

A Clinician's Guide to Bioinformatics for Next-Generation Sequencing



Nicholas Bradley Larson, PhD,^{a,*} Ann L. Oberg, PhD,^b Alex A. Adjei, MD, PhD,^c Ligu Wang, PhD^b

^a*Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo Clinic College of Medicine and Science, Rochester, Minnesota*

^b*Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic College of Medicine and Science, Rochester, Minnesota*

^c*Taussig Cancer Institute, Cleveland Clinic, Cleveland, Ohio*

Received 3 November 2021; revised 31 October 2022; accepted 5 November 2022
Available online - 12 November 2022

ABSTRACT

Next-generation sequencing (NGS) technologies are high-throughput methods for DNA sequencing and have become a widely adopted tool in cancer research. The sheer amount and variety of data generated by NGS assays require sophisticated computational methods and bioinformatics expertise. In this review, we provide background details of NGS technology and basic bioinformatics concepts for the clinician investigator interested in cancer research applications, with a focus on DNA-based approaches. We introduce the general principles of presequencing library preparation, postsequencing alignment, and variant calling. We also highlight the common variant annotations and NGS applications for other molecular data types. Finally, we briefly discuss the revealed utility of NGS methods in NSCLC research and study design considerations for research studies that aim to leverage NGS technologies for clinical care.

© 2022 International Association for the Study of Lung Cancer. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Keywords: Bioinformatics; Next-generation sequencing; DNA; Review

Introduction

Genetic and genomic assays have become increasingly important in biomedical research, especially in the context of cancer diagnosis and treatment. Many of these assays rely on some form of DNA sequencing, which is the process of characterizing the base nucleotide-resolution information of given DNA target sequence(s), consisting of canonical bases adenine (A),

guanine (G), thymine (T), and cytosine (C). Reliably capturing the inherited germline or acquired somatic variants in a patient's genome can provide critical diagnostic, prognostic, or predictive information for a given disease.

DNA sequencing technologies have been available since the 1970s, with Sanger sequencing emerging as the definitive standard approach.¹ Although this "first generation" technology remains in practice as an accurate sequencing solution, the scope of Sanger sequencing in terms of target genomic content is highly limited. In 2004, the Roche 454 FLX Pyrosequencer introduced a new age of commercially available sequencing technologies that have now been collectively referred to as next-generation sequencing (NGS).^{2,3} The "next" in NGS refers to the revolutionary technological leaps that permit massive parallelization of the DNA fragment sequencing process, analogous to millions of individual Sanger sequencing experiments running simultaneously. This high-throughput "shotgun" solution is capable of sequencing entire genomes in a rapid fashion. The magnitude of this throughput also comes at substantially reduced financial costs, making personalized genomics

*Corresponding author.

Disclosure: The authors declare no conflict of interest.

Address for correspondence: Nicholas Bradley Larson, PhD, Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo Clinic College of Medicine and Science, 200 First Street SW, Rochester, MN 55905. E-mail: Larson.Nicholas@mayo.edu

© 2022 International Association for the Study of Lung Cancer. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

ISSN: 1556-0864

<https://doi.org/10.1016/j.jtho.2022.11.006>

and precision medicine a modern reality. The large amount of data generated by NGS, however, requires extensive computational resources and sophisticated bioinformatics software to yield informative and actionable results.

In this review, we aim to broadly familiarize the reader with fundamental bioinformatics concepts related to NGS, targeting a clinical audience possessing modest familiarity with genomics and an interest in leveraging these technologies in research studies. First, we will outline the distinguishing characteristics of NGS as a technology with respect to DNA sequencing, define relevant terminology, and highlight key elements that require consideration for the downstream bioinformatics procedures. Next, we will discuss the raw NGS output data formats and primary bioinformatics procedures of alignment, variant calling, and annotation. Finally, we

briefly summarize how NGS technologies can be applied to various other molecular types and how study design considerations apply to experiments involving NGS. A glossary of common NGS bioinformatics terms can be found in Table 1. We also make note that although the covered concepts similarly lend to clinical sequencing, the highly regulated nature of Clinical Laboratory Improvement Amendments–certified laboratories necessary for clinical decision-making bears its own unique considerations for NGS bioinformatics, which we consider out of scope for this review.

Sample Processing for NGS

Sample Collection and Storage

Massively parallel NGS technologies require several initial sample preparation steps before sequencing.

Table 1. Glossary of Common NGS Bioinformatics Terms

Term	Definition
Alignment or mapping	The bioinformatics process of mapping sequencing reads to a reference genome.
Barcode	Short unique oligonucleotide sequence that is used in multiplexing to uniquely label DNA fragments from a specific sample. These barcode sequences can then be used to demultiplex sequencing output from the instrument.
Base quality	The Phred-scaled confidence that the output base in a given read reflects the true nucleotide status of the sequenced fragment.
cDNA	cDNA is produced from input RNA as the final library for RNA sequencing.
cfDNA	cfDNA, typically in reference to DNA fragments circulating in the bloodstream.
CNV	CNV, sometimes referred to as CNA in the instance of somatically acquired mutations.
Coverage	The average sequencing depth of targeted genetic bases. Sometimes also used as a synonym for depth. It is common to use "fold" to measure coverage. Fold = (mapped read count * read length) / total genome size. 10-fold is also called 10X.
Demultiplexing	The process of sorting sequencing reads and assigning them back to individual multiplexed samples
Depth	The number of sequencing reads overlapping a particular nucleotide position
Exome	The exome consists of all exons in the genome that can be transcribed into RNAs and comprises approximately 1% of the total human genome.
Frameshift mutation	An INDEL mutation that alters the ORF of a protein-coding gene.
Genotyping	The process of detecting genetic differences between individuals.
GRCh38 (hg38)	The latest version of the human reference genome.
INDEL	A relatively short (<10 kb) INDEL of nucleotide(s) in the genome
Insert	A fragment of DNA that is inserted between adapters as part of a DNA library
Library	A collection of DNA fragments that is prepared for sequencing.
Minor allele frequency	The population frequency of a heritable (i.e., germline) allele for the least frequent allele for a biallelic variant.
Multiplexing	The process of pooling multiple sample libraries together for sequencing in a single run using unique barcode labels.
Oligonucleotide	A short (approximately 10-25nt) sequence of DNA or RNA
ORF	The string of trinucleotide codons that can be translated into a protein.
Read	The oligonucleotide string and corresponding base qualities that are output by a sequencing instrument. A read may be single or have a corresponding paired read in paired-end sequencing.
Single- or paired-end	Refers to whether DNA inserts are sequenced from one or both ends in a sequencing experiment.
SNP	A heritable single-base change in the genome.
SV	Structural variations, which are large genomic rearrangements such as translocations and inversions.
Targeted sequencing or panel sequencing	A rapid and cost-effective way to identify known and novel variants in selected sets of genes (i.e., gene panel) or genomic regions
VAF	The prevalence of the variant allele at a given position in a given sample.
Variant calling	The process of identifying a change in the genome compared with some reference.

cDNA, complementary DNA; cfDNA, cell-free DNA; CNA, copy number alteration; CNV, copy number variant; INDEL, insertion or deletion; NGS, next-generation sequencing; ORF, open-reading frame; SNP, single-nucleotide polymorphism; SV, structural variant; VAF, variant allele frequency.

Although these steps do not directly involve bioinformatics per se, they may have downstream consequences on the bioinformatics algorithms used. First, nucleic acid extraction and purification must be performed on the input tissue sample to isolate the DNA to be sequenced. The amount of sample necessary for DNA extraction varies by sequencing application and tissue type. For solid tissue samples, methods of tissue acquisition and preservation (i.e., fresh frozen versus formalin-fixed, paraffin-embedded [FFPE]) are also relevant. Generally, a tissue volume of 8 mm³ from either preservation method is sufficient for most sequencing applications,^{4,5} where typical DNA input requirements range from 10 to 1000 ng. This extracted DNA is then evaluated for various characteristics, including quality, yield, and concentration, to ensure that it is adequate for sequencing. DNA from FFPE tissue is more susceptible to damage from the fixation and preservation processes compared with fresh-frozen tissue. Nevertheless, comparative studies have found good overall concordance in NGS output between these tissue preservation processes.⁶ We refer the interested reader to guidelines⁷ published by the College of American Pathologists for further discussion of specimen acquisition and processing considerations for molecular profiling.

In the instance of tumor sequencing, pathologic characteristics of the source sample are important, particularly some approximate estimates of tumor cell purity. This has implications on other sequencing experiment parameters for detecting somatic alterations, as lower tumor purity consequently leads to lower somatic mutation prevalence in the sample. Sufficient thresholds may vary by application and sequencing conditions, so pathology-based estimation of purity is a critical presequencing quality assurance step. Tumor purity itself may also be estimated from sequencing output using *in silico* approaches,⁸ which may be compared with pathologist estimates and leveraged to improve more complex bioinformatics analyses. Nevertheless, these estimates may be susceptible to error under varying conditions (e.g., high genomic instability) and should be interpreted with caution.

Library Preparation

Once DNA is extracted and isolated, it must be further processed to make it amenable to NGS, which is broadly referred to as “library preparation.”⁹ First, the input DNA is fractionated into smaller fragments suitable for sequencing, which may be performed either mechanically or by enzymatic reaction. Special short adapter sequences, or oligomers, are then ligated to each end of the DNA fragments, which in this context are now

referred to as inserts (i.e., the genetic sequence “inserted” between the adapters). Additional size selection typically follows this step to ensure uniform and appropriate insert size for the NGS application of interest and to reduce the presence of any adapter dimers. Finally, polymerase chain reaction (PCR)-based amplification is typically applied to increase the overall DNA concentration before sequencing. The result of these processing steps is referred to as the input library, which is then ready to be sequenced.

Targeted Sequencing and Multiplexing

In contrast to untargeted whole genome sequencing (WGS), often there is interest in only sequencing selected regions of the genome, such as targeted gene panels. The exome consists of the coding regions of all protein-coding genes (i.e., exons) and comprises approximately 1.5% of the total genome. Consequently, whole exome sequencing (WES) can be a highly efficient strategy for capturing potentially high-impact genetic variation for discovery-based research applications. Alternatively, when there is a large amount of prior biological knowledge about relevant genes of interest, gene panels can be highly efficient and often provide much deeper sequencing coverage. A natural limitation of targeted sequencing in general is that the regions outside the design are not characterized and their genetic content remains unknown. In addition, identification of other types of DNA alterations, such as structural variation, is often much more difficult.

Two main strategies for this targeted sequencing include hybridization capture and amplicon-based sequencing. Capture-based enrichment involves the use of designed oligonucleotide probes, also known as baits, that bind to complementary DNA sequences present in inserts from the sequencing library to enrich the DNA fragments of interest. In contrast, amplicon-based enrichment is based on the design of flanking PCR primer sequences that lead to specific genomic regions being amplified for sequencing.

For many applications, it is efficient and cost-effective to also pool multiple sample libraries together and sequence them all simultaneously. This process is known as multiplexing, which also requires some mode of preserving source sample identities during the sequencing experiment. This is achieved using additional small oligomers known as sample barcodes or indexes (typically 8–12 bases in length) that are also ligated to the inserts and are unique to the individual sample. The presence of the index sequences provides a mechanism for assigning raw sequencing output back to individual samples by demultiplexing.

NGS Technologies

There are a wide variety of specific NGS technologies currently available from multiple companies, and appropriate platforms are often dependent on the sample characteristics and ultimate analytical goals (Supplementary Table 1). For purposes of illustration, we go into greater detail for the Illumina sequencing-by-synthesis (SBS) technology (Illumina, San Diego, CA), as this is one of the most widely adopted NGS platforms and has broad applicability. The Illumina SBS process involves loading the prepared library onto a solid substrate, or flow cell, which is coated with small oligomers complementary to the adapter sequence used in library preparation. The physical design of the flow cell typically includes multiple lanes that can accommodate different sequencing experiments. Once libraries are loaded on to the flow cell, bridge PCR amplifies the bound DNA fragments, leading to clonal sequence clusters consisting of thousands of DNA fragment copies.

Illumina SBS chemistry adopts similar concepts of Sanger sequencing, which is based on random chain termination PCR (Fig. 1 [left]). In Sanger sequencing, the target DNA sequence is denatured and cooled to allow a designed primer attachment to a single DNA strand. DNA polymerase then extends the complementary strand of the template DNA by adding one deoxynucleotide (dNTP) at a time until completion. Characterizing the sequence itself is achieved by including a small amount of fluorescently labeled di-dNTPs (ddNTPs), which prevent DNA polymerase from further extending the complement strand. Multiple PCR cycles and random chain termination from the ddNTPs produce varying fragment lengths terminating at each nucleotide position, which can be separated out and read by gel electrophoresis and subsequent analysis.

NGS performed using Illumina SBS similarly uses fluorescently labeled ddNTPs to block further synthesis by DNA polymerase (Fig. 1 [right]). The cluster fluorescence intensities are then detected by the autofocus laser system, representing the initial base of the clonal DNA fragment. Nevertheless, in contrast to Sanger sequencing, the chain termination in SBS is reversible, permitting the controlled continuation of the single-base synthesis of the DNA template. The fluorescent tag and blocker are removed, and sequencing proceeds in a stepwise fashion in what are referred to as cycles, with the number of cycles corresponding to the number of bases sequenced in the fragment (typically 75–150 bases).

Sequencing may be performed as single-end or paired-end, which refers to whether one or both ends of the insert are sequenced. Paired-end sequencing

provides multiple substantial benefits compared with single-end sequencing, including improved read mapping accuracy, increased genomic coverage, and the potential ability to detect genomic rearrangements such as gene fusions. Thus, paired-end sequencing is the standard in most NGS applications. If there is interest in identifying complex genomic structural variation, the paired ends of larger insert sizes (e.g., 2–5 kilobases [kb]) may be sequenced (referred to as mate pair sequencing), although the library preparation and bioinformatics analyses vary considerably from the standard paired-end sequencing.

Because the flow cell has a fixed number of lanes, the overall throughput of NGS sequencing is controlled by the yield characteristics of the libraries, the degree of multiplexing, and the number of cycles. Often, this throughput is referenced with respect to expected sample coverage, which is the average number of unique reads that overlap a given target base nucleotide. This coverage number is often accompanied by “X” in technical documentation (e.g., 30X coverage). Coverage has important bioinformatics consequences, as a larger number of sequencing reads overlapping a given genomic position will result in higher sensitivity and specificity for genetic variation. In contrast, lower coverage can accommodate a greater number of unique samples or larger amount genomic content to be sequenced at a comparable cost. For example, Illumina recommends 30X to 50X coverage for WGS and 100X coverage for WES. Targeted gene panels may target much higher depths to improve confidence in variant calling, especially for somatic mutations (e.g., >500X).

Sequencing Output Data

The primary output of the Illumina sequencing instrument is the binary base call (BCL) image files, which are massive files that contain base calls and corresponding qualities on the basis of the cluster fluorescence intensities. The base calls (i.e., the nucleotides A, C, T, and G) and base qualities (Q-scores) contained in BCL files are demultiplexed and converted into DNA nucleotide sequences, or “reads,” and corresponding base quality strings, which are then saved to the structured plain-text FASTQ file format. BCL files are typically only stored temporarily, and most downstream analyses require FASTQ files as input; thus, FASTQ files are often considered to be the “raw” sequencing output data format.

Each read in a FASTQ file is represented by four lines, including “read identifier,” “nucleotide sequence,” “separator,” and “string of Q-scores” (Fig. 2). These base quality Q-scores are presented in terms of the Phred scale, which is a logarithmic mapping of base error

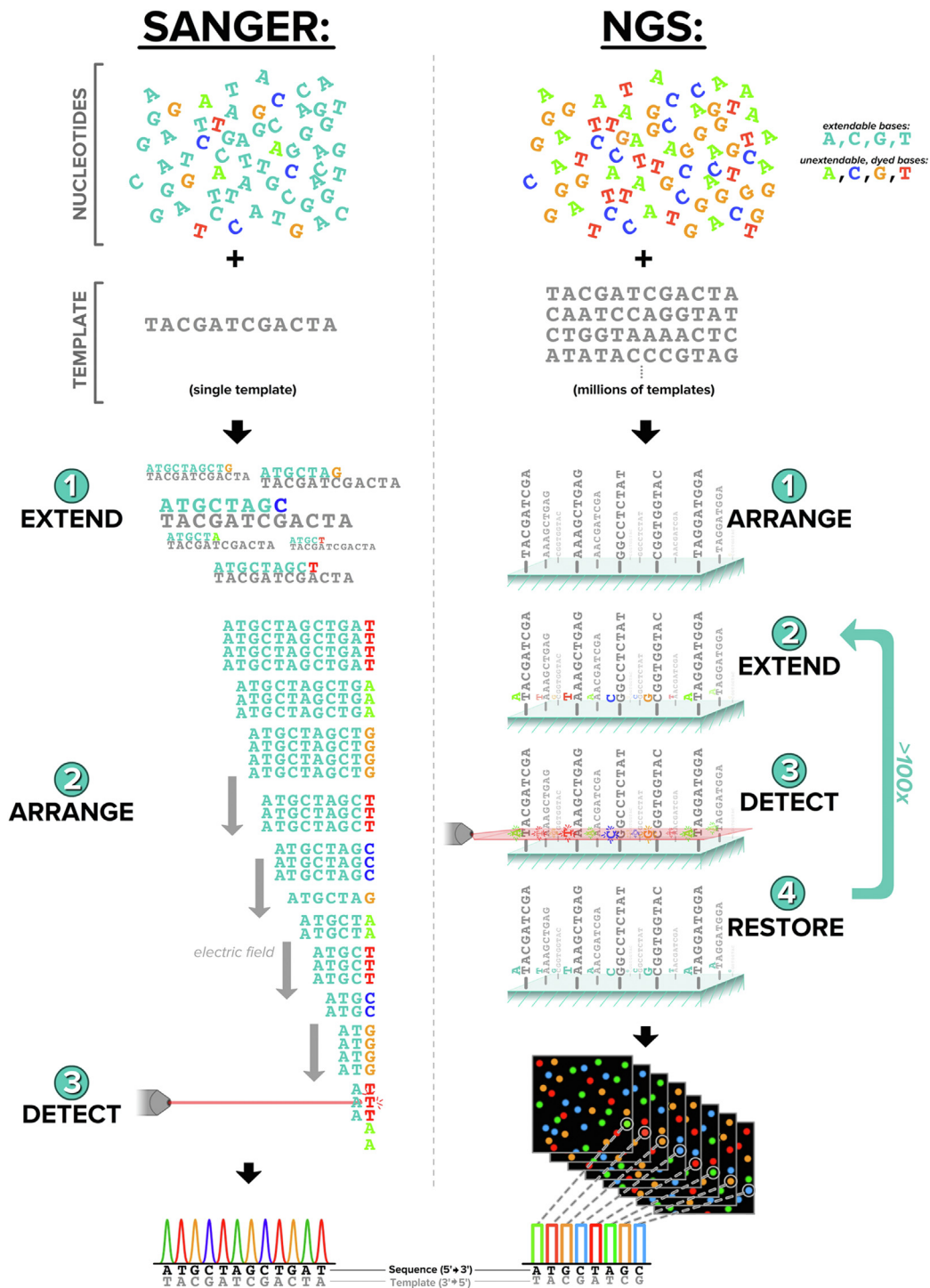


Figure 1. Comparison of traditional Sanger sequencing (left) versus next-generation sequencing (right). Both methods leverage fluorescently labeled ddNTPs for chain termination. Nevertheless, although Sanger sequencing uses subsequent size selection to characterize the sequence of a single template, NGS leverages reversible chain termination to characterize sequences one base at a time in sequential order for millions of templates. This image was reproduced from Figure 1 in Muzzey et al.¹⁶ licensed under Creative Commons Attribution 4.0 International License. ddNTP, dideoxynucleotide; NGS, next-generation sequencing.

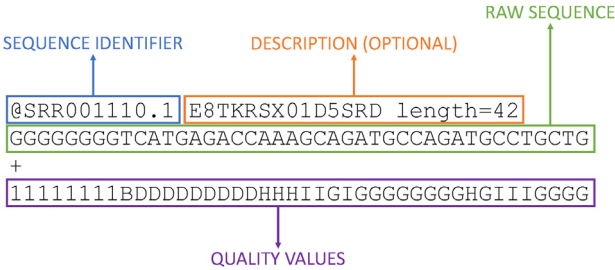


Figure 2. Example entry for sequencing read stored in a FASTQ file from platinum genome NA12878, illustrating the various components of the format. The FASTQ file was retrieved from NCBI sequence read archive (SRX000194). NCBI, National Center for Biotechnology Information.

probabilities. For a defined base error probability P , the Phred quality score Q is defined as

$$Q = -10 \cdot \log_{10}(P)$$

Phred quality scores can range from 0 to infinity, such that larger values indicate higher base call accuracy; for example, a Phred quality score of $Q = 30$ is equivalent to a base call accuracy of 99.9%. Phred scaling has since been widely adopted for other NGS bioinformatics quality metrics, such as read mapping and genotype quality. To make the FASTQ files more compact, each Q-score is encoded as a single character (instead of 2- or 3-digit numbers) with an ASCII code.

One FASTQ file is generated from single-end sequencing, and two FASTQ files are generated from paired-end sequencing. FASTQ files have become the standard format for storing NGS data, and most short reads aligners accept the FASTQ files as input. FASTQs can be readily analyzed and visualized using tools such as FastQC¹⁰ to evaluate the overall sequencing quality. See Table 2 for a summary of these and other most often encountered bioinformatics file types.

Sequencing Alignment

Routine sequencing-based analyses include the identification of genomic variants and the quantification of genomic features. Before performing these analyses, the sequencing reads in the FASTQ files need to be aligned to a reference genome, a process referred to as “read mapping.” This amounts to searching the reference genome for the most likely source sequence of a given read, while flexibly accommodating natural genetic variation and sequencing errors. The most recent version of the human reference genome is the Genome Reference Consortium Human Build 38 (GRCh38 or hg38), although many clinical laboratories still use the previous hg19 build.

Owing to the short length of sequencing reads in relation to the complete human genome, the probability

Table 2. Common Bioinformatics File Formats Along With Their File Extensions and Brief Descriptions

File Type	File Extension	Description
FASTA	.fa,.fasta	FASTA files contain text-based representation of sequence information. Typically used for reference sequence data storage (e.g., human reference genome).
FASTQ	.fastq,.fq	FASTQ files contain text-based representation of sequencing read information and corresponding base qualities. This is the typical raw output delivered from most NGS experiments.
SAM	.sam	SAM files contain tab-delimited information output from the read alignment process.
BAM	.bam	BAM files are a binary compressed version of corresponding SAM files and contain the exact same information in a smaller file size.
CRAM	.cram	CRAM is a reference-based compressed alignment file that leverages a given reference genome for additional file-size reduction.
VCF	.vcf	VCF files contain text-based genetic variant call data. They consist of a header with various metadata, along with 8 mandatory data columns. Each row corresponds to a unique variant, and VCF files can be either single- or multi-sample.
gVCF	.gvcf	Genomic VCF files are single-sample variant call files that also include information on same-as-reference regions of an individual sample. These are common intermediate files that are used to create multisample VCF files.
GTF and GFF	.gff,.gff2,.gff3,.gtf	GFF files are tab-delimited text files that typically are used for representing gene structure. GFF3 is the most current version of this format. GTF is highly similar to GFF files while also containing grouping information to accommodate gene-transcript identifier pairs.
MAF	.maf	MAF files are tab-delimited text files that list mutation information from a VCF file. These are often a filtered subset of variants identified through paired tumor-normal sequencing or putative functional impact.
BCL	.bcl	BCL files are the raw intensity files that are generated by Illumina sequencing instruments. These are demultiplexed and converted to FASTQ files for further bioinformatics processing.

BAM, binary alignment map; BCL, binary base call; GFF, general feature format; GTF, gene transfer format; MAF, mutation annotation format; NGS, next-generation sequencing; SAM, sequence alignment map; VCF, variant call format.

of inaccurate read alignment is nontrivial. To reduce the false-positive alignments, low-quality bases and exogenous sequences (such as sequencing adapters) need to be trimmed. Decoy sequences (e.g., mitochondrial and viral sequences that are integrated into the human genome) can also be added to the reference genome to “absorb” reads that do not truly originate from human chromosomes. Short-read alignment algorithms^{11,12} then efficiently map potentially hundreds of millions of reads to the reference genome.

The sequence alignment output is usually saved in plain-text SAM (Sequence Alignment Map) format¹³ or its binary version (BAM). Since first published in 2009, BAM has quickly become the most popular file format to store short-read alignments and is generally considered as the starting point for most NGS analysis tasks, such as variant calling and gene expression quantification. The BAM format has several advantages compared with SAM and other plain-text formats. First, it is compressed, making it convenient for transfer and storage. Second, it is line-oriented with all the alignment information of a read arranged into one row, making it easy to process. Third, the BAM file is indexed (creating a .bai file) and supports random access so that regional information can be “sliced” without loading the whole BAM file into memory. Finally, BAM files can be visualized using tools such as the Integrative Genomics Viewer¹⁴ or the University of California Santa Cruz Genome Browser (<https://genome.ucsc.edu/>), which can be helpful for manual qualitative assessment of a given variant call.

Another compressed form of SAM is CRAM, which is a reference-based compression file format where only differences between the sequencing reads and the reference genome are stored. CRAM is becoming increasingly popular for data storage, as it has all the advantages of BAM but is more compact in size (i.e., 50%–80% file size reduction). Nevertheless, some CRAM files use lossy compression and thus cannot be completely faithfully restored back to the original BAM source data.

Variant Calling

Variant calling is the process of detecting genetic differences between the aligned reads of a given sample and a corresponding reference genome sequence, and the respective algorithms are generally referred to as variant callers. The most common types of variants of interest are single-nucleotide polymorphism and single-nucleotide variants (SNPs and SNVs), short (<20 base pair) insertions and deletions (INDELs), and copy number variants and copy number alterations (CNVs and CNAs). SNPs specifically refer to single-base

substitutions (e.g., C changed to a T) in germline DNA that are most often observed in a given population. Most SNPs are biallelic, such that there are two of the four possible bases present in the population; although less common, multiallelic SNPs also exist where three or all four bases are represented, and a given subject may carry any combination of alleles. SNV is a more general term that can refer to any point mutation. SNVs and INDELs are often called together by the same bioinformatics algorithms and are sometimes collectively referred to as “short variants.”

A CNV is a type of variant where larger sections of the genome are amplified or deleted. Definitions have varied with respect to distinguishing CNVs from INDELs in terms of segment length, mechanism of alteration, and sequence content.¹⁵ Although the term CNV generically refers to changes in DNA copy number, CNA is more often applied in the context of acquired somatic copy number changes, particularly in cancer-based applications. CNVs and CNAs belong to a larger class of structural variants, which also includes inversions and translocations. Larger copy number events may include partial or whole chromosomal duplication or deletions.

The relationship between sequencing depth and variant call confidence is highly related to the variant allele frequency (VAF), defined as the proportion of DNA that harbors the variant allele in a given sample.¹⁶ Germline heterozygous variants correspond to an expected VAF of 50% under typical conditions and can often be confidently called at 20X to 30X coverage. Nevertheless, higher sequencing coverage is necessary to detect somatic variants, as VAFs tend to be lower than 50% owing to tumor tissue impurity. Similarly, subclonal variation that is only present in a subset of all tumor cells may be even more difficult to detect. To illustrate this concept in the context of a binomial sampling problem, consider a simple criterion of more than or equal to five variant-containing reads to be a sufficient evidence that a variant is present. Ignoring the additional complexity of sequencing error, the respective probabilities for the variant read support and different levels of coverage are presented for a range of VAFs in [Figure 3](#). We observe that even with 500X coverage, identifying evidence of a variant with VAF equals to 1% is not much higher than 50% under these conditions.

Preprocessing

BAM files generated by the short-read aligners are not directly usable for variant discovery, and some preprocessing is needed to prepare the BAM file for variant calling. First, duplicate reads (i.e., reads originated from the same original DNA fragments through some artifactual processes such as PCR and sequencing)

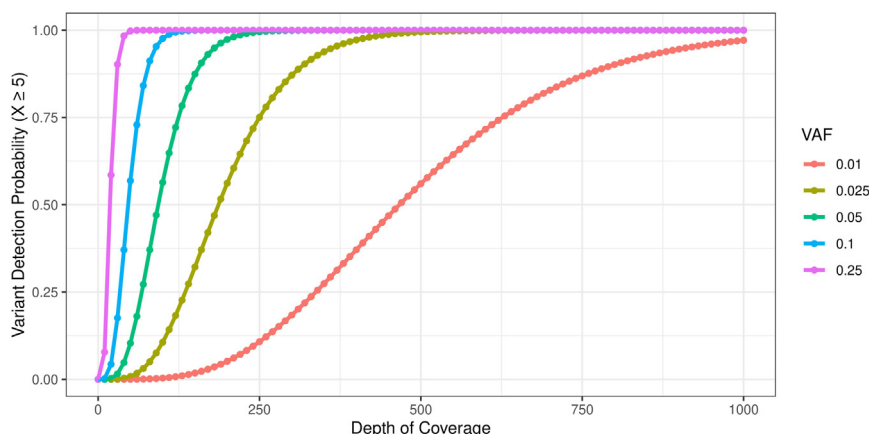


Figure 3. Illustration of relationship between sequencing depth and variant call confidence as a function of VAF. This simplified representation considers a variant to be detected under the criterion that at least five unique reads support the variant allele to be detected using a binomial probability model with success probability equals to the VAF. VAF, variant allele frequency.

need to be marked out. Marking out duplicates is necessary because they are nonindependent measurements of the original sequence, and variants (or sequencing errors) may be propagated to all the copies and influence VAF estimates. Deduplication is recommended in WGS or WES data analyses, but it is not recommended in PCR-based amplicon sequencing applications. Second, systematic bias or errors in the base quality scores need to be recalibrated. After these pre-processing steps, the BAM file will be ready for SNV, INDEL, and CNV discovery. The BAM file can also be used to assess sample contamination using tools such as VerifyBamID, which can use external variant calling data or population allele frequency information for quality assessment.

Germline Variant Calling

For germline variant calling of SNVs, basic bioinformatics utilities such as Samtools¹³ can be applied. More sophisticated variant callers, such as GATK HaplotypeCaller,¹⁷ also apply local realignment algorithms to improve variant calling in the genome regions of low complexity. For biallelic SNPs, there are two possible alleles: the reference allele (REF) defined in the corresponding reference genome and the alternate or variant allele (ALT). Note that these may differ from definitions of major and minor alleles, which relate to the prevalence of an allele in a given population (i.e., the major being more common than the minor). Because humans are diploid, nucleated cells carry two homologous copies of each chromosome, leading to three possible SNP genotypes: homozygous reference (REF, REF), heterozygous (REF, ALT), and homozygous alternate allele genotypes (ALT, ALT), respectively, corresponding to VAFs of 0%, 50%, and 100%. Variant genotypes are then

assigned a genotype quality (GQ), which is the genotype error probability and, similar to base quality Q-scores, is also Phred scaled.

Variant calling algorithms for CNVs using NGS data are usually based on sequencing coverage profiles, but they may also incorporate VAFs of overlapping variants. Identifying CNVs in targeted sequencing data, such as WES or gene panels, is complicated by breaks in coverage information and coverage irregularities induced by capture-based biases. This makes within-sample CNV detection difficult, and many algorithms for targeted NGS CNV detection rely on a reference set of normal samples for purposes of comparison.^{18,19}

Somatic Variant Calling

To identify acquired somatic variants in the tumor tissue, NGS data from the matched benign tissue (or blood) are highly useful, and many algorithms have been developed for paired tumor-normal sequencing. Somatic variant calling itself involves identifying all potential variants from the tumor tissue and then filtering out inherited germline variants using the matched normal sample as a reference. Nevertheless, corresponding normal samples are not always available and sequencing both tumor and normal samples for a given subject can be costly; in this situation, a “panel of normals” may be used to serve as a reference for germline variants and technical artifacts. In addition, some form of classifier is trained from respective somatic and germline variant databases to predict somatic status of variants detected from tumor tissues. Nevertheless, these algorithmic solutions for identifying somatic mutations are not without limitations, especially given the Euro-centric bias of many population-based allele frequency databases. Consequently, accuracy may be diminished for

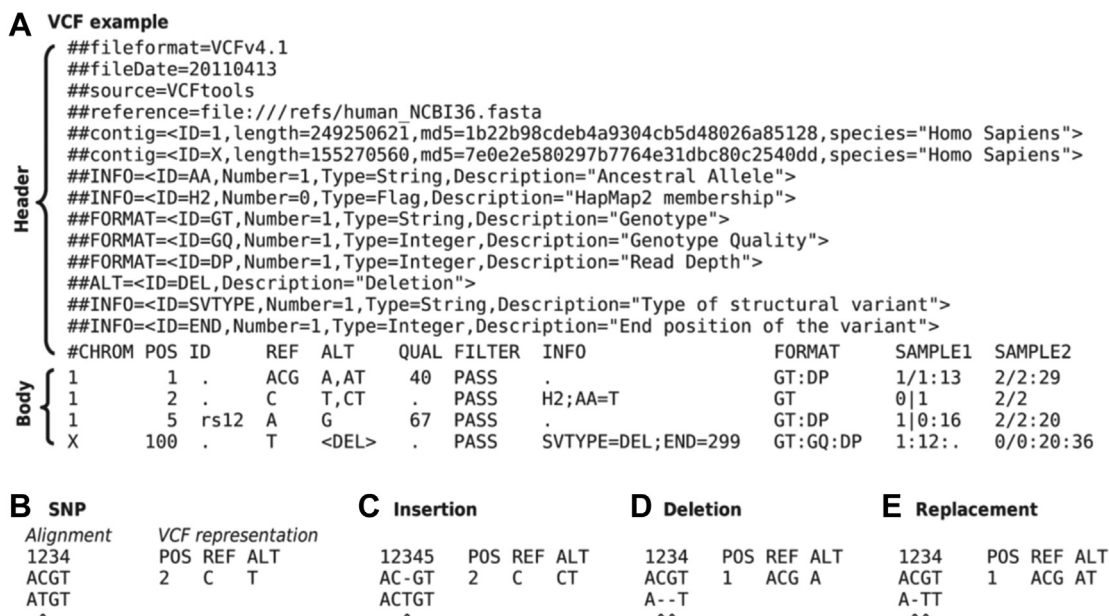


Figure 4. (A) Example of a valid VCF file with header and a few variant site records. The header includes multiple pieces of information relevant to the data set, including the file format, reference data, and details on format and annotation. The body includes variant records where rows indicate individual variants. (B-E) These illustrate representations of sequence alignments and corresponding VCF entries for various variant types. This figure is adapted from Figure 1 from Danecek et al.²¹ under the Creative Commons Attribution Non-Commercial License. VCF, variant call format.

underrepresented minorities where allele frequency data are more limited.

Because many variant callers have been developed with different algorithms, the choice of appropriate variant caller largely depends on the data type and the specific goals of the project. For further details on somatic variant calling algorithms, we refer the interested reader to a review by Xu.²⁰

Output Files

Called short-variant genotypes are typically stored in variant call format (VCF) files.²¹ Variants detected from one or more samples can be saved in a single VCF file. We illustrate the basic structure of a VCF file in Figure 4, including the overall header and body format (panel A) and representations of different variant types (panels B-E). Similar to the VCF, the single-sample genomic VCF (gVCF) file was developed to store both variant and nonvariant genomic regions. Because VCF files only contain variant positions relative to a reference, gVCF files permit the rapid merging of single-sample data with accurate same-as-reference genotype calls. To facilitate downstream analyses, VCF and gVCF files can also be indexed such as BAM files. The mutation annotation format (MAF) file is a tab-delimited text file developed by The Cancer Genome Atlas project to store somatic mutation data. MAF files are produced by aggregating mutation information from one or more VCF files generated from a project.

Quality Control

Although NGS methods have been found to be highly accurate compared with traditional Sanger sequencing, there are sources of technical artifacts that can lead to both false-positive and false-negative errors in variant calling. These include low coverage, base sequencing errors, and read misalignment. Postprocessing of variant call sets using tools such as GATK Variant Quality Score Recalibration can aid in germline variant filtering using various variant call characteristics along with highly validated variant resources. These quality control steps lead to a balancing of sensitivity and specificity, although modern bioinformatics germline variant calling pipelines tend to be highly accurate (F1 score > 0.99).²²

The process of somatic variant calling is more notably error prone than germline variant calling, and many factors influence the quality of variant calling (Supplementary Table 2). Thus, variant filtering is generally necessary before any downstream analysis, and individual variants that are clinically meaningful may be visualized using tools such as Integrative Genomics Viewer.

Downstream Analysis and Tumor Mutation Burden

Somatic mutation profiles may be used for various downstream analyses, including identification of significantly mutated genes, which are putative drivers of

cancer initiation, or calculating tumor mutation burden (TMB). TMB is broadly defined as the number of mutations per megabase (Mb) of DNA in a tumor, particularly in the context of WES. Theoretically, a higher TMB can result in a greater number of neoantigens, and therefore, it is used to predict the efficacy of immune checkpoint inhibitors. It has also been reported that higher non-synonymous TMB is associated with a better prognosis in patients with resected NSCLC.²³ Nevertheless, targeted gene panels can bias estimates of TMB relative to global WES-derived estimates on the basis of the limited content they capture (e.g., enrichment for likely driver genes), and industry sequencing vendors can differ dramatically in how they calculate this measure. This makes comparisons across studies challenging, and recent harmonization efforts have aimed to reduce this heterogeneity.²⁴

Variant Annotation and Interpretation

NGS can often lead to an overwhelming number of called variants, some of which may be of variants of unknown significance (VUSs), and it may not be immediately clear which variants are relevant to clinical conditions or to prioritize for follow-up study. Consequently, variant annotation has become an important bioinformatics process to aid in variant interpretation. For genetic variants in the protein-coding regions of genes, in silico prediction tools such as REVEL²⁵ have been developed to assign predicted functional impact of missense variants on the resultant protein structure. Similarly, variants in noncoding regions that are more likely to be regulatory in function may be annotated with scores from CADD,²⁶ FunSeq2,²⁷ and RegulomeDB²⁸ or overlapping epigenomic annotation from the ENCODE²⁹ and Roadmap Epigenomics³⁰ projects. Information can also be pulled from various external resources, including population allele frequencies from large sequencing databases (e.g., 1000 Genomes Project,³¹ gnomAD³²) and information from disease knowledge bases (e.g., ClinVar,³³ HGMD,³⁴ COSMIC,³⁵ OncoKB³⁶). Comprehensive functional annotation software packages such as ANNOVAR³⁷ can leverage a diverse array of external resources to append variant-level details to input VCF files in a rapid and high-throughput fashion.

In the context of cancer, it is common for tumor-only sequencing to be performed, which is cost-effective but has disadvantages than paired tumor-normal sequencing. Although variant calling itself is generally conducted in the same manner, the output is an unknown mixture of tumor and germline variation. Isolation of somatic mutations therefore requires that germline variation be inferred and filtered out, typically by a combination of variant characteristics (e.g., VAF)

and population allele frequency thresholds. These filtering approaches have tradeoffs with respect to sensitivity and specificity of true somatic alterations, and allele-frequency thresholds may be inaccurately applied for underrepresented populations in reference databases.³⁸ Similarly, TMB estimates can be biased in tumor-only sequencing experiments that leverage even highly sophisticated filtering strategies, as limited information can lead to an increase in false-positive somatic variants and inflate TMB estimates for underrepresented minorities.^{39,40}

Different Molecular Data Types

NGS technologies have rapidly extended from DNA sequencing to various other molecular types. Although the general steps described above still apply, the specifics of how each step is implemented can vary considerably depending on the biospecimen and molecular datatype of interest. We briefly describe some other common applications of NGS, highlighting a few of the relevant bioinformatics considerations.

RNA Sequencing

In addition to genomic variation captured by DNA NGS, gene expression profiling (either targeted or transcriptome wide) can also provide valuable information. In contrast to the genome, the transcriptome is highly dynamic and levels of expression are cell-type dependent and heavily influenced by in vivo conditions. The NGS platform can be similarly used to study the transcriptome by RNA sequencing (RNA-Seq). RNA-Seq is a powerful approach for identifying novel transcripts such as small regulatory RNAs or long noncoding RNAs, antisense transcripts, gene fusions, and aberrant splicing variants that may be implicated in tumor development and progression. When DNA is unavailable, RNA-seq data can also be used to identify genomic variants from the expressed coding regions using specialized variant callers, although this has limitations.⁴¹ RNA-based measurements have the potential for cancer diagnosis, prognosis, and therapeutic selection. For example, the *EML4-ALK* gene fusion was originally reported in a subset of NSCLC in 2007,⁴² and the ALK inhibitors crizotinib and ceritinib were approved by the Food and Drug Administration to treat ALK rearrangement-positive NSCLC in 2011⁴³ and 2015,⁴⁴ respectively.

RNA from the sample is first isolated. Because ribosomal RNA (rRNA) is the predominant form of cellular RNA found in most cells, additional steps such as poly-A selection or ribosomal RNA depletion are needed to remove rRNA and enrich mRNA. Recall that polyadenylation is a post-transcriptional RNA processing

step that adds a long chain of A nucleotides (100–250) to improve mRNA stability. This makes mRNA easily identifiable, and protocols that can select molecules on the basis of this long poly-A tail remove rRNAs, all smaller RNAs, and most of the long intergenic noncoding RNAs that have no polyadenylation signal. In contrast, ribosomal RNA depletion approaches selectively remove the rRNA molecules. Therefore, poly-A selection-based RNA sequencing is called mRNA-seq and rRNA depletion-based RNA sequencing is called total RNA-seq.

The RNA library is prepared by conversion of the single-stranded RNA to its complementary DNA followed by sequencing. The goals of the experiment will dictate the alignment strategy and bioinformatics algorithms to use for aligning the reads with the genome,⁴⁵ and specialized alignment algorithms are typically required.^{46–48} Alignment itself can be performed with respect to a genome reference or transcriptome reference. Alternatively, reads may be processed by reference-free de novo assembly, such that reads with overlapping content are aligned to each other without any prior information. Use of a genome reference enables discovery of novel transcripts but has the added task of correct identification of splice junctions, whereas use of a transcriptome reference leverages known splice junctions but does not allow the discovery of novel transcripts.

Expression quantification from the aligned RNA reads produces an output expression matrix containing counts for the number of reads (or fragments) observed for each gene or transcript. Although the abundance measure is relative rather than absolute, lower counts indicate lower levels of expression and higher counts indicate higher expression. Owing to the relative nature of the counts, normalization is required to remove experimental shifts.^{49,50} Differential expression between the study groups can then be assessed by statistical tools appropriate for count data.^{49,50}

Circulating DNA and Tumor Cells

An increasingly popular assay in cancer applications is the so-called liquid biopsy, which leverages the existence of either (1) circulating tumor cells (CTCs) to directly characterize the tumor genome or (2) cell-free DNA (cfDNA) in the bloodstream to detect and describe sequence characteristics of circulating tumor DNA (ctDNA). The motivation for liquid biopsies is predicated on tumor cells or DNA being released into the bloodstream during tumor growth or tumor cell damage, which has practical appeal owing to the asymptomatic cancer screening potential and noninvasive nature. The results from ctDNA analysis are promising for disease progression monitoring and guidance of targeted

therapies; for example, in detection of *EGFR* gene alterations or *ALK* rearrangements.^{51,52} Finally, there has been an increased interest in exosomes, which are small microvesicles released by cells that carry various proteins and RNA species. Recent studies have identified potential micro-RNA signatures related to diagnosis, prognosis, and treatment response, although efficient exosome isolation technologies are in an early stage of development.⁵³

A major challenge for NGS cfDNA analysis is the typically low proportion of ctDNA present in cfDNA, often yielding VAFs of 1% or lower. This substantially exacerbates issues with variant-calling confidence already discussed for low-purity tumor sequencing. The necessary sequencing depth for accurate mutation detection is generally orders of magnitude higher than other sequencing applications to avoid false negatives and false positives, with target coverages as high as 10,000X. Thus, smaller targeted gene panels (total genomic content < 300 kb) are typical for liquid biopsy assays designed to accurately identify somatic mutations.⁵⁴ Direct isolation of CTCs can enrich a given sample for tumor DNA; however, this requires accurate detection of CTCs and sufficient number of intact tumor cells in circulation. There are also notable limitations with respect to potential false positives from other somatic events, including clonal hematopoiesis,⁵⁵ which may inaccurately be treated as tumor-derived mutations. We refer the interested reader to Chen and Zhao⁵⁶ for a more detailed discussion of targeted cfDNA NGS techniques.

DNA Methylation Sequencing

DNA methylation is an epigenetic process where methyl groups are added to the fifth carbon of the cytosines forming 5-methylcytosine, and almost exclusively occurs in the sequence contexts of CG dinucleotides (CpGs). DNA methylation is one of key epigenetic mechanisms to silence gene expression and plays pivotal roles in tumorigenesis; for example, promoter hypermethylation of *MLH1* is frequently observed in NSCLC and associated with poor prognosis.^{57,58} During the library preparation, DNA is first subjected to a bisulfite treatment, during which unmethylated cytosines are converted to uracil whereas methylated cytosines are unchanged. Uracil is read as thymine during sequencing (called a C-T conversion), which has implications for the alignment process. As described in Sun et al.,⁵⁹ aligners generally differ in their strategy for handling the C-T conversion. After sequencing, DNA methylation for a given CpG is usually quantified as the percent of methylated cytosines out of all cytosines (cytosines + thymine), a value ranging from 0 to 1

referred to as a beta value. Differentially methylated CpGs or differentially methylated regions can be identified using the beta-binomial regression; incorporating CpG island information is advantageous as means of variable reduction.

Third-Generation DNA Sequencing

In contrast to the short-read sequencing of NGS, a new generation of sequencing methods is already beginning to mature. Sometimes referred to as “third-generation” sequencing, these methods aim to address one of the major limitations of NGS methodology—short-read length. Shorter reads are more difficult to align, particularly in repetitive regions of the genome, and phase information of genetic variants detected across reads is generally lost. Moreover, the reliance of NGS on PCR makes it difficult to characterize regions of high GC content bias. Technologies such as PacBio single-molecule, real-time sequencing (Pacific Biosystems, Menlo Park, CA) and ONT nanopore sequencing (Oxford Nanopore Technologies, Oxford, United Kingdom) can produce read lengths from 1 kb to more than 1 Mb. Current limitations of these long-read sequencing technologies relative to NGS include higher error rates, lower throughput, and overall cost, although these continue to improve over time.

NGS and Study Design Considerations

Factors to consider when planning a research study using NGS technology are myriad, including the type of specimen, data storage, cost, and specific aims. Hypotheses involving disease risk generally focus on germline DNA characteristics and use nontumor specimens, such as blood or buccal swabs. Hypotheses involving tumor molecular characteristics generally focus on somatic DNA and use tumor block specimens. As noted previously, NGS methods are available for fresh frozen or FFPE specimens for most molecular datatypes, though different performance characteristics are associated with each. A prospective study affords some control over sample purity, quality, and handling; in contrast, these are out of the investigators’ control in a retrospective study using banked specimens. NGS assays should also be conducted with biological effects of interest distributed throughout the assay run process to ensure that biological and experimental effects can be distinguished.

Output data files can often range from 100 to 560 gigabytes in size per specimen, depending on sequencing depth, length, and targeted versus the whole genome, and quickly generate terabytes of data for all specimens being studied. Transferring such large data sets requires specialized information technology expertise and makes cloud data storage a practical solution. Although the

production cost to sequence an entire genome has fallen dramatically and is currently typically less than \$1000 US dollar,⁶⁰ this generally does not reflect all expenses related to NGS-based research (e.g., data management, analytics, storage), and it remains costly to perform a large sequencing study.

Another study design consideration is the overall study sample size. Nevertheless, the study goals for which NGS can be used are so vast that it is not possible to provide an overview of power and sample size planning thoroughly. We highlight some considerations here. In general, the required sample size can be determined as a function of the hypothesis, expected differences, desired power and type I error rate, and variation in the data. Additional considerations in NGS experiments include sequencing depth, expected population minor allele frequency (for germline variation), and minimum VAF for somatic mutations.⁶¹ Owing to the sheer quantity of hypothesis tests performed with most NGS technologies, it is expected that a more strict significance criterion be used to penalize for performing multiple comparisons. Accepted multiple comparison strategies include use of Bonferroni correction and control of the false discovery rate (i.e., the expected proportion of false-positive findings in the set of genes declared significant), with the preferred strategy depending on the NGS assay and research objectives.

Publicly Available Resources

Research may also be augmented by (or completely conducted with) data previously generated by other sequencing studies. Investigators may apply for permission to use controlled access data from multiomic profiling initiatives, such as The Cancer Genome Atlas,^{62,63} or smaller individual studies deposited in the database of Genotypes and Phenotypes^{64,65} (dbGaP, <https://www.ncbi.nlm.nih.gov/gap>). Data in these repositories have enabled comprehensive molecular profiling studies and identification of potential therapeutic targets in various cancer types, including lung cancer.^{66,67} Processed data (such as mutation, CNA, RNA or protein abundance, DNA methylation) are available to the general public through the Genomic Data Commons^{68,69} (<https://gdc.cancer.gov/>), Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>), or cBioPortal⁷⁰ (<https://www.cbioportal.org/>). Similarly, large-scale genomic profiling projects with medical records linkage, such as the UK 100,000 Genomes,⁷¹ UK Biobank,²⁴ and the National Institutes of Health All of Us Research Program,⁷² also aim to empower genomic research through massive data sets. Direct access to individual-level data typically involves an application and review process along with institutional

commitments to data security or restriction to access by cloud-based data platforms.

Conclusions

Technological advancements in NGS have had a profound impact on basic and translational research, pharmaceutical and biotechnology development, and individualized patient care. In this review, we have discussed the basics of NGS data generation, storage, processing, annotation, and interpretation with the intent to equip the reader with general familiarity of these concepts.

CRedit Authorship Contribution Statement

Nicholas Bradley Larson, Ann L. Oberg, Ligu

Wang: Writing—original draft.

Alex A. Adjei: Writing—review and editing.

Acknowledgments

This work was supported by the National Cancer Institute (grant numbers U10CA180882, P50CA136393, P50CA102701, and P30CA15083).

Useful Web Links

Variant and Mutation Databases

COSMIC: <https://cancer.sanger.ac.uk/cosmic>

HGMD: <http://www.hgmd.cf.ac.uk/ac/index.php>

ClinVar: <https://www.ncbi.nlm.nih.gov/clinvar/>

gnomAD: <https://gnomad.broadinstitute.org/>

OncoKB: <https://www.oncokb.org/>

1000 Genomes: <https://www.internationalgenome.org/>

GWAS Catalog: <https://www.ebi.ac.uk/gwas/>

Variant Annotation Tools

SNPnexus: <https://www.snp-nexus.org/v4/>

SNPsnap: <https://data.broadinstitute.org/mpg/snp-snap/about.html>

VEP: <https://useast.ensembl.org/info/docs/tools/vep/index.html>

Variant Annotation Integrator: <https://genome.ucsc.edu/cgi-bin/hgVai>

GVS: <https://gvs.gs.washington.edu/GVS150/>

Data Repositories

NCI Genomic Data Commons (GDC): <https://gdc.cancer.gov/>

dbGaP: <https://www.ncbi.nlm.nih.gov/gap/>

EGA: <https://ega-archive.org/>

GEO: <https://www.ncbi.nlm.nih.gov/geo/>

SRA: <https://www.ncbi.nlm.nih.gov/sra/>

Data Browsers

cBioPortal: <https://www.cbioportal.org/>

NIH GDC Portal: <https://portal.gdc.cancer.gov/>

Integrative Genomics Viewer (IGV): <https://software.broadinstitute.org/software/igv/>

Supplementary Data

Note: To access the supplementary material accompanying this article, visit the online version of the *Journal of Thoracic Oncology* at www.jto.org and at <https://doi.org/10.1016/j.jtho.2022.11.006>.

References

1. Sanger F, Air GM, Barrell BG, et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*. 1977;265:687-695.
2. Shendure J, Porreca GJ, Reppas NB, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*. 2005;309:1728-1732.
3. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;437:376-380.
4. Austin MC, Smith C, Pritchard CC, Tait JF. DNA yield from tissue samples in surgical pathology and minimum tissue requirements for molecular testing. *Arch Pathol Lab Med*. 2016;140:130-133.
5. Cho M, Ahn S, Hong M, et al. Tissue recommendations for precision cancer therapy using next generation sequencing: a comprehensive single cancer center's experiences. *Oncotarget*. 2017;8:42478-42486.
6. Spencer DH, Sehn JK, Abel HJ, Watson MA, Pfeifer JD, Duncavage EJ. Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. *J Mol Diagn*. 2013;15:623-633.
7. Roy-Chowdhuri S, Dacic S, Ghofrani M, et al. Collection and handling of thoracic small biopsy and cytology specimens for ancillary studies: Guideline from the College of American Pathologists in Collaboration with the American College of Chest Physicians, Association for Molecular Pathology, American Society of Cytopathology, American Thoracic Society, Pulmonary Pathology Society, Papanicolaou Society of Cytopathology, Society of Interventional Radiology, and Society of Thoracic Radiology. *Arch Pathol Lab Med*. 2020;144(8):933-958. doi:10.5858/arpa.2020-0119-CP
8. Yadav VK, De S. An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Brief Bioinform*. 2015;16:232-241.
9. Head SR, Komori HK, LaMere SA, et al. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*. 2014;56:61-77.
10. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed 06/01/22.

11. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754-1760.
12. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357-359.
13. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25:2078-2079.
14. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29:24-26.
15. Pös O, Radvanszky J, Buglyó G, et al. DNA copy number variation: main characteristics, evolutionary significance, and pathological aspects. *Biomed J*. 2021;44:548-559.
16. Muzzey D, Evans EA, Lieber C. Understanding the basics of NGS: from mechanism to variant calling. *Curr Genet Med Rep*. 2015;3:158-165.
17. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491-498.
18. Sathirapongsasuti JF, Lee H, Horst BA, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*. 2011;27:2648-2654.
19. Straver R, Weiss MM, Waisfisz Q, Sistermans EA, MJT Reinders. WISExome: a within-sample comparison approach to detect copy number variations in whole exome sequencing data. *Eur J Hum Genet*. 2017;25:1354-1363.
20. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput Struct Biotechnol J*. 2018;16:15-24.
21. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156-2158.
22. Koboldt DC. Best practices for variant calling in clinical sequencing. *Genome Med*. 2020;12:91.
23. Devarakonda S, Rotolo F, Tsao MS, et al. Tumor mutation burden as a biomarker in resected non-small-cell lung cancer. *J Clin Oncol*. 2018;36:2995-3006.
24. Backman JD, Li AH, Marcketta A, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*. 2021;599:628-634.
25. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet*. 2016;99:877-885.
26. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019;47:D886-D894.
27. Fu Y, Liu Z, Lou S, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol*. 2014;15:480.
28. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22:1790-1797.
29. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57-74.
30. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518:317-330.
31. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526:68-74.
32. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581:434-443.
33. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46:D1062-D1067.
34. Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum Genet*. 2020;139:1197-1207.
35. Bamford S, Dawson E, Forbes S, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer*. 2004;91:355-358.
36. Chakravarty D, Gao J, Phillips SM, et al. OncoKB: a precision oncology knowledge base. *JCO Precis Oncol*. 2017;2017:PO:17.00011.
37. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164.
38. Garofalo A, Sholl L, Reardon B, et al. The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. *Genome Med*. 2016;8:79.
39. Asmann YW, Parikh K, Bergsagel PL, et al. Inflation of tumor mutation burden by tumor-only sequencing in under-represented groups. *NPJ Precis Oncol*. 2021;5:22.
40. Parikh K, Huether R, White K, et al. Tumor mutational burden from tumor-only sequencing compared with germline subtraction from paired tumor and normal specimens. *JAMA Netw Open*. 2020;3:e200202.
41. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet*. 2013;93:641-651.
42. Soda M, Choi YL, Enomoto M, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*. 2007;448:561-566.
43. Malik SM, Maher VE, Bijwaard KE, et al. U.S. Food and Drug Administration approval: crizotinib for treatment of advanced or metastatic non-small cell lung cancer that is anaplastic lymphoma kinase positive. *Clin Cancer Res*. 2014;20:2029-2034.
44. Khozin S, Blumenthal GM, Zhang L, et al. FDA approval: ceritinib for the treatment of metastatic anaplastic lymphoma kinase-positive non-small cell lung cancer. *Clin Cancer Res*. 2015;21:2436-2439.
45. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 2016;17:13.
46. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25:1105-1111.
47. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7:562-578.

48. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15-21.
49. Love MI, Anders S, Kim V, Huber W. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Res*. 2015;4:1070.
50. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*. 2012;13:204-216.
51. Vendrell JA, Mau-Them FT, Béganton B, Godreuil S, Coopman P, Solassol J. Circulating cell free tumor DNA detection as a routine tool for lung cancer patient management. *Int J Mol Sci*. 2017;18:264.
52. Rolfo C, Mack PC, Scagliotti GV, et al. Liquid biopsy for advanced non-small cell lung cancer (NSCLC): A statement paper from the IASLC. *J Thorac Oncol*. 2018;13:1248-1268.
53. Li W, Liu JB, Hou LK, et al. Liquid biopsy in lung cancer: significance in diagnostics, prediction, and treatment monitoring. *Mol Cancer*. 2022;21:25.
54. Christensen E, Nordentoft I, Vang S, et al. Optimized targeted sequencing of cell-free plasma DNA from bladder cancer patients. *Sci Rep*. 2018;8:1917.
55. Young SJ, Fuhlbrück F, Peterson M, et al. Clonal hematopoiesis in late-stage non-small-cell lung cancer and its impact on targeted panel next-generation sequencing. *JCO Precis Oncol*. 2020;4:1271-1279.
56. Chen M, Zhao H. Next-generation sequencing in liquid biopsy: cancer screening and early detection. *Hum Genomics*. 2019;13:34.
57. Safar AM, Spencer H 3rd, Su X, et al. Methylation profiling of archived non-small cell lung cancer: a promising prognostic system. *Clin Cancer Res*. 2005;11:4400-4405.
58. Seng TJ, Currey N, Cooper WA, et al. DLEC1 and MLH1 promoter methylation are associated with poor prognosis in non-small cell lung carcinoma. *Br J Cancer*. 2008;99:375-382.
59. Sun Z, Cunningham J, Slager S, Kocher JP. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics*. 2015;7:813-828.
60. NHGRI. The cost of sequencing a human genome. <https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>. Accessed October 11, 2021.
61. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher JP. Calculating sample size estimates for RNA sequencing data. *J Comput Biol*. 2013;20:970-978.
62. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*. 2015;19:A68-A77.
63. Wang Z, Jensen MA, Zenklusen JC. A practical guide to The Cancer Genome Atlas (TCGA). *Methods Mol Biol*. 2016;1418:111-141.
64. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;39:1181-1186.
65. Tryka KA, Hao L, Sturcke A, et al. NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Res*. 2014;42:D975-D979.
66. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489:519-525.
67. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511:543-550.
68. Heath AP, Ferretti V, Agrawal S, et al. The NCI genomic data commons. *Nat Genet*. 2021;53:257-262.
69. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data Commons as an engine for precision medicine. *Blood*. 2017;130:453-459.
70. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multi-dimensional cancer genomics data. *Cancer Discov*. 2012;2:401-404.
71. Peplow M. The 100,000 Genomes project. *BMJ*. 2016;353:i1757.
72. Murray J. The "All of Us" research program. *N Engl J Med*. 2019;381:1884.