Basic Principles of Bioinformatics for Next-Generation Sequencing Molecular Testing in Oncology

Simona De Summa and Stefania Tommasi

Contents

A. Russo et al. (eds.), *Practical Medical Oncology Textbook*, UNIPA Springer Series,
https://doi.org/10.1007/978-3-030-56051-5_17

Learning Objectives

By the end of the chapter, the reader will:

- Have learned the meaning of bioinformatic pipeline for next-generation sequencing and its key steps
- Have learned the differences among the most important output of a pipeline
- Have reached the knowledge of variant annotation and guidelines helpful for clinical reporting

17.1 Introduction

Cancer is a complex class of diseases affecting the genome. Thus, which is a better way to study it if not through the comprehension of DNA complexity? Revolutionary technological advances have been made since the completion of the first genome sequencing to date. High-throughput technologies pose many steps forward since the identification of tumor suppressor genes and oncogenes to the uncovering of the genomic landscape of many tumors. In particular, the advent of next-generation sequencing (NGS) platforms in the first decade of 2000 made possible to shed light in the taxonomy of cancers. Nevertheless, many questions arise from deeper knowledge deriving from these advances, starting from technical issues, e.g., depth/breadth of sequencing, to biological interpretation, e.g., how to distinguish variants with pathological significance from biological neutral ones, and ethical problems, e.g., management of incidental findings.

Precision oncology and genome-driven clinical trials [1] are the direct consequences of the introduction of NGS in the routine laboratory activity. Moreover, we are now able to exploit the tumor heterogeneity and acquired tumor resistance. However, we are still far from the real patient-tailored therapies [2]. To gain such a knowledge, the creation of consortia, e.g., The Cancer Genome Atlas, with the aim of data sharing and the creation of bioinformatic algorithms able to handle and integrate such amount of data, are mandatory.

17.2 A Brief History of Sequencing: From Sanger to Third-Generation Sequencing Platforms

A step forward in molecular biology was the development in 1983 of polymerase chain reaction (PCR) by Kary Mullis, awarded with Nobel Prize in Chemistry in 1993. Such a method, which seems to be very far from the present technologies, is still fundamental in the new sequencing platforms. Indeed, Sanger DNA sequencing, also known as chain terminator sequencing, developed in 1997, relies on PCR. It was automated through the introduction of capillary electrophoresis and was considered the gold standard until almost the first decade of 2000 [3]. In the meantime, human genome project was launched in 1990, and it requires 13 years to complete the first almost-complete sequence of human genome. However, different technological advances started to be implemented. During 1996, the first NGS platform was developed, and in 2004 it was commercialized: Roche 454. Thus, the possibility to fully sequence an individual's genome at the cost of $1000 dollars was not considered so utopistic [4]. Roche 454 was just the beginning because since then, new platforms continued to be implemented with different chemistries, lowering, by late 2015, the cost to obtain a high-quality human genome to $1500 dollars.

To date, two major companies, Illumina and Thermo Fisher, are the vendor of the most important NGS platforms. Both of them are short-read sequencer producing reads shorter than 300 bp. Both Illumina protocols and Thermo Fisher ion semiconductor sequencing (Ion Torrent) are cheap sequencing methods extensively used in clinical laboratory. The last-born sequencer from Qiagen also produces short reads: 100–150 bp length. Two new platforms are available only for research purposes, also known as third-generation sequencing platforms. They are able to produce long reads. The PacBio SMRT (single-molecule real time) technology could sequence reads longer than 2.5 Kb, while the Oxford Nanopore Technologies MinION, through the use of single-stranded pore technology, is able to sequence molecules >10 Kb.

Despite the possibility of sequencing the whole genome or the whole exome, the most used approach in molecular testing is targeted gene panels, including a discrete number of genes both as coding regions and hotspots, that are very small regions to detect a single-specific mutation. Gene panels are cost-effective and allow to obtain data with very high depth. Whole genome/exome sequencing are not routinely used in laboratories being time-consuming and with still elevated costs. Moreover, they pose ethical problems regarding incidental findings and their management.

17.3 From Wet-to-Dry Methods

17.3.1 NGS Intrinsic Errors

NGS technologies lead to the spread of bioinformatic efforts to appropriately analyze and manage data. Indeed, a clear separation between wet phase, namely the bench procedures, and data analysis exists. However, to be able to be appropriate in such a purpose, it is man-

datory to deeply know intrinsic errors related to sequencing methods. All NGS technologies primary consist of preparing a "library," which is the creation of a collection of small fragments of DNA which in turn will be sequenced. During library preparation, the fragments of DNA are linked to molecular barcodes to perform multiple sample sequencing, PCR primers and linkers which binds molecules to surface where molecules have to be sequenced. Then library have to be enriched for targeted sequencing (e.g., gene panels or whole exome sequencing). Enrichment could be performed through sequence capture which uses hybridization to complementary sequences (capture-based approach) or by PCR (amplicon-based approach). After these steps, sequencing could be performed. Illumina (e.g., HiSeq, MiSeq, NextSeq) and Ion Torrent (e.g., IonPGM, IonProton, S5) have different chemistries and thus biases. In detail, each DNA fragment is immobilized to a flow cell for Illumina and to a bead for Ion Torrent in order to clonally amplify each fragment. Sequencing by synthesis is the methods of Illumina sequencer, which uses fluorescently labelled reversible terminator-bound dNTPs. At each step, before to be washed way, the fluorophore bound to the added base is illuminated by a laser. The issue regards the similar emission spectra of fluorophores of A and C as well as G and T (red and green light, respectively, and separated by filters). Moreover, phasing (incomplete 3′ terminator removal due to erroneous enzyme kinetics) and pre-phasing (the skipping of incorporation of 3′ terminator caused by too fast synthesis) are further problems, which makes miscalls the type of error typical of Illumina platforms.

Ion Torrent chemistry is related to variation of pH due to H+ release after base incorporation, sensed by a solid-state pH sensor. When a stretch of homopolymers has to be sequenced, it was observed that AA stretch corresponds to a twofold increase in the pH with respect to single A. AAA stretch reaches only 1.5-fold increase of AA stretch and so on. Thus, reduction of increase of pH changes with the increase of the number of bases in homopolymers stretch results in incorrectly called homopolymer regions.

17.3.2 Alignment and Coverage Evaluation

Notwithstanding all the issues depicted above, sequencing run ended and data have to be correctly analyzed (Fig. 17.1).

At the end of sequencing, raw data, namely the DNA short fragment amplified, are in the format of FASTQ files, which include not only the sequences of the fragments (reads) but also quality scores of each base. They are used to check quality of FASTQ files (e.g., FASTQC),

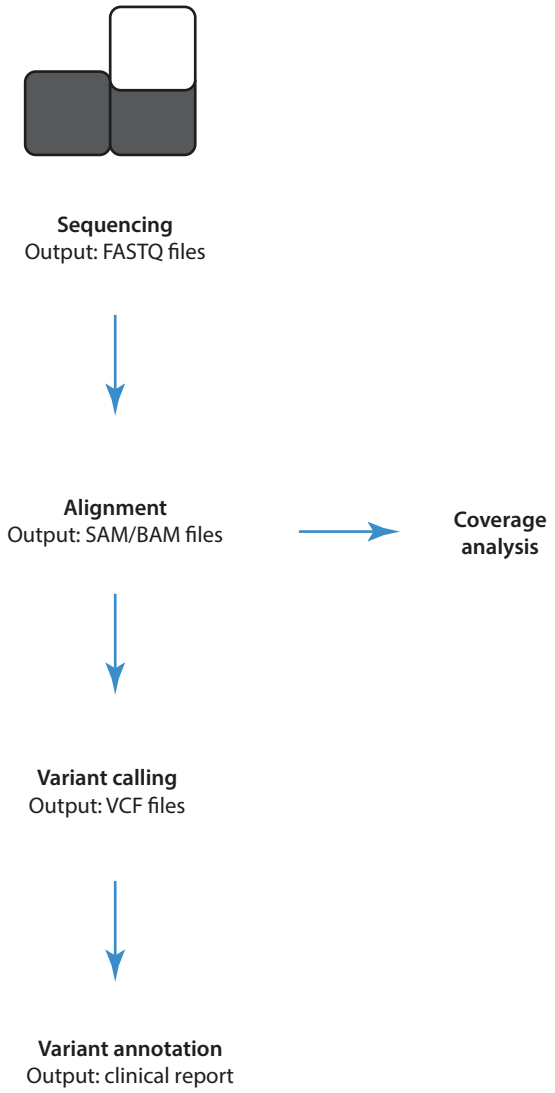

Fig. 17.1 Description of a typical bioinformatic pipeline for next-generation sequencing variant calling

which in turn could be trimmed to maintain only high-quality bases (e.g., Trimmomatic, CutAdapt). Trimming is a bioinformatic step which allows to cut low-quality bases.

A typical bioinformatic pipeline (namely, the series of steps to perform a bioinformatic analysis) includes the alignment of reads against a reference genome whose version has to be always specified to contextualize the genomic coordinate (e.g., hg19/Grch37). Different algorithms are used to perform this step. BWA [5] and Bowtie [6] are considered the best algorithms to manage short reads coming from Illumina platform. Ion Torrent has developed a "proprietary" aligner, TMAP, which is able to perform alignment for reads including also information of the flows that are the pH changes due to the incorporation of a specific base. The output is a SAM or BAM file.

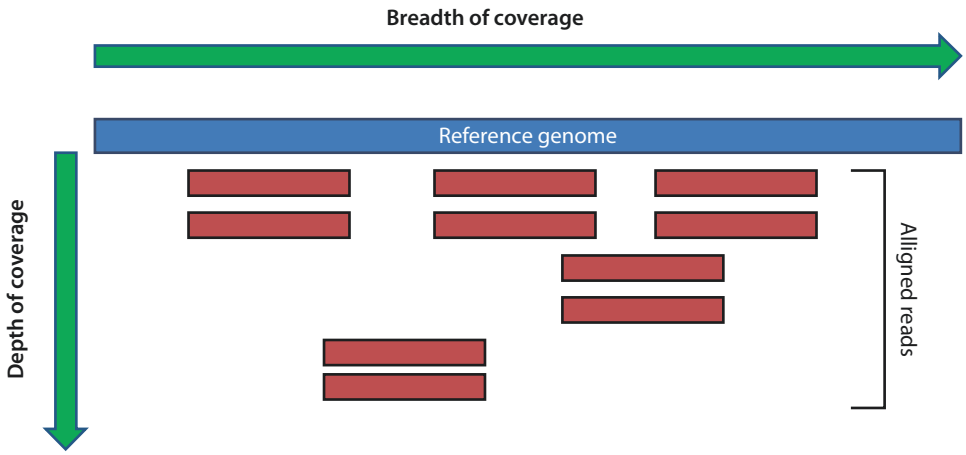

Fig. 17.2 Representation of the concept of breadth and depth of coverage

To evaluate the performance of a sequencing run to be confident on results, coverage should be checked. The term "coverage" often is misinterpreted. It is important to be able to distinguish two aspects: per-base coverage and breadth of coverage (Fig. 17.2).

Their definitions, as reported in ▶ http://www.metagenomics.wiki/, are:

"Per-base coverage is the average number of times a base of a genome is sequenced. The coverage depth of a genome is calculated as the number of bases of all short reads that match a genome divided by the length of this genome. It is often expressed as 1X, 2X, 3X,... (1, 2, or, 3 times coverage)."

"Breadth of coverage is the percentage of bases of a reference genome that are covered with a certain depth. For example: 90% of a genome is covered at 1X depth; and still 70% is covered at 5X depth."

Practically, in clinical reports coverage could be reported as average indicating percentage of targeted bases covered over the cutoff (e.g., average coverage panel of 2.5X with 99% of targeted bases covered >200X). For germline mutations, a coverage of 80X could be sufficient to confidently call variants; somatic alterations, often present at subclonal level, require higher coverage (at least 500X).

17.3.3 Variant Calling

The crucial step is variant calling, which is the identification of DNA alterations. Ion Torrent has an integrated plugin to call variants (Torrent Variant Caller). Regarding Illumina platforms, many variant caller algorithms have been implemented, as GATK HaplotypeCaller or VarScan2, each of them with different performances and with tunable options to gain confidence in variant calling process. In oncology testing, somatic variants, the so-called actionable variants, have to be reported to clinicians. Somatic alteration calling could be performed by the "tumor-normal" pipelines (e.g., MuTect, Strelka), referring to algorithms which analyze tumor samples coupled to germline control. In such a way, confounding factors related to the noise present in the germline samples are used to handle variants identified in tumor sample. Results coming from this type of approach are more reliable in particular regarding specificity.

The output is a variant call format (VCF) file, which includes not only the genomic coordinates and the type of called variants but also information about quality. In particular, variant read number, strand bias and variant allele frequency are important values to be taken into consideration when raw VCF variant filtering has to be performed to retain as much as possible true positive variants.

Variant reads are the number of reads supporting the presence of a variant. Generally, calls supported by fewer than five variant reads are typically considered to be likely false-positive calls.

Strand bias is a statistics measure of the deviation of the probability of a variant to be sequenced both on minus and plus strands. Higher values are associated with a probable sequencing artifact [7].

Variant allele frequency (VAF) is the number of reads linked to a variant divided by the overall coverage at the same locus. For germline testing, VAF is a measure of zygosity (50% VAF indicates heterozygous alterations, while 100% VAF is associated with homozygous alterations). For somatic testing, which is the most frequent in an oncology setting, VAFs are related to clonality that is the number of clones carrying a mutation. Somatic VAFs show a very high variability. For example,

in mutation related to resistance, e.g. EGFR T790M in non-small cell lung cancer patients, even very low VAF variants are reported in order to set a correct therapeutic approach.

Moreover, visual inspection of variants is allowed by the Integrative Genome Viewer [7], which works in a desktop-friendly manner (Fig. 17.3).

17.3.4 Variant Annotation

Filtered VCFs, containing as much as possible reliable variants, have to be annotated. Variant annotation, generally speaking, gives sense to the list of mutations present in VCF file in terms of biological impact in the transcription and translation of the gene. It is an impor-

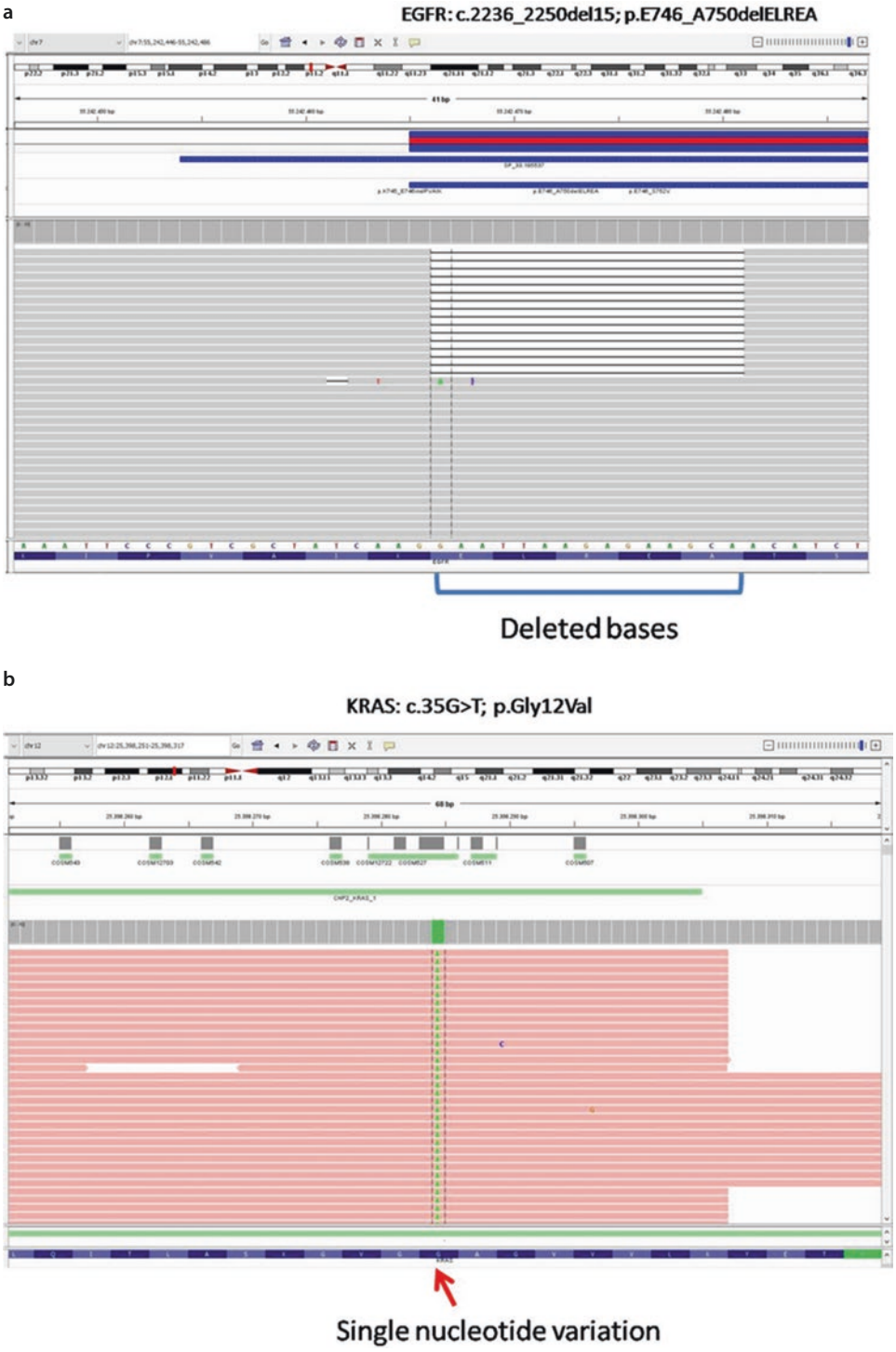

Fig. 17.3 Visual inspection of (**a**) a pathogenic deletion in EGFR gene indicating responsiveness to tyrosine kinase inhibitors in non-small cell lung cancer patients and (**b**) a pathogenic single-nucleotide variation in KRAS for patients affected by colorectal cancer which could benefit of targeted therapy

Table 17.1 Important databases used in variant annotation

Annotation databases		
Prediction databases	dbNSFP	► http://varianttools.sourceforge.net/Annotation/dbNSFP
Population databases	ExAC 1000genomes Exome sequencing project	► http://exac.broadinstitute.org/ ► http://www.internationalgenome.org/ ► http://evs.gs.washington.edu/EVS/
Oncology databases	COSMIC TCGA My cancer genome	► https://cancer.sanger.ac.uk/cosmic ► https://portal.gdc.cancer.gov/ ► https://www.mycancergenome.org/

tant step to filter germline variants and to retain only somatic ones when it is required to set a therapeutic strategy (e.g., KRAS alteration in codons 12, 13 and 61 in colon cancer; BRAF V600 alteration in melanoma, etc.) or for diagnostic or prognostic purpose. Indeed, in clinical setting tumor-normal pipelines, considering tumor and healthy genetic cell assessment in each individuous, generally could not be applied due to the lack of blood specimens. In detail, this step involves the use of several databases (Table 17.1):

- Database helpful in the prediction of deleteriousness of variants, including several in silico algorithms, e.g., SIFT and Polyphen. These tools allow to predict pathogenicity of a variants through the analysis of conserved amino acids in homologous proteins.
- Genetic population databases, reporting allele frequency of variants detected in general population or in specific-population, e.g., Caucasian. Population databases reports allele frequencies of alternative alleles in healthy individuals. In such a way, it could be possible to infer a biological impact because low frequencies could be associated to a pathology;
- Somatic databases, reporting allele frequency and, eventually, pathogenicity of cancer alterations. In such a way, it could be possible to know the penetrance of a somatic alteration in the onset of a malignancy.

17.3.5 CNV Detection

Detection of copy number variations (CNVs) is a clinical need for some malignancies (e.g., HERB2 amplification in breast cancer). NGS allows to detect CNVs, even if it is still challenging for amplicon-based panels. CNV calling requires algorithms different from tools used for variant calling. Generally laboratory confirms results through an alternative wet (e.g., MLPA) or bioinformatic method. Three main classes of method are at the basis of the algorithms.

- Depth of coverage method: bioinformatic tools detect increase or decrease of coverage in genomic region. The miscalling is due to the nonuniformity of coverage between samples or runs. The advantage is the possibility to identify large CNV (e.g., using EXCAVATOR2 tool [8]).
- Read pair analysis: this method requires paired-end sequencing and can detect only small CNV (e.g., using BreakDancer tool [9]).
- Split read: similarly, to read pair analysis, this analysis requires paired-end sequencing, but it is also able to detect breakpoint because it uses reads failing or partially failing to map (e.g., using Pindel tool [10]).

Ion Torrent platforms use a proprietary algorithm. The core of such a method is the creation of Variability Correction Informatics Baseline, including at least 48 samples. The baseline allows to perform correction on log2 ratio of amplicons. Moreover, baseline includes information about sex of samples (important for X chromosome because only a copy is present in male subjects) and tumor cellularity.

17.4 Liquid Biopsy

In 1869, the first evidence of circulating tumor cells in the blood of metastatic patients has been provided by Thomas Ashworth. Circulating tumor cells and cell-free DNA could be analyzed in all liquid compartment of the body (e.g., blood serum and plasma, urine, liquor, sputum, etc.). For diagnostic purpose, plasma is still the most used [11]. The concept underlying liquid biopsy is the monitoring of disease (e.g., minimal residual dis-

ease) and of the response to treatment (e.g., detection of resistance mutation T790M in EGFR gene to tyrosine kinase inhibitors in NSCLC patients) in a cost-effective and noninvasive fashion.

Cell-free DNA could be detected in many body fluids as a consequence of release from dying cells and circulating tumor DNA (ctDNA) is a part of the total amount, spanning from 0.01% to 90% in relation to stage of disease, tumor burden, and vascularity [12]. ctDNA could be deep sequenced with bias introduced during library preparation (e.g., 8-oxoG pairing with adenine and not citosine) and sequencing with 0.1–1% of miscalling depending of the platform used for NGS. Bioinformatic analyses are responsible in particular for false-positive calling in repetitive sequences, but the development of appropriate tools is overcoming such a problem.

The major issue is the very low allele frequency of alterations to be detected from experimental noise.

The bioinformatic pipelines are similar to those illustrated above, but some steps are performed by algorithms optimized for ctDNA (AfterQC [13], MrBam (► https://github.com/OpenGene/MrBam) and MutScan (► https://github.com/OpenGene/MutScan)).

In detail:

- AfterQC allows a better preprocessing of FASTQ files.
- MrBam improve supporting read number counting for mutations.
- MutScan is a visualization tool for interactive analysis.

Molecular barcoding sequencing [14] and CAPP-Seq [15] methods improve variant identification in ctDNA. Molecular barcodes (Unique Identifiers, UID, or Unique Molecular Identifier, UMI) are strings of complete random nucleotides, ligated to templates through ligation or through primers during PCR. Data analysis could be summarized into three steps:

1. UID extraction: the advantage of molecular tagging is the introduction of a fixed short sequence (five to seven nucleotides) between UID and DNA sequence, avoiding issues related to synthesis errors which could be responsible for alterations in the length of the barcode (FASTQ files).
2. Clustering of the reads with the same UID from BAM files.
3. Generation of a consensus read for each cluster and scoring of each position to call mutations.

CAPP-Seq is another approach for the detection of alterations in ctDNA. Basically, it is based on the definition of a "selector" from bioinformatic analyses of publicly available data to determine the most frequent mutations, ranked by their recurrence in samples. Selector is used to design biotinylated probes to reduce library to the region of interest. Then, variant calling is performed through different statistical approaches against the background of other ctDNA mutations through Bonferroni-adjusted Z-test.

17.5 Bioinformatic Pipeline Validation

Validation of an NGS process is critical because it involves both the wet methods and the subsequent bioinformatic analyses. Validation of wet procedures could be performed through the use of other laboratory techniques (e.g., Sanger sequencing, fluorescent-based method, or droplet digital PCR) or by samples with known genotype. Institutions as the National Institute of Standards and Technology (NIST) are able to certify reference standards. Their main feature is commutability, which is the "ability of a reference standard to perform comparably to treated samples" [16] in library preparation, sequencing, and analysis (e.g., FFPE samples could reduce commutability of a reference). In one established reference standards, the uncertainty could be established from the differences from the expected and the observed values. NA12878, that is the genome of a healthy European female, is one of the most used reference standard in many clinical laboratories. In detail, reference standards offer a "truth set" to evaluate NGS performance in terms of sensitivity, specificity, accuracy, and precision. Bioinformatic analyses are also a complex step to be evaluated and validated, and they require also a "ground truth." Generally, in silico datasets are generated through several available tools. FASTQ or BAM files could be easily created, providing not only datasets with known genotype but also with profile error of the used platform, and some algorithms could simulate heterogeneity in tumor samples. Of note, simulated datasets are able to validate bioinformatic step, but they could not replace the complexity of real samples and could not control wet phase of NGS testing.

17.6 Variant Interpretation and Clinical Reporting of Bioinformatic-Related Information

Clinical interpretation of variants is the most important step in the workflow of a molecular testing, even if it could be time-consuming due to the difficulty in its automation. Due to the large use of multigene targeted

sequencing panels, many variants could be detected and the reporting of results is not so simple. Variants need to be prioritized and logically interpreted in a clinical sense. For instance, it could be possible to identify variants which could be targeted by a drug not specific for the malignancy under evaluation or mutations whose consequences do not fit with the mechanism of action of a drug.

Regarding germline variants, the American College of Medical Genetics and Genomics, the Association of Molecular Pathologists, and the College of American Pathologists [16] wrote guidelines to assign clinical relevance of variants combining different approaches. Criteria include minor allele frequency reported in databases, frequency in affected individuals, prediction of the effect of the mutations, segregation, and inheritance information. Population-specific minor allele frequency (e.g., European-specific minor allele frequency) is another important factor to be taken into consideration. Minor allele frequency is the frequency observed in healthy population of the alternative allele. Population database reports data from almost 12,000 individuals, generally not affected by severe diseases; thus rare variants have great probability to have been detected and then reported. It is clear that without automation, following these recommendations could be influenced by operators. Indeed, it has been measured that the application of these guidelines to the same group of alterations reached 71% of consensus between different laboratories.

Similarly, in 2017 guidelines for somatic alterations have been drafted by American Society of Clinical Oncology, Association for Molecular Pathology, and College of American Pathologists (ASCO/AMP/CAP) [17]. They suggested to group variants into four categories based on four levels of evidence (◘ Table 17.2). Guidelines are helpful to better know a variant comparing them with knowledge-based databases, even if a deep know-how is requested to be able to manage this particular step. To date, many research groups and companies focused on the implementation of knowledge-based databases (e.g., OncoKB). Generally, they are developed by a group of specialists, from molecular biologists to clinicians, also known as curator, that "enrich" variants with information. Curators link to a variant several levels of information regarding the biology of the gene, the prediction of pathogenicity, and all detail regarding its involvement in prognosis and/or therapeutic approach.

ASCO/AMP/CAP guidelines recommend to report methodology details in the report, such as limit of detection and minimal coverage. Moreover, sequenced genomic regions (e.g., full gene or codon position) have to be clearly specified at the end of the clinical report.

More specific recommendations regarding bioinformatic analyses should be drafted, and, to date, many tools are available and under development. Thus, it would require more time to gain a consensus on bioinformatic algorithms.

17.7 Reproducibility in Bioinformatics

Given the acquired importance of bioinformatics in the last years, issues related to reproducibility regard also the analysis steps. The presence of several algorithms, each of them with different versions, to perform each part of a bioinformatic pipeline, changes in the library used to compile and install packages are responsible for such an issue. Moreover, it has been observed that only 10 papers out of 50 selected reported BWA parameters used to perform alignment, and only 11% of studies made available source code and data of simulated datasets.

Sandve [18] proposed ten rules for good practice in bioinformatic analyses (◘ Table 17.3). Many solutions are available to deal with reproducibility. For instance, Galaxy (▶ https://galaxyproject.org/) are cloud solutions that do not completely fulfill the suggested rules because pipelines are not customizable and privacy and ethical issues exist. To date the most promising approach is the container technology that is the virtualization, the so-called image of a bioinformatic pipeline. Softwares and dependencies are packed together avoiding problems related to versions and upgrading of operating system. To date, Docker (▶ http://www.docker.com) is considered the best environment to fit the rules of good practice. However, the use of such environment requires programming skills to be able to customize bioinformatic workflows. To make it easier, recently the Reproducible Bioinformatics Project (▶ http://reproducible-bioinformatics.org) has been proposed not only for the distribution of docker images but also for the implementation of a framework to build up pipelines fulfilling the ten rules.

Thus, many efforts to make reproducible bioinformatics are being made, and, at the very first instance, bioinformaticians have to be clear in the description of the workflow to allow other scientist to reproduce their results.

Table 17.2 Categories identified by ASCO/AMP/CAP guidelines and their levels of evidence useful for clinical report of variants

Categories	Evidence	Therapy	Diagnosis	Prognosis
Tier I: Variants of strong clinical significance	Level A	1. Biomarkers that predict response to FDA-approved treatments 2. Biomarkers included in professional guidelines that predict response or resistance to therapies for a specific type of tumor	Biomarkers included in professional guidelines as diagnostic for a specific type of tumor	Biomarkers included in professional guidelines as prognostic for a specific type of tumor
	Level B	Biomarkers that predict response or resistance to therapies for a specific type of tumor based on well-powered studies with consensus from experts in the field	Biomarkers of diagnostic significance for a specific type of tumor based on well-powered studies with consensus from experts in the field	Biomarkers of prognostic significance for a specific type of tumor based on well-powered studies with consensus from experts in the field
Tier II: Variants of potential clinical significance	Level C	1. Biomarkers that predict response or resistance to therapies approved by the FDA or professional societies for a different type of tumor 2. Biomarkers that serve as inclusion criteria for clinical trials 2. Biomarkers that serve as inclusion criteria for clinical trials	Biomarkers of diagnostic significance based on the results of multiple small studies	Biomarkers of prognostic significance based on the results of multiple small studies
	Level D	Biomarkers that show plausible therapeutic significance based on preclinical studies	Biomarkers that may assist disease diagnosis themselves or along with other biomarkers based on small studies or a few case reports	Biomarkers that may assist disease prognosis themselves or along with other biomarkers based on small studies or a few case reports
Tier III: Variants of unknown clinical significance	Not observed a significant allele frequency in population databases or pan-cancer/ tumor-specific databases			
Tier IV: Benign or likely benign variants	Observed at high allele frequency in population and no significant association with cancer			

Table 17.3 List of the ten rules suggested by Sandve et al. [18] for good practice in bioinformatic analyses

	Ten rules for reproducibility	
1	For every result keep track of how it was produced	Almost records of programs, version, and parameters to reproduce analyses
2	Avoid manual data manipulation steps	Manual manipulation if data is error-prone, thus it is preferable to use specific commands
3	Archive the exact versions of all external programs used	Different versions of the same program could not output the same results; thus version has to be recorded
4	Version controls all custom scripts	Workflow and modification of analysis steps have to be recorded
5	Record all intermediate results, when possible in standardized formats	Intermediate results allow to re-run an analysis step, and debugging could be simplified
6	For analyses that include randomness, note underlying random seeds	Some prediction analysis or simulations require the inclusion of a quote of causality. To be able to reproduce results, recording of the seed number is a very good practice
7	Always store raw data beyond plots	Plots summarize results, and data used to generate them have to be always stored
8	Generate hierarchical analysis output, allowing layers of increasing detail to be inspected	Plots or table is summarized results, but the presence of HTML links, for example, leading to data underlying results could be appropriate
9	Connect textual statements with underlying results	Statements result from an interpretation; thus to include link to analysis help reproducibility
10	Provide public access of scripts, runs, and results	It is a good practice to make available executables used to run an analysis workflow

17.8 Conclusions

NGS approaches posed a step forward into the deep knowledge of the human genome. The assessment of the presence of specific alterations is widely applied in the oncological clinical settings. The use of multigenic panel is both time- and cost-effective. Thus, the field of clinical bioinformatics is going to have a widespread diffusion. Data analysis is now considered the dry phase of an experimental protocol because of the importance to correctly tune parameters linked to sequencing data. A pipeline could be considered as validated not only when is the best "combination" when compared to "ground truth" but also when it is reproducible. In conclusion, there is an urgency to draw shared and unique good practices to grant "true" and reproducible results.

Key Points

- The knowledge of the intrinsic bias of the used NGS platform is the first step to perform a correct data analysis.
- Quality check of data, alignment, variant calling and variant annotation are the key steps of a variant calling pipeline.
- Pipeline has to be validated through "ground truth," which could be a simulated dataset or samples with known mutational status.
- Variant interpretation is the last step involving the use of specific databases and specific rules for clinical reporting.
- Reproducibility is an important issue in bioinformatics and there is an effort to grant it.

References

1. Hyman DM, Taylor BS, Baselga J. Implementing genome-driven oncology. Cell. 2017;168(4):584–99. https://doi.org/10.1016/j.cell.2016.12.015. Review. PubMed PMID: 28187282; PubMed Central PMCID: PMC5463457.
2. Russo A, Incorvaia L, Malapelle U, et al. The tumor-agnostic treatment for patients with solid tumors: a position paper on behalf of the AIOM-SIAPEC/IAP-SIBIOC-SIF italian scientific societies [published online ahead of print, 2021 Aug 6]. Crit Rev Oncol Hematol. 2021;103436. https://doi.org/10.1016/j.critrevonc.2021.103436.
3. Marziali A, Akeson M. New DNA sequencing methods. Annu Rev Biomed Eng. 2001;3:195–223. Review. PubMed PMID: 11447062.
4. Service RF. Gene sequencing. The race for the $1000 genome. Science. 2006;311(5767):1544–6. PubMed PMID: 16543431.
5. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. [PMID: 19451168].
6. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25. https://doi.org/10.1186/gb-2009-10-3-r25. Epub 2009 Mar 4. PubMed PMID:19261174; PubMed Central PMCID: PMC2690996.
7. Guo Y, Li J, Li CI, Long J, Samuels DC, Shyr Y. The effect of strand bias in illumina short-read sequencing data. BMC Genomics. 2012;13:666. https://doi.org/10.1186/1471-2164-13-666. PubMed PMID: 23176052; PubMed Central PMCID: PMC3532123.

8. D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A. Enhancedcopy number variants detection from whole-exome sequencing data using EXCAVATOR2. Nucleic Acids Res. 2016;44(20):e154. Epub 2016 Aug 9. PubMed PMID:27507884; PubMed Central PMCID: PMC5175347.
9. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER. Break dancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6(9):677–81. https://doi.org/10.1038/nmeth.1363. Epub 2009 Aug 9. PubMed PMID: 19668202; PubMed Central PMCID: PMC3661775.
10. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25(21):2865–71. https://doi.org/10.1093/bioinformatics/btp394. Epub 2009 June 26.
11. Russo A, Incorvaia L, Del Re M, et al. The molecular profiling of solid tumors by liquid biopsy: a position paper of the AIOM-SIAPEC-IAP-SIBioC-SIC-SIF Italian Scientific Societies [published online ahead of print, 2021 Jun 3]. ESMO Open. 2021;6(3):100164.https://doi.org/10.1016/j.esmoop.2021.100164.
12. Heitzer E, Ulz P, Geigl JB. Circulating tumor DNA as a liquid biopsy forcancer. Clin Chem. 2015;61(1):112–23. https://doi.org/10.1373/clinchem.2014.222679. Epub 2014 Nov 11. Review. PubMed PMID: 25388429.
13. Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J. AfterQC: automatic filtering,trimming, error removing and quality control for fastq data. BMC Bioinformatics. 2017;18(Suppl 3):80. https://doi.org/10.1186/s12859-017-1469-3. PubMed PMID:28361673; PubMed Central PMCID: PMC5374548.
14. Meldrum C, Doyle MA, Tothill RW. Next-generation sequencing for cancerdiagnostics: a practical perspective. Clin Biochem Rev. 2011;32(4):177–95. PubMed PMID: 22147957; PubMed Central PMCID: PMC3219767.
15. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, Liu CL, Neal JW, Wakelee HA, Merritt RE, Shrager JB, Loo BW Jr, Alizadeh AA, Diehn M. Anultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. 2014;20(5):548–54. https://doi.org/10.1038/nm.3519. Epub 2014 Apr 6. PubMed PMID: 24705333; PubMed Central PMCID: PMC4016134.
16. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. Nat Rev Genet. 2017;18(8):473–84. https://doi.org/10.1038/nrg.2017.44. Epub 2017 June 19. Review. PubMed PMID: 28626224. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, HegdeM, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–24. https://doi.org/10.1038/gim.2015.30. Epub 2015 Mar 5. PubMed PMID: 25741868; PubMed Central PMCID: PMC4544753.
17. Amendola LM, Jarvik GP, Leo MC, McLaughlin HM, Akkari Y, Amaral MD, Berg JS, Biswas S, Bowling KM, Conlin LK, Cooper GM, Dorschner MO, Dulik MC, Ghazani AA, Ghosh R, Green RC, Hart R, Horton C, Johnston JJ, Lebo MS, Milosavljevic A, Ou J, Pak CM, Patel RY, Punj S, Richards CS, Salama J, Strande NT, Yang Y, Plon SE, Biesecker LG, Rehm HL. Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. Am J Hum Genet. 2016;98(6):1067–76. https://doi.org/10.1016/j.ajhg.2016.03.024. Epub 2016 May 12. Erratum in: Am J Hum Genet. 2016;99(1):247. PubMed PMID: 27181684; PubMed Central PMCID: PMC4908185.
18. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013;9(10):e1003285. https://doi.org/10.1371/journal.pcbi.1003285. Epub 2013 Oct 24.