Bioinformatics, 35(20), 2019, 4159–4161
doi: 10.1093/bioinformatics/btz193
Advance Access Publication Date: 19 March 2019
Applications Note

Genome analysis

ScanNeo: identifying indel-derived neoantigens using RNA-Seq data

Ting-You Wang[1], Li Wang[1], Sk Kayum Alam[1], Luke H. Hoeppner[1] and Rendong Yang[1,2,*]

[1]The Hormel Institute, University of Minnesota, Austin, MN 55912, USA and [2]Masonic Cancer Center, University of Minnesota, Minneapolis, MN 55455, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on November 21, 2018; revised on February 28, 2019; editorial decision on March 12, 2019; accepted on March 15, 2019

Abstract

Summary: Insertion and deletion (indels) have been recognized as an important source generating tumor-specific mutant peptides (neoantigens). The focus of indel-derived neoantigen identification has been on leveraging DNA sequencing such as whole exome sequencing, with the effort of using RNA-seq less well explored. Here we present ScanNeo, a fast-streamlined computational pipeline for analyzing RNA-seq to predict neoepitopes derived from small to large-sized indels. We applied ScanNeo in a prostate cancer cell line and validated our predictions with matched mass spectrometry data. Finally, we demonstrated that indel neoantigens predicted from RNA-seq were associated with checkpoint inhibitor response in a cohort of melanoma patients.

Availability and implementation: ScanNeo is implemented in Python. It is freely accessible at the GitHub repository (https://github.com/ylab-hi/ScanNeo).

Contact: yang4414@umn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Insertion and deletion (indels) in cancer genomes may lead to frameshift that can generate novel open reading frames (ORFs), often referred as neo-ORF (Hacohen *et al.*, 2013). These indels contribute significantly to produce novel, tumor-specific peptides bound by the patient's HLA molecules and potentially highly immunogenic that can be exploited for personalized cancer immunotherapy (Turajlic *et al.*, 2017). To date, multiple studies have successfully discovered tumor-specific neoantigens by utilizing next-generation sequencing data (Carreno *et al.*, 2015; Gubin *et al.*, 2014). DNA sequencing [e.g. whole exome sequencing (WES)] has been the major source for discovering indel-derived neoantigens, as indel detection is primarily based on DNA-Seq. However, detected indels from DNA may not express and thus it is essential to detect the indels directly from RNA-Seq data to get a high confidence neoantigen call set. RNA-Seq has already been utilized to identify gene fusion and alternative splicing-derived neoantigens (Kahles *et al.*, 2018; Smart *et al.*, 2018; Zhang *et al.*, 2017). Here, we developed an open source pipeline, named ScanNeo, for indel-derived neoantigen discovery through analyzing RNA-Seq data. ScanNeo leverages transIndel (Yang *et al.*, 2018) to detect indels from RNA-Seq and identify the indel-derived neoantigens by inferring the HLA alleles and MHC-binding prediction. We applied ScanNeo to analyze cancer cell line and patient samples to demonstrate its utility for identifying neoantigens based on RNA-Seq data and the association of RNA-Seq-derived indel neoantigens with checkpoint immunotherapy response.

2 The ScanNeo pipeline

ScanNeo is composed of three steps: (i) Indel discovery, (ii) annotation and filtering and (iii) neoantigen prediction (Fig. 1).

The first step requires RNA-Seq data in BAM format aligned by a splice-aware aligner (e.g. HISAT2; Kim et al., 2015) as input to call indels. In this step, ScanNeo first removed duplicated reads with Picard tools (http://broadinstitute.github.io/picard/). Next, spliced reads with inferred transcriptional directions but not carrying indels

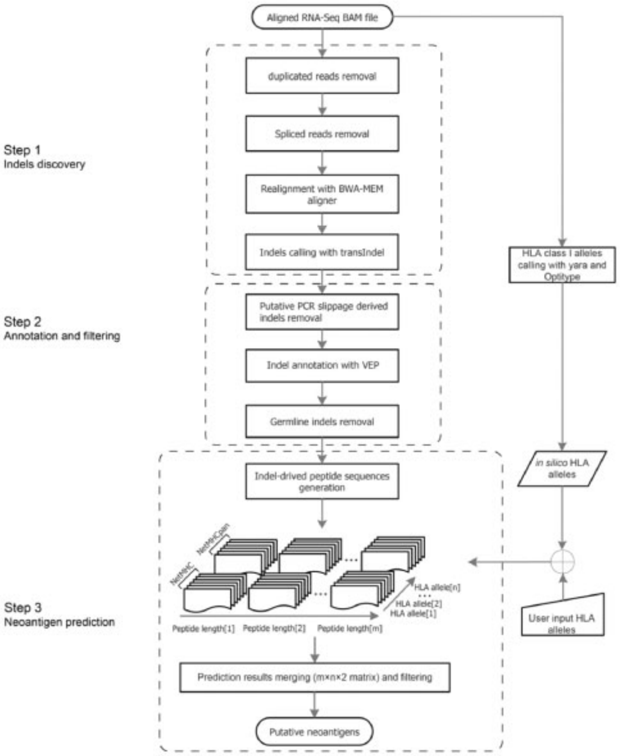

Fig. 1. Workflow of ScanNeo. Starting with aligned RNA-Seq BAM file, ScanNeo consists of three steps—(i) indels discovery, (ii) annotation and filtering and (iii) neoantigen prediction

within their alignment (e.g. 'I' and 'D' in CIGAR string) were removed using sambamba (Tarasov *et al.*, 2015) to avoid the false positive deletion calls that are actually splicing junctions (e.g. exitrons; Marquez et al., 2015).

Retained reads were converted to single-end reads in FASTQ file using BEDTools (Quinlan and Hall, 2010) and re-aligned using BWA-MEM (Li, 2013) which reports chimeric alignments with an 'SA' tag in the alignment records. Finally, ScanNeo applied transIndel (Yang *et al.*, 2018) to determine small to large-sized indels from the RNA-Seq reads. Indel calls were reported in variant call format (VCF) file.

In the second step, ScanNeo first added corresponding reference and alternate allele sequences to each identified indel, and then indels with short tandem repeats and homopolymer will be removed by default; they may result from slippage during polymerase chain reaction amplification. Next, these indels were annotated with variant effect predictor (VEP) (McLaren *et al.*, 2016). Finally, germline indels annotated by Genome Aggregation Database (Lek *et al.*, 2016) and 1000 Genomes project (http://www.internationalgenome.org/) will be removed if their variant allele frequency is larger than an user-defined threshold (by default 0.01).

The third step for neoantigen prediction used VEP annotated indels in VCF format and RNA-Seq read alignment in BAM format as input to predict neoantigens. The HLA class I typing of the input sample was either inferred with an integrated HLA Class I caller, including yara aligner (Siragusa *et al.*, 2013) and Optitype tool (Szolek *et al.*, 2014) or provided by users with their own predicted HLA alleles. We adopted the method used in pVAC-Seq (Hundal *et al.*, 2016) to generate FASTA file of peptides sequences that included wild-type (WT) and mutant (MT) peptide sequences with ten flanking amino acids on each side of the indels (Supplementary Fig. S1A–C). ScanNeo employed NetMHC (Lundegaard *et al.*, 2008) and NetMHCpan (Nielsen and Andreatta, 2016) to predict high-affinity peptides that bind to the HLA Class I molecule with the identified HLA haplotype and generated WT and MT sequences. By default, epitopes with their length in 8–11 amino acids were used to by ScanNeo to calculate their binding affinity score and those with the score <500 nM were reported as neoantigens. The predicted neoantigens were then ranked based on defined metrics including binding affinity, variant allele frequency and agretopicity index value (Supplementary Material).

Each step of ScanNeo was implemented as a standalone module with their own parameters. This allows users to integrate ScanNeo functions into their pipelines. Notably, existing neoantigen prediction pipelines for analyzing sequencing data are very time-consuming. To address this issue, ScanNeo implemented a parallel computing architecture to predict and screen neoantigens for each combination of HLA alleles, peptide k-mer length and MHC-binding prediction algorithms (Fig. 1), which allows evaluating neoepitopes in a faster turnaround time compared with existing neoantigen prediction pipelines (Supplementary Tables S1 and S2), making it applicable for both research and clinical utility.

3 Applications

To validate our pipeline, we first chose to analyze RNA-Seq data of the LNCaP (ATCC, no. CRL-1740) prostate cancer cell line from Cancer Cell Line Encyclopedia (Barretina *et al.*, 2012) to discover indels and neoantigens with ScanNeo. We identified 48 indels and 58 neoantigens at length 11 AA in the LNCaP cell line (Supplementary Dataset S1). Three of the predicted neoantigen were confirmed by the mass spectrometry (MS) data (peptideatlas.org, Accession: PAe003664) of LNCaP cells using SearchGUI (Vaudel *et al.*, 2011) at a false discovery rate of 1%. One of the MS-validated neoantigen was generated by a deletion event identified in RNA-Seq but not identified in matched WES data due to the low coverage in WES data (Supplementary Fig. S2A and B). Moreover, we made a comprehensive comparison of indels identified from RNA-Seq and WES data of LNCaP and found that 11 out of 14 RNA-Seq only indels have supporting evidence from whole genome sequencing data (SRA accession: SRX994870) of LNCaP (Supplementary Fig. S3), suggesting a false discovery rate of 0.06 for indel detected by ScanNeo from RNA-seq.

It has been well known that neoantigen burden is associated with clinical benefit from immune checkpoint inhibitor therapy (Ribas and Wolchok, 2018). We next examined whether indel-derived neoantigens from RNA-Seq identified by ScanNeo might be associated with checkpoint immunotherapy response. We analyzed a cohort of melanoma patients ($n = 27$) with clinical outcomes of immunotherapy (Hugo *et al.*, 2016). Interestingly, neoantigens arising from both indels (RNA-based), generated via our pipeline, and nsSNV (DNA-based), generated via pVAC-Seq, are associated with treatment responses ($P = 0.042$ and 0.045, respectively), but indels exhibited increased neoantigen load than nsSNV (Supplementary Fig. S4, Supplementary Dataset S2). In addition, we also observed that patients who received benefit from immune checkpoint blockade therapy tended to have higher frameshift indels load and neoantigen-yielding indels load than those who did not respond, which is consistent with the notion that frameshift events may be more immunogenic than in-frame mutations as they generated completely novel epitopes (Hacohen *et al.*, 2013).

4 Conclusion

We presented ScanNeo, a computational pipeline that allows identification of indel-derived neoantigens from RNA-Seq. We demonstrated that it can efficiently identify neoantigens from RNA-Seq data in cell lines or patient samples that might be missed from analyzing DNA-Seq data. ScanNeo is a complementary tool to existing DNA-Seq-based neoantigen discovery methods that allow robustly neoantigen predictions from expressed indel mutations.

Funding

This work was supported by Young Investigator Award from the Prostate Cancer Foundation and Research Starter Grant from PhRMA foundation.

Conflict of Interest: none declared.

References

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Carreno,B.M. *et al.* (2015) Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science*, **348**, 803–808.

Gubin,M.M. *et al.* (2014) Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature*, **515**, 577–581.

Hacohen,N. *et al.* (2013) Getting personal with neoantigen-based therapeutic cancer vaccines. *Cancer Immunol. Res.*, **1**, 11–15.

Hugo,W. *et al.* (2016) Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell*, **165**, 35–44.

Hundal,J. *et al.* (2016) pVAC-Seq: a genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.*, **8**, 11.

Kahles,A. *et al.* (2018) Comprehensive analysis of alternative splicing across tumors from 8, 705 patients. *Cancer Cell*, **34**, 211–224.e216.

Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

Lek,M. *et al.* (2016) Analysis of protein-coding genetic variation in 60, 706 humans. *Nature*, **536**, 285–291.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. eprint arXiv:1303.3997, 2013:arXiv:1303.3997.

Lundegaard,C. *et al.* (2008) NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res.*, **36**, W509–W512.

Marquez,Y. *et al.* (2015) Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity. *Genome Res.*, **25**, 995–1007.

McLaren,W. *et al.* (2016) The ensembl variant effect predictor. *Genome Biol.*, **17**, 122.

Nielsen,M. and Andreatta,M. (2016) NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.*, **8**, 33.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Ribas,A. and Wolchok,J.D. (2018) Cancer immunotherapy using checkpoint blockade. *Science*, **359**, 1350–1355.

Siragusa,E. *et al.* (2013) Fast and accurate read mapping with approximate seeds and multiple backtracking. *Nucleic Acids Res.*, **41**, e78.

Smart,A.C. *et al.* (2018) Intron retention is a source of neoepitopes in cancer. *Nat. Biotechnol.*, **36**, 1056–1058.

Szolek,A. *et al.* (2014) OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*, **30**, 3310–3316.

Tarasov,A. *et al.* (2015) Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, **31**, 2032–2034.

Turajlic,S. *et al.* (2017) Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.*, **18**, 1009–1021.

Vaudel,M. *et al.* (2011) SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X! Tandem searches. *Proteomics*, **11**, 996–999.

Yang,R. *et al.* (2018) Indel detection from DNA and RNA sequencing data with transIndel. *BMC Genomics*, **19**, 270.

Zhang,J. *et al.* (2017) INTEGRATE-neo: a pipeline for personalized gene fusion neoantigen discovery. *Bioinformatics*, **33**, 555–557.