Next-Generation Sequencing in Clinical Oncology

Subjects: Oncology

Contributor: Simon Cabello-Aguilar , Julie A. Vendrell , Jérôme Solassol

Next-generation sequencing (NGS) has taken on major importance in clinical oncology practice. With the advent of targeted therapies capable of effectively targeting specific genomic alterations in cancer patients, the development of bioinformatics processes has become crucial. Thus, bioinformatics pipelines play an essential role not only in the detection and in identification of molecular alterations obtained from NGS data but also in the analysis and interpretation of variants, making it possible to transform raw sequencing data into meaningful and clinically useful information.

bioinformatics clinical oncology pipeline next-generation sequencing (NGS)

1. Introduction

Progress in next-generation sequencing (NGS), including an increase in its accessibility and cost effectiveness, has enabled comprehensive genetic testing in many cancer centers and transformed cancer treatment. In particular, NGS has permitted the advancement of precision oncology focused on identifying genetic changes in tumors that include single-nucleotide variants (SNVs), copy number variations (CNVs), small insertions and deletions (indels), structural variants (SVs), and microsatellite instability (MSI) [1][2]. Such valuable insights into the molecular characteristics of tumors provided by NGS have made it an essential tool for the diagnosis and treatment of cancer [3].

Robust and reliable bioinformatics pipelines able to organize, interpret, and accurately identify these molecular alterations from within sequencing datasets are crucial in the treatment decision-making process. The robustness ensures that the pipeline can handle variations in the data and produce consistent results, while the reproducibility ensures that the same results can be obtained when the pipeline is run multiple times. In addition, the comprehensive traceability and understanding of how the pipeline works ensure that others are able to reproduce the results. To this end, a well-designed and well-documented bioinformatics pipeline can provide reliable and accurate guidance for oncologists.

2. Workflow Management

In clinical oncology, the rapid evolution of high-throughput sequencing technologies has increased data generation, necessitating robust and efficient bioinformatic pipelines for analysis. Command-line tools [4][5] offer a flexible and efficient means to handle these data. These tools enable bioinformaticians to construct intricate pipelines that encompass various stages of analysis. The command-line interface, with its text-based interaction, allows for precise control over parameters, facilitating the customization and optimization of workflows to suit the specific requirements of clinical oncology research. However, command-line tools rely solely on text-based interfaces, requiring users to input commands in a terminal or console, while workflow management tools commonly provide users with a graphical or text-based interface to design workflows, offering a more visually intuitive experience. Workflow management tools [6] also ensure the automation and standardization of the bioinformatics process and allow the user to define the order, parameters, and input data for a sequence of tools. They directly take care of the correct execution and documentation of the intermediate steps. Several workflow managers are available, including Snakemake and Nextflow, among others [7][8][9][10][11]. Such systems help bioinformaticians save time, reduce errors, and ensure the accuracy and reliability of their analyses. In cancer genomics, a bioinformatics pipeline is executed by the workflow manager such as that described in **Figure 1** and comprises different steps: (i) quality control, (ii) adapter trimming, (iii) alignment, (iv) variant calling, (v) variant annotation, (vi) variant filtering, (vii) CNV calling, (viii) MSI status calling, and (ix) interface generation.

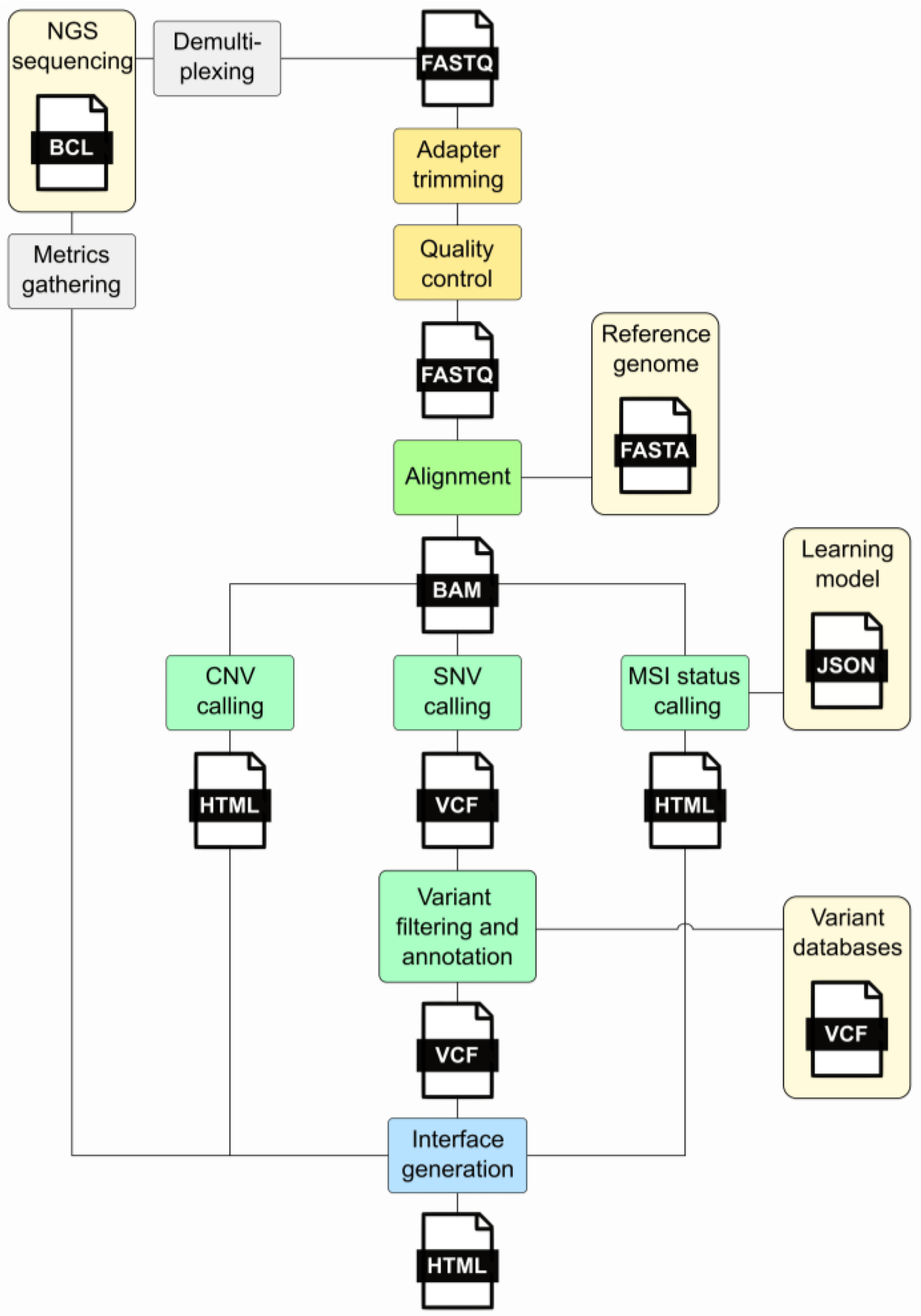


Figure 1. Major steps of an NGS bioinformatics pipeline. This diagram illustrates the processes forming the pipeline and the files generated during its execution. The gray segments denote processes that exist independently of the pipeline. Light yellow signifies external prerequisites, while yellow represents the initial pipeline stages

involving FastQ processing. The alignment stage is highlighted in green, while light green indicates the analyses conducted, encompassing SNV, CNV, and MSI status calling. The final step, interface generation, is illustrated in blue. Acronyms: FASTQ—a text-based file storing nucleotide sequences and corresponding quality scores; BAM—Binary Alignment Map; VCF—Variant Call Format; CNV—Copy Number Variation; SNV—Single-Nucleotide Variant; MSI—Micro Satellite Instability.

An up-to-date compilation of available tools for each step of the pipeline is provided in **Table 1**. It is important to mention that the Broad Institute provides a Genome Analysis Toolkit (GATK) [12], which contains a wide variety of tools designed for variant discovery and genotyping that covers the steps described in **Figure 1**. Moreover, the nf-core community project [13] has assembled a curated collection of analysis pipelines constructed with Nextflow including a somatic variant calling workflow, SAREK [14][15], available at "https://nf-co.re/sarek/3.4.0 (accessed on 1 December 2023)". Nf-core offers portable and reproducible analysis pipelines and the support of an active community.

Galaxy [16] and Taverna [17] are both noteworthy platforms in the field of bioinformatics analysis. Galaxy, as an open-source platform, offers a web-based interface for analyzing high-throughput genomics data, especially NGS data. It accommodates users with varying levels of bioinformatics expertise, allowing them to create, execute, and share workflows for diverse bioinformatics analyses. Featuring a user-friendly graphical interface, Galaxy is accessible to a broad audience, providing tools and workflows for tasks such as sequence alignment, variant calling, and various genomic analyses. The platform emphasizes reproducibility, enabling users to systematically save and share their analyses. Taverna serves as a distinct workflow management system designed for various scientific applications, including bioinformatics. It facilitates the design and execution of workflows, providing a flexible environment for scientific analysis and automation. Additionally, Tavaxy [18] shortens the workflow development cycle by incorporating workflow patterns to streamline the creation process. It facilitates the reuse and integration of existing (sub-) workflows from Taverna and Galaxy, while also supporting the creation of hybrid workflows.

Noteworthy, private solutions also exist. For example, the DRAGEN secondary analysis pipeline ensures all the steps from sequencing files to annotated and filtered genetic alterations. It was recently benchmarked, and the authors claim its value in a preprint that came out this year [19].

Table 1. List of commonly used bioinformatic tools.

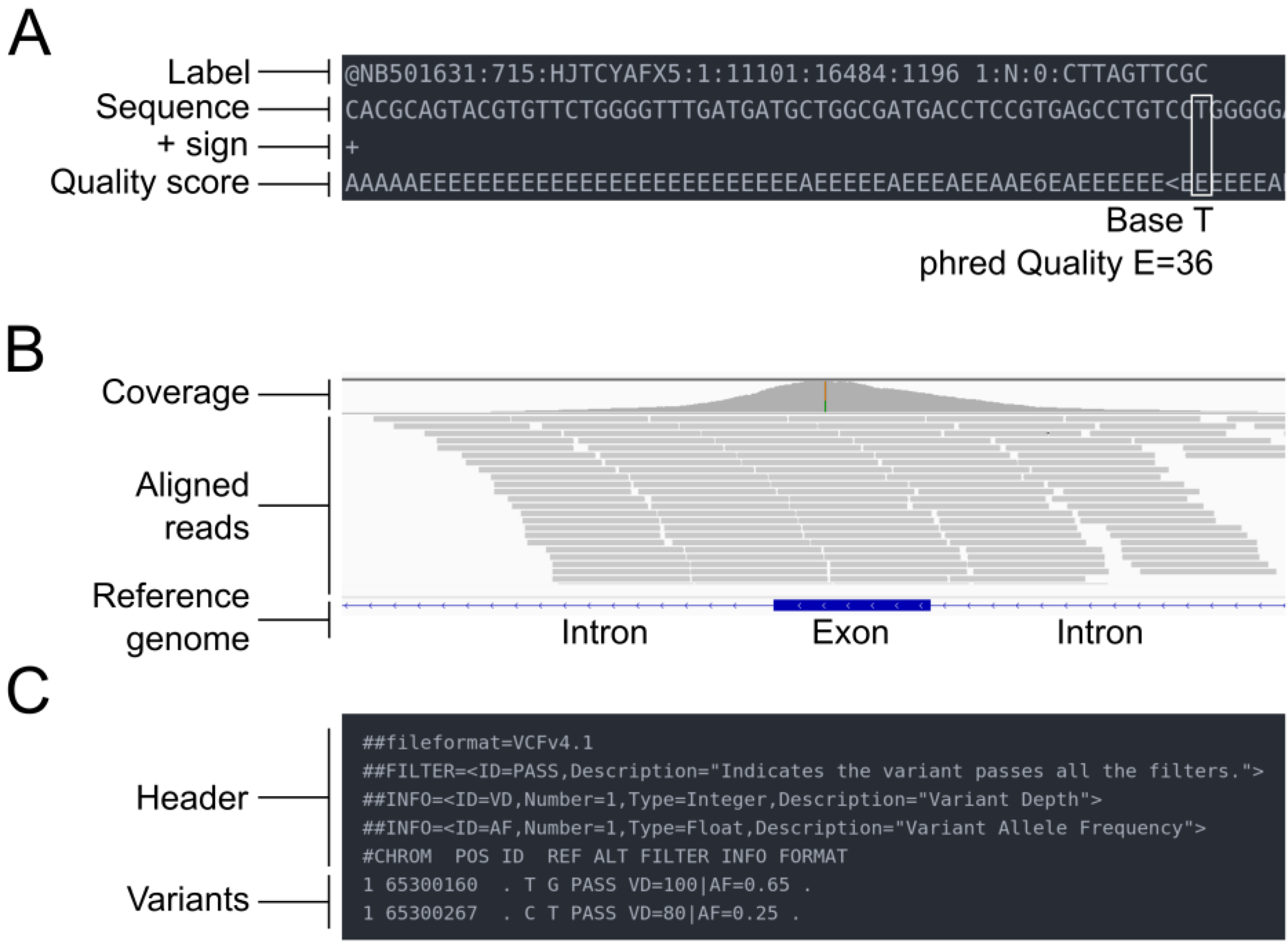
Process	Tools	References
Workflow managers	Nextflow, Snakemake	[7][8]
Quality control	fastp, FastQC *, Picard, MultiQC	[20][21][22][23]
Adapter trimming	fastp, trimmomatic, cutadapt *, BBDuk	[20][24][25][26]

Process	Tools	References
Reads alignment	BWA *, Bowtie, HISAT2, STAR	[27][28][29][30]
Variant calling	HaplotypeCaller, freebayes, mutect2, verdict *	[31][32][33][34]
Variant filtering	dbSNP, 1000G, GnomAD *	[35][36][37]
Variant annotation	VEP *, MobiDetails, ANNOVAR, SnpEff	[38][39][40][41]
CNV calling	CNV-LOF, CoverageMaster, CNV-RF, DeepCNV, CNV_IFTV, HBOS-CNV, CNV-MEANN, ControlFREEC, ifCNV *, mcna	[42][43][44][45][46][47] [48][49][50][51]
MSI status calling	MIAmS *, MSIsensor, MSIdetect, deltaMSI	[52][53][54][55]

* Used in in-house bioinformatics pipeline.

3.1. Quality Control

NGS sequencing produces binary base call sequence files (BCL) that are demultiplexed into FASTQ format sequencing files for each sample. The FASTQ format is a text-based format designed to store nucleotide sequences, along with their corresponding quality scores (**Figure 2**A). The initial stage of all bioinformatics pipelines is to assess the quality of the data. Indeed, sequence quality control is an essential step in the analysis of NGS data, which are generated in large volumes and can be prone to various types of errors, such as sequencing errors, adapter contamination, and sample cross-contamination. Sequence quality control aims to ensure that the sequencing data are accurate, reliable, and free from technical artifacts that could affect downstream analysis. It aims to identify low-quality bases, sequence bias, and over-representation of certain sequences. Quality assessment can be performed using tools such as fastp [20] or FastQC [21], a flexible and widely used tool for quality control, developed at the Babraham Institute to assess the quality of sequencing data in fastq files. This tool is robust, can be used on all operating systems, and offers both a graphical user interface and a command line interface. It is commonly incorporated by bioinformaticians as a quality control step in customized pipelines. The latest versions of FastQC include Picard [22], a tool developed by the Broad Institute that can manage SAM, BAM, and VCF files and perform quality control at different stages of the bioinformatics pipeline. An example of good and bad sequence quality profiles (i.e., the mean quality value across each base position in the read) obtained using FastQC is provided in **Figure 3**A. Moreover, MultiQC [23] consolidates data from various QC tools to create a cohesive report, complete with interactive plots, spanning multiple samples.



```
@NB501631:715:HJTCYAFX5:1:11101:16484:1196 1:N:0:CTTAGTTCGC
CACGCAGTACGTGTTCTGGGGTTTGATGATGCTGGCGATGACCTCCGTGAGCCTGTCCTGGGG
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEAEEEAEEAAE6EAEEEEEEE<EEEEEEA
```

```
##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="Indicates the variant passes all the filters.">
##INFO=<ID=VD,Number=1,Type=Integer,Description="Variant Depth">
##INFO=<ID=AF,Number=1,Type=Float,Description="Variant Allele Frequency">
#CHROM  POS ID  REF ALT FILTER INFO FORMAT
1 65300160  . T G PASS VD=100|AF=0.65 .
1 65300267  . C T PASS VD=80|AF=0.25 .
```

Figure 2. Overview of the different file types mentioned in the pipeline. (**A**) FASTQ file. (**B**) SAM/BAM file. (**C**) VCF file.

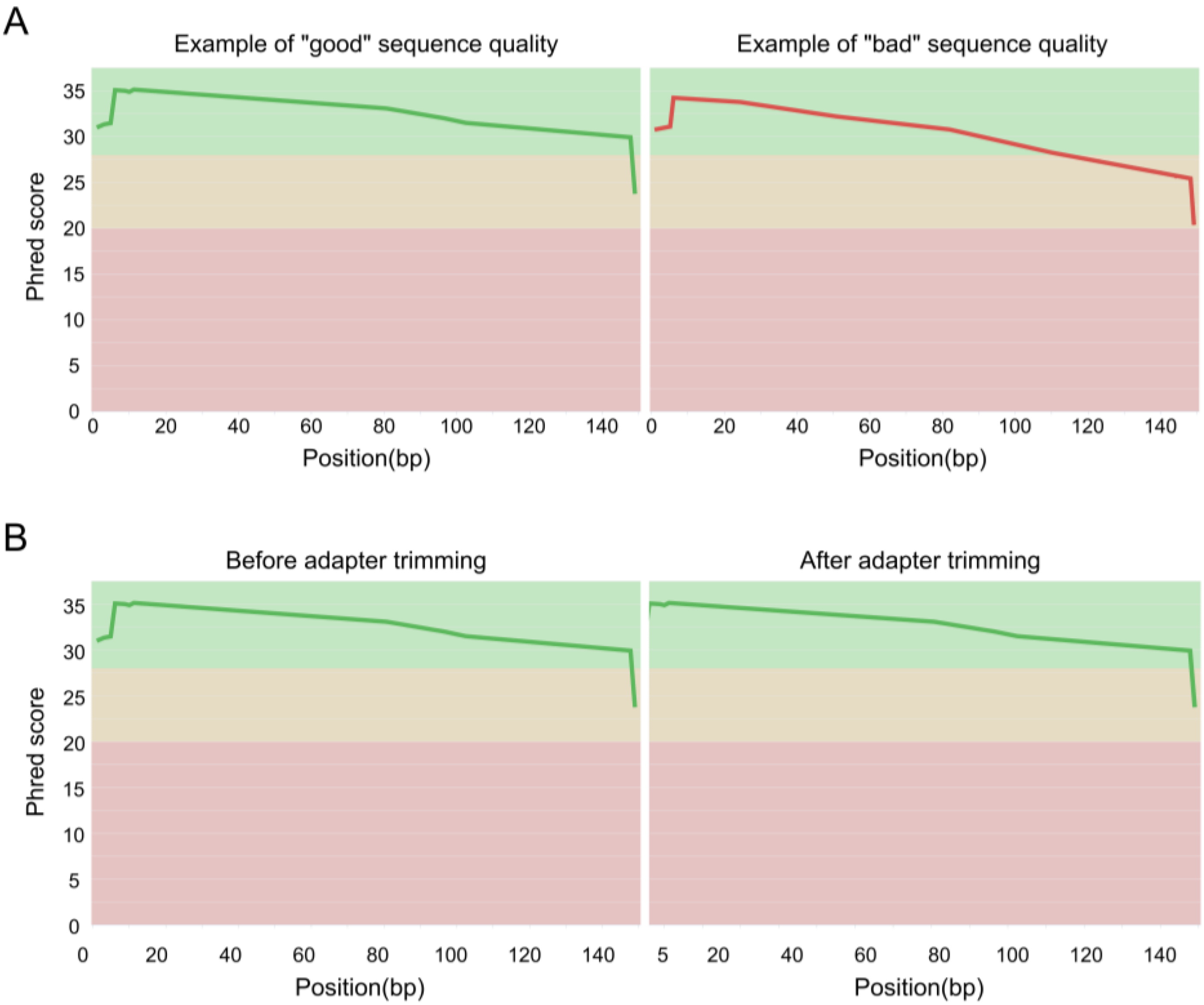


Figure 3. FastQC mean quality scores. (**A**) Examples of “good” and “bad” sequence quality. (**B**) Overview of the adapter trimming impact.

3.2. Adapter Trimming

Another preprocessing step is the adapter trimming, which involves removing adapter sequences, low-quality reads, and contaminating sequences from the raw sequencing data. The most widely used tools for data preprocessing are fastp [20], Trimmomatic [24], Cutadapt [25], and BBDuk [26]. In **Figure 3**B, researchers present quality profiles obtained using FastQC, illustrating the impact of adapter trimming with Cutadapt.

4. Alignment of the Nucleotide Sequence on a Reference Genome

After adapter trimming, the next step is to align the reads to a reference genome. The Genome Reference Consortium introduced the latest human reference genome, GRCh38 [56], in 2017, followed by subsequent improvements, the latest being GRCh38.p14 in March 2022, which remarkably reduced the number of gaps in the assembly to 349 compared to the initial version's approximately 150,000 gaps. Notably, these gaps were predominantly found in regions like telomeres, centromeres, and long repetitive sequences. Last year, the Telomere-to-Telomere (T2T) Consortium presented the first fully assembled reference genome [57], T2T-CHM13, eliminating all gaps.

The alignment step is performed by read mapper software, which assigns a location on the reference genome to each read based on its sequence. Since the reads do not contain information about their location in the genome, the mapper infers this information by comparing the read sequence to the reference genome. Essentially, it checks which parts of the reference genome match the sequences in the reads, determining where these reads originated in the genome. However, this seemingly straightforward task is computationally intensive and time-consuming because the software must meticulously compare each read to the entire reference genome and assign a precise position for it. The computational demand arises from the need for high accuracy and reliability in determining the origin of each read, a fundamental step in understanding the genetic information contained within the sequenced sample. There are many different read mappers available, each with its own strengths and weaknesses. Common examples include BWA [27] for genome and Bowtie2 [28] for transcriptome. These tools employ a Burrows–Wheeler transform, a computational method invented by Michael Burrows and David Wheeler in 1994. This method involves rearranging character strings into sequences of similar characters, which offers significant computational benefits. Indeed, strings with repeated characters are easily compressible using techniques like move-to-front transform and run-length encoding. Various aligners employ distinct strategies; for instance, HISAT2 [29] is a graph-based genome alignment tool. The utilization of a graph-based approach allows leveraging theoretical advancements in computer science, resulting in a rapid and memory-efficient search algorithm. In transcriptome alignment, STAR [30] is also widely employed, using the Maximal Exact (Unique) Match concept for seed searching, it proves particularly advantageous for aligning long reads (>200 bp), such as those generated by third-generation sequencing.

The results of the read mapping step are usually provided in SAM format files, which can be converted to BAM format for more efficient storage and processing. SAM/BAM files can be accessed through the Integrative Genomics Viewer (IGV), allowing visualization of the reads (**Figure 2**B). The BAM files undergo different modifications during the alignment post-processing step, which includes tasks such as sorting, marking duplicate reads, and recalibrating base quality scores. The goal of these post-processing steps is to improve the accuracy and reliability of the final variant calls.

After the read mapping step, the resulting SAM/BAM files are sorted according to their genomic coordinates. This sorting is important because downstream analysis often relies on the order of the aligned reads. PCR duplicates are then commonly removed using tools such as Picard [22][58] or SAMtools [5]. PCR duplicates are identical copies of the same genomic fragment and can be introduced during sample preparation and PCR amplification steps. They can bias the analysis and lead to overrepresentation of certain regions of the genome. However, it is important to note that duplicated reads can also be biological copies originating from the same genomic location of chromosomes of different cells. For deep-coverage targeted sequencing approaches the probability of a duplicate read to be a biological copy increases with coverage, and therefore, the removal of duplicates is typically not performed in these cases.

References

1. Prasad, V.; Fojo, T.; Brada, M. Precision oncology: Origins, optimism, and potential. Lancet Oncol. 2016, 17, e81–e86.
2. Buermans, H.P.J.; Den Dunnen, J.T. Next generation sequencing technology: Advances and applications. Biochim. Biophys. Acta (BBA)—Mol. Basis Dis. 2014, 1842, 1932–1941.
3. Arora, N.; Chaudhary, A.; Prasad, A. Editorial: Methods and applications in molecular diagnostics. Front. Mol. Biosci. 2023, 10, 1239005.
4. Brandies, P.A.; Hogg, C.J. Ten simple rules for getting started with command-line bioinformatics. PLoS Comput. Biol. 2021, 17, e1008645.
5. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, 25, 2078–2079.
6. Leipzig, J. A review of bioinformatic pipeline frameworks. Brief. Bioinform. 2016, 18, bbw020.
7. Di Tommaso, P.; Chatzou, M.; Floden, E.W.; Barja, P.P.; Palumbo, E.; Notredame, C. Nextflow enables reproducible computational workflows. Nat. Biotechnol. 2017, 35, 316–319.
8. Mölder, F.; Jablonski, K.P.; Letcher, B.; Hall, M.B.; Tomkins-Tinch, C.H.; Sochat, V.; Forster, J.; Lee, S.; Twardziok, S.O.; Kanitz, A.; et al. Sustainable data analysis with Snakemake. F1000Research 2021, 10, 33.
9. Sadedin, S.P.; Pope, B.; Oshlack, A. Bpipe: A tool for running and managing bioinformatics pipelines. Bioinformatics 2012, 28, 1525–1526.
10. Crusoe, M.R.; Abeln, S.; Iosup, A.; Amstutz, P.; Chilton, J.; Tijanić, N.; Ménager, H.; Soiland-Reyes, S.; Gavrilović, B.; Goble, C.; et al. Methods included: Standardizing computational reuse and portability with the Common Workflow Language. Commun. ACM 2022, 65, 54–63.
11. Voss, K.; der Auwera, G.V.; Gentry, J. Full-stack genomics pipelining with GATK4 + WDL + Cromwell. F1000Research 2017, 6.
12. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010, 20, 1297–1303.

13. Ewels, P.A.; Peltzer, A.; Fillinger, S.; Patel, H.; Alneberg, J.; Wilm, A.; Garcia, M.U.; Di Tommaso, P.; Nahnsen, S. The nf-core framework for community-curated bioinformatics pipelines. Nat. Biotechnol. 2020, 38, 276–278.
14. Hanssen, F.; Garcia, M.U.; Folkersen, L.; Pedersen, A.S.; Lescai, F.; Jodoin, S.; Miller, E.; Wacker, O.; Smith, N.; Community, N.-C.; et al. Scalable and efficient DNA sequencing analysis on different compute infrastructures aiding variant discovery. bioRxiv 2023, 549462.
15. Garcia, M.; Juhos, S.; Larsson, M.; Olason, P.I.; Martin, M.; Eisfeldt, J.; DiLorenzo, S.; Sandgren, J.; Ståhl, T.D.D.; Ewels, P.; et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. F1000Research 2020, 9, 63.
16. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Res. 2022, 50, W345–W351.
17. Oinn, T.; Addis, M.; Ferris, J.; Marvin, D.; Senger, M.; Greenwood, M.; Carver, T.; Glover, K.; Pocock, M.R.; Wipat, A.; et al. Taverna: A tool for the composition and enactment of bioinformatics workflows. Bioinformatics 2004, 20, 3045–3054.
18. Abouelhoda, M.; Issa, S.A.; Ghanem, M. Tavaxy: Integrating Taverna and Galaxy workflows with cloud computing support. BMC Bioinform. 2012, 13, 77.
19. Scheffler, K.; Catreux, S.; O'Connell, T.; Jo, H.; Jain, V.; Heyns, T.; Yuan, J.; Murray, L.; Han, J.; Mehio, R. Somatic small-variant calling methods in Illumina DRAGEN™ Secondary Analysis. bioRxiv 2023, 534011.
20. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018, 34, i884–i890.
21. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 1 December 2023).
22. Broad Institute Picard Toolkit. 2019. Available online: http://broadinstitute.github.io/picard/ (accessed on 1 December 2023).
23. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016, 32, 3047–3048.
24. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 2014, 30, 2114–2120.
25. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011, 17, 10–12.
26. Bushnell, B. BBDuk. 2018. Available online: https://sourceforge.net/projects/bbmap/ (accessed on 1 December 2023).

27. Jung, Y.; Han, D. BWA-MEME: BWA-MEM emulated with a machine learning approach. Bioinformatics 2022, 38, 2404–2413.
28. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 2012, 9, 357–359.
29. Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 2019, 37, 907–915.
30. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 2013, 29, 15–21.
31. Poplin, R.; Ruano-Rubio, V.; DePristo, M.A.; Fennell, T.J.; Carneiro, M.O.; der Auwera, G.A.V.; Kling, D.E.; Gauthier, L.D.; Levy-Moonshine, A.; Roazen, D.; et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 2018, 201178.
32. Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. arXiv 2012, arXiv:1207.3907.
33. Benjamin, D.; Sato, T.; Cibulskis, K.; Getz, G.; Stewart, C.; Lichtenstein, L. Calling Somatic SNVs and Indels with Mutect2. bioRxiv 2019, 861054.
34. Lai, Z.; Markovets, A.; Ahdesmaki, M.; Chapman, B.; Hofmann, O.; McEwen, R.; Johnson, J.; Dougherty, B.; Barrett, J.C.; Dry, J.R. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016, 44, e108.
35. Sherry, S.T.; Ward, M.; Sirotkin, K. dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. Genome Res. 1999, 9, 677–679.
36. Auton, A.; Abecasis, G.R.; Altshuler, D.M.; Durbin, R.M.; Abecasis, G.R.; Bentley, D.R.; Chakravarti, A.; Clark, A.G.; Donnelly, P.; Eichler, E.E.; et al. A global reference for human genetic variation. Nature 2015, 526, 68–74.
37. Karczewski, K.J.; Francioli, L.C.; Tiao, G.; Cummings, B.B.; Alföldi, J.; Wang, Q.; Collins, R.L.; Laricchia, K.M.; Ganna, A.; Birnbaum, D.P.; et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020, 581, 434–443.
38. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.S.; Thormann, A.; Flicek, P.; Cunningham, F. The Ensembl Variant Effect Predictor. Genome Biol. 2016, 17, 122.
39. Baux, D.; Van Goethem, C.; Ardouin, O.; Guignard, T.; Bergougnoux, A.; Koenig, M.; Roux, A.-F. MobiDetails: Online DNA variants interpretation. Eur. J. Hum. Genet. 2021, 29, 356–360.
40. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010, 38, e164.

41. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. Fly 2012, 6, 80–92.
42. Breunig, M.M.; Kriegel, H.-P.; Ng, R.T.; Sander, J. LOF: Identifying Density-Based Local Outliers. ACM SIGMOD Rec. 2000, 29, 93–104.
43. Rapti, M.; Zouaghi, Y.; Meylan, J.; Ranza, E.; Antonarakis, S.E.; Santoni, F.A. CoverageMaster: Comprehensive CNV detection and visualization from NGS short reads for genetic medicine applications. Brief. Bioinform. 2022, 23, bbac049.
44. Onsongo, G.; Baughn, L.B.; Bower, M.; Henzler, C.; Schomaker, M.; Silverstein, K.A.T.; Thyagarajan, B. CNV-RF Is a Random Forest-Based Copy Number Variation Detection Method Using Next-Generation Sequencing. J. Mol. Diagn. 2016, 18, 872–881.
45. Glessner, J.T.; Hou, X.; Zhong, C.; Zhang, J.; Khan, M.; Brand, F.; Krawitz, P.; Sleiman, P.M.A.; Hakonarson, H.; Wei, Z. DeepCNV: A deep learning approach for authenticating copy number variations. Brief. Bioinform. 2021, 22, bbaa381.
46. Yuan, X.; Yu, J.; Xi, J.; Yang, L.; Shang, J.; Li, Z.; Duan, J. CNV_IFTV: An Isolation Forest and Total Variation-Based Detection of CNVs from Short-Read Sequencing Data. IEEE ACM Trans. Comput. Biol. Bioinf. 2021, 18, 539–549.
47. Guo, Y.; Wang, S.; Yuan, X. HBOS-CNV: A New Approach to Detect Copy Number Variations From Next-Generation Sequencing Data. Front. Genet. 2021, 12, 642473.
48. Huang, T.; Li, J.; Jia, B.; Sang, H. CNV-MEANN: A Neural Network and Mind Evolutionary Algorithm-Based Detection of Copy Number Variations From Next-Generation Sequencing Data. Front. Genet. 2021, 12, 700874.
49. Boeva, V.; Popova, T.; Bleakley, K.; Chiche, P.; Cappo, J.; Schleiermacher, G.; Janoueix-Lerosey, I.; Delattre, O.; Barillot, E. Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 2012, 28, 423–425.
50. Cabello-Aguilar, S.; Vendrell, J.A.; Van Goethem, C.; Brousse, M.; Gozé, C.; Frantz, L.; Solassol, J. ifCNV: A novel isolation-forest-based package to detect copy-number variations from various targeted NGS datasets. Mol. Ther.—Nucleic Acids 2022, 30, 174–183.
51. Viailly, P.-J.; Sater, V.; Viennot, M.; Bohers, E.; Vergne, N.; Berard, C.; Dauchel, H.; Lecroq, T.; Celebi, A.; Ruminy, P.; et al. Improving high-resolution copy number variation analysis from next generation sequencing using unique molecular identifiers. BMC Bioinform. 2021, 22, 120.
52. Escudié, F.; Van Goethem, C.; Grand, D.; Vendrell, J.; Vigier, A.; Brousset, P.; Evrard, S.M.; Solassol, J.; Selves, J. MIAmS: Microsatellite instability detection on NGS amplicons data. Bioinformatics 2019, 36, btz797.

53. Niu, B.; Ye, K.; Zhang, Q.; Lu, C.; Xie, M.; McLellan, M.D.; Wendl, M.C.; Ding, L. MSIsensor: Microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics 2014, 30, 1015–1016.
54. Swaerts, K.; Dedeurwaerdere, F.; De Smet, D.; De Jaeger, P.; Martens, G.A. DeltaMSI: Artificial intelligence-based modeling of microsatellite instability scoring on next-generation sequencing data. BMC Bioinform. 2023, 24, 73.
55. Marques, A.C.; Ferraro-Peyret, C.; Michaud, F.; Song, L.; Smith, E.; Fabre, G.; Willig, A.; Wong, M.M.L.; Xing, X.; Chong, C.; et al. Improved NGS-based detection of microsatellite instability using tumor-only data. Front. Oncol. 2022, 12, 969238.
56. Schneider, V.A.; Graves-Lindsay, T.; Howe, K.; Bouk, N.; Chen, H.-C.; Kitts, P.A.; Murphy, T.D.; Pruitt, K.D.; Thibaud-Nissen, F.; Albracht, D.; et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017, 27, 849–864.
57. Aganezov, S.; Yan, S.M.; Soto, D.C.; Kirsche, M.; Zarate, S.; Avdeyev, P.; Taylor, D.J.; Shafin, K.; Shumate, A.; Xiao, C.; et al. A complete reference genome improves analysis of human genetic variation. Science 2022, 376, eabl3533.
58. Robinson, J.T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative Genomics Viewer. Nat. Biotechnol. 2011, 29, 24–26.

Retrieved from https://encyclopedia.pub/entry/history/show/119321