Bioinformatics, 2023, **39(11)**, btad659
https://doi.org/10.1093/bioinformatics/btad659
Advance Access Publication Date: 26 October 2023
Applications Note

Genome analysis

ScanNeo2: a comprehensive workflow for neoantigen detection and immunogenicity prediction from diverse genomic and transcriptomic alterations

Richard A. Schäfer [1], **Qingxiang Guo** [1], **Rendong Yang** [1,2,*]

[1]Department of Urology, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, United States
[2]Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, United States

*Corresponding author. Department of Urology, Northwestern University Feinberg School of Medicine, 303 E Superior St, Chicago, IL 60611, United States. E-mail: rendong.yang@northwestern.edu (R.Y.)

Associate Editor: Christina Kendziorski

Abstract

Motivation: Neoantigens, tumor-specific protein fragments, are invaluable in cancer immunotherapy due to their ability to serve as targets for the immune system. Computational prediction of these neoantigens from sequencing data often requires multiple algorithms and sophisticated workflows, which are currently restricted to specific types of variants, such as single-nucleotide variants or insertions/deletions. Nevertheless, other sources of neoantigens are often overlooked.

Results: We introduce `ScanNeo2` an improved and fully automated bioinformatics pipeline designed for high-throughput neoantigen prediction from raw sequencing data. Unlike its predecessor, ScanNeo2 integrates multiple sources of somatic variants, including canonical- and exitron-splicing, gene fusion events, and various somatic variants. Our benchmark results demonstrate that `ScanNeo2` accurately identifies neoantigens, providing a comprehensive and more efficient solution for neoantigen prediction.

Availability and implementation: `ScanNeo2` is freely available at https://github.com/ylab-hi/ScanNeo2/ and is accompanied by instruction and application data.

1 Introduction

Neoantigens are foreign protein fragments that originate from genetic mutations within cancer cells and are entirely absent in normal tissue. When presented on major histocompatibility complex (MHC) molecules, these neoepitopes can be recognized by $CD4^+$ or $CD8^+$ T-cells. This can trigger an anti-cancer immune response in patients. However, cancer cells have developed resistance to anti-cancer immunity (Zhu *et al.* 2021). This effect can be reversed by cancer immunotherapies, which e.g. improve the presentation of neoepitopes. For that, tumor-specific neoantigens need to be identified to improve adoptive T-cell therapies. In the past, `ScanNeo` (Wang *et al.* 2019) has been developed for detecting insertion and deletion (indel)-derived neoantigens and later combined with `ScanExitron` (Wang *et al.* 2021) to detect neoantigens from exitron-derived events (Wang and Yang 2021). While several tools have been created to identify neoantigens from sequencing data, most focus mainly on finding neoantigens that come from single-nucleotide variants (SNVs), or indel events. But, they often overlook other sources of neoantigens like gene fusions or alternative splicing. In that regard, `pVACtools` (Hundal *et al.* 2020) provides a toolkit for neoantigen prediction and visualization, but requires a list of identified non-synonymous somatic variations identified by other pipelines. However, this requires the integration of multiple tools that need to be coordinated with one another. This can be simplified using workflow management systems, such as the `Nextflow` workflow language (Di Tommaso *et al.* 2017) or the `snakemake` workflow management system (Mölder *et al.* 2021). For instance, `nextNEOpi` (Rieder *et al.* 2022) presents a comprehensive workflow implemented in `Nextflow` but uses most routines from `pVACtools`. Features of existing workflows for the prediction of neoantigens are listed in Supplementary Table S1. To better address the complexity of coordinating tools and overcome these limitations, we present `ScanNeo2`. This advanced `snakemake` workflow is designed to discover neoantigens from a broader array of sources, offering a holistic and streamlined method to predict neoantigens from sequencing data.

2 Results

2.1 Neoantigen detection from multiple sources

To date, various methods have been established for detecting genomic mutations and transcriptomic variants. We took advantage of those tools and developed a `snakemake`-based workflow for the detection of neoantigens from multiple sources, termed `ScanNeo2` (Fig. 1). It takes as input both raw and processed WGS/WES data and/or RNA-seq data from normal/tumor samples. In addition, it also accepts a list of already identified variants in "Variant Call Format."

Received: 31 July 2023; **Revised:** 10 October 2023; **Editorial Decision:** 20 October 2023; **Accepted:** 24 October 2023

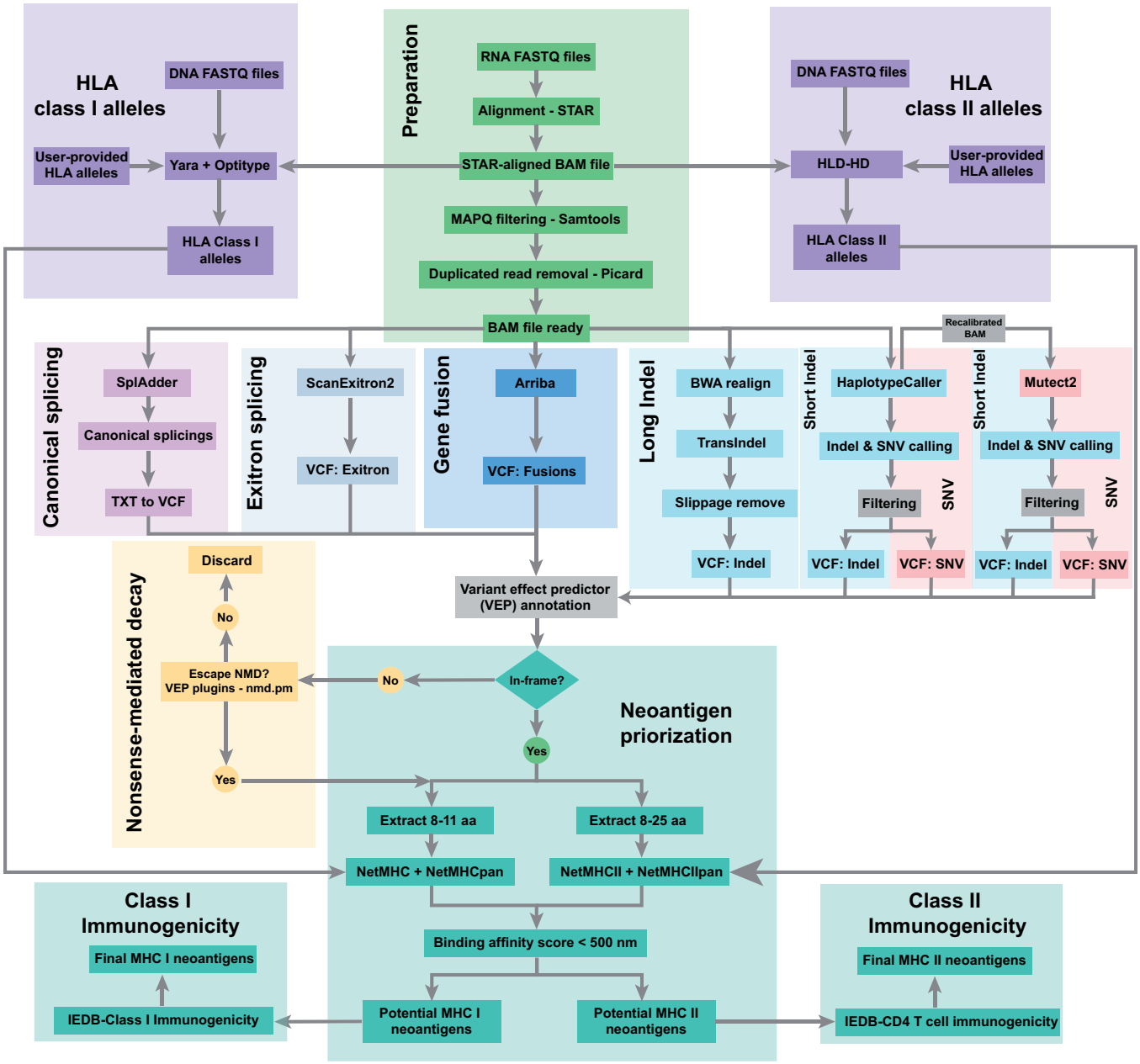

Figure 1. The workflow of ScanNeo2 is divided into three parts. In the first part, the input reads are prepared for the subsequent analysis, which includes the genotyping of the HLA alleles. In the next step, the variation calling is performed using the different modules canonical splicing, exitron, gene fusion, and indel/snp detection. This is concluded with the neoantigen prioritization.

In principle, `ScanNeo2` consists of different steps described in the following. Based on the provided data, the sequencing reads can be pre-processed, which includes quality filtering or removal of adapter sequences. For that, `fastp` (Chen *et al.* 2018) is applied, and the reads are then subjected to the genotyping of HLA alleles and the mutation calling. In the former, DNA-seq and RNA-seq data are filtered for HLA reads using `yara` (Siragusa *et al.* 2013) and subsequently aligned against a panel of HLA class I alleles using `Optitype` (Szolek *et al.* 2014). This results in a list of detected 4-digit HLA genotypes that are later used in the prediction MHC class I antigen presentation. In a similar manner, `HLA-HD` (Kawaguchi *et al.* 2017) is used for HLA class II genotyping. In the following, the RNA-seq data are aligned against the reference genome using `BWA` (Li and Durbin 2009), and subsequently filtered for mapping quality, duplicates, and discarded of unmapped reads. In the mutation calling, the processed reads are subjected to `SplAdder` (Kahles *et al.* 2016) for the detection of alternative splicing events, and `ScanExitron` for identifying exitron-splicing events. In addition, `GATK` (Van der Auwera and O'Connor 2020) is utilized for detection of short indels, as well as SNVs. Here, the modules `mutect2`, and `haplotypecaller` are used for identifying somatic and germline variants, respectively. This is combined with `transIndel` (Yang *et al.* 2018) to detect long variants. `ScanNeo2` also incorporates gene fusion events. For that, the aligned reads are first realigned using `STAR` (Dobin *et al.* 2013), and then subjected to `Arriba` (Dobin *et al.* 2013). For the individual

outputs of the respective tools, the (VCF) is used to determine the downstream effects of the variants. In the case of a frameshift, only the transcripts are further considered that escape nonsense-mediated decay. The next step is to extract the amino acid sequence of the transcripts surrounding the variant. For that, peptides of different lengths are extracted that span the variants. This include peptides of length 8–11 and 13–25 amino acids, corresponding to MHC class I and MHC class II peptides, respectively. In the following, the Immune Epitope Database and Analysis Resource (IEDB) (Dhanda *et al.* 2019) is used to predict the T-cell epitope for MHC class I and II as well as the immunogenicity (Calis *et al.* 2013). This results in a tab-delimited list of predicted neoantigens.

2.2 Benchmarking using available datasets

`ScanNeo2` was assessed using the `TESLA` (Wells *et al.* 2020) dataset that contains data obtained from subjects ($n = 8$) in metastatic melanoma and non-small cell lung cancer. In that study, the authors tested 608 peptides for immunogenicity, of which 37 were found to be immunogenic. We analyzed the data using `ScanNeo2` (see Supplementary Data for details), which included the data preparation (Supplementary Table S2) and identified in each patient 9152–96 819 putative HLA-binding peptides that account for unique peptides (Supplementary Table S3). Among those, 5671–34 981 candidate neoantigens originate from a single source, such as alternative splicing, exitron-splicing, gene fusion, indels, or SNVs. `ScanNeo2` detects 35 of the 37 experimentally validated immunogenic peptides but also captures 476 non-immunogenic peptides. In the following, we applied `ScanNeo2` with more stringent parameters (immunogenicity score ≥ 0.5, TPM ≥ 2, ranking score ≥ 1000, and VAF ≥ 0.02), which reduces the number of non-immunogenic peptides to 391 while retaining 34 of the validated peptides (Supplementary Table S4). Similar settings in `nextNEOp` detects less immunogenic peptides. In addition, we looked at the runtime and memory requirements of `ScanNeo2`. In comparison to `nextNEOpi`, `ScanNeo2` requires on average ~22.5% more CPU time. The main reason for that is the extensive variant calling in which multiple sources are used with high sensitivity.

3 Conclusion

We introduce `ScanNeo2`, a comprehensive `snakemake`-based pipeline for predicting tumor neoepitopes from sequencing data. It has been implemented in `snakemake` to ensure ease of installation, usage as well as high portability. In contrast to other workflows, `ScanNeo2` incorporates multiple sources, thereby providing a means to decipher the neoantigen landscape to an unprecedented degree.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported in part by funds from the National Institutes of Health [NIH/NCI: R01CA259388].

Data availability

WES and RNA-seq data used in this article are available in the Synapse platform, at https://www.synapse.org/#!Synapse:syn21048999.

References

Calis JJA, Maybeno M, Greenbaum JA *et al.* Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput Biol* 2013;**9**:e1003266.

Chen S, Zhou Y, Chen Y *et al.* Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90.

Dhanda SK, Mahajan S, Paul S *et al.* IEDB-AR: immune epitope database—analysis resource in 2019. *Nucleic Acids Res* 2019;**47**:W502–6.

Di Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;**35**:316–9.

Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21.

Hundal J, Kiwala S, McMichael J *et al.* pVACtools: a computational toolkit to identify and visualize cancer neoantigens. *Cancer Immunol Res* 2020;**8**:409–20.

Kahles A, Ong CS, Zhong Y *et al.* SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics* 2016;**32**:1840–7.

Kawaguchi S, Higasa K, Shimizu M *et al.* HLA-HD: an accurate HLA typing algorithm for next-generation sequencing data. *Hum Mutat* 2017;**38**:788–97.

Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 2009;**25**:1754–60.

Mölder F, Jablonski KP, Letcher B *et al.* Sustainable data analysis with snakemake. *F1000Res* 2021;**10**:33.

Rieder D, Fotakis G, Ausserhofer M *et al.* nextNEOpi: a comprehensive pipeline for computational neoantigen prediction. *Bioinformatics* 2022;**38**:1131–2.

Siragusa E, Weese D, Reinert K *et al.* Fast and accurate read mapping with approximate seeds and multiple backtracking. *Nucleic Acids Res* 2013;**41**:e78.

Szolek A, Schubert B, Mohr C *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* 2014;**30**:3310–6.

Van der Auwera GA, O'Connor BD. *Genomics in the Cloud*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2020.

Wang T-Y, Liu Q, Ren Y *et al.* A pan-cancer transcriptome analysis of exitron splicing identifies novel cancer driver genes and neoepitopes. *Mol Cell* 2021;**81**:2246–60.e12.

Wang T-Y, Wang L, Alam SK *et al.* ScanNeo: identifying indel-derived neoantigens using RNA-Seq data. *Bioinformatics* 2019;**35**:4159–61.

Wang T-Y, Yang R. Integrated protocol for exitron and exitron-derived neoantigen identification using human RNA-seq data with ScanExitron and ScanNeo. *STAR Protoc* 2021;**2**:100788.

Wells DK, van Buuren MM, Dang KK *et al.*; Tumor Neoantigen Selection Alliance. Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* 2020;**183**:818–34.e13.

Yang R, Van Etten JL, Dehm SM *et al.* Indel detection from DNA and RNA sequencing data with transIndel. *BMC Genomics* 2018;**19**:270.

Zhu S, Zhang T, Zheng L *et al.* Combination strategies to maximize the benefits of cancer immunotherapy. *J Hematol Oncol* 2021;**14**:156.