German Medical Data Sciences 2023 — Science. Close to People.
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI230694

Onkopipe: A Snakemake Based DNA-Sequencing Pipeline for Clinical Variant Analysis in Precision Medicine

Jingyu Yang[a], Tim Beißbarth[a,b,c], and Jürgen Dönitz[a,b,c,1]
[a] *Department of Medical Bioinformatics, University of Göttingen, Göttingen, Germany*
[b] *Campus-Institute Data Science (CIDAS), University of Göttingen, Göttingen, Germany*
[c] *Comprehensive Cancer Center Niedersachsen (CCCN), Göttingen, Germany*

Abstract. NGS is increasingly used in precision medicine, but an automated sequencing pipeline that can detect different types of variants (single nucleotide - SNV, copy number - CNV, structural - SV) and does not rely on normal samples as germline comparison is needed. To address this, we developed Onkopipe, a Snakemake-based pipeline that integrates quality control, read alignments, BAM pre-processing, and variant calling tools to detect SNV, CNV, and SV in a unified VCF format without matched normal samples. Onkopipe is containerized and provides features such as reproducibility, parallelization, and easy customization, enabling the analysis of genomic data in precision medicine. Our validation and evaluation demonstrate high accuracy and concordance, making Onkopipe a valuable open-source resource for molecular tumor boards. Onkopipe is being shared as an open source project and is available at https://gitlab.gwdg.de/MedBioinf/mtb/onkopipe.

Keywords. Next-generation sequencing, Molecular tumor board, DNA-sequencing pipeline

1. Introduction

With the established next-generation sequencing technology (NGS) and greatly reduced sequencing costs, the acquisition of a large number of genetic profiles became feasible. The availability of patient-specific biomarkers provides new opportunities in precision oncology for patients where standard guideline therapies did not lead to remission. In molecular tumor boards (MTBs) a multidisciplinary team identifies and discusses potential therapies based on genetic analysis [1]. Several tools exist to support the interpretation of genetic variants for a treatment recommendation [2,3,4], e. g. Perea-Bel et al. [2] presented an automated MTB tool that produces patient-specific reports with treatment recommendations for actionable variants of the patient (Fig. 1). However, these tools require a list of valid and reliable variants as input.

Various pipelines for whole genome sequencing (WGS) or whole exome sequencing (WES) have been published, such as Omics-Pipe [5] and Unipro UGENE NGS pipeline [6], which can automate sequencing analysis from raw data to annotated variants, but lack features required for clinical analysis of MTB, such as structural variant (SV) and

[1] Corresponding Author, Jürgen Dönitz, University Medical Center Göttingen (UMG), Goldschmidtstr. 1, 37077 Göttingen, Germany; E-mail: juergen.doenitz@bioinf.med.uni-goettingen de.

copy number variant (CNV) calling. Previous works have focused on comparing different tools and pipelines, while Sentieon [7] DNASeq pipeline shows good performance in germline variant calling but is not open source and lacks containerization and CNV/SV calling functionalities. DNAp [8] offers a comprehensive WGS/WES analysis solution but does not cover CNV calling and requires matched control data. Oh and Zhao's [9,10] approaches addresses the tumor-only data issue but still requires control data from other normal samples and focuses on biomarkers like tumor mutation burden (TMB) and microsatellite instability (MSI) rather than variant detection. Examples of pipelines based on Snakemake are NGS-pipe and V-Pipe [11,12], which provide powerful features but lack dockerization for easy portability and do not support the calling of structural variants.

Concerning all of these above, we built Onkopipe, an automated and customizable pipeline based on Snakemake framework [13] for the variant calling for analysis in a MTB. The goal of Onkopipe is to create a containerized workflow for the steps between raw sequences to uniform variants as input for a MTB tool, even without matched controlled data.

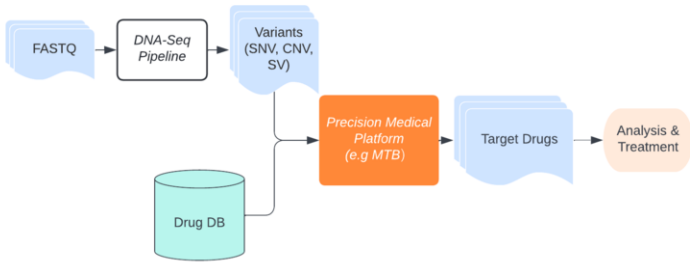

Figure 1. Workflow of Precision Medicine (e.g. MTB). Overview of the precision medicine workflow from raw data (FASTQ) to actionable personalized treatment. Raw data sequenced from patients always requires a DNA-Seq-pipeline to detect genetic biomarker like SNV, CNV, and SV. Genetic biomarkers and corresponding drug databases can then be used to identify actionable variants and matched drugs. Finally, depending on the level of evidence for the target drug, personalized treatment or medical analysis can be done.

2. Methods

2.1. Onkopipe is built using Snakemake and Docker

2.1.1. Pipeline based on Snakemake

Onkopipe is a flexible and customizable pipeline for variant detection analysis, implemented using Snakemake and encapsulated within a Docker image. The rule-based design allows for easy modification of analysis steps through configuration files, and intermediate files and error logs are stored to ensure documentation and reproducibility. Onkopipe also uses an isolated Conda environment configuration file to store the parameters for each tool, which allows for the automatic deployment of pre-defined software tools in the specified versions. To maximize computing resource utilization and minimize runtime, Onkopipe provides parallelization of single steps. Snakemake identifies which tasks can run in parallel and which need to be run in order, and provides an overview of the pipeline structure in directed acyclic graph (DAG).

2.1.2. *Dockerization of Onkopipe*

To achieve cross-platform functionality, we encapsulated Onkopipe into a Docker container image. Once the Onkopipe Docker image is built, the same conda based Snakemake environment is provided for processing genome sequencing data. By specifying the directory for the raw input data as docker volume, input files can be accessed in the container and outputs will be automatically generated under the same folder. The source code of Onkopipe is freely available at https://gitlab.gwdg.de/MedBioinf/mtb/onkopipe.

2.2. *Onkopipe is built using Snakemake and Docker*

2.2.1. *Raw data preprocessing*

Raw data (FASTQ) can not directly be used for variant discovery analysis. The first stage of the workflow includes pre-processing steps that are necessary to convert data from FASTQ files into BAM files, suitable for downstream analysis.

- Quality control: In Onkopipe, Trim Galore [14] is used for quality control, which incorporates Cutadapt and FastQC [14]. Trim Galore trims low-quality base calls and removes adapter sequences from the 3' end of reads using Cutadapt, with the option to customize adapter trimming. FastQC generates a quality control report with important information about the raw data, such as sequence length and quality per base, which guides tool configuration adjustment.
- Alignment: Alignment of raw reads to a reference genome is a crucial step in DNA sequencing. Several mapping tools have been benchmarked [15], including bowtie, Bowtie2, BWA, etc [15,34]. Among these, bowtie showed the best throughput, while BWA performed better for longer read lengths. Recent studies [16,17] have shown that BWA-MEM achieved better sensitivity and false positives rate for DNA sequencing data, making it the optimal choice for Onkopipe, which focuses on DNA sequencing with an average read length of more than 70bp, such as FFPE materials that usually have a length larger than 100bp. (Table 1)
    - BAM preprocessing: After the raw data is successfully aligned, a SAM (Sequence Alignment Map) file, a text-based format for storing biological sequences aligned to a reference genome [18], is created and will be processed to BAM (Binary Alignment Map) file, a binary representation of SAM [18], which is used for the following steps: 1. Sort, index and mark duplicates. 2. Base quality score recalibration.

Table 1. DNA sequence alignment tools

Tool	Description	Advantages	Disadvantages
BWA-MEM	Suitable for DNA sequencing	Good sensitivity & false-positive rate, especially for long reads	Lower throughput
Bowtie2	Improved version of Bowtie	Flexible and fast	Less optimal for long reads
Stampy	Aligns short and long reads	Handles high mutation rates	Slower than other tools

2.2.2. Variant Discovery

After the BAM file has been generated, the pipeline will continue with the variant discovery phase. All three types of variants (SNV, CNV, SV) can be detected and are stored in VCFv4.2 files. To facilitate the subsequent analysis, in addition to keep the separated three types of variants files, a concatenated VCFv4.2 file is created as well.

- SNV calling involves identifying single nucleotide variants from NGS data. We carefully selected GATK Mutect2 as our SNV caller, a decision guided by both established best practice guidelines and pertinent literature studies [19,20]. Mutect2 remains a top contender in the field of bioinformatics pipelines, with its robust performance and unique tumor-only feature distinguishing it from other callers [19,20].
- CNV calling refers to detecting events where sections of the genome are duplicated or deleted, with CNVkit performing better for smaller CNV sizes and cn.MOPS being more suitable for larger CNV sizes [21]. Onkopipe integrates CNV detection using CNVkit with a "flat" reference of neutral copy number, without control data.
- SV calling: Structural variants (SVs) are alterations in genomic regions due to insertions, deletions, inversions and translocations. LUMPY showed the best performance for coverage between 8 and 32x among all callers in Sarwal's benchmark study [22].

2.3. Material

2.3.1. Analysis tools

Onkopipe is based on Snakemake [13] pipeline management framework and Docker [23]. Sequencing data analyses were performed using Trim Galore v0.6.6 [14], BWA mem v0.7.17 [16], Samtools v1.10 [18], bcftools v1.10 [24], GATK Mutect2 v4.1.9.0 [25], Picard v2.23.8 [25], Lumpy-SV v0.3.1 [26] and the CNVkit v0.9.9 [27], and vcf-compare from NEAT v3.0 [28]. We used the R package VariantAnnotation [29] to annotate somatic variants data for validation.

2.3.2. Equipments

All of our experiments were conducted on a server with Intel (R) Haswell Xeon CPU E5- 2698 processors with 64 cores, 2.30GHz. The total size of RAM and HDD are 768GB and 29TB respectively. The operating system is Debian Linux 5.10.84-1. The pre-installed Docker version is v20.10.6 and the Bioconda version is v4.11.0. Onkopipe can automatically take advantage of multiple CPU cores.

2.3.3. Data

Structural Multiplex Reference Standard gDNA HD753 was used for variant calling validation. Whole genome sequencing (WGS) NA12878 was downloaded from Illumina basespace for evaluation. The NA12878 Illumina Platinum variant calls were used as the truth set to evaluate variant calling accuracy. The May 2022 GATK bundle was used for the human reference (hg38), dbSNP (build 146), and the Mills and 1000G gold standard indels.

3. Results

3.1. Implementation of Onkopipe

The Onkopipe was developed using Docker and the Snakemake framework, making it easy to run identical analyses on different machines and platforms. The selection of tools for read alignment and variant calling was based on a combination of literature study, recommendations from Broad Institute Best Practices guide, and the bioinformatics pipeline of The NCI's Genomic Data Commons (GDC) [18,34]. The pipeline was customized with different callers suitable for SNV, CNV, and SV variant detection. The input to Onkopipe is a patient's tumor sample FASTQ file and the output is VCF files with the identified variants after full end-to-end processing. Our solution can detect three types of variants at once without the need for germline controlled data. Quality control reports and aligned BAM files are stored in the predefined output folder. The YAML files have already configured the runtime environment of tools like Trim Galore, BWA-MEM, GATK, Picard, Mutect2, CNVkit, and LUMPY-SV, which are automatically downloaded and set up on the first run. Figure 2. shows the implementation of Onkopipe based on the process described in the Methods section and the tools selected above.

Figure 2. Workflow of Onkopipe. Onkopipe analysis steps generally follow the National Cancer Institute Genomic Data Commons (GDC) bioinformatics pipeline and the best practice guidelines published by the Broad Institute. It includes the use of Trim Galore for quality control, BWA mem for alignment, Picard MarkDuplicates to remove duplicates, and GATK for base quality score recalibration. The SNV, CNV and SV variants were called using GATK mutect2, CNVkit, and LUMPY-SV, respectively and concatenated into a unified VCF as output. Additional SNP calling (HaplotypeCaller, GenotypeGVCFs) and annotation (VariantAnnotation) steps are only used for evaluation, thus they are not shown in the flowchart.

3.2. Validation with HD753

Following the development of Onkopipe, a reference standard sample HD753 which was produced by Horizon DiagnosticsTM in Waterbeach, United Kingdom with known

somatic variants was used to validate the accuracy of the pipeline. HD753 with a read depth of 500X contains 8 SNVs, 6 INDELs, 2 CNVs, and two gene fusions.

Table 2. Validation results of variants (SNV, INDEL, CNV and fusion) in reference gDNA HD753.

Variant Type	Gene	Variant	Detected Gene	Detected Variant
SNV High GC	GNA11	Q209L	GNA11	Q209L
SNV High GC	AKT1	E17K	AKT1	E17K
SNV High GC	NOTCH1	P668S	NOTCH1	P668S
SNV Low GC	PIK3CA	E545K	PIK3CA	E545K
SNV Low GC	KRAS	G13D	KRAS	G13D
SNV	EGFR	G719S	EGFR	G719S
SNV	BRAF	V600E	BRAF	V600E
SNV	KRAS	H1047R	KRAS	H1047R
Short Deletion	MET	V237fs	MET	V237fs
Short Deletion	FLT3	S985fs	FLT3	S985fs
Short Deletion	BRCA2	A1689fs	BRCA2	A1689fs
Short Deletion	FBXW7	G667fs	FBXW7	G667fs
Long Insertion	EGFR	V769_D770insASV	EGFR	V769_D770insASV
Long Insertion	EGFR	ΔE746-A750	EGFR	ΔE746-A750
CNV	MET	amplification (4.5 copies)	MET	3 copies
CNV	MYC-N	amplification (9.5 copies)	MYC-N	8 copies
Fusion	ROS1	SLC34A2/ROS1	ROS1	N]chr6:117337156] /N]chr4:25665007] (SLC34A2/ROS1)
Fusion	RET	CCDC6/RET	RET	[chr10:43114500[N /[chr10:59878853[N (CCDC6/RET)

3.3. Calling Accuracy Evaluation

We evaluated the accuracy of Onkopipe by comparing its results to the truth sets using the vcf-compare tool [28]. Haplotypecaller and GenotypeGVCFs [25] were integrated for the evaluation, and the comparison was limited to the Illumina platinum confidence region. The resulting precision, recall, and F-score were calculated based on the counts of TP, FP, and FN obtained from the vcf-compare results. The evaluation results showed that Onkopipe had high accuracy on both the synthetic chr 20-22 dataset and the Illumina Platinum reference data NA12878. Notably, NA12878 has been extensively used as a standard reference dataset in numerous studies, including the PrecisionFDA truth challenge [32,33]., Remarkably, Onkopipe's performance surpassed that of the GATK and Sentieon's pipeline [7], which were among the winners of the PrecisionFDA truth challenge in both 2016 and 2020. The detailed evaluation results are presented in Table 3.

Table 3. Variant detection accuracy: F1 scores.

Dataset	synthetic WGS, chr 20-22	NA12878
Onkopipe vs Truthset	0.96	0.98
Sentieon's DNASeq Pipeline vs Truthset	0.96	0.96
GATK4 vs Truthset	0.95	0.96

4. Discussion

We developed Onkopipe, a bioinformatics pipeline designed for precision medical analysis and MTB [1,2,3]. Onkopipe provides simultaneous variant calling (SNV, CNV, SV), docker containerization, and tumor-only processing features. We achieved the tumor-only analysis feature in Onkopipe by utilizing public germline data such as gnomAD and dbSNP in SNV calling and "flat" reference data in CNV detection [27]. This approach helps to minimize the inclusion of common germline variants and improves the distinction between somatic and germline variants. Onkopipe is the only open-source end-to-end pipeline for three types of variant detection in precision medicine that does not require control data. Patients' SNV, CNV, and SV can be called by Onkopipe in one single run and stored in uniform VCF4.2 format for further therapeutic analysis in accordance with MTB's clinical variants detection needs [1,2].

All the software tools included in Onkopipe have been carefully selected to meet the needs of precision medical molecular oncology research. One downside of the dockerized version of Onkopipe is that modifying the configuration or changing tools to suit specific needs can be time-consuming as new container images need to be created and rebuilt. Debugging directly in a docker-based pipeline can also be more challenging. On the other hand, local Snakemake pipelines allow for easy modification of read aligners, variant callers, and annotators to suit specific research needs. Nevertheless, the dockerized Onkopipe allows for consistent, repeatable results across multiple platforms, making it a valuable tool for project collaborators.

Pipeline validation with the gold standard reference HD753 showed that all 18 confirmed variants including SNVs, INDELs, and fusion mutations can be correctly detected. The small deviation in copy numbers detection may be the result of a lack of matched control data. As a result, sensitivities of SNVs, INDELs, CNVs were measured at 100% (Table 2). Compared to Sentieon DNASeq [11], Onkopipe demonstrated similar or even better performance. One possible explanation for this could be that we utilized an updated caller version and reference data such as a panel of normal and known dbsnp. We acknowledge the potential limitations of the tumor-only approach and its implications for MTB decision-making and patient care. Therefore, any variant of interest identified by our tool should ideally be validated through orthogonal methods before being used to make clinical decisions. In summary, Onkopipe, which is a novel open-source end-to-end DNA sequencing analysis pipeline designed for MTB clinical analysis, was implemented to aid in precise oncological diagnosis and treatment.

Declarations

Ethical vote: Not applicable

Conflict of Interest: The authors declare that they have no competing interests.

Authors contributions: JY, JD and TB designed the study. JY developed and implemented the pipeline. JY, JD nad TB wrote the manuscript. All authors critically reviewed the content and approved the final manuscript.

Funding: This work was supported by the Volkswagen Foundation within research project MTB-Report (ZN3424).

References

[1] Luchini C, Lawlor RT, Milella M, Scarpa A. Molecular tumor boards in clinical practice. Trends in Cancer. 2020;6(9):738–44. doi: 10.1016/j.trecan.2020.05.008
[2] Perera-Bel J, Hutter B, Heining C, Bleckmann A, Fröhlich M, Fröhling S, et al. From somatic variants towards Precision Oncology: Evidence-driven reporting of treatment options in molecular tumor boards. Genome Medicine. 2018;10(1). doi: 10.1186/s13073-018-0529-2
[3] Kurz NS, Perera-Bel J, Höltermann C, Tucholski T, Yang J, Beissbarth T, et al. Identifying actionable variants in cancer – the dual web and batch processing tool MTB-Report. Studies in Health Technology and Informatics. 2022; doi: 10.3233/SHTI220806
[4] Li Q, Ren Z, Cao K, Li MM, Wang K, Zhou Y. Cancervar: An artificial intelligence–empowered platform for clinical interpretation of somatic mutations in cancer. Science Advances. 2022;8(18). doi: 10.1126/sciadv.abj1624
[5] Fisch KM, Meißner T, Gioia L, Ducom J-C, Carland TM, Loguercio S, et al. OMICS pipe: A community-based framework for reproducible multi-omics data analysis. Bioinformatics. 2015;31(11):1724–8. doi: 10.1093/bioinformatics/btv061
[6] Golosova O, Henderson R, Vaskin Y, Gabrielian A, Grekhov G, Nagarajan V, et al. Unipro Ugene NGS pipelines and components for variant calling, RNA-seq and chip-seq data analyses. PeerJ. 2014;2. doi: 10.7717/peerj.644
[7] Kendig KI, Baheti S, Bockol MA, Drucker TM, Hart SN, Heldenbrand JR, et al. Sentieon dnaseq variant calling workflow demonstrates strong computational performance and accuracy. Frontiers in Genetics. 2019;10. doi: 10.3389/fgene.2019.00736
[8] Causey JL, Ashby C, Walker K, Wang ZP, Yang M, Guan Y, et al. DNAp: A pipeline for DNA-seq data analysis. Scientific Reports. 2018;8(1). doi: 10.1038/s41598-018-25022-6
[9] Oh S, Geistlinger L, Ramos M, Morgan M, Waldron L, Riester M. Reliable analysis of clinical tumor-only whole-exome sequencing data. JCO Clinical Cancer Informatics. 2020;(4):321–35. doi: 10.1200/CCI.19.00130
[10] Zhao C, Jiang T, Hyun Ju J, Zhang S, Tao J, Fu Y, et al. TruSight Oncology 500: Enabling comprehensive genomic profiling and biomarker reporting with targeted sequencing. 2020; doi: 10.1101/2020.10.21.349100
[11] Singer J, Ruscheweyh H-J, Hofmann AL, Thurnherr T, Singer F, Toussaint NC, et al. NGS-pipe: A flexible, easily extendable and highly configurable framework for NGS Analysis. Bioinformatics. 2017;34(1):107–8. doi: 10.1093/bioinformatics/btx540
[12] Posada-Céspedes S, Seifert D, Topolsky I, Jablonski KP, Metzner KJ, Beerenwinkel N. V-pipe: A computational pipeline for assessing viral genetic diversity from high-throughput data. Bioinformatics. 2021;37(12):1673–80. doi: 10.1093/bioinformatics/btab015
[13] Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28(19):2520–2. doi: 10.1093/bioinformatics/bts480
[14] Krueger F. Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. Babraham Institute. 2015. doi: 10.5281/zenodo.7598955
[15] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60. doi: 10.1093/bioinformatics/btp324
[16] Jäger N. Bioinformatics workflows for clinical applications in precision oncology. InSeminars in cancer biology 2022 Sep 1 (Vol. 84, pp. 103-112). Academic Press. doi: 10.1016/j.semcancer.2020.12.020
[17] Zanti M, Michailidou K, Loizidou MA, Machattou C, Pirpa P, Christodoulou K, Spyrou GM, Kyriacou K, Hadjisavvas A. Performance evaluation of pipelines for mapping, variant calling and interval padding, for the analysis of NGS germline panels. BMC bioinformatics. 2021 Apr 28;22(1):218. doi: 10.1186/s12859-021-04144-1
[18] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. doi: 10.1093/bioinformatics/btp352
[19] Chen Z, Yuan Y, Chen X, Chen J, Lin S, Li X, Du H. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. Scientific reports. 2020 Feb 26;10(1):3501. doi: 10.1038/s41598-020-60559-5

[20] Pei S, Liu T, Ren X, Li W, Chen C, Xie Z. Benchmarking variant callers in next-generation and third-generation sequencing analysis. Briefings in Bioinformatics. 2021 May;22(3):bbaa148. doi: 10.1093/bib/bbaa148
[21] Zhao L, Liu H, Yuan X, Gao K, Duan J. Comparative study of whole exome sequencing-based copy number Variation Detection Tools. BMC Bioinformatics. 2020;21(1). doi: 10.1186/s12859-020-3421-1
[22] Sarwal V, Niehus S, Ayyala R, Kim M, Sarkar A, Chang S, et al. A comprehensive benchmarking of WGS-based deletion structural variant callers. Briefings in Bioinformatics. 2022;23(4). doi: 10.1093/bib/bbac221
[23] Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux j. 2014 Mar 2;239(2):2 doi: 10.5555/2600239.2600241
[24] Li H. A statistical framework for SNP Calling, mutation discovery, association mapping and Population Genetical Parameter Estimation from Sequencing Data. Bioinformatics. 2011;27(21):2987–93. doi: 10.1093/bioinformatics/btr509
[25] Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FASTQ data to high-confidence variant calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics. 2013;43(1). doi: 10.1002/0471250953.bi1110s43
[26] Layer RM, Chiang C, Quinlan AR, Hall IM. Lumpy: A probabilistic framework for structural variant discovery. Genome Biology. 2014;15(6). doi: 10.1186/gb-2014-15-6-r84
[27] Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-wide copy number detection and visualization from targeted DNA sequencing. PLOS Computational Biology. 2016;12(4). doi: 10.1371/journal.pcbi.1004873
[28] Stephens ZD, Hudson ME, Mainzer LS, Taschuk M, Weber MR, Iyer RK. Simulating next-generation sequencing datasets from empirical mutation and sequencing models. PLOS ONE. 2016;11(11). doi: 10.1371/journal.pone.0167047
[29] Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. Bioinformatics. 2014;30(14):2076–8. doi: 10.1093/bioinformatics/btu168
[30] Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology. 2013 Mar;31(3):213-9. doi: 10.1038/nbt.2514
[31] Van der Auwera GA, O'Connor BD. Genomics in the cloud: using Docker, GATK, and WDL in Terra. O'Reilly Media; 2020 Apr 2.
[32] Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, De La Vega FM. An open resource for accurately benchmarking small variant and reference calls. Nature biotechnology. 2019 May;37(5):561-6. doi: 10.1038/s41587-019-0074-6
[33] Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O, Jáspez D. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. Cell Genomics. 2022 May 11;2(5):100129. doi: 10.1016/j.xgen.2022.100129
[34] Oliva A, Tobler R, Cooper A, Llamas B, Souilmi Y. Systematic benchmark of ancient DNA read mapping. Briefings in Bioinformatics. 2021 Sep;22(5):bbab076. doi: 10.1093/bib/bbab076