World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

In silico neoantigen prediction in glioblastoma for personalised immunotherapy

Ananya Nitin Jadhav * and Elamathi Natarajan

Biotecnika, Bengaluru, India.

World Journal of Advanced Research and Reviews, 2026, 30(03), 2150-2156

Publication history: Received on 26 June 2026; revised on 27 June 2026; accepted on 29 June 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.30.3.1789

Abstract

Glioblastoma multiforme is a type IV glioma and is one of the most fatal cancers affecting the Brain and the Spinal cord. Even with medications, patients only survive for about 15 months. Therefore, personalised immunotherapy has become an avid area of research. Neoantigen based vaccines are gaining further recognition because herein the immune system specifically identifies cancer cells and not the healthy ones. Neoantigens are peptides that form within cancer cells when they undergo somatic mutations. These neoantigens then bind to an HLA molecule causing them to be picked up by a T cell which then triggers an immune response.

In this article I showcase a bioinformatic pipeline to identify peptides that can be further analysed for personalised immunotherapy in glioblastoma patients. A publicly available RNA sequencing data from NCBI SRA database was used for this project. The pipeline includes quality control, read alignment, variant calling, annotation, HLA typing and binding prediction. The neoantigen binding predictions were carried out using HLA-A*02:01.

Out of all the peptides analysed, VSDPGQLEHV showed the strongest binding to HLA-A*02:01 and a percentile rank of 2.4. Overall, this article gives a workflow that can be followed to identify neoantigens and therefore showcases the importance of computational analysis in personalised immunotherapeutics.

Keywords: Glioblastoma; Neoantigens; Immunotherapy; Bioinformatics; HISAT2; GATK Mutect2; IEDB

1. Introduction

Glioblastoma multiforme is one of the most aggressive and common tumours found in adults. The glioma tumour belongs to the cold tumour group meaning it has a low immune cell count. Even after years of study in chemotherapy, there is not much progress in survival rates. [1] There are several reasons for this. The glioblastoma tumour shows a lot of variation within the tumour mass. The tumour is also known to adapt quickly once the treatment begins. One more reason is that the tumour grows in an infiltrative manner, that is, it spreads into the brain tissue such that it becomes almost impossible to surgically remove the tumour. The blood brain barrier also prevents certain drugs from reaching the tumour site. [3][4] Therefore, active research is going on in personalised approaches that target tumour specific antigens.

Neoantigens are peptides that are specific to the tumour caused by somatic mutations within the cancer cells. Neoantigens unlike tumour associated antigens are completely foreign to the body and therefore are perfect for eliciting an anti-tumour immune response. These mutant peptides are processed inside the cell and then displayed on the surface alongside MHC I molecules. The CD8+ cytotoxic T cells then recognise them as non self and kill them. The clinical relevance of neoantigens can be backed by several evidences across different types of cancers. In melanoma and non-small cell lung cancer, higher neoantigen count is associated with better response to immune checkpoint blockade therapy. Personalised neoantigen vaccines have performed well in the early trials of melanoma and glioblastoma

* Corresponding author: Ananya Jadhav.

showcasing that these vaccines are able to elicit a neoantigen specific T cell response. [5] Glioblastoma is especially hard to treat because all of the mutated peptides displayed on the outside of the cancer cells have a unique immune profile. These peptides are formed due to driver mutations in genes such as EGFR, PTEN, TP53. [6] Besides driver mutations there are passenger mutations which are larger, much more variable and differ drastically from person to person and sometimes also between different regions within the same tumour. Therefore, personalised neoantigen immunotherapy along with a multi-region targeting strategy is used to tackle glioblastoma. [7][8]

The computational workflow for identifying neoantigens begins with next generation sequencing data which can be Whole exome sequencing (WES) data or RNA sequencing data. They are used to detect somatic mutations which are then annotated for their functional effects. The mutations that cause a change in the amino acid sequence, that is, cause missense mutations are selected to build peptides. Then tools like IEDB are used to rank the peptides which show highest binding affinity with patient specific HLA alleles. RNA seq is used as it has a major advantage of offering both the mutation as well as the gene expression data at the same time. This helps prioritise peptides that are both mutated and transcribed within the tumour. This is important because peptides can only be displayed on the cell surface if the gene is being expressed. [9][10][11]

For this project, a publicly available anonymised Glioblastoma RNA-seq data was used to check if biologically usable neoantigen candidates could be predicted.

1.1. Objectives

The objectives of this project were as follows:

- To collect raw data from RNA seq data from the NCBI SRA database.
- To perform quality control and adapter trimming of the sequencing reads using Falco and Trim Galore.
- To align the processed reads to the human reference genome (GRCh37) using HISAT2.
- To perform variant calling using GATK Mutect2.
- To identify variants using SnpEff.
- To perform HLA allele using OptiType.
- To generate peptides from missense variants and predict binding affinity towards MHC I using IEDB.

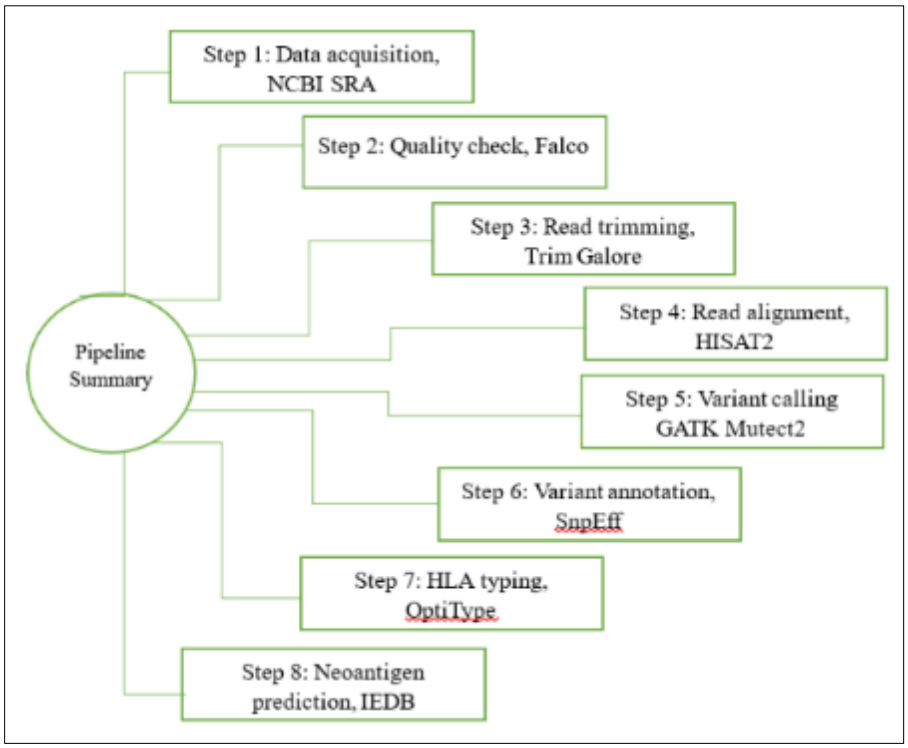


Figure 1 An overview of the computational workflow for neoantigen prediction

2. Methodology

2.1. Data acquisition

The data was acquired by downloading the RNA seq data of glioblastoma patients from NCBI SRA. This is a public database that presents high throughput sequencing data submitted by researchers from around the world. The raw FASTQ files were obtained using the SRA Toolkit (fastq-dump). The dataset had paired end RNA seq reads sequenced on an Illumina platform therefore it gave two FASTQ files (forward and reverse reads) which matched the requirements of the downstream bioinformatic tools. Since the data used was publicly available and anonymised beforehand, no ethical clearance was required for this project.

2.2. Quality control

The quality assessment of the FASTQ files was done using Falco. This is a tool that is widely used for assessing the quality of high throughput data. [12] The following criteria were checked: Per base sequence quality scores with Q30 used as threshold of acceptable quality. Q30 is the error probability of approximately 1 in 1000 bases. [13] Per sequence GC content was assessed against the theoretical GC distribution for the human transcriptome. Sequence duplication analysis was done to check the presence of duplicated reads. Adapter contamination and overrepresentation of sequences was checked.

2.3. Read trimming

Adapter sequences and low-quality bases were removed using Trim Galore. This is a tool that combines Cutadapt algorithm with FastQC quality reporting. [14] Trimming was performed keeping the default settings of base quality threshold Q20, minimum read length of 20 base pairs and automatic adapter detection.

2.4. Read alignment

The trimmed reads were aligned to the human reference genome (GRCh37) using HISAT2. This tool is a graph-based data structure specifically designed for alignment of sequences showing alternatively spliced transcripts. [15] HISAT2 uses a hierarchical index scheme that is based on the FM index which helps in rapid and more accurate alignment. Paired end mode was used for alignment and the SAM files received were converted to sorted and indexed BAM form using SAMtools. BAM format is used by GATK because of its smaller size and as it can be. [16]

2.5. Variant calling

Somatic variant calling was done using GATK Mutect2. Mutect2 is used to detect somatic nucleotide variants (SNVs) and small indels (insertion/deletions) from tumour sequencing data. [17] Mutect2 was run in tumour only form because the matching normal RNA seq dataset was not present.

2.6. Variant annotation

Annotation of the VCF files was done using SnpEff. This tool is mainly used for annotating variants based on their genomic location and then it predicts coding effects such as frameshifts or start/stop codon losses/gains. [18] Each variant was annotated as missense variant, synonymous variant, frameshift variant along with a predicted classification of high, moderate, low or modifier.

2.7. HLA typing

HLA typing was attempted using OptiType and Seq2HLA however both tools failed to give results. The errors were due to technical issues. Additionally, WES data is more accurate for HLA typing than RNA seq data. [19] Therefore, the allele HLA-A*02:01 which is said to be the most common class I allele and widely used in immunological research was taken for the further IEDB analysis. [20]

2.8. Neoantigen prediction

The mutant peptides were built manually from the missense annotated variants. The amino acids surrounding the mutation were noted and a single amino acid substitution as caused by the mutant allele was introduced to form the 10 amino acid long mutant peptide. To check the binding affinity of these mutant peptides to the MHC-I, IEDB tool was used. The binding affinity was given in the form of a percentile rank. The lower the percentile rank, better is the binding affinity. [21]

3. Results and discussion

Quality assessment on the raw reads showed good quality as all the data points were above the Q30 threshold. The dataset had 20426433 sequencing reads of 51-64 sequence length. Per sequence GC content was analysed which was found to be 40% and even though a warning was given the GC pattern was similar to the expected pattern therefore no GC contamination was seen. Sequence duplication analysis showed a total duplicated percentage of 50.8668 % and a warning was given but this is commonly observed in RNA seq experiments because highly expressed transcripts generate multiple identical sequencing reads. [22] Although adapter contamination was seen, after the use of Trim Galore the adapter sequences were removed along with low quality bases. No overrepresented sequences were seen. The sequences were then ready for alignment.

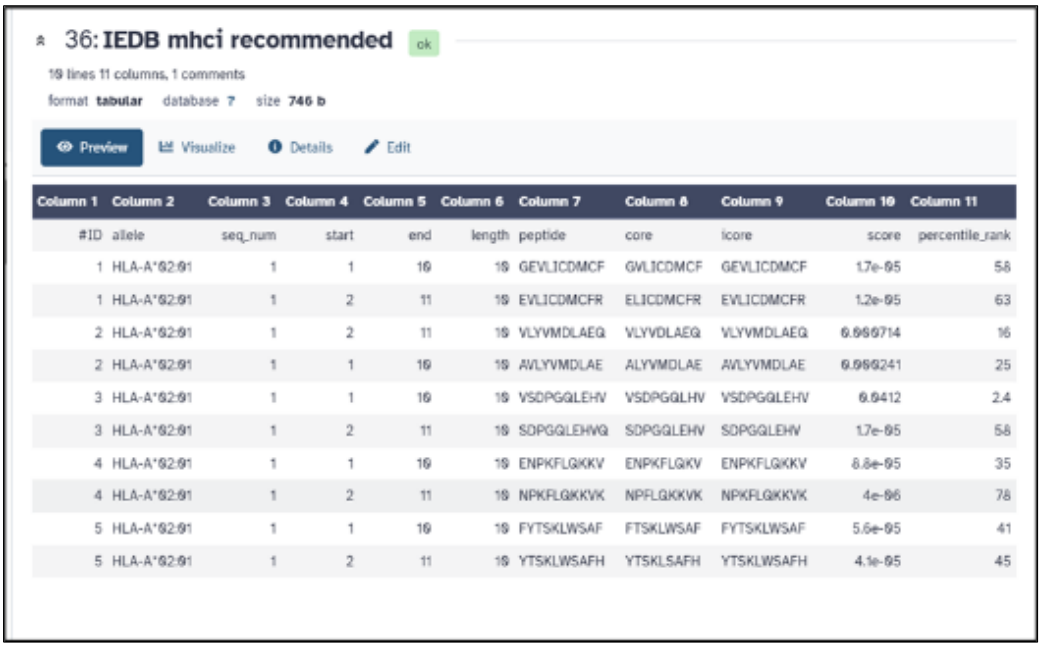
HISAT2 was used to align the sequences to GRCh37(hg19). The results were viewed in Integrative Genomics Viewer to check the mapping quality. Most reads aligned well with the reference genome showing a total alignment rate of 87.17%.

Somatic variants were identified using GATK Mutect2. The VCF file showed the presence of single nucleotide variants (SNVs) and indels (insertion/deletions) in the sample. SnpEff was used to annotate the variants. Out of 142,095 variants, SnpEff identified 3209 high impact variants, 33528 moderate impact variants and 13,006 low impact variants. Functional classification showed 33,785 missense variants, 1,824 nonsense variants and 12,552 silent variants. For this project only the variants that were classified as missense mutations with moderate impact were chosen. This is because missense mutations do not derange the reading frame while altering the amino acid sequence. These mutations cause tumour specific peptides to differ from wild type sequences by a single amino acid. Therefore, the variants annotated as missense with a moderate functional impact were used to form mutated peptides for further analysis.

As OptiType and Seq2HLA failed to give results, HLA-A*02:01 was used since it is the most common HLA class I allele.[20] MHC I binding predictions were checked against this allele. Lastly the mutant peptides that were constructed manually were analysed using IEDB tool. The variant VSDPGQLEHV showed the lowest percentile rank of 2.4 and is therefore the one showing the strongest binding to the MHC I allele. This result makes it a good neoantigen candidate for further investigation.

The study revealed that getting a candidate neoantigen from RNA seq data using a bioinformatics pipeline is feasible. The variant VSDPGQLEHV showed a percentile rank of 2.4. This is significant because when a sequence is submitted to IEDB, it is broken down into multiple peptides and the binding affinity for each is determined. Since the binding affinity is compared to a large set of random peptides, lower rank means that there are very few other peptides that bind better than the peptide under study. [21] Therefore, a percentile rank of 2.4 indicates that VSDPGQLEHV peptide binds stronger than 97.6% of all random peptides. So even though a rank of <=1% is considered a strong binder, a rank of 2.4 proves to be a promising candidate for more research.

The study is consistent with the previous studies that show neoantigen based vaccines to be promising targets in several cancer types including melanoma and non-small cell lung cancer, in which a greater number of mutations in the tumour provide many neoantigens. Studies have shown that personalised neoantigens elicit tumour specific responses which in turn helps in immunotherapy. In glioblastoma, although there is lower immunogenicity and high immunosuppressive microenvironment, neoantigens are proving to be a good therapeutic potential. [1] All these studies have emphasized on a computational workflow to predict neoantigens as therapeutic targets. Accordingly, this current project has identified the peptide VSDPGQLEHV as a potential candidate neoantigen with sufficient binding abilities. Therefore, this suggests that a bioinformatics-based approach can be used to discover new immunotherapeutic targets for glioblastoma.



36: IEDB mhci recommended ok

19 lines 11 columns, 1 comments

format tabular database ? size 746 b

Preview | Visualize | Details | Edit

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11
#ID	allele	seq_num	start	end	length	peptide	core	icore	score	percentile_rank
1	HLA-A*02:01	1	1	19	19	GEVLICDMCF	GVLICDMCF	GEVLICDMCF	1.7e-05	58
1	HLA-A*02:01	1	2	11	19	EVLICDMCFR	ELICDMCFR	EVLICDMCFR	1.2e-05	63
2	HLA-A*02:01	1	2	11	19	VLYVMDLAEQ	VLYVDLAEQ	VLYVMDLAEQ	6.6e-06	16
2	HLA-A*02:01	1	1	19	19	AVLYVMDLAE	ALYVMDLAE	AVLYVMDLAE	6.9e-06	25
3	HLA-A*02:01	1	1	19	19	VSDPGQLEHV	VSDPGQLHV	VSDPGQLEHV	6.9e-04	2.4
3	HLA-A*02:01	1	2	11	19	SDPGQLEHVQ	SDPGQLEHV	SDPGQLEHV	1.7e-05	58
4	HLA-A*02:01	1	1	19	19	ENPKFLQKKV	ENPKFLQKV	ENPKFLQKKV	8.8e-05	35
4	HLA-A*02:01	1	2	11	19	NPKFLQKKVK	NPFLQKKVK	NPKFLQKKVK	4e-06	78
5	HLA-A*02:01	1	1	19	19	FYTSKLWSAF	FTSKLWSAF	FYTSKLWSAF	5.6e-05	41
5	HLA-A*02:01	1	2	11	19	YTSKLWSAFH	YTSKLSAFH	YTSKLWSAFH	4.1e-05	45

Figure 2 10 Peptides ranked by IEDB.

Limitations

Although this project used RNA seq data and was successful in giving a potential neoantigen, a whole exome sequencing data is more favoured for variant calling since it gives more coverage. [19] RNA seq data was used here as it gives the mutational and gene expression data simultaneously. This ensures peptide presentation on the cell surface which is especially useful in glioblastoma where the tumour is varying in different parts within the same tumour. However, RNA seq presents problems too. Allele specific expression, RNA editing and varying expression levels is what makes RNA seq tricky. So, while it worked well in this project, more investigation and validation is needed for future projects. [23]

Another limitation of this study is the absence of a patient specific HLA allele. Due to technical limitations, OptiType and Seq2HLA both failed to give functional results. Therefore, a more general and extensively distributed HLA-A*02:01 allele was used for downstream analysis. So even though VSDPGQLEHV peptide showed a strong binding to HLA-A*02:01, this study cannot certainly confirm that it is a strong binder for the actual patient. Therefore, it is a promising candidate but without the patient specific HLA typing, its biological significance is unresolved.

Lastly, this study is not experimentally validated for neoantigen prediction. The binding of peptide to the HLA allele was predicted only by computational methods and no in vivo or in vitro experiments were done to confirm the neoantigen as a therapeutic target. Therefore, more rigorous biological experimentation will be required to verify the capabilities of the neoantigen.

4. Conclusion

This project showcased a bioinformatics workflow to predict neoantigens from a glioblastoma RNA seq data. From raw sequencing reads and working through a set of protocols that included quality assessment, read trimming, alignment, variant calling, annotation and binding prediction, a small set of resulting peptides were obtained that displayed binding affinity towards HLA-A*02:01.

Out of these, VSDPGQLEHV depicted a percentile rank of 2.4 therefore showcasing a satisfactorily strong binding with HLA-A*02:01. Future lab-based experimentation on this peptide can give promising results. This project hence supports the current ongoing research in neoantigens as a new contributor to advanced immunotherapeutics. It is directly aligned with how personalised immunotherapies are developed and as technology and computational tools improve, these kinds of bioinformatic pipelines will become more significant in understanding and eventually treating cancers like glioblastoma.

Compliance with ethical standards

Acknowledgments

This study was supported by Biotecnika, India.

Disclosure of conflict of interest

The authors declare there was no financial or non-financial conflict of interest in completing this study.

Statement of ethical approval

Since the data was collected from a publicly available source and was anonymised, there are no ethical issues.

Statement of informed consent

Since the data was collected from a publicly available source and was anonymized, the informed consent requirement was not needed.

References

[1] Segura-Collar B, Hiller-Vallina S, de Dios O, et al. Advanced immunotherapies for glioblastoma: tumor neoantigen vaccines in combination with immunomodulators. *Acta Neuropathol Commun*. 2023;11(1):79. Published 2023 May 10. doi:10.1186/s40478-023-01569-y

[2] Hu Y, Zhou T, Cai P, He Z. Neoantigens: new hope for cancer therapy. *Front Oncol*. 2025;15:1531592. Published 2025 Mar 11. doi:10.3389/fonc.2025.1531592

[3] Yalamarty SSK, Filipczak N, Li X, et al. Mechanisms of Resistance and Current Treatment Options for Glioblastoma Multiforme (GBM). *Cancers (Basel)*. 2023;15(7):2116. Published 2023 Apr 1. doi:10.3390/cancers15072116

[4] Cruz JVR, Batista C, Afonso BH, et al. Obstacles to Glioblastoma Treatment Two Decades after Temozolomide. *Cancers (Basel)*. 2022;14(13):3203. Published 2022 Jun 30. doi:10.3390/cancers14133203

[5] Fang X, Guo Z, Liang J, Wen J, Liu Y, Guan X, Li H. Neoantigens and their potential applications in tumor immunotherapy. Oncol Lett. 2022 Mar;23(3):88. doi: 10.3892/ol.2022.13208. Epub 2022 Jan 21. PMID: 35126730; PMCID: PMC8805178.

[6] Yang C, Tan Y, Li S, Zhou J, Wang Q, Wang Y, Xie Y, Chen L, Li J, Fang C, Kang C. Genomic landscapes by multiregion sequencing combined with circulation tumor DNA detection contribute to molecular diagnosis in glioblastomas. Aging (Albany NY). 2019 Dec 10;11(23):11224-11243. doi: 10.18632/aging.102526. Epub 2019 Dec 10. PMID: 31822636; PMCID: PMC6932900.

[7] Johanns TM, Garfinkle EAR, Miller KE, Livingstone AJ, Roberts KF, Rao Venkata LP, Dowling JL, Chicoine MR, Dacey RG, Zipfel GJ, Kim AH, Mardis ER, Dunn GP. Integrating Multisector Molecular Characterization into Personalized Peptide Vaccine Design for Patients with Newly Diagnosed Glioblastoma. Clin Cancer Res. 2024 Jul 1;30(13):2729-2742. doi: 10.1158/1078-0432.CCR-23-3077. PMID: 38639919; PMCID: PMC11215407.

[8] Tarabini RF, Fioravanti Vieira G, Rigo MM, de Souza APD. Mutations in glioblastoma proteins do not disrupt epitope presentation and recognition, maintaining a specific CD8 T cell immune response potential. Sci Rep. 2024 Jul 19;14(1):16721. doi: 10.1038/s41598-024-67099-2. PMID: 39030304; PMCID: PMC11271619.

[9] Hu H, Xiong Y, Lu D, Sun W, Su X, Mo D, Chen L, Wang G, Wang J, Zhang X, Lu M, Huang G. RNA sequencing enables neoantigen discovery and vaccine validation in breast and lung cancer. Front Immunol. 2025 Oct 8;16:1682312. doi: 10.3389/fimmu.2025.1682312. PMID: 41132668; PMCID: PMC12540381.

[10] Rubinsteyn A, Kodysh J, Hodes I, Mondet S, Aksoy BA, Finnigan JP, Bhardwaj N, Hammerbacher J. Computational Pipeline for the PGV-001 Neoantigen Vaccine Trial. Front Immunol. 2018 Jan 18;8:1807. doi: 10.3389/fimmu.2017.01807. PMID: 29403468; PMCID: PMC5778604.

[11] Wert-Carvajal C, Sánchez-García R, Macías JR, Sanz-Pamplona R, Pérez AM, Alemany R, Veiga E, Sorzano CÓS, Muñoz-Barrutia A. Predicting MHC I restricted T cell epitopes in mice with NAP-CNB, a novel online tool. Sci Rep. 2021 May 24;11(1):10780. doi: 10.1038/s41598-021-89927-5. PMID: 34031450; PMCID: PMC8144223.

[12] de Sena Brandine G, Smith AD. Falco: high-speed FastQC emulation for quality control of sequencing data. F1000Res. 2019 Nov 7;8:1874. doi: 10.12688/f1000research.21142.2. PMID: 33552473; PMCID: PMC7845152.

[13] Tan G, Opitz L, Schlapbach R, Rehrauer H. Long fragments achieve lower base quality in Illumina paired-end sequencing. Sci Rep. 2019 Feb 27;9(1):2856. doi: 10.1038/s41598-019-39076-7. PMID: 30814542; PMCID: PMC6393434.

[14] Overbey EG, Saravia-Butler AM, Zhang Z, Rathi KS, Fogle H, da Silveira WA, Barker RJ, Bass JJ, Beheshti A, Berrios DC, Blaber EA, Cekanaviciute E, Costa HA, Davin LB, Fisch KM, Gebre SG, Geniza M, Gilbert R, Gilroy S, Hardiman G, Herranz R, Kidane YH, Kruse CPS, Lee MD, Liefeld T, Lewis NG, McDonald JT, Meller R, Mishra T, Perera IY, Ray S, Reinsch SS, Rosenthal SB, Strong M, Szewczyk NJ, Tahimic CGT, Taylor DM, Vandenbrink JP, Villacampa A, Weging S, Wolverton C, Wyatt SE, Zea L, Costes SV, Galazka JM. NASA GeneLab RNA-seq consensus pipeline: standardized processing of short-read RNA-seq data. iScience. 2021 Mar 26;24(4):102361. doi: 10.1016/j.isci.2021.102361. PMID: 33870146; PMCID: PMC8044432.

[15] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019 Aug;37(8):907-915. doi: 10.1038/s41587-019-0201-4. Epub 2019 Aug 2. PMID: 31375807; PMCID: PMC7605509.

[16] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. Epub 2010 Jul 19. PMID: 20644199; PMCID: PMC2928508.

[17] López-Cade I, Gómez-Sanz A, Sanvicente A, Díaz-Tejeiro C, Manzano A, Pérez-Segura P, Győrffy B, Ocaña A, de la Hoya M, García-Barberán V. Comparative Evaluation of Mutect2, Strelka2, and FreeBayes for Somatic SNV Detection in Synthetic and Clinical Whole-Exome Sequencing Data. Biomolecules. 2025 Oct 30;15(11):1532. doi: 10.3390/biom15111532. PMID: 41301450; PMCID: PMC12650410.

[18] Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;6(2):80-92. doi:10.4161/fly.19695

[19] Yi J, Chen L, Xiao Y, Zhao Z, Su X. Investigations of sequencing data and sample type on HLA class Ia typing with different computational tools. *Brief Bioinform*. 2021;22(3):bbaa143. doi:10.1093/bib/bbaa143

[20] Chen KY, Liu J, Ren EC. Structural and functional distinctiveness of HLA-A2 allelic variants. Immunol Res. 2012 Sep;53(1-3):182-90. doi: 10.1007/s12026-012-8295-5. PMID: 22434516.

[21] Fleri W, Paul S, Dhanda SK, Mahajan S, Xu X, Peters B, Sette A. The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design. Front Immunol. 2017 Mar 14;8:278. doi: 10.3389/fimmu.2017.00278. PMID: 28352270; PMCID: PMC5348633.

[22] Deschamps-Francoeur G, Simoneau J, Scott MS. Handling multi-mapped reads in RNA-seq. Comput Struct Biotechnol J. 2020 Jun 12;18:1569-1576. doi: 10.1016/j.csbj.2020.06.014. PMID: 32637053; PMCID: PMC7330433.

[23] Sarita Maurya, Chakresh Kumar Jain, A comprehensive review of variant calling tools for RNA-seq: Challenges and advances, Mutation Research - Reviews in Mutation Research, Volume 798, 2026, 108591, ISSN 1383-5742, https://doi.org/10.1016/j.mrrev.2026.108591.