neoepiscope improves neoepitope prediction with multi-variant phasing

Mary A. Wood, Austin Nguyen, Adam Struck, Kyle Ellrott, Abhinav Nellore, and Reid F. Thompson
Computational Biology Program, Oregon Health and Science University, Portland, OR, USA

Abstract

Cancer neoepitope prediction depends upon accurate reconstruction of somatic variation and its effects on resulting peptides. **Phasing of germline and somatic variants is an often-neglected aspect of neoepitope prediction.** Co-occurrence of two or more variants within the length of a putative MHC Class I or Class II epitope can affect the resulting peptide sequence. Further, upstream frameshifting variants can change the peptide context of downstream variants in a shared haplotype. Disregarding these phenomena can lead to many false positive and false negative results in neoepitope prediction. We developed neoepiscope to address these issues. This tool also allows for the inclusion of background germline variants for greater patient specificity. **neoepiscope is more performant, flexible, and accurate than alternative neoepitope prediction tools.**

Motivation

- Cancer-specific variants and their corresponding neoepitopes appear central to adaptive anti-tumor immune response[1]
- Numerous neoepitope prediction pipelines exist, with their own **unique sets of features and limitations (See Figure 1)**
- Most neoepitope prediction tools ignore interactions between variants in the same haplotype
- Somatic and germline mutations frequently co-occur[2]
- Somatic mutations may also cluster closely with each other[3]
- Frameshifting indels upstream of a SNV can alter the peptide context, and thus the consequences of the SNV itself
- **We developed neoepiscope to incorporate germline context and address variant phasing for SNVs and indels**
- How do these features affect neoepitope prediction across multiple patient datasets?
- How does neoepiscope perform compared to other tools?

		pVACseq	pVACfuse	INTEGRATE-neo	Epi-Seq	CloudNeo	TSNAD	MuPeXI	neoantigenR	NeoepitopePred	neoantigen_calling (Van Allen)	retained-intron (Van Allen)	Epidisco	antigen.garnish	NeoPredPipe	Neopepsee	Timiner	ScanNeo	neoANT-HILL	Neoepiscope
Variant Phenomena	Missense SNV	✓	X	X	✓	✓	✓	✓	X	✓	✓	X	✓	✓	✓	✓	✓	X	✓	✓
	Nonstop SNV	X	X	X	X	✓	✓	X	X	X	✓	X	✓	X	✓	?	?	X	?	✓
	Nonstart SNV	X	X	X	X	X	X	X	X	X	X	X	X	X	X	?	?	X	?	✓
	Nonsense SNV	X	X	X	X	X	✓	X	X	X	X	X	✓	X	X	?	?	X	?	✓
	In-frame Insertion	✓	X	X	X	X	✓	✓	X	X	✓	X	✓	✓	X	✓	✓	✓	X	✓
	In-frame Deletion	✓	X	X	X	X	✓	✓	X	X	✓	X	✓	✓	X	✓	✓	✓	X	✓
	Frame-shift Insertion	✓	X	X	X	X	✓	✓	X	X	✓	X	✓	✓	X	✓	✓	✓	X	✓
	Frame-shift Deletion	✓	X	X	X	X	✓	✓	X	X	✓	X	✓	✓	X	✓	✓	✓	X	✓
	Fusion	X	✓	✓	X	X	X	X	✓	X	X	X	✓	X	X	X	X	X	X	X
	Retained intron	X	X	X	X	X	X	X	△	X	X	✓	X	X	X	X	X	X	X	X
	Novel splice junction	X	X	X	X	X	X	X	✓	X	X	X	△	X	X	X	X	X	X	X
	Phased variants (DNA-seq)	△	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	✓
	Phased variants (RNA-seq)	X	X	X	△	X	X	X	X	X	X	X	△	X	X	X	X	X	X	X
	Germline context	△	X	X	X	X	X	X	X	X	X	X	△	X	X	X	X	X	X	✓
MHC Binding Prediction	MHC class I peptides	✓	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	MHC class II peptides	✓	✓	X	X	✓	✓	✓	X	X	X	X	✓	X	X	X	X	✓	✓	✓
	NetMHC integration	✓	✓	✓	✓	✓	X	X	X	X	X	X	✓	✓	X	X	X	✓	✓	X
	NetMHCpan integration	✓	✓	X	X	✓	✓	✓	X	X	✓	✓	✓	✓	X	✓	✓	✓	✓	✓
	NetMHCIIpan integration	✓	✓	X	X	X	X	X	X	X	X	X	✓	✓	X	X	X	X	✓	✓
	NetMHCcons integration	✓	✓	X	X	X	X	X	✓	X	X	✓	X	X	X	X	X	✓	✓	X
	MHCflurry integration	X	X	X	X	X	X	X	X	X	X	✓	✓	X	X	X	X	✓	✓	
	MHCnuggets integration	X	X	X	X	X	X	X	X	X	X	X	✓	X	X	X	X	X	✓	
	IEDBtools integration	X	X	X	X	X	X	X	X	X	X	✓	X	X	✓	X	✓	✓	X	
	Other tools integration §	✓	✓	X	X	X	X	X	X	X	X	X	✓	X	✓	X	✓	✓	X	
Code	Installable with package manager	✓	✓	X	X	X	X	X	X	X	X	X	X	X	X	X	✓	X	✓	
	Under active maintenance *	✓	✓	✓	X	✓	✓	✓	✓	✓	X	✓	X	✓	✓	✓	X	✓	✓	✓
	Performant ‡	X	?	?	?	?	✓	✓	?	?	?	?	?	?	✓	?	?	?	?	✓

Figure 1: *A feature comparison of pVACseq and pVACfuse[4], INTEGRATE-neo[5], Epi-Seq[6], CloudNeo[7], TSNAD[8], MuPeXI[9], neoantigenR[10], NeoepitopePred[11], neoantigen_calling_pipeline (Van Allen laboratory)[12], retained-intron-neoantigen-pipeline (Van Allen laboratory)[13], Epidisco[14], antigen.garnish[15], NeoPredPipe[16], Neopepsee[17], Timiner[18], ScanNeo[19], neoANT-HILL[20], and neoepiscope. Each column corresponds to a software tool, with software features listed by row. A green check mark indicates that the tool possesses or processes the indicated feature, while a red "X" indicates that the tool does not possess or process the indicated feature. A yellow warning symbol ( ) indicates that a tool incompletely supports the corresponding feature. Gray question marks ("?") denote unknown or unassessed values. * A tool was considered to be under activate maintenance if a new release or GitHub commit had occurred within 6 months prior to submission of this manuscript. § Other MHC binding prediction tools used include NNalign, PickPocket, SMM, SMMPMBEC, and SMM align for pVACseq and pVACfuse; NetMHCII for antigen.garnish, and NetCTLpan for Neopepsee. ‡ A tool was considered to be performant if neoepitope prediction averaged less than 10 minutes per sample in our benchmarking.*

Interested in using neoepiscope? It's easy to install with pip! Learn more on our GitHub repository and by reading our recently accepted manuscript in *Bioinformatics*:
Repository: https://github.com/pdxgx/neoepiscope
Manuscript: Wood et al., 2019. *Bioinformatics*. http://bit.ly/neoepiscope_manuscript

Co-occurrence of variants

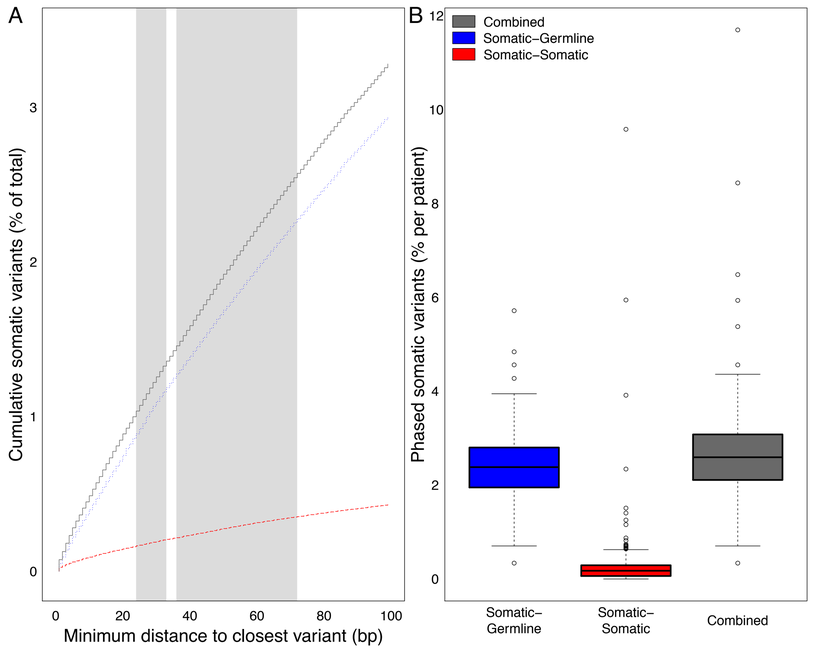
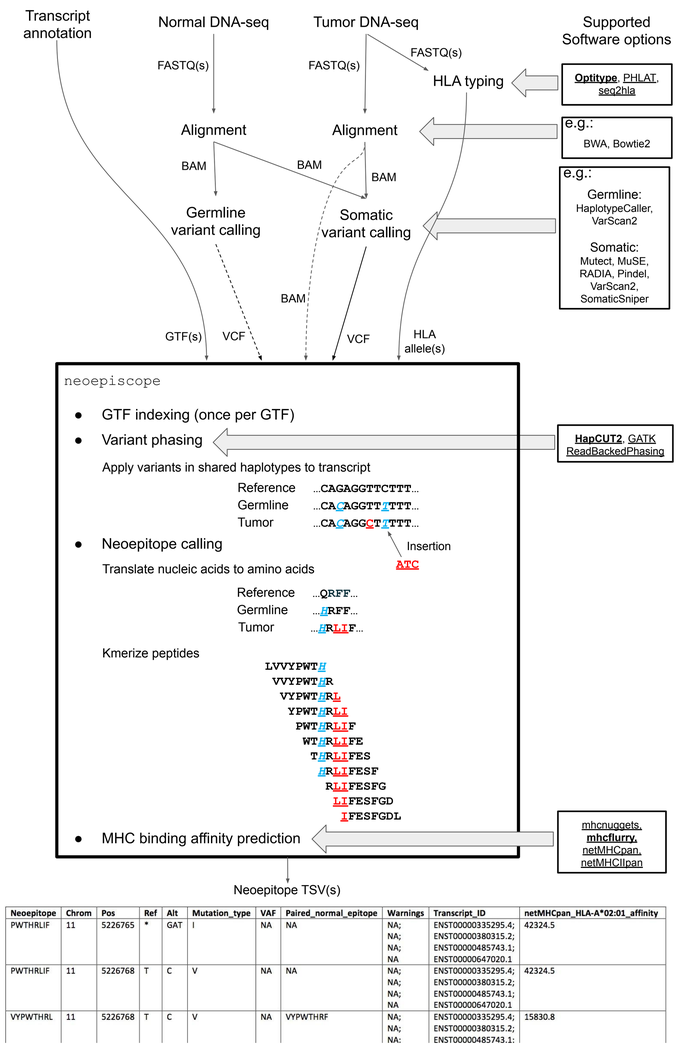


Figure 2: *Variant co-occurrence among 285 melanoma patients, 34 NSCLC patients, and 28 colon, endometrial, and thyroid cancer patients[12,21-29]. A) The cumulative average percentage (y-axis) of somatic variants across all tumors that co-occur with germline variants (blue), other somatic variants (red), or either type of variant (black) is shown as a function of increasing nucleotide span (x-axis). Canonical MHC Class I and Class II epitope size ranges are shaded in light gray (24-33bp and 36-72bp, respectively). B) Box plots demonstrating per-patient percentage of somatic variants (y-axis) across all tumors that co-occur with germline variants (blue), other somatic variants (red), or either variant type (dark gray).*

Neoepitope-calling pipeline



Neoepitope	Chrom	Pos	Ref	Alt	Mutation_type	VAF	Paired_normal_epitope	Warnings	Transcript_ID	netMHCpan_HLA-A*02:01_affinity
PWTHRLIF	11	5226765	*	GAT	I	NA	NA	NA; NA; NA; NA	ENST00000335295.4; ENST00000380315.2; ENST00000485743.1; ENST00000647020.1	42324.5
PWTHRLIF	11	5226768	T	C	V	NA	NA	NA; NA; NA; NA	ENST00000335295.4; ENST00000380315.2; ENST00000485743.1; ENST00000647020.1	42324.5
VYPWTHRL	11	5226768	T	C	V	NA	VYPWTHRF	NA; NA; NA:	ENST00000335295.4; ENST00000380315.2; ENST00000485743.1:	15830.8

Figure 3: *Neoepitope prediction pipeline diagram describing canonical neoepiscope workflow. Global inputs are shown at the top of the figure, with connecting arrows demonstrating interim inputs and outputs between preprocessing and processing steps. Direct inputs to and outputs from neoepiscope are shown directly entering or leaving the outlined box listing neoepiscope functionality. Multiple potential software options are shown at right for each relevant processing step as indicated by horizontal arrows (tools that are directly compatible with neoepiscope are underlined, with those in bold implemented as default). Direct neoepiscope functionality is depicted within the outlined box, with example sequences showing both somatic and background germline variants in a mock transcript sequence, and their translation and kmerization into short peptides (8mers). An example of the resulting neoepiscope output is shown at bottom.*

Performance

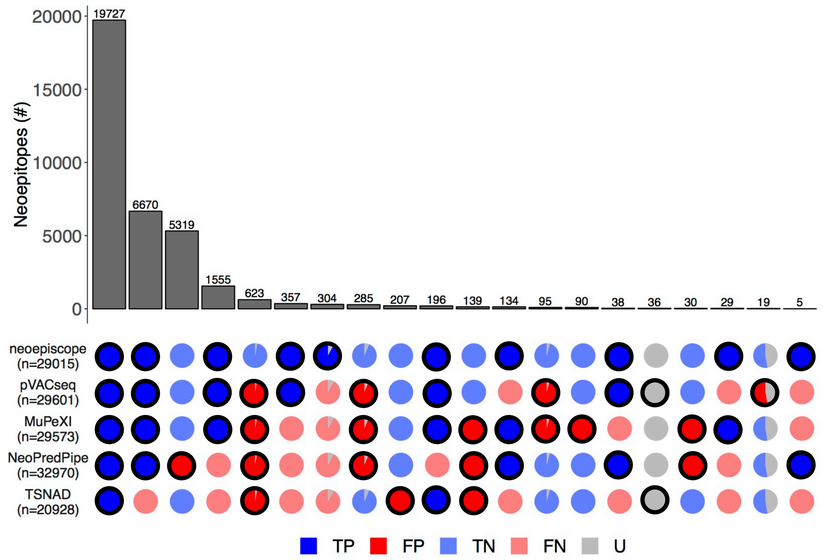


Figure 4: *Detailed comparison of the complete set of neoepitope predictions from neoepiscope, MuPeXI, pVACseq, NeoPredPipe, and TSNAD. Patterns of agreement or disagreement among groups of neoepitopes predicted by different combinations of tools across 5 melanoma patients are shown along each column (e.g. the first column corresponds to neoepitopes predicted by all tools). Each row indicates the neoepitope predictions associated with the indicated tool, with the total number of neoepitopes predicted by each tool shown as n. The number of neoepitopes in each column (bar in upper pane) corresponds to the size of the subset predicted by the indicated combination of tools (outlined circles in the bottom pane). The veracity of predictions corresponding to each group of neoepitopes are shown as pie charts, with colors corresponding to true positive ("TP", dark blue), false positive ("FP", dark red), true negative ("TN", light blue), false negative ("FN", light red), and uncertain ("U", gray) predictions. Uncertain predictions are considered a result of one or more factors including origin from unassembled contig regions, the presence of RNA edits, or inconsistencies in variant phasing predictions between HapCUT2 and GATK's ReadBackedPhasing.*

Conclusions

- Germline context and variant phasing are important for accurate and comprehensive neoepitope prediction
- neoepiscope is a novel, performant, and flexible tool for neoepitope prediction from DNA-seq data
- **neoepiscope improves sensitivity and specificity** compared with existing software tools (which incorrectly or incompletely predict ~5% of neoepitopes)
- The neoepiscope framework can accommodate numerous variant types, nonsense-mediated decay products, and epitope prediction across different genomes

References

1. Efremova et al., 2017. *Front Immunol*.
2. Koire et al., 2017. *PloS One*.
3. Muino et al., 2014. *Comput Biol Chem*.
4. Hundal et al., 2016. *Genome Med*.
5. Zhang et al., 2017. *Bioinformatics*.
6. Duan et al., 2014. *J Exp Med*.
7. Bais et al., 2017. *Bioinformatics*.
8. Zhou et al., 2017. *J Royal Soc Med*.
9. Bjerregaard et al., 2017. *Cancer Immunol Immunother*.
10. Tang and Madhavan., 2017. Preprint.
11. Chang et al. 2017., *Genome Med*.
12. Van Allen et al., 2015. *Science*.
13. Smart et al., 2018. Preprint.
14. Rubinsteyn et al., 2017. *Front Immunol*.
15. Rech et al., 2018. *Cancer Immunol Res*.
16. Schenck et al., 2019. *BMC Bioinformatics*.
17. Kim et al., 2018. *Ann Oncol*.
18. Tappeiner et al., 2017. *Bioinformatics*.
19. Wang et al., 2019. *Bioinformatics*.
20. Coelho et al., 2019. Preprint.
21. Amaria et al., 2018. *Nat Med*.
22. Carreno et al., 2015. *Science*.
23. Gao et al., 2016. *Cell*.
24. Hugo et al., 2017. *Cell*.
25. Le et al., 2017. *Science*.
26. Rizvi et al., 2015. *Science*.
27. Roh et al., 2017. *Sci Transl Med*.
28. Snyder et al., 2014. *N Engl J Med*.
29. Zaretsky et al., 2016. *N Engl J Med*.