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

Ana Carolina Coelho
Universidade Federal do Rio Grande do Norte https://orcid.org/0000-0003-1511-3844
Andre Fonseca
UFRN
Danilo Martins
Universidade Federal do Rio Grande do Norte
Paulo Lins
Universidade Federal do Rio Grande do Norte
Lucas da Cunha
Universidade Federal do Rio Grande do Norte
Sandro de Souza (sandro@imd.ufrn.br)

Software

Keywords: neoantigens; cancer; immunogenomic analyses

Posted Date: February 7th, 2020

DOI: https://doi.org/10.21203/rs.2.16149/v5

Version of Record: A version of this preprint was published on February 22nd, 2020. See the published version at https://doi.org/10.1186/s12920-020-0694-1.

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.

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], TIminer [10] and TSNAD [11], which predict potential neoantigens produced by non- synonymous 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 1).

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 (Figure 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 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], NetMHCstabpan [27], PickPocket [28], SMM [29] and SMMPMBEC [30], and the MHCflurry 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 quanTIseq [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 (Figure 2) implemented on Flask framework. The interface comprises three main sections: (i) Home (Figure 2A), (ii) Processing (Figure 2B), and (iii) Results (Figure 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. 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) >= 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 2). 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 under 500 nM) (Supplementary Table 3). 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 (F**S**GEYIPTV), 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 (Figure 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 (F**L**GEYIPTV) and was found in 5 samples (1.07%). As another example, we can also highlight the potential shared neoantigen (LSMIVLLPN**K**) related to mutation E250K in the SERPINB3 gene (Figure 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].

Conclusions

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.

Declarations

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

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

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

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/
(ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA12812/exome_alignment/NA12812.mapped.SOLID.bfast.CEU.exome.20110411.bam,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA12749/exome_alignment/NA12749.mapped.illumina.mosaik.CEU.exome.20110521.bam,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA20510/exome_alignment/NA20510.mapped.SOLID.bfast.TSI.exome.20110521.bam,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA19119/exome_alignment/NA19119.mapped.illumina.mosaik.YRI.exome.20110411.bam,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA19204/exome_alignment/NA19204.mapped.illumina.mosaik.YRI.exome.20110411.bam,

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA18498/exome_alignment/NA18498.mapped.illumina.mosaik.YRI.exome.20110411.bam,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA12489/exome_alignment/NA12489.mapped.SOLID.bfast.CEU.exome.20110411.bam,

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA20752/exome_alignment/NA20752.mapped.illumina.mosaik.TSI.exome.20110521.bam,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA18517/exome_alignment/NA18517.mapped.illumina.mosaik.YRI.exome.20110521.bam,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA11992/exome_alignment/NA11992.mapped.SOLID.bfast.CEU.exome.20110411.bam,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA19144/exome_alignment/NA19144.mapped.illumina.mosaik.YRI.exome.20110411.bam,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA20759/exome_alignment/NA20759.mapped.illumina.mosaik.TSI.exome.20110521.bam,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA19137/exome_alignment/NA19137.mapped.illumina.mosaik.YRI.exome.20110411.bam,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA19257/exome_alignment/NA19257.mapped.illumina.mosaik.YRI.exome.20110521.bam,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/NA12006/exome_alignment/NA12006.mapped.SOLID.bfast.CEU.exome.20110411.bam).
The melanoma TCGA mutation data was downloaded from the cBio datahub
(https://github.com/cBioPortal/datahub/blob/master/public/skcm_tcga/data_mutations_extended.txt).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by a CAPES grant (23038.004629/2014-19). 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.

Acknowledgments

Not applicable

Author Information

Affiliations

1. **Bioinformatics Multidisciplinary Enviroment (BioME), Institute Metropolis Digital, Federal University of Rio Grande do Norte, UFRN, Brazil**

Ana Carolina MF Coelho, André L Fonseca, Danilo L Martins, Lucas M da Cunhas, Paulo R B Lins & Sandro J de Souza

2. **PhD Program in Bioinformatics, UFRN, Natal, Brazil**

Lucas M da Cunha

3. **Brain Institute, Federal University of Rio Grande do Norte, UFRN, Brazil**

Sandro J. de Souza

4. **Institutes for Systems Genetics, West China Hospital, Sichuan University, Chengdu, China**

Sandro J de Souza

Author Contributions

ACMFC, 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.

Abbreviations

NGS: Next Generation Sequencing

HLA: Human Leukocyte Antigens

RNA-Seq: RNA-Sequencing

GUI: Graphical User Interface

DAI: Differential Agretopicity Index

TPM: Transcripts per Million

TCGA: The Cancer Genome Atlas

SKCM: Skin Cutaneous Melanoma

DP: Read Depth

References

1. Efremova M, Finotello F, Rieder D, Trajanoski Z. Neoantigens generated by individual mutations and their role in cancer immunity and Frontiers in immunology. 2017; doi: 10.3389/fimmu.2017.01679
2. Kato T, Matsuda T, Ikeda Y, Park JH, Leisegang M, Yoshimura S, Hikichi T, Harada M, Zewde M, Sato S, Hasegawa K, Kiyotani K, Nakamura Y. Effective screening of T cells recognizing neoantigens and construction of T-cell receptor-engineered T cells. Oncotarget. 2018; doi: 18632/oncotarget.24232

3. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, Walsh LA, Postow MA, Wong P, Ho TS, Hollmann TJ, Bruggeman C, Kannan K, Li Y, Elipenahli C, Liu C, Harbison CT, Wang L, Ribas A, Wolchok JD, Chan TA. Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. New England Journal of Medicine. 2014; doi: 10.1056/nejmoa1406498
4. Bailey P, Chang DK, Forget M, Lucas FAS, Alvarez HA, Haymaker C, Chattopadhyay C, Kim S, Ekmekcioglu S, Grimm EA, Biankin AV, Hwu P, Maitra A, Roszik J. Exploiting the neoantigen landscape for immunotherapy of pancreatic ductal adenocarcinoma. Scientific Reports. 2016; doi: 10.1038/srep35848
5. Riaz N, Morris L, Havel JJ, Makarov V, Desrichard A, Chan TA. The role of neoantigens in response to immune checkpoint blockade. International Immunology. 2016; doi: 10.1093/intimm/dxw019
6. Lu Y, Robbins PF. Cancer immunotherapy targeting neoantigens. Seminars in Immunology. 2016;28(1):22-27.
7. Liu XS, Mardis ER. Applications of Immunogenomics to Cancer. Cell. 2017;168(4):600-612.
8. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, Griffith M. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Medicine. 2016;8(1).
9. Bjerregaard A, Nielsen M, Hadrup SR, Szallasi Z, Eklund AC. MuPeXI: prediction of neo-epitopes from tumor sequencing data. Cancer Immunology, Immunotherapy. 2017;66(9):1123-1130.
10. Tappeiner E, Finotello F, Charoentong P, Mayer C, Rieder DE, Trajanoski Z. TIminer: NGS data mining pipeline for cancer immunology and immunotherapy. Bioinformatics. 2017;33(19):3140-3141.
11. Zhou Z, Lyu X, Wu J, Yang X, Wu S, Zhou J, Gu X, Su Z, Chen S. TSNAD: an integrated software for cancer somatic mutation and tumour-specific neoantigen detection. Royal Society Open Science. 2017;4(4):170050.
12. Lappalainen T, Sammeth M, Friedländer MR, Hoen PAC 't, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donelly P, McCarthy MI, Flicek P, Strom TM, The Geuvadis Consortium, Lehrach H, Schreiber S, Sudbrak R, Carracedo Á, Antonarakis SE, Häsler R, Syvänen AC, van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501(7468), 506.
13. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics. 2013;45(10):1113-1120.
14. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43(5):491-498.
15. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics. 2013;:11.10.1-11.10.33.
16. Cibulskis K, Lawrence MS, Carter SL, Sivachenko 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. Nature Biotechnology. 2013;31(3):213-219.
17. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118}; iso-2; iso-3. Fly. 2012;6(2):80-92.
18. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, Dicuccio M, Kitts P, Murphy, TD, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2015;44(D1):D733-D745.
19. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014;30(23):3310-3316.
20. Duan F, Duitama J, Sahar AIS, Cory MA, Steven AC, Arpita PP, Tatiana B, David M, John S, Alessandro S, Brian MB, Ion IM, Pramod KS. Genomic and bioinformatic profiling of mutational neoepitopes reveals rules to predict anticancer immunogenicity. Journal of Experimental Medicine. 2014;211(11):2231-2248.
21. Richman LP, Vonderheide RH, Rech Aj. Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade. Cell Syst. 2019; 9:375-382.
22. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JRWheeler DK, Sette A, Peters B. The ImmuneEpitope Database (IEDB): 2018 update. Nucleic Acids Research. 2018..
23. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics. 2015;32(4):511-517.

24. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, Brunak S, Lund O. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Science. 2003;12(5):1007-1017.
25. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. The Journal of Immunology. 2017;199(9):3360-3368.
26. Karosiene E, Lundegaard C, Lund O, Nielsen M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. Immunogenetics. 2011;64(3):177-186.
27. Rasmussen M, Fenoy E, Harndahl M, Kristensen AB, Nielsen IK, Nielsen M, Buus S. Pan-Specific Prediction of Peptide–MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. The Journal of Immunology. 2016;197(4):1517-1524.
28. Zhang H, Lund O, Nielsen M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. Bioinformatics. 2009;25(10):1293-1299.
29. Peters B, Sette A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. BMC Bioinformatics. 2005; 6(1):132.
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, Rubinsteyn 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. NetMHCIIpan-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):711-724.
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).
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).
35. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F, Hammer J. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nature Biotechnology. 1999;17(6):555-561.
36. Bray N, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016;34(5):525-527.
37. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, Krogsdam A, Loncova Z, Posch W, Wilflingseder D, Sopper S, Ijsselsteijn M, Brouwer TP, Johnson D, Xu Y, Wang Y, Sanders ME, Estrada MV, Ericsson-Gonzalez P, Charoentong P, Balko J, de Miranda NFDCC, Trajanoski Z. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Medicine. 2019;11(1).
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. The American Journal of Human Genetics. 2013;93(4):641-651.
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 Research. 2015;28(5):590-598.
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. Nature Genetics. 2016;48(11):1327-1329.

Table

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	FSGEYITV	HLA-A*02:01	17/466
KLHDC7A	E635K	HTATVRAKK	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	GTMALWASQRK	HLA-A*11:01	8/466
CRNKL1	S128F	LQVPLPVPRF	HLA-A*15:01	7/466
IL37	S202L	FLFQPVCKA	HLA-A*02:01	7/466
SERPINB3	E250K	LSMIVLLPNK	HLA-A*11:01	6/466
DNAJC5B	E22K	STTGEALYK	HLA-A*11:01	6/466
MYO7B	E512K	MSIISLLDK	HLA-A*11:01	6/466
MORC1	E878K	IQNTYMVQYK	HLA-A*11:01	6/466
SCN7A	S445F	IEMKKRSPIF	HLA-A*15:01	6/466
PSG9	E404K	KISKSMTVK	HLA-A*11:01	6/466
RAC1	P29L	FLGEYIPTV	HLA-A*02:01	5/466
NUTF2	Q20K	SSFIQHYYK	HLA-A*11:01	5/466

Figures

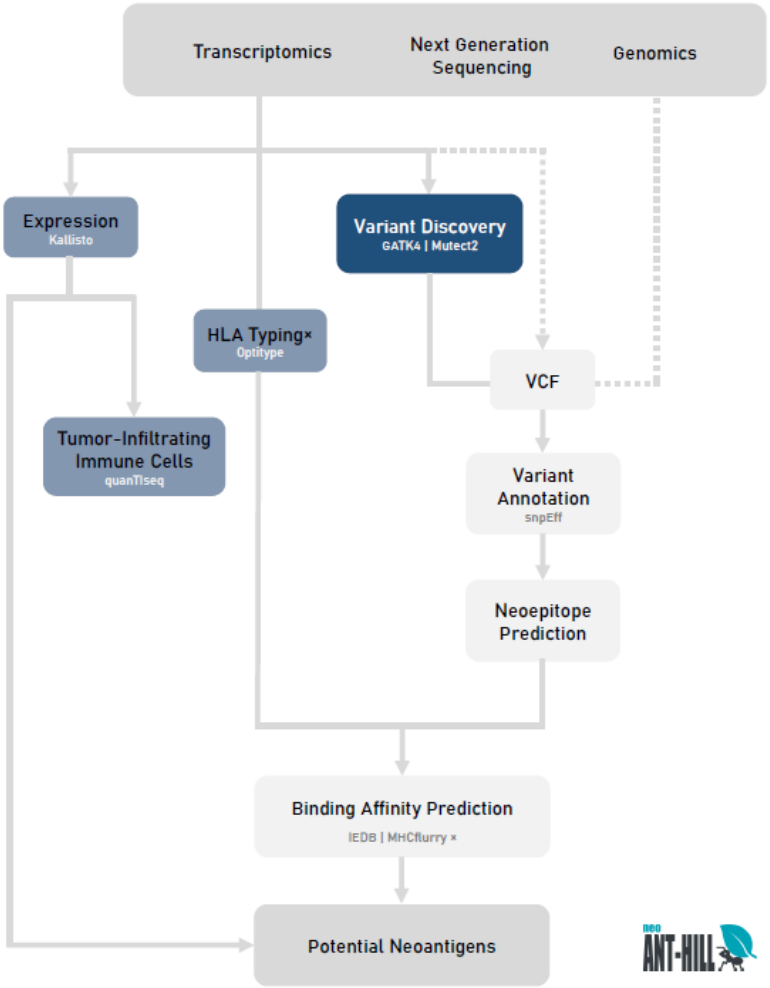

Figure 1

Overall workflow of neoANT-HILL. The neoANT-HILL was designed to analyze NGS data, such as genome (WGS or WES) and transcriptome (RNA-Seq) data. Basically, it takes as input distinct data types, including raw and pre-aligned sequences from RNA-Seq, as well as, variant calling files

(VCF) from genome or transcriptome data (dotted lines indicate that the VCF must be previously created by the user). The blue boxes represent the transcriptome analyses, which should be carried out using data in either BAM format (variant calling) or fastq format (expression, HLA typing and tumor-infiltrating immune cells). The neoANT-HILL can perform gene expression (Kallisto), variant calling (GATK4 | Mutect2), HLA typing (Optitype), and Tumor-infiltrating immune cells (quanTIseq). The gene expression quantification is used as input to identify molecular signatures associated with immune cell diversity into the tumor samples. On the other hand, the gray boxes represent common steps to genome and transcriptome data. NeoANT-HILL uses the variant calling data to reconstruct the proteins sequences using as reference the NCBI RefSeq database. The VCF files can be either generated by using our pipeline or by external somatic variant-calling software. Next, reconstructed proteins are submitted to neoepitope binding prediction using HLA alleles from Optitype results or defined by the user. Finally, all steps and results are shown into a user-friendly interface.

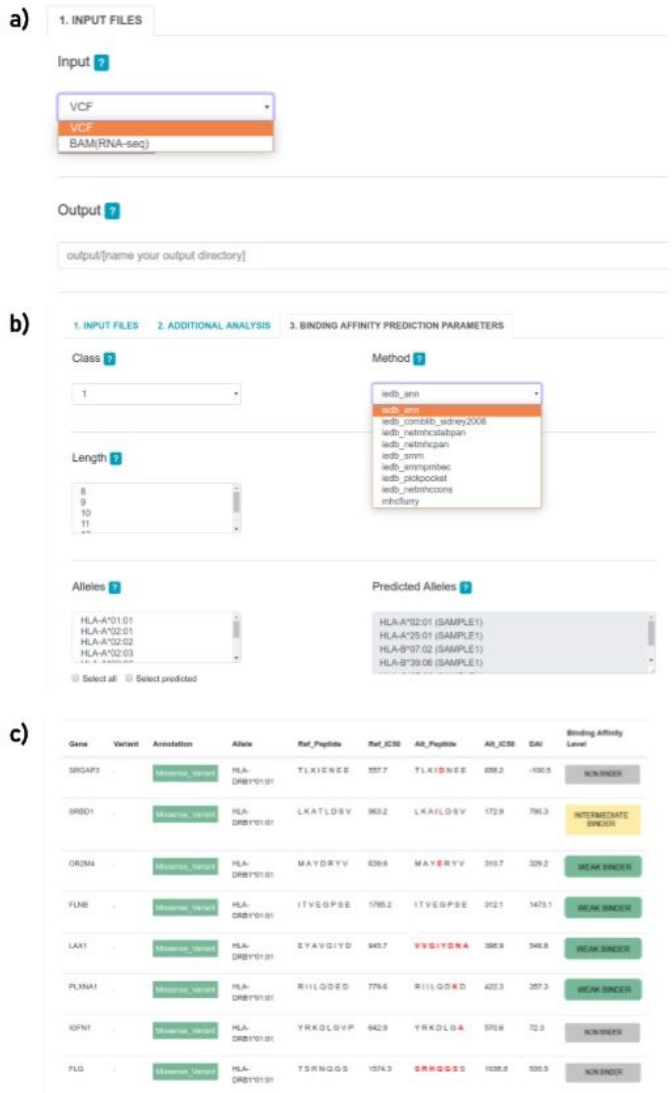

Figure 2

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).

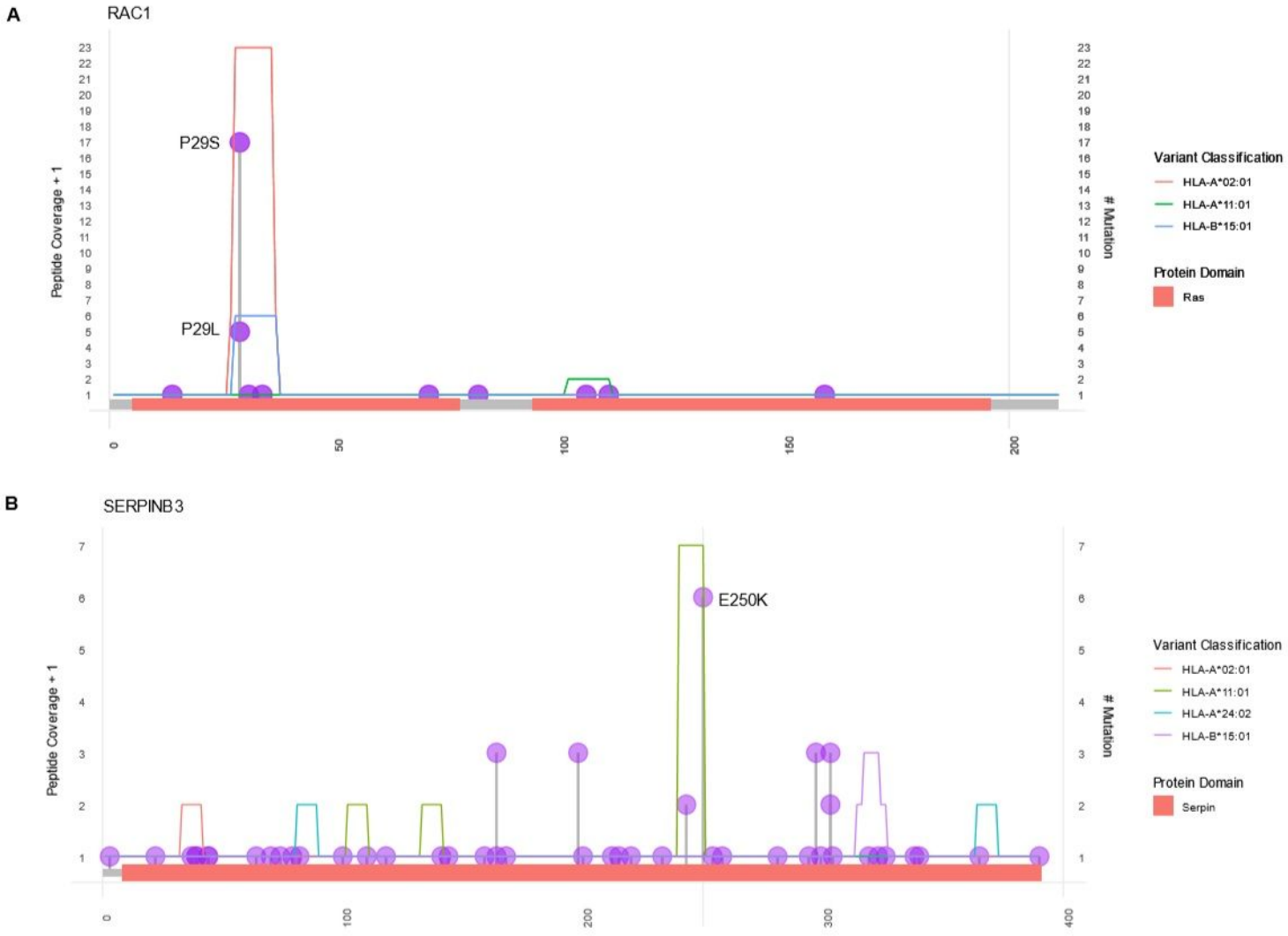

Figure 3

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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile03.xlsx
- AdditionalFile01.xlsx
- AdditionalFile02.xlsx