DNA sequences in blocks of three "codon." Each codon codes for an amino acid which is added to the protein during translation
1965	Robert Holley and colleagues	Sequenced yeast tRNA
1970	Ray Wu	Used primer extension to read a short sequence of DNA for the first time
1972	Walter Fiers	Sequenced first whole gene coding for a MS2 virus protein
1973	Walter Gilbert and Allan Maxam	Developed a method to sequence DNA using chemicals to cut DNA at certain bases

TABLE 6.1 (continued)

Year	Lead researcher/ association	Highlight
1975	Frederick Sanger	Introduced "plus and minus" method for DNA sequencing using gels to separate DNA by size
1977	Frederick Sanger	Establishes dideoxy sequencing methodology
1983	Kary Mullis	Developed polymerase chain reaction (PCR)
1984	Fritz Pohl	Developed nonradioactive sequencing platform
1985	Alec Jeffreys	Developed a method for DNA profiling
1986	Leray Hoad and Applied Biosystem (ABI)	Developed first automated sequencer
1990	Human Genome Project (world's largest collaborative biological project)	Human Genome Project is launched
1995	Fleischmann RD and colleagues	Bacterial genome sequenced (*Haemophilus influenzae)*
	Fraser CM and colleagues	Bacterial genome sequenced (*Mycoplasma genitalium)*
1996	Mostafa Ronaghi	Introduced pyrosequencing, next-generation "sequencing by synthesis" method
	Applied Biosystem (ABI)	Introduced first commercial sequencing using capillary electrophoresis
	International collaboration	Sequenced the genome of yeast, *Saccharomyces cerevisiae*

(continued)

Table 6.1 (continued)

Year	Lead researcher/ association	Highlight
1998	John Sulston and Bob Waterston	Published the genome of the nematode worm, *Caenorhabditis elegans*
	Solexa Inc.	Developed sequencing by synthesis method that uses fluorescent dye
1999	Part of Human Genome Project	First human chromosome 22 is sequenced
2000		Genome of *Drosophila melanogaster* sequenced
	University of California, Santa Cruz	Launch of the UCSC Genome Browser
2001	Human Genome Project	Human Genome Project publishes first draft human genome sequence
2002	International Mouse Genome Sequencing Consortium	Mouse genome published
	The International HapMap Project	Project is launched to generate a “catalogue” of common human genetic variations and their locations
2003	Human Genome Project	Completed and confirmed humans have approximately 20,000–25,000 genes
	National Human Genome Research Institute	Launched the ENCODE project with the aims to identify and characterize all the genes in the human genome
2005	454 Life Sciences	The 454 system, based on pyrosequencing becomes the first commercially available next-generation sequencer

Table 6.1 (continued)

Year	Lead researcher/ association	Highlight
	The International HapMap Project	Map of human genetic variations published
2007	SOLid Systems	Launched a sequencing technology based on ligation
2008	1000 Genomes Project	Aims to sequence the whole genomes of a large number of people (2500)
	Cancer Genome Consortium	Comprehensive analysis of cancer genome
	Ley TJ	Sequences first cancer (AML) genome characterized by NGS
2009		Third-generation sequencing with single-molecule fluorescence technology is launched with Helicos sequencer
2011	Pacific Biosciences	Launched first commercial single-molecule real-time technology
2012	Oxford Nanopore Technologies	Commercialization of the portable nanopore sequencing methods

In 1976, Nowell highlighted the strong relationship between genetic instability and tumorigenesis [6], which provided the foundation for studying precise genetic alterations and their association with cancer. The majority of genetic analyses for cancer have been conducted using traditional cytogenetic techniques, such as karyotyping, and cytogenetic markers have played a major role in the diagnosis and classification of leukemias. The field of cytogenetics was initiated in 1956 [14] with the discovery and description of the number of chromosomes in a diploid human cell. There are several different techniques within the cytogenetic field that have been previously reviewed [15]. Overall, sensitivity and specificity are optimized when multiple cytogenetic methods are performed concurrently to overcome the limitations of any single method. Therefore, it is essential to have broad and precise methods to integrate multiple data sources for characterizing MRD to facilitate risk stratification and therapeutic selection [16].

History of DNA Sequencing

In 1910, Albrecht Kossel discovered the five nucleotide bases: adenine, cytosine, guanine, thymine and uracil, as the fundamental building blocks of nucleic acids [17]. Four decades later, Erin Chargaff recognized the pairing pattern of these nucleotides in DNA and RNA [17]. Robert Holley and colleagues (1965) were accredited for sequencing the first ever full nucleic acid molecule, 77 nucleotides of the yeast, *Saccharomyces cerevisiae*, alanine tRNA with a proposed cloverleaf structure [18]. It took more than 5 years to extract enough tRNA from the yeast to identify the sequence of nucleotide residues using specific ribonucleases, two-dimensional chromatography, and spectrophotometric procedures [18]. Initially, scientists focused their sequencing efforts on the readily available populations of RNA species because of the following properties: (i) bulk production in culture, (ii) not complicated by a complementary strand, and (iii) considerably shorter than DNA [19, 20]. The laborious and expensive nature of the sequencing drove the continuous development and refinement of subsequent sequencing methods.

Fred Sanger and colleagues at Cambridge were also actively working on methods for sequencing DNA molecules. They developed a technique based on the detection of radiolabeled partially-digested fragments after two-dimensional fractionation [21], allowing addition of nucleotides to the growing pool of ribosomal and transfer RNA sequences. Using a primer extension method in year 1968, Ray Wu and Dale Kaiser sequenced a short sequence of DNA for the first time [22]. However, the actual determination of bases was still restricted to small sequences of DNA because of the requirement for radioactive and hazardous chemicals. These continuous efforts resulted in generating the first complete protein-coding gene sequence, which was the coat protein of bacteriophage MS2 in 1972 [23], and the first complete 3569-nucleotide-long genome sequence of the bacteriophage MS2 RNA in 1976 [24].

Two influential techniques in the mid-1970s emerged which later gave a new dimension to the field of molecular biology. The two techniques were Alan Coulson and Sanger's "plus and minus" technique, using DNA polymerase to sequentially add radiolabeled nucleotides, and Allan Maxam and Walter Gilbert's chemical cleavage technique [25–27]. Both of these techniques moved away from 2D fractionation toward polyacrylamide gel electrophoresis, which provided better base resolution. The development of these two methods is often described as the foundation of modern sequencing but was supplanted in 1977 with Sanger's "chain termination" or "dideoxy technique," which quickly became the most widely used sequencing method over the next several decades.

The full potential of Sanger sequencing was not realized until the integration of a series of seminal improvements occurred. First, radioactive isotope labels were replaced with variably colored fluorescent tags for each nucleotide, which enabled the reaction to occur in a single vessel instead of four. A second key improvement was the use of capillary tube-based electrophoresis which provided better resolution, required less equipment space, and decreased the time required. Following these improvements, Smith et al. (1986) at Applied Biosystems Instruments™ (ABI) designed the first automated capillary sequencing system and later introduced the first commercial automated DNA sequencer [28].

These retrospectively named "first-generation" sequencers were the first to incorporate computer-based data acquisition and analysis and were capable of producing reads >300 bp. However, to analyze longer DNA molecules, "shotgun sequencing" was developed by separately cloning and sequencing overlapping DNA fragments. Coinciding with the discovery of polymerase chain reaction (PCR) and the launch of the Human Genome Project, a series of enhancements allowed machine cycle times to decrease from 18 h to 3 h [29].

In 1992, the Institute for Genomic Research (TIGR) in Rockville, Maryland, founded by J. Craig Venter, pioneered the industrialization of an automated sequencer, with a focus on studying various genomes [2, 30]. With the establishment of both the first Affymetrix® and GeneChip® microarrays in 1996, expression studies involving various genes in prokaryotes and eukaryotes were possible [31]. By the end of 1999, TIGR had generated 83 million nucleotides of cDNA sequence, 87,000 human cDNA sequences, and the complete genome sequences of *Haemophilus influenzae* [32] and *Mycoplasma genitalium* [33]. The platform resulted in the early completion of the Human Genome Project in 2003.

Next-Generation Sequencing Application

With the completion of the human genome sequence, the clinical and research appetite for comparative sequencing data expanded overnight, rapidly overwhelming the capacity and cost structure of dideoxy base sequencing. Various groups sought to bring new instruments to market (Fig. 6.1) that offered various strategies for (i) the parallelization of many sequencing reactions, (ii) the preparation of amplified sequencing libraries prior to sequencing, (iii) library amplification on miniature surfaces (solid surfaces, beads, emulsion droplets), (iv) direct monitoring of the nucleotides via advanced microfluidics and imaging, (v) reduced per-nucleotide costs, and (vi) decreased machine cycle times. However, early NGS platforms were designed to sequence entire genomes from single subjects rather than selected regions from multiple subjects. Thus, targeted sequencing for the coding regions of the genome (i.e., the exome) or regions of interest was facilitated via probe hybridization of fragmented DNA or by customized PCR amplification. Figure 6.1 highlights common library preparation protocols, sequencing platforms, and bioinformatic considerations for performing an NGS project.

Short-read sequencing (SRS) typically produces reads that are 50–600 bp in length and has become the dominant type of NGS available today through multiple vendors (ref 34). SRS often results in scaffolding gaps due to bias from high GC content, repeat sequences, and missing insertions. There are several advantages of SRS such as high throughput, low cost

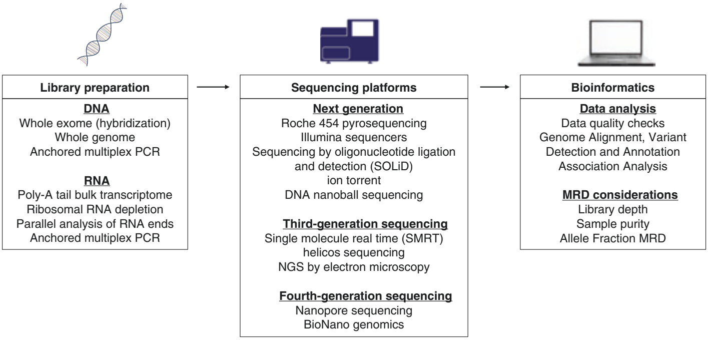


FIGURE 6.1 Overview of next-generation sequencing. *Library preparation box*: Library preparations are specific for DNA or RNA sequencing applications. Different capture techniques are available and determine what type of mutations can be captured for sequencing. *Sequencing platforms box*: The first instruments capable of performing massive parallelization of sequencing were termed next-generation sequencers. Third-generation instruments focus on long-read sequencing techniques. Fourth-generation sequencers involve unique technologies that preserve the spatial localization of the DNA/RNA molecules. *Bioinformatics box*: Raw next-generation sequencing data is unstructured and massive, requiring special analytical pipelines. When detecting minimal residual disease, it is essential to consider characteristics about the sample (purity and allele fraction MRD), sequencing technology (library depth requirements), and capture technique

per base, and a low raw read rate [35]. However, the short-read length complicates genome alignment leading to false-positive and false-negative variant calling [36, 37]. The error rate of approximately 1% primarily occurs due to dephasing of nucleotide additions (most frequently due to adding an erroneous base but can also occur due to missing base addition or adding an extra base inappropriately) to random sequences at random clusters across the flow cell, more so in the later sequencing cycles. Furthermore, de novo assembly approaches can be challenging with SRS and require enhanced algorithms for performing these operations, such as SOAPdenovo [38]. Assemble of a large genome, especially for non-model organisms, generated from SRS are limited as long-range linking information is not available [39].

In contrast, newer long-range sequencing (LRS) techniques produce reads between 10 Kb and 40 Kb [3, 4, 40, 41]. While the long read makes alignment and phasing more tractable, these platforms have historically suffered from lower total output, relatively high error rates and cost. Unfortunately, these approaches are not presently sufficient to detect very rare mutations in heterogeneous nucleic acid samples, but that may change with additional improvements.

There are several variant algorithm detection methods, including FreeBayes [42], that are specific for SRS data. The advantages for SRS for MRD include low error rate and the ability to generate deep coverage for a specific region of the genome. Therefore, SRS has dominated the field for cancer genomics as variant detection is more accurate with SRS over LRS techniques that have a higher error rate and less sensitive limit of detection. Furthermore, as error-corrected sequencing (Overview Fig. 6.2) and single-cell sequencing continue to develop, the advantages of SRS increase.

More than 70% of genetic variations seen in humans are non-SNP variations and can be missed easily with short-read sequencing [34]. Long-read sequencing enables reads longer than 10 kb, which improves alignment to the reference genome, high consensus accuracy, uniform coverage, and detection of epigenetic modifications. In addition, long-read sequencing is beneficial in transcriptomic analyses as it allows

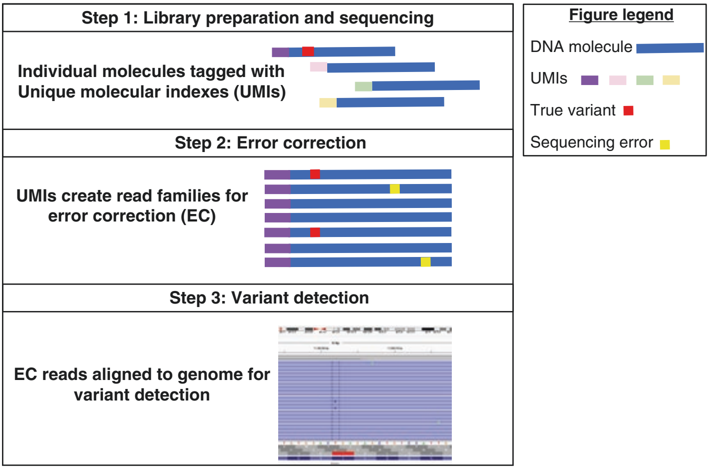

Step 1: Library preparation and sequencing
Individual molecules tagged with Unique molecular indexes (UMIs)
Step 2: Error correction
UMIs create read families for error correction (EC)
Step 3: Variant detection
EC reads aligned to genome for variant detection
Figure legend
DNA molecule
UMIs
True variant
Sequencing error

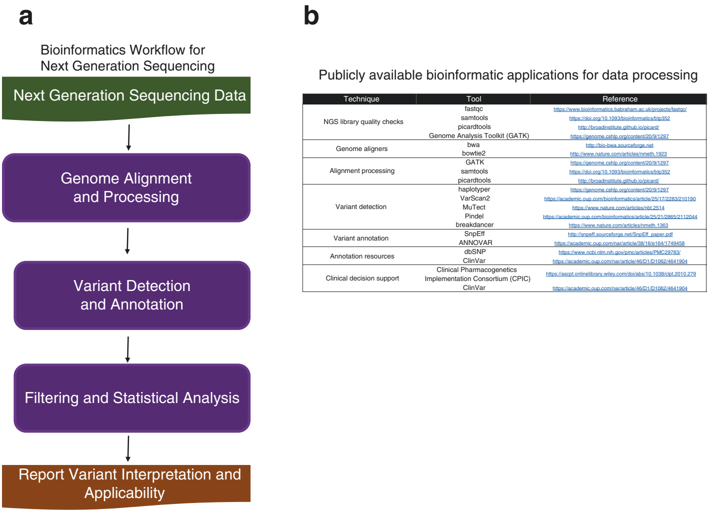
FIGURE 6.2 Workflow for error-corrected sequencing. By tagging each DNA molecule with a unique molecular index (Step 1), read families can be generated, by aligning reads with the same UMI, and used to determine sequencing errors versus low allelic variants (Step 2). Error-corrected sequences are then aligned to a reference genome and used to calculate variants (Step 3)

detection of splice isoforms with a high level of confidence without requiring assembly. High costs of long-read sequencing and high error rates are the major hurdle for adopting these platforms as a global sequencing platform.

Bioinformatics and NGS

Bioinformatics is an interdisciplinary field focused on developing methods for translating one or more large biological datasets, inherent in NGS, into applicable knowledge. Through translational computational discoveries, clinicians have gained a better understanding of genetic alterations associated with many disorders, as a priori knowledge is not required. With the publication of several best practices guidelines for NGS, basic processing steps have been well established in the scientific community (Overview Fig. 6.3).

However, as the field continues to leverage more complex NGS strategies in cancer genetics, bioinformatics has become the bottleneck in terms of expertise, infrastructure, and time to results. Currently, there is no gold standard computational pipeline that will work for all analysis, and oftentimes each project needs specific tailoring of the algorithms, which requires rigorous validation of computational processes and biological results.


a
Bioinformatics Workflow for Next Generation Sequencing
Next Generation Sequencing Data
Genome Alignment and Processing
Variant Detection and Annotation
Filtering and Statistical Analysis
Report Variant Interpretation and Applicability
b

Publicly available bioinformatic applications for data processing

Technique	Tool	Reference
NGS library quality checks	fastqc	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
	samtools	https://doi.org/10.1093/bioinformatics/btp352
	picardtools	http://broadinstitute.github.io/picard/
	Genome Analysis Toolkit (GATK)	https://genome.cshlp.org/content/20/9/1297
Genome aligners	bwa	http://bio-bwa.sourceforge.net
	bowtie2	http://www.nature.com/articles/nmeth.1923
Alignment processing	GATK	https://genome.cshlp.org/content/20/9/1297
	samtools	https://doi.org/10.1093/bioinformatics/btp352
	picardtools	http://broadinstitute.github.io/picard/
Variant detection	haplotyper	https://genome.cshlp.org/content/20/9/1297
	VarScan2	https://academic.oup.com/bioinformatics/article/25/17/2283/210190
	MuTect	https://www.nature.com/articles/nbt.2514
	Pindel	https://academic.oup.com/bioinformatics/article/25/21/2865/2112044
	breakdancer	https://www.nature.com/articles/nmeth.1363
Variant annotation	SnpEff	http://snpeff.sourceforge.net/SnpEff_paper.pdf
	ANNOVAR	https://academic.oup.com/nar/article/38/16/e164/1749458
Annotation resources	dbSNP	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC29783/
	ClinVar	https://academic.oup.com/nar/article/46/D1/D1062/4641904
Clinical decision support	Clinical Pharmacogenetics Implementation Consortium (CPIC)	https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1038/clpt.2010.279
	ClinVar	https://academic.oup.com/nar/article/46/D1/D1062/4641904

FIGURE 6.3 Bioinformatics workflow for next-generation sequencing assays. (**a**) NGS generates massive raw unstructured sequence data (green box) that is used for downstream processing. The most common file format for this data is a fastq file. The NGS reads are aligned to a reference genome and processed for variant detection annotation (purple boxes). The final output from a computational pipeline is a variant call file (VCF) that contains all of the relevant metadata per variant (dark orange box). (**b**) Numerous publicly available tools resources are available for data analysis. *maybe do a supplemental table linked here with full webpage and references? [43]

Quality Assessment of NGS Libraries

In 2012, the US Centers for Disease Control and Prevention (CDC) published the guidelines assembled by a national working group, termed Next-Generation Sequencing: Standardization of Clinical Testing or Nex-StoCT, to lead an initiative for defining platform-independent guidelines for using NGS in clinical practice [44]. The Supplementary Guidelines published by Nex-StoCT highlight key quality metrics that should be considered when establishing and validating a clinical NGS workflow. There are several publicly available tools for performing these types of quality control assessment. For example, fastqc, a platform-independent NGS quality tool (Babraham Institute, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), can import data from alignment files or raw NGS data and reports an overview of quality statistics that may indicate problems or biases in the NGS data. There are ten key statistical modules within the fastqc pipeline that report a value of “pass,” “warning,” or “fail” for the NGS library, consistent with the guidelines published by the CDC. We have briefly summarized each of these ten metrics below.

The sequence length, or insert size, is a basic analysis, and skewing from the expected insert size can indicate poor library construction, which needs to be carefully considered when analyzing data for novel InDels. This is a relatively simple calculation that is considered along with other basic statistics such as the total number of sequences and the maximum and minimum sequence length. The library depth is key for understanding limits of detection and should be taken into consideration during data analysis and experimental design phases. One of the remarkable aspects of genomics is its scalability. Depending upon the mutation frequency threshold (e.g., 5% vs 1% vs 0.1% or lesser) being interrogated, the sequencing strategy and library depth for an MRD assay will be quite different compared to germline sequencing requirements.

When analyzing a sample for MRD, it is essential to examine the per sequence quality scores to determine if the library has a portion of reads with low-quality values that might skew results. Ideally, poor quality reads should be a very small percentage of the total raw data. Occasionally the 3′ end of NGS reads can be of poor quality when sequencing by synthesis because sequences in a cluster can elongate at slightly different rates, which will slowly lead to desynchronizationand quality issues [45]. These low-quality bases should therefore be "trimmed" or "clipped" to help with accurate alignment and variant detection. Furthermore, overrepresented sequences are any sequences that may be overrepresented in the NGS reads, i.e., adaptors, and it is important to trim these types of sequences from the raw NGS data to improve genome alignment. Some alignment software, such as bwa-mem, enable soft clipping and can "trim" these adapter sequences during genome alignment. However, it is good practice to determine the level of adapter contamination in a library file prior to downstream analysis. Several publicly available tools, such as cutadapt [46], are established for just these purposes and improve the efficiency of downstream NGS analysis [47].

Analyzing the per base sequence content statistic is another key metric to analyze. Typically, a sample is flagged if the difference between A, T, G, or C is greater than 10% in any position. However, some targeted or enriched NGS datasets, such as exomes from hybridization capture, are known to display discreet, position-specific compositional biases [48]. In addition, nucleotide composition is known to more likely be G + C rich in first exons, and general variation basis within the genome and at the gene level are noted [49]. These known skews within coding regions can cause a targeted sequencing library to be inappropriately flagged as poor quality, so it is important to recognize the difference between targeted sequencing statistics compared to random bulk.

The per base GC content statistics calculates the GC content across the length of each sequence and compares it to a modeled normal distribution of GC content. Typically, it is important for all libraries to have the sum of the deviations from the normal distribution represented <30% of the reads.

The per base N content statistic calculates the percentage of base calls at each position for which N, or no base call, appears. A base is called "*N*" when the sequencing quality is too low to accurately call a base. The number of "*N*" bases should be limited and typically only at the 5′ or 3′ end of the reads. A library with a large portion of bases called as N indicates overall poor sequencing, and the library should be flagged and re-sequenced.

The sequence duplication level statistics counts the degree of duplication for every sequence in a library. This is important as libraries with high sequence duplication levels indicate low complexity (i.e., not sampling enough total molecules) along with a likely enrichment bias during library preparation. However, certain types of library preparations, such as RNA-seq, are known to have some bias that can lead to high duplication levels, whereas for exomes, the number of duplicated reads should be minimal. Duplicate reads can be flagged and eliminated or normalized during computational analysis to avoid false-positive variant calling and is part of the best practice guidelines for downstream processing.

The Kmer statistics measures the number of each 7-mer at each position in the library. It then uses a binomial test to look for significant deviations from an even coverage at all positions [50]. Any Kmers with a positional biased enrichment are reported. Libraries which derive from random priming can show Kmer bias at the start of the library due to an incomplete sampling of the possible random primers (Babraham Institute).

Genome Alignment and Processing

High-quality genome alignment of NGS reads is essential for accurate detection of germline, somatic, and MRD variants. There are numerous publicly available genome aligners, including but not limited to bwa-mem [51], bowtie2 [52], and FreeBayes [42], as well as de novo assembly packages that do not require forced reference mapping. Genome alignment works well for single-nucleotide variants (SNVs), small InDels, and structural variant (SV) detection when combined with pair-end data. DNA and RNA can be aligned to a genome, although RNA mapping requires an aligner capable of gapped alignment, such as tophat [53]/bowtie2 [52], due to the structure of DNA (introns) compared to RNA molecules.

The Broad Institute has developed a publicly available software package, Genome Analysis Toolkit (GATK), that provides a suite of algorithms required for processing and analyzing alignment files. GATK best practices [54] include several steps for improving genome alignment. The first step in processing the alignment files is to mark duplicate reads because they potentially represent a clonal amplification rather than a randomly sheared DNA fragment.

Next, the alignment files are analyzed for potential intervals that need realignment based on known genetic alterations. These regions undergo a realignment step for optimization. Finally, the alignment file undergoes a recalibration of base quality scores based on the adjustments from the proceeding steps. After recalibration, the quality scores, in theory, should be more accurate because the new score is closer to the actual probability of mismatching to the reference genome.

After the alignment files are appropriately processed, it is important to assess the quality of the overall genome alignment. There are several different criteria that can be used, such as percentage of reads mapped to the genome, but it is important to consider the depth and breadth of coverage. Coverage calculations are especially important when trying to understand variant calling capabilities and limitations. The depth of coverage is how many times a nucleotide was sequenced, whereas the breadth of coverage is the average coverage per base per the number of total bases queried or the sequence interval.

Variant Detection and Annotation

Single-nucleotide polymorphisms (SNPs) naturally occur between healthy individuals with estimates ranging from 1 in

1000 to 1 in 1500 nucleotides [55–57]. Collectively, SNPs result in ~3 million nucleotide differences using the estimated genome size of ~3 billion nucleotides (haploid). Additionally, somatic mutations are difficult to detect because they occur at low frequencies in the genome and might only be present in a small fraction of the DNAmolecules [58]. Often, tools used for detecting germline SNPs are not recommended for detecting SNVs. The sensitivity and specificity of an algorithm to detect a somatic mutation are dependent on several characteristics such as sequencing depth, local sequencing error rate, and allelic fraction.

There are several publicly available algorithms, including MuTect [59] and VarScan [60], for detecting somatic variants. Algorithms like MuTect simultaneously analyze both a noncancer and cancer alignment file from the same patient, consisting of four key steps: (1) removal of low-quality sequence data, (2) variant detection in the tumor sample using a Bayesian classifier or probabilistic classifier, (3) filtering to remove false positives resulting from correlated sequencing artifacts that are not captured by the error model, and (4) designation of the variants as somatic or germline by a second Bayesian classifier.

There are multiple approaches for detecting SV using NGS data. For example, de novo assembly, with either the complete dataset or unmapped reads, is one strategy for detecting large SV [61]. One limitation to this approach is that it can only detect homozygous SV because detecting heterozygous SV requires assembly of haplotype sequences, which is a complex problem that is not fully resolved.

Reference mapping strategies are another approach and include concepts around split pair-end read mapping, read coverage depth analysis, or analysis of inconsistent insert size of paired-end reads. These approaches first require the NGS reads to be mapped to a reference genome, and then the alignment files are analyzed for genomic variants. The detection of SV using NGS data requires accurate prediction of copy, content, and structure. Often algorithms developed for detecting SVs are specific for a class of SVs, making it a necessity to incorporate multiple SV algorithms into the workflow.

Common algorithms include Pindel, BreakDancer, and VarScan2. Pindel can detect breakpoints of large deletions, medium-sized insertions, inversions, and tandem duplications by leveraging a pattern growth approach [62]. Previously, Pindel has been cited for detecting an internal tandem duplication (ITD) in the *FLT3* gene by using a pattern growth approach to analyze NGS data misaligned to the reference genome due to biological differences.

BreakDancer predicts five types of structural variants: insertions, deletions, inversions, inter- and intrachromosomal translocations, and the results from BreakDancer can be directly feed into Pindel to help enhance the analysis as a whole. VarScan2 is another package capable of detecting SVs, including copy number variations (CNVs) and InDels [63]. Ultimately, the analysis pipeline requires the integration of all of these variant callers, because leukemias are genetically heterogeneous diseases that are not characterized by a few variants or even the same class of variants.

The majority of the variant detection algorithms output a variant call file (VCF), which needs further annotation. There are several publicly available tools for performing robust somatic variant annotation, and recently the Association for Molecular Pathology published standards and guidelines for interpreting such annotations [64]. The majority of these algorithms are capable of annotating SNVs with putative functional consequences, reporting functional importance scores, and identifying previously reported SNVs and allele frequencies [64]. Oftentimes these annotations are applied in downstream filtering strategies to prioritize relevant variants. SnpEff is another publicly available algorithm for annotating VCFs [65] and provides a suite for predicting the effect of variants. Typically, researchers are interested in variants that alter the sequence of proteins, such as a missense or frameshift. However, it is noticeable that predicting whether or not a variant is damaging is still a complex issue that needs further refinement, which was recently highlighted by the Critical Assessment of Genome Interpretation (CAGI) experiments [66].

Variant Filtering and Association Analyses

NGS assays generate massive amounts of variant data, including normal population heterogeneity, sequencing artifacts, and potentially disease-associated variation. Ultimately, an effective filtering strategy for identifying disease-associated variation is required, which must be accompanied by appropriate false-negative and false-positive rates. VCF annotation packages (reviewed above) enable a researcher to build an effective filtering strategy to focus on variants that are high quality and of clinical relevance [67, 68].

Typically, filtering strategies are implemented for small sample sizes, and often more complex association statistics, like genome-wide association studies (GWAS), are not possible. Recently, several algorithms have been designed for performing associations between a set of rare variants and phenotypes from NGS data, including SNP-set (sequence) kernel association test (SKAT) [69]. Of interest, SKAT is capable of analyzing the cumulative effect of rare and common variants and is well suited for associating NGS data to a phenotype of interest.

Due to the rarity of the mutations in question, filtering strategies for MRD are complex and different compared to the filtering strategies employed to detect high-frequency somatic variants and germline variants or SNPs. Thus, while not mandatory, in order to effectively analyze MRD, it is essential to quantify the cumulative profile of low AF variants at time of diagnosis if at all possible, akin to ΔN flow cytometry, and track the presence of these and other new mutations at multiple time points post therapy.

DNA Sequencing and Applications in MRD

The three most common NGS DNA sequencing approaches in oncology are whole genome sequencing (WGS), whole exome sequencing (WES), or a targeted gene panel approach (Table 6.2). Each has strengths and weaknesses. WES enriches for sequences encoding proteins, which represents ~1–2% of the human genome [70]. WES works well for common sequence changes in coding regions, such as germline variants and high-frequency somatic mutations in cancer, but is not adequate for MRD below 2% VAF as the error rates of NGS preclude identification below that threshold (Table 6.2). More importantly, WES ignores the noncoding, regulatory regions of the genome and is incapable of detecting the breadth of mutational diversity common in cancer, such as cryptic gene fusions and complex structural variants.

WGS generates the sequence of the entire genome, not just the 1–2% in protein-coding regions. This is of obvious benefit for detecting cancer-related mutations [71]. WGS generates massive amounts of data per individual and is still fairly expensive for a clinical assay. Furthermore, the computational infrastructure required to analyze WGS is complex and beyond the majority of most clinical labs, even in the commercial space. Similar to WES, this technique works well for germline variants and detection of common somatic mutations. Beyond WES, WGS will identify duplications,

Table 6.2 NGS and MRD

NGS library preparation	Analyte	Advantages MRD	Disadvantage MRD	Recommended for MRD
Hybridization	DNA	Targeted assay Compatible with error correction Cheap	Limitations in variant capture for structural variants, cryptic gene fusions, and large insertions/deletions	Yes

Table 6.2 (continued)

NGS library preparation	Analyte	Advantages MRD	Disadvantage MRD	Recommended for MRD
Whole genome	DNA	Bulk sequencing no a priori knowledge required	Expensive Variant capture limitations Limit of detection not suited for MRD High error rate	No
Anchored multiplex PCR (AMP)	DNA / RNA	Targeted assay Compatible with error correction Cheap Diverse variant capture	High sequencing depths required	Yes
Poly-A tail bulk transcriptome	RNA	Bulk sequencing no a priori knowledge required	High error rate Expensive for depth requirements Limitations in variant capture	No
Ribosomal RNA depletion	RNA	Bulk sequencing no a priori knowledge required	High error rate High sequencing depth required Expensive Limitations in variant capture	No
Single-cell sequencing	RNA	Robust clonal analysis	High error rate Sequencing depth requirements are cost prohibitive Limitations in variant capture	Yes

DNA fusions, inversions, large InDels, and other SVs that would not be visible by WES. However, genomes typically don't get high depth of sequencing as a cost-saving measure, which obviates their utility for MRD (Table 6.2). For both WES and WGS, two critical considerations in the clinical space are whether they are reimbursed by third-party payors and what is the obligation of the individual ordering the test to relate incidental findings. For these reasons, many third-party payors refuse payment, and many clinicians are reticent to order tests that are much more broad in scope than the cancer-related question at hand. However, some academic centers have established genetic counseling services or other protocols to relate the transfer of incidental genetic findings to patients and families.

For MRD, targeted panels are currently better suited due to being more customizable for specific disease-related loci and cost-efficacy. Targeted panels use a variety of strategies to enrich for target sequences such as nucleic acid hybridization, rolling circle amplification, molecular inversion probes, and various PCR-based enrichment strategies. Each has advantages and disadvantages, but the main metric is "on/off target" percentage. At the best of times, these various strategies (other than plain PCR) only "capture" somewhere between 2 and 7% of the molecules available. For MRD, this needs to be taken into careful consideration. For example, to detect a mutation at 0.0001 (1:10,000), one needs to query at least 10,000 different molecules. If one starts with 5,000,000 molecules (~8.2 μg) and 95% are not captured, that leaves 250,000 molecules, so a mutation at 0.0001 should be seen 25 times. Obviously, the amount of starting material becomes substantial, and often rate-limiting, quite quickly.

Error-Corrected DNA Sequencing

Given that leukemias are a heterogenous mixture of subclones [72], error-corrected sequencing (ECS) enables the tagging of a single DNA molecule with a unique molecular index [72–75]. Utilizing this approach, stochastic errors are introduced by the sequencing platforms. As demonstrated by Young et al. [73], targeted gene panels incorporating a unique molecular index are capable of detecting clonal hematopoiesis involving known oncogenes in healthy adults [73].

The main aspects of error-corrected sequencing as described by Young et al. are (i) aggregating all of the reads arising from a single molecule as demonstrated by sharing the same random index (e.g., "read family") to computationally subtract stochastic sequencing artifacts and (ii) an analysis of the error rate at each base by establishing a negative binomial distribution. Read family aggregation into a single "error-corrected consensus sequence" is done prior to genome alignment. After variant detection, a second analysis can be done to calculate the error rate at each position and further filter variants to remove false positives that are actually sequencing artifacts [73].

Different ECS strategies must be implemented to identify SNVs and/or InDels at low allelic frequencies [73, 76] compared to SVs. Precise quantification of SVs in DNA involves amplifying a target locus with many different amplification primers on each side of the putative lesion or breakpoint. This results in many possible amplicons, which can then be aligned via de novo assembly rather than forced reference alignment. A major aspect of applying this technology for MRD (Table 6.2) is the ability to link the data to either an earlier time point from the same subject, and to connect the variants of interest to external resources for rigors filtering strategies.

RNA Sequencing and Applications in MRD

Various library techniques exist for RNA sequencing, including poly-A tail capture for mRNA, ribosomal RNA depletion, parallel analysis of RNA ends (PARE-seq), and targeted gene panels [77, 78]. Bulk RNA sequencing techniques process thousands of cells at once and represent the "average" of all of the molecules within the mixed population sequenced.

The traditional analysis for RNA-seq was differential gene expression, but over the last several years, the community has developed extensive computational pipelines for detecting variants [79].

RNA sequencing enables the detection of several variant classes that are not easily detectable via DNA sequencing, including cryptic gene fusions, exon usage, and allele-specific expression. These types of variants are quickly being associated with various cancers [80] and are of interest for MRD applications [81] either alone or in combination with other diagnostic tools. Poddighe et al. (2018) recently developed an MRD assay for CBFB-MYH11 gene fusion. Interestingly, the fusion was also detected at time of birth in the same subject when analyzing cord blood [81] highlighting the growing realization that "cancer-related" mutations are far more common in the general population than previously appreciated (Young AL et al., *Nat Commun* 2016) because only ill people have historically undergone such careful analyses. As capture techniques and computational pipelines continue to improve our ability to detect these types of variants, our understanding of their biology is quickly changing.

Similar to gene panels for DNA sequencing, targeted RNA gene panel sequencing strategies enable an in-depth analysis of transcripts of interest. A PCR amplification strategy, single or opposing primers, enables the capture of cryptic gene fusions and will hopefully help to improve our understanding of intron retention in cancer patients [82]. Targeted RNA panels also have the advantage, compared to bulk sequencing methods, of incorporating unique molecular indexes, which enables error correction.

Error-Corrected RNA Sequencing

Similar to DNA-ECS, RNA-ECS is currently best applied to targeted gene panels. In contrast, the RNA-ECS consensus read family sequence is aligned to the genome using a gap reference aligner and analyzed for small InDels and SNVs.

For more complex variants, including cryptic gene fusions, alternative exon usage, structural variants, and novel transcript structure, such as retained intron, a de novo assembly approach from multiple, variably sized amplicons across the breakpoint is required [83], compared to a force reference alignment. There are several publicly available de novo assemblers available for this type of work, such as ABySS [84]. One major limitation of ECS, whether for RNA or DNA, is the inability to co-localize mutations within the same cellular background. The identified clonal mutations are often so rare that (a) the overall difference in the abundance and genes that are mutated between heathy and diseased individuals is minimal, (b) making the likelihood of multiple mutations co-occurring to be statistically highly unlikely, but biologically critical. One could imagine new mutations being identified serially in a patient treated for leukemia. The effect size of these new mutations may range from negligible, if occurring alone at a low frequency, to catastrophic, if co-occurring in a clone or subclone of the original leukemia. ECS can only make inferences as to the relatively likelihood of mutational co-occurrence.

Single-Cell Sequencing

Against this context, the field of cancer genomics will quickly transition to sequencing techniques that enable the genomic or epigenomic characterization of an individual cell, providing higher resolution of co-occurring mutations within the same cell and have even lead to the discovery of new cell types [85]. Processing for single-cell RNA sequencing (sc-RNA-seq) is quite different than traditional transcriptome sequencing. Currently, RNA preservation is key such that the first step is isolation of viable, single cells from the tissue of interest. For dissociation of single cells from solid tumors, this can effect RNA quality and requires special handling procedures compared to storage for DNA sequencing projects.

The application of sc-RNA-seq for MRD has not yet been fully realized because current RNA sequencing on these platforms only queries dozens of bases at the 3′ end of an mRNA molecule. Thus, it is quite good at quantifying transcripts but cannot uniformly quantify mutations residing further upstream in the mRNA molecule. As these technologies mature and become more amenable to MRD analyses, it will be important to consider sequencing depth requirements and number of cells assayed [86].

References

1. BBC NEWS | science/nature | what they said: genome in quotes.
2. Kulski JK. Next-generation sequencing — an overview of the history, tools, and "Omic" applications; 2016
3. Levy SE, Myers RM. Advancements in next-generation sequencing. Annu Rev Genomics Hum Genet. 2016;17(1):95–115.
4. Srinivasan S, Batra J. Four generations of sequencing- is it ready for the clinic yet? J Next Gener Seq Appl. 2014;1:107.
5. Ari Ş, Arikan M. Next-generation sequencing: advantages, disadvantages, and future. In: Plant omics: trends and applications. Cham: Springer; 2016. p. 109–35.
6. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23–8.
7. Fan J, Han F, Liu H. Challenges of big data analysis. Natl Sci Rev. 2014;1(2):293–314.
8. Kruglyak KM, Lin E, Ong FS. Next-generation sequencing and applications to the diagnosis and treatment of lung Cancer. Adv Exp Med Biol. 2016;890:123–36.
9. https://dx.doi.org/10.1093%2Fnar%2Fgks1443.
10. Walter FM, Emery JD. Genetic advances in medicine: has the promise been fulfilled in general practice? Br J Gen Pract. 2012;62(596):120–1.
11. Landau DA, Carter SL, Getz G, Wu CJ. Clonal evolution in hematological malignancies and therapeutic implications. Leukemia. 2014;28(1):34.
12. Niederhuber J, Armitage J, Doroshow J, Kastan M, Tepper J. Abeloff's clinical oncology. 5th ed. Philadelphia: Saunders; 2013. p. 2224.

13. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions | nature medicine.
14. THE CHROMOSOME NUMBER OF MAN - TJIO - 1956 - Hereditas - Wiley Online Library.
15. Speicher MR, Carter NP. The new cytogenetics: blurring the boundaries with molecular biology. Nat Rev Genet. 2005;6(10):782.
16. Pui C-H, Carroll WL, Meshinchi S, Arceci RJ. Biology, risk stratification, and therapy of pediatric acute Leukemias: an update. J Clin Oncol. 2011;29(5):551–65.
17. Admin. The history of DNA timeline. DNA worldwide. 2014.
18. Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, et al. Structure of a ribonucleic acid. Science (80-). 1965;147(3664):1462–5.
19. Buermans HPJ, den Dunnen JT. Next generation sequencing technology: advances and applications. Biochim Biophys Acta Mol basis Dis. 2014;1842(10):1932–41.
20. Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. Genomics. 2016;107(1):1–8.
21. Sanger F, Brownlee GG. Barrell BG. A two-dimensional fractionation procedure for radioactive nucleotides. J Mol Biol. 1965;13(2):373–98.
22. Wu R, Kaiser AD. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. J Mol Biol. 1968;35(3):523–37.
23. Jou WM, Haegeman G, Ysebaert M, Fiers W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature. 1972;237(5350):82.
24. Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature. 1976;260(5551):500.
25. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94(3):441–8.
26. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, et al. Nucleotide sequence of bacteriophage φX174 DNA. Nature. 1977;265(5596):687.
27. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977;74(2):560–4.

28. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, et al. Fluorescence detection in automated DNA sequence analysis. Nature. 1986;321(6071):674.
29. Ansorge WJ. Next-generation DNA sequencing techniques. New Biotechnol. 2009;25(4):195–203.
30. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science (80-). 1991;252(5013):1651–6.
31. Bumgarner R. Overview of DNA microarrays: types, applications, and their future. Curr Protoc Mol Biol. 2013; Chapter 22:Unit 22.1. doi: https://doi.org/10.1002/0471142727.mb2201s101.
32. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science (80-). 1995;269(5223):496–512.
33. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al. The minimal gene complement of mycoplasma genitalium. Science (80-). 1995;270(5235):397–404.
34. A map of human genome variation from population scale sequencing. Nature. 2010;467(7319):1061–73.
35. Caspar SM, Dubacher N, Kopps AM, Meienberg J, Henggeler C, Matyas G. Clinical sequencing: from raw data to diagnosis with lifetime value. Clin Genet:n/a–a.
36. Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, et al. Fast computation and applications of genome Mappability. PLoS One. 2012;7(1):e30377.
37. Mandelker D, Schmidt RJ, Ankala A, Gibson KM, Bowser M, Sharma H, et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. Genet Med. 2016;18(12):1282.
38. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20(2):265–72.
39. The Long and the Short of DNA Sequencing. GEN.
40. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. Bio Med Res Int. 2012;2012:251364.
41. Illumina | sequencing and array-based solutions for genetic research.
42. Garrison E. Freebayes: Bayesian haplotype-based genetic polymorphism discovery and genotyping; 2018.

43. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
44. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol [Internet]. 2012;30(11):1033–6. Available from: http://www.nature.com/articles/nbt.2403.
45. https://doi.org/10.1038/nbt.1585.
46. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.
47. Del FC, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One. 2013;8(12):e85024.
48. Kozlowski P, de Mezer M, Krzyzosiak WJ. Trinucleotide repeats in human genome and exome. Nucleic Acids Res [Internet]. 2010;38(12):4027–39. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq127.
49. Louie E, Ott J, Majewski J. Nucleotide frequency variation across human genes. Genome Res. 2003;13(12):2594–601.
50. Babraham Institute, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
51. Burrows-Wheeler Aligner.
52. Bowtie 2: fast and sensitive read alignment.
53. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.
54. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet [Internet]. 2011;43(5):491–8. Available from: http://www.nature.com/articles/ng.806.
55. Schneider JA, Pungliya MS, Choi JY, Jiang R, Sun XJ, Salisbury BA, et al. DNA variability of human genes. Mech Ageing Dev [Internet]. 2003;124(1):17–25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12618002.
56. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature [Internet]. 2001;409(6822):928–33. Available from: http://www.nature.com/doifinder/10.1038/35057149.

57. Jorde LB, Wooding SP. Genetic variation, classification and "race". Nat Genet [Internet]. 2004;36(11s):S28–33. Available from: http://www.nature.com/doifinder/10.1038/ng1435.
58. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol [Internet]. 2013;31(3):213–9. Available from: http://www.nature.com/doifinder/10.1038/nbt.2514.
59. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213.
60. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568–76.
61. Li Y, Zhang Q, Yin X, Yang W, Du Y, Hou P, et al. Generation of iPSCs from mouse fibroblasts with a single gene, Oct4 and small molecules. Cell Res [Internet]. 2011;21(1):196–204. Available from: http://www.nature.com/articles/cr2010142.
62. 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 [Internet]. 2009;25(21):2865–71. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp394.
63. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in Cancer by exome sequencing. Genome Res [Internet]. 2012;22(3):568–76. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.129684.111.
64. https://doi.org/10.1016/j.jmoldx.2016.10.002.
65. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) [Internet]. 2012;6(2):80–92. Available from: http://www.tandfonline.com/doi/abs/10.4161/fly.19695.
66. https://doi.org/10.1002/humu.23290.
67. Carson AR, Smith EN, Matsui H, Brækkan SK, Jepsen K, Hansen J-B, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC Bioinf. 2014;15:125.

68. 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(14):e107.
69. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013;92(6):841–53.
70. Wang Z, Liu X, Yang B-Z, Gelernter J. The role and challenges of exome sequencing in studies of human diseases. Front Genet. 2013;4:160.
71. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis. 2010;31(1):27–36.
72. Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, et al. Sequencing small genomic targets with high efficiency and extreme accuracy. Nat Methods. 2015;12(5):423.
73. Young AL, Challen GA, Birmann BM, Druley TE. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat Commun. 2016;7:12484.
74. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci. 2011;108(23):9530–5.
75. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci. 2012;109(36):14508–13.
76. Zheng Z, Liebers M, Zhelyazkova B, Cao Y, Panditi D, Lynch KD, Chen J, Robinson HE, Shim HS, Chmielecki J, Pao W, Engelman JA, Iafrate AJ, Le LP. Anchored multiplex PCR for targeted next-generation sequencing. Nat Med. 2014;20(12):1479–84. https://doi.org/10.1038/nm.3729.
77. German MA, Pillay M, Jeong D-H, Hetawal A, Luo S, Janardhanan P, et al. Global identification of microRNA–target RNA pairs by parallel analysis of RNA ends. Nat Biotechnol. 2008;26(8):941.
78. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. Comparison of RNA-Seq by poly (a) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics. 2014;15:419.
79. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-Seq data. Am J Hum Genet. 2013;93(4):641–51.
80. https://www.medscape.com/viewarticle/555206.

81. Poddighe PJ, Veening MA, Mansur MB, Loonen AH, Westers TM, Merle PA, et al. A novel cryptic CBFB-MYH11 gene fusion present at birth leading to acute myeloid leukemia and allowing molecular monitoring for minimal residual disease. Hum Pathol Case Reports. 2018;11(Supplement C):34–8.
82. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. Genome Med. 2015;7:45.
83. 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(11):909.
84. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, et al. De novo transcriptome assembly with ABySS. Bioinformatics. 2009;25(21):2872–7.
85. Villani A-C, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, et al. Single-cell RNA-seq reveals new types of human blood dendritic cell, monocytes, and progenitors. Science (80-). 2017;356(6335):eaah4573.
86. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9:75.