**neoANT-HILL: an integrated tool for identification of potential neoantigens**

Ana Carolina M. F. Coelho¹, André L. Fonseca¹, Danilo L. Martins¹, Paulo B. R. Lins¹, Lucas M. da Cunha¹² and Sandro J. de Souza¹,³,⁴*

**Abstract**

**Background:** Cancer neoantigens have attracted great interest in immunotherapy due to their capacity to elicit antitumoral responses. These molecules arise from somatic mutations in cancer cells, resulting in alterations on the original protein. Neoantigens identification remains a challenging task due largely to a high rate of false-positives.

**Results:** We have developed an efficient and automated pipeline for the identification of potential neoantigens. neoANT-HILL integrates several immunogenomic analyses to improve neoantigen detection from Next Generation Sequence (NGS) data. The pipeline has been compiled in a pre-built Docker image such that minimal computational background is required for download and setup. NeoANT-HILL was applied in The Cancer Genome Atlas (TCGA) melanoma dataset and found several putative neoantigens including ones derived from the recurrent RAC1:P29S and SERPINB3:E250K mutations. neoANT-HILL was also used to identify potential neoantigens in RNA-Seq data with a high sensitivity and specificity.

**Conclusion:** neoANT-HILL is a user-friendly tool with a graphical interface that performs neoantigens prediction efficiently. neoANT-HILL is able to process multiple samples, provides several binding predictors, enables quantification of tumor-infiltrating immune cells and considers RNA-Seq data for identifying potential neoantigens. The software is available through github at https://github.com/neoanthill/neoANT-HILL.

**Keywords:** Neoantigens, Cancer, Immunogenomic analyses

**Background**

Recent studies have demonstrated that T cells can recognize tumor-specific antigens that bind to human leukocyte antigens (HLA) molecules at the surface of tumor cells [1, 2]. During tumor progression, accumulating somatic mutations in the tumor genome can affect protein-coding genes and result in mutated peptides [1]. These mutated peptides, which are present in the malignant cells but not in the normal cells, may act as neoantigens and trigger T-cell responses due to the lack of thymic elimination of autoreactive T-cells (central tolerance) [3–5]. As result, these neoantigens appear to represent ideal targets attracting great interest for cancer immunotherapeutic strategies, including therapeutic vaccines and engineered T cells [1, 6].

In the last few years, advances in next-generation sequencing have provided an accessible way to generate patient-specific data, which allows the prediction of tumor neoantigens in a rapid and comprehensive manner [7]. Several approaches have been developed, such as pVAC-Seq [8], MuPeXI [9], Tlminer [10] and TSNAD [11], which predict potential neoantigens produced by non-synchronous mutations. However, none of these proposed tools considers tumor transcriptome sequencing data (RNA-seq) for identifying somatic mutations. Moreover, only one of these tools provides quantification of the fraction of tumor-infiltrating immune cell types (Supplementary: Table S1).

Here we present a versatile tool with a graphical user interface (GUI), called neoANT-HILL, designed to identify potential neoantigens arising from cancer somatic mutations. neoANT-HILL integrates complementary features to prioritizing mutant peptides based on predicted binding affinity and mRNA expression level (Fig. 1). We used datasets from GEUVADIS...
RNA sequencing project [12] to demonstrate that RNA-seq is also a potential source of mutation detection. Finally, we applied our pipeline on a large melanoma cohort from The Cancer Genome Atlas [13] to demonstrate its utility in predicting and suggesting potential neoantigens that could be used in personalized immunotherapy.

**Implementation**

neoANT-HILL requires a variant list for potential neoantigen prediction. Our pipeline is able to handle a VCF file (single- or multi-sample) for the genome data or a tumor transcriptome sequence data (RNA-seq) in which somatic mutation will be called following GATK best practices [14, 15] with Mutect2 [16] on tumor-only mode. However, the RNA-seq data must be previously aligned to the reference genome (BAM) by the user. The size of corresponding BAM files from the RNA-Seq can be a limiting factor in the analysis. Since neoANT-HILL is run locally, the user must guarantee that enough space and memory are available for a proper execution of the program. In the current implementation, neoANT-HILL supports VCF files generated using the human genome version GRCh37. The variants are properly annotated by
Fig. 2 Screenshots of neoANT-HILL interface. 

a) Processing tab for submitting genome or transcriptome data. b) Processing tab for parameters selection to run neoepitope binding affinity prediction. On this tab, all the parameters can be defined by the users through selection boxes, ranging from the MHC class, corresponding prediction methods, to parallelization settings. The length and HLA alleles parameters allow multiple selections, although that might interfere in the processing time.

c) Binding prediction results tab shows an interactive table which reports all predicted neoepitopes and information about each prediction, respectively. The interactive table shows several columns, such as the donor gene, HLA allele, mutation type, reference (Ref_Peptide) and altered (Alt_peptide) peptides sequences, reference (Ref_IC50) and altered (Alt_IC50) binding affinity scores, binding affinity category (High, Moderate, Low, and Non-binding) and differential agretopicity index (DAI).
snpEff [17] to identify non-synonymous mutations (missense, frameshift and inframe).

Once the VCF files have been annotated, the resulting altered amino acid sequences are inferred from the NCBI Reference Sequence database (RefSeq) [18]. For frameshift mutations, the altered amino acid sequence is inferred by translating the resulting cDNA sequence. Altered epitopes (neoepitopes) are translated into a 21-mer sequence where the altered residue is at the center. If the mutation is at the beginning or at the end of the transcript, the neoepitope sequence is built by taking the 20 following or preceding amino acids, respectively. The neoepitope sequence and its corresponding wild-type are stored in a FASTA file. Non-overlapping neoepitopes can be derived from frameshift mutations.

A list of HLA haplotypes is also required. If this data had not been provided by the user, neoANT-HILL includes the Optitype algorithm [19] to infers class-I HLA molecules from RNA-Seq. The subsequent step is the binding affinity prediction between the predicted neoepitopes and HLA molecules. This can be executed on single or multi-sample using parallelization with the custom configured parameters. The correspondent wild-type sequences are also submitted at this stage, which allows calculation of the fold change between wild-type and neoepitopes binding scores, known as differential agretopicity index (DAI) as proposed by [20]. This additional neoantigen quality metric contributes to a better prediction of neoantigens that can elicit an antitumor response [21].

neoANT-HILL employs seven binding prediction algorithms from Immune Epitope Database (IEDB) [22], including NetMHC (v. 4.0) [23, 24], NetMHCpan (v. 4.0) [25], NetMHCcons [26], NetMHCstapban [27], Pick Pocket [28], SMM [29] and SMMPMBEC [30], and the MHChBlurry algorithm [31] for HLA class I. The user is able to specify the neoepitope lengths to perform binding predictions. Each neoepitope sequence is parsed through a sliding window metric. Our pipeline also employs four IEDB-algorithms for HLA class II binding affinity prediction: NetMHCIIpan (v. 3.1) [32], NN-align [33], SMM-align [34], and Sturniolo [35].

Moreover, when the unmapped RNA-seq reads are available (fastq), neoANT-HILL can quantify the expression levels of genes carrying a potential neoantigen. Our pipeline uses the Kallisto algorithm [36] and the output is reported as transcripts per million (TPM). Potential neoantigens arising from genes showing an expression level under 1 TPM are excluded. In addition, neoANT-HILL also offers the possibility of estimating quantitatively, via deconvolution, the relative fractions of tumor-infiltrating immune cell types through the use of quan-Tlseq [37].

Our software was developed under a pre-built Docker image. The required dependencies are packed up, which simplify the installation process and avoid possible incompatibilities between versions. As previously described, several analyses are supported and each one relies on different tools. Several scripts were implemented on Python to complete automate the execution of these single tools and data integration.

**Results**

neoANT-HILL was designed through a user-friendly graphical interface (Fig. 2) implemented on Flask framework. The interface comprises three main sections: (i) Home (Fig. 2a), (ii) Processing (Fig. 2b), and (iii) Results (Fig. 2c). neoANT-HILL stores the outputs in sample-specific folders. Our pipeline provides a table of ranked predicted neoantigens with HLA alleles, variant information, binding prediction score (neoepitope and wild-type) and binding affinity classification. When optional analyses are set by the user, the outputs are stored in separated tabs. Gene expression is provided as a list with corresponding RNA expression levels and it is used to filter the neoantigens candidates.

**Variant identification on RNA-Seq**

We evaluate the utility of RNA-seq for identifying frameshift, indels and point mutations by using samples \(n=15\) from the GEUVADIS RNA sequencing project.

**Table 1** Top 15 potential shared neoantigens based on TCGA-SKCM cohort. Recurrent mutations observed on TCGA-SKCM cohort. The amino acid (AA) residue changes caused by somatic mutations are highlighted in the (neo) epitopes sequences. The frequency represents the number of samples showing the corresponding mutation.

| Gene   | AA change | Neoepitope | HLA haplotype | Frequency |
|--------|-----------|------------|---------------|-----------|
| RAC1   | P29S      | FSGEYTIV   | HLA-A*02:01   | 17/466    |
| KLCD7A | E63SK     | HTATVRKK    | HLA-A*11:01   | 12/466    |
| INMT   | S212F     | YMVGKREFFCV | HLA-A*02:01   | 9/466     |
| CDH6   | S524L     | FLFSLAPEAA  | HLA-A*02:01   | 8/466     |
| ZBED2  | E157K     | G7MTALWASQK | HLA-A*11:01   | 8/466     |
| CRNL1  | S128F     | LQVPLPVPRF  | HLA-A*15:01   | 7/466     |
| IL37   | S202L     | FLFQPVCVA   | HLA-A*02:01   | 7/466     |
| SERPINB3 | E250K    | LSMILPLPNK  | HLA-A*11:01   | 6/466     |
| DNAJC5B | E22K      | STTGEALYK   | HLA-A*11:01   | 6/466     |
| MYO7B  | E512K     | MSISLSDLK   | HLA-A*11:01   | 6/466     |
| MORC1  | E878K     | IQNYMQVKY   | HLA-A*11:01   | 6/466     |
| SCN7A  | S445F     | IEKMKRSPIF  | HLA-A*15:01   | 6/466     |
| PSG9   | E404K     | KISKSMTVK   | HLA-A*11:01   | 6/466     |
| RAC1   | P29L      | FLEGEYIPTV  | HLA-A*02:01   | 5/466     |
| NUTF2  | Q20K      | SSFDQHYYK   | HLA-A*11:01   | 5/466     |
Although these samples are not derived from tumor cells, the goal of these analysis was to benchmark the efficiency of our pipeline to detect somatic mutations from RNA-Seq data. We limited our analysis to variants with read depth (DP) $\geq 10$ and supported by at least five reads. The overall called variants were then compared to the corresponding genotypes (same individuals) provided by the 1000 Genomes Project Consortium (1KG) [38]. We found that on average 71% of variants in coding regions detected by RNA-seq were confirmed by the genome sequencing (concordant calls) (Supplementary Table S2). Variants in genes that are not expressed cannot be detected by RNA-seq and RNA editing sites could partially explain the discordant calls. Furthermore, some of the discrepancies can be also due to low coverage in the genome sequence, which generated a false-negative in the calling. Although calling variants from RNA-Seq data has been shown to be more challenging, it is an interesting alternative for genome sequencing and a large amount of tumor RNA-seq samples do not have normal matched data [39, 40].

**Use case**

We applied our pipeline on a large melanoma cohort (SKCM, $n = 466$) from TCGA to demonstrate its utility in identifying potential neoantigens. We found approximately 198,000 instances of predicted neoantigens binding to HLA-I. It is important to note that the large number of mutant peptides is due to: i) the larger cohort size, ii) the high mutational burden of melanoma and iii) the large set of HLA alleles that was used to run the binding prediction. These neoepitopes were classified as strong (IC50 under 50 nM), intermediate (IC50 between 50 nM and 250 nM) or weak binders (IC50 over 250 nM and

![Fig. 3](image-url)  
*Fig. 3* Distribution of recurrent missense mutations that generated high-affinity neoantigens. The y-axis shows peptide coverage based on the number of epitope binding predictions in each region. The coverage was calculated by increasing the overall frequency of each amino acid by one, including non-high-affinity regions. The allele classification is shown as colored lines. The x-axis shows the protein length, and also contains information about conserved domains for each protein. (a) P29S and RAC1 gene generated recurrent mutant peptides with strong affinity to HLA-A*02:01 and P29L generated peptides with strong affinity to HLA-A*02:01 or HLA-A*11:01, depending on peptide length. (b) E250K in SERPINB3 gene generate a recurrent potential neoantigen that binds to HLA-A*11:01
under 500 nM) (Supplementary Table S3). We limited our analyses to high binding affinity candidates to reduce potential false positives.

We observed that the majority of strong binder mutant peptides are private and unique, which is likely linked to the high intratumor genetic diversity. However, we observed that frequent mutations may be likely to generate recurrent mutant peptides (Table 1). These recurrent neoantigens are interesting since they could be used as a vaccine for more than one patient. Figure 3 shows potential neoepitopes arising from recurrent mutations. The potential neoantigen (FSGYEYIPTV), which was predicted to form a complex with HLA-A*02:01 allele, was found to be shared among 17 samples (3.65%). It was generated from the P29S mutation in gene RAC1 (Fig. 3a). RAC1 P29S have been described as a candidate biomarker for treatment with anti-PD1 or anti-PD-L1 antibodies [41]. Another mutation (P29L) in the same gene formed a recurrent potential neoantigen (FLGYEYIPTV) and was found in 5 samples (1.07%). As another example, we can also highlight the potential shared neoantigen (LSMIVLLPNK) related to mutation E250K in the SERPINB3 gene (Fig. 3b). This was found in 6 samples (1.29%) and it was likely to form a complex with the HLA-A*11:01 allele. Mutations in SERPINB3 have also been related to response to immunotherapy [42].

Conclusion
We present neoANT-HILL, a completely integrated, efficient and user-friendly software for predicting and screening potential neoantigens. We have shown that neoANT-HILL can predict neoantigen candidates, which can be targets for immunotherapies and predictive biomarkers for immune responses. Our pipeline is available through a user-friendly graphical interface which enables its usage by users without advanced programming skills. Furthermore, neoANT-HILL offers several binding prediction algorithms for both HLA classes and can process multiple samples in a single running. Unlike the majority of existing tools, our pipeline enables the quantification of tumor-infiltrating lymphocytes and considers RNA-Seq data for variant identification. The source code is available at https://github.com/neoanthill/neoANT-HILL.

Availability and requirements
Project name: neoANT-HILL.
Project home page: https://github.com/neoanthill/neoANT-HILL.
Operating system(s): Unix-based operating system, Mac OS, Windows.
Programming language: Python 2.7.
Other requirements: Docker.
License: Apache License 2.0.

Any restrictions to use by non-academics: None.

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1186/s12920-020-0694-1.

Additional file 1.
Additional file 2.
Additional file 3.

Abbreviations
DAI: Differential Agretnicity Index; DP: Read Depth; GUi: Graphical User Interface; HLA: Human Leukocyte Antigens; NGS: Next Generation Sequencing; RNA-Seq: RNA-Sequencing; SKCM: Skin Cutaneous Melanoma; TCGA: The Cancer Genome Atlas; TPM: Transcripts per Million

Acknowledgments
Not applicable.

Author’s contributions
ACMF, DLM and PRBL designed and carried out the implementation of the computational pipeline. DLM led debugging efforts, LMC contributed to design the computational pipeline. ACMFC and ALF analyzed the data. SJS supervised the project. ACMFC and ALF discussed the results and commented on the manuscript in consultation with SJS. SJS reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by a CAPES grant (2308.004692/2014–19). ACMFC, DLM, ALF, LMC and PRBL were supported by CAPES. The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The RNA-Seq dataset from Geuvadis RNA sequencing project analyzed during the current study are available in the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/) under the accession number E-GEUV-1. The corresponding genotyping data (Phase I) from each sample are available from the 1000 Genomes Project and was downloaded from the FTP site hosted at the EBI.

Publications
DLM, ALF, LMC and PRBL were supported by CAPES. The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Any restrictions to use by non-academics: None.
Caracedo Á, Antonarakis SE, Häcker R, Syvánen AC, van Ommen GJ, Brazma A, Meijting T, Rosenberg P, Guigó R, Gut IG, Estivill X, Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variations in humans. Nature. 2013;501(7468):506.

11. Zhou Z, Lyu X, Wu J, Yang X, Wu S, Zhou J, Gu X, Su Z, Chen S. TSNAD: an extended.txt).

12. DeFrusto MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kerteszky AM, Svakchenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(3):491–4.

13. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shikir Koazen D, Barth B, Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variations in humans. Nature. 2013;500(7468):506.

14. Weinstein JN, Collison ES, Mills GB, Shaw KR, Ozmenberger BA, Elliott K, Shumulevich I, Sander C, Stuart JM. The Cancer genome atlas pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20.

15. Bjerregaard A, Nielsen M, Hadrup SR, Szallasi Z, Eklund AC. MuPeXI: an extended.txt).

16. Cibulskis K, Lawrence MS, Carter SL, Svakchenko 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. Nat Biotechnol. 2013;31(3):213–9.

17. Tappa Reiner, Tappeiner E, Finotello F, Charoentong P, Mayer C, Rieder DE, Trajanoski Z. Neantigenome: prediction of individual mutations and their role in cancer immunity and immunotherapy. Front Immunol. 2017. https://doi.org/10.3389/fimmu.2017.01679.

18. van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shikir Koazen D, Barth B, Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variations in humans. Nature. 2013;500(7468):506.

19. Zhou Z, Lyu X, Wu J, Yang X, Wu S, Zhou J, Gu X, Su Z, Chen S. TSNAD: an extended.txt).

20. DeFrusto MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kerteszky AM, Svakchenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(3):491–4.

21. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shikir Koazen D, Barth B, Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variations in humans. Nature. 2013;500(7468):506.

22. Bjerregaard A, Nielsen M, Hadrup SR, Szallasi Z, Eklund AC. MuPeXI: an extended.txt).

23. Zhou Z, Lyu X, Wu J, Yang X, Wu S, Zhou J, Gu X, Su Z, Chen S. TSNAD: an extended.txt).

24. DeFrusto MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kerteszky AM, Svakchenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(3):491–4.
30. Kim Y, Sidney J, Pinilla C, Sette A, Peters B. Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a Bayesian prior. BMC Bioinformatics. 2009;10(1):394.
31. O'Donnell TJ, Rubinstein A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell Systems. 2018;7(1):129–132.e4.
32. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics. 2013;65(10):711–24.
33. Nielsen M, Lund O. NN-align: An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics. 2009;10(1):296.
34. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics. 2007;8(1):238.
35. Stunzio T, Bono E, Ding J, Raddazzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Pinti MP, Sinigaglia F, Hammer J. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnol. 1999;17(6):555–61.
36. Bray N, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7.
37. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, Krogsdam A, Loncova Z,Posch W, Willflingseder D, Sopper S, Ijselsteijn M, Brouwer TP, Johnson D, Xu Y, Wang Y, Sanders ME, Estrada MV, Ericson-Gonzalez P, Charoentong P, Balko J, NFDC, d M, Trajanosi Z. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. 2019;11(1):34.
38. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
39. Piskol R, Ramaswami G, Li J. Reliable identification of genomic variants from RNA-Seq data. Am J Hum Genet. 2013;93(4):641–51.
40. Coudray A, Battenhouse AM, Bucher P, Iyer VR. Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. PeerJ. 2018;6:e5362.
41. Vu HL, Rosenbaum S, Purwin TJ, Davies MA, Aplin AE. RAC1 P29S regulates PD-L1 expression in melanoma. Pigment Cell Melanoma Res. 2015;28(5):590–8.
42. Riaz N, Havel JJ, Kendall SM, Makarov V, Walsh LA, Desrichard A, Weinhold N, Chan TA. Recurrent SERPINB3 and SERPINB4 mutations in patients who respond to anti-CTLA4 immunotherapy. Nat Genet. 2016;48(11):1327–9.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:
• fast, convenient online submission
• thorough peer review by experienced researchers in your field
• rapid publication on acceptance
• support for research data, including large and complex data types
• gold Open Access which fosters wider collaboration and increased citations
• maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.
Learn more biomedcentral.com/submissions