Automating biomarker identification for immunotherapies:

Non-canonical peptides presented on MHC molecules

Shane Ian Pullens

Supervisor: C. Kesmir

Theoretical Biology & Bioinformatics
Utrecht University

Utrecht University August 2020

Automating biomarker identification for immunotherapies:

Non-canonical peptides presented on MHC molecules.

S.I. PULLENS[1]

THEORETICAL BIOLOGY & BIOINFORMATICS GROUP - UTRECHT UNIVERSITY

ABSTRACT

Neo-antigens are a promising area of research in the development of immunotherapies against cancer. The neo-antigens arise due to mutations in cancerous cell, which often helps the cancer cell to hide from the surveillance of the immune system. In the last decade, the amount of mass spectrometry data has been growing exponentially. Researchers often found that the origin of all peptides eluted from cancer cells could not be mapped, which suggest that the tumor alters the translation process to generate new peptides that are presented in the MHC-complex on the cell surface. Obviously, this finding opens up a totally new area for cancer specific biomarkers. Here we present our pipeline to identify these non-canonical (cryptic) peptide candidates from RNA count data.

Introduction

Cancer is still one off the most leading causes of death in the world. Due to the global population growth and increasing age, the number of cancer casualties is increasing rapidly (1–3). Even though new treatment methods are being developed, the success of a treatment is mostly limited due to multiple aspects. First, the physician must determine the best tumor repressive treatment, considering factors such as the potential side effects and toxicity for the patient. This is challenging as there are no clinical predictive tests available to assess the suitability of specific chemotherapy regimens, requiring that they be empirically evaluated for each patient. In addition, the chosen chemotherapy should not be de-novo resistant and should induce multidrug resistance (4). These considerations make it difficult to establish an effective, personalized treatment plan for each patient.

Recently, patient-specific biomarker methods have been developed and have been cleared for use in clinical trials (5,6). The main limitation of this approach is the uncertainty that the treatment will take effect on target cells only, as there is always a chance that these markers are found on healthy cells. If that is the case, the therapy would cause possible harm to the surrounding cells. Moreover, the identification of patient-specific biomarkers often requires a significant amount of manual labor, which can be botch costly and time consuming. It is important to initiate treatment quickly in order to improve the patient's life expectancy (7).

An effective treatment would specifically targets only cancer cells, be effective for multiple patients and have minimal side effects. One way this could be achieved is by encapsulating chemotherapeutic-molecules in a liposome that will only fuse with the target tumor cells. This makes it safe for the surrounding cells, which in return will reduce side effects for the patient. This will, in theory, be possible if the targeting of the cancer cells is very specific.

Tumors can proliferate, if they escape the immune surveillance system. Normally, healthy cells present self-peptides that inhibit T cell-mediated immune response. This is achieved by generating self-peptides, that are presented on the cell surface (8). Cancer cells, on the other hand, express neo-antigens (9–11), which are mutated self-peptides. These neo-antigens could be a biomarker to use in several immunotherapies against cancer. These neo-antigens unfortunately, are typically cancer- and patient-specific, making them not suited for affordable, patient-wide therapies.

Due to protein identification by Tandem MS becoming a standard (12,13), more peptide-MHC complexes were being detected that could not be explained by canonical protein translation. Already in 1989, a hypothesis named 'pepton hypothesis', arose that stated: "antigenic peptides are derived from the cellular genome, and are not a degradation product of cellular proteins, but can be generated directly by the autonomous transcript and translation of short sub genic regions" (14). Due to lack of evidence however, this hypothesis was quickly disregarded. Now that we are finding evidence, this theory is being explored more thoroughly. (15–19). These peptides are believed to be translated in a non-traditional manner, originating from the same genomic data as canonical peptides.

Because all cryptic peptide research has been performed on De Novo Peptide Sequencing (20–23), we wanted to explore the possibilities of cryptic peptide detection in not-peptide-specific-sequencing data that is publicly available.

This research has been done in collaboration with the Molecular Targeted Therapies research group from Utrecht University (https://cellbiology.science.uu.nl/research-groups/sabrina-oliveira-molecular-targeted-therapies/). The objective was to find non-canonical, neo-antigen peptides that were derived from open source data. To achieve this goal, we developed a simple to use, user-flexible pipeline to predict cryptic peptide candidates.

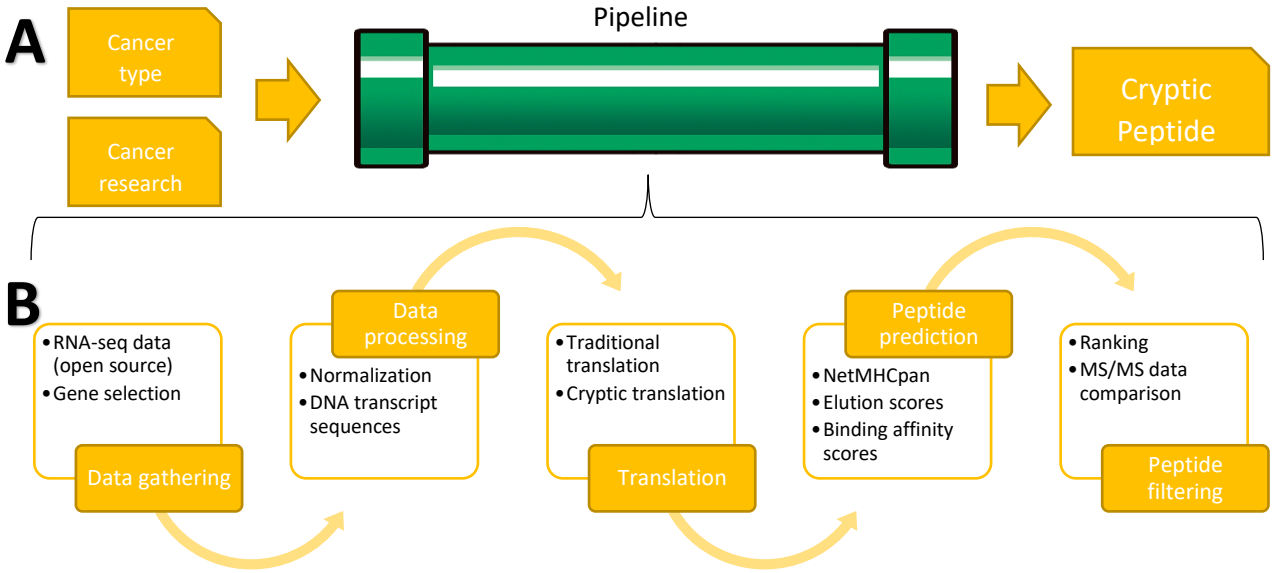


Figure 1, Overview of developed pipeline. A, basic overview of input for the pipeline and expected output. The user inputs the cancer type they want analyze, and the research project that can be found in the GDC dataset, more can be found in the methods. Note, output can change based on tweaked parameters (Not shown in figure, can be found in methods). B, more detailed flowchart inside the pipeline. Here, 5 high-level steps are shown, with their own internal steps. These steps will be discussed in detail in the methods.

RESULTS

DATA AND THE DEVELOPMENT OF THE PIPELINE

The RNA-seq count data was fetched from the GDC Data portal (24) on 26-11-2022. Open data corresponding to a total of 1699 patients, 473 and 1226 for skin and breast cancers respectively, was downloaded. RNA-seq count data has been processed using the STAR algorithm, which provides gene expression values for 60617 genes. 40.85% of these genes were RNA related, giving us a large pool of sequences to search for cryptic peptides, as most of those are expected to be found in the long non-coding RNA (lncRNA) (25,26).

The first step in the pipeline is to select for the RNA related genes that are highly expressed (see Figure 1). To this end, we calculated the median expression (counts) of each gene in the patient cohort and defined the genes with top 1% medians as the "highly expressed genes". After stabilization (which removes the dependence of the variance on the mean), we found 1125 (83% lncRNA) and 1251 (82.8% lncRNA) highly expressed genes for skin and breast cancers, respectively (Table 1). To our surprise, RNA related genes are expressed in higher amounts than coding genes. Moreover, we found that 62% of the genes that are identified as highly expressing in the skin and breast cancer data set are overlapping (of which 81% is lncRNA), promising some pan-cancer RNA based biomarkers.

The second step in the pipeline is to convert transcripts for every gene and exclude the coding regions by double checking the annotations. To this end, we used transcriptomics data from the Ensemble database (27). The average amount of transcripts, with known coding regions (CDS) (according to the database) per gene was 2.20 and 2.38 for skin and breast cancer data set respectively, with outliers of 14 and 17 transcripts per gene (Table 1). Among the RNA-related genes, we could only find a single transcript with a known CDS in the skin dataset (ENST00000598322). This is probably a wrong annotated transcript, mainly because both the transcript and related gene (ENSG00000269825) are annotated as 'novel'. Also, the transcript does not have a Transcript Support Level (TSL), which should be either one or two for a transcripts with a known CDS (28).

The third step in the pipeline, was to decide how to handle the transcripts that do not contain any coding regions. This was particularly important due the Ensemble database only holding sequences for coding regions, meaning no direct sequences were available for the non-coding transcripts. To tackle this, we looked at the earliest start codon on the genome that we could find for all transcripts linking to our gene (Sup. Fig 1.). We did the same for the end, by taking the latest associated stop codon. By saving the coordinates for these start- and stop codons, we were able to fetch a sequence from the human genome. The main issue that arises with this method, is that taking the whole sequence also contaminates the read with canonical peptides, originating from coding regions. In order to solve this problem, we translated the whole sequence, including the transcripts that are known to contain a CDS. Afterwards the peptides that originated from the known CDS transcripts were removed, leaving only theoretical, cryptic peptides.

The next step of the pipeline is the translation. Unfortunately, there is not yet an available

Table 1, Overview of data that was gathered and used for this project.

Cancer Tissue	*No.* Genes	*No.* Patients	Percentage RNA-related	*No.* top 1% high-expressed genes	Percentage of lncRNA in top 1%	*Avg. No.* transcripts with known CDS per gene
Skin	60617	473	41%	1125	83,0%	2,20
Breast	60617	1226	41%	1251	82,8%	2,38

software package to predict the translation of sequences outside of CDS regions. Due to lack of RNA-seq sequences, we were not able to perform alternate splicing simulation. Recent research (29,30), including a recently published paper on the standardization of alternative open reading frames (31), suggest that cryptic peptides may originate from alternative open reading frames (aORFs). In response to this, we have developed an in-house algorithm that analyzes messenger RNA (mRNA) (including untranslated regions and introns) and predicts aORFs based on the presence of non-traditional start and stop codons in other parts of the sequence. This cryptic peptide prediction is performed with 6-frame translation; 3 open reading frames in the forward direction, and 3 on the complementary strand in the reverse direction. With this algorithm (further explained in Supp. Fig 1), we used 5 of the 6 newly defined alternate ORFs (31) (see methods).

The final step of the pipeline is the predicted binding to MHC molecules. The predicted peptide binding of the translated sequences was done with NetMHCpan (32). (Used parameters can be found in the methods)

Cryptic peptides as possible MHC ligands

Filtering the possible MHC molecule binders, we obtained (on average) 216 and 219 strong binders peptides for RNA-related genes in the skin and breast dataset respectively. These numbers are significantly higher than the average amount of canonical peptide binders that were predicted: 42 for skin and 41 for the breast dataset. This main difference is hypothesized to arise from the 6-frame translation that is performed in the in-house algorithm to cover all possible peptides.

To test whether or not the high amount of predicted peptides is very unusual, we compared the number of peptides that have been found per read (for alleles HLA-A01:01 and HLA-A02:01), with the expected number of peptides (Fig. 2). Since we are using a 1% threshold on the ranking system of NetMHCpan, the expected number of peptides was calculated by taking the top 1% of the expected number of peptides. for a single read this is approximately 1% of the read length (see Equation 1).

In Figure 2, the predicted and expected MHC peptide binders are plotted in a scatter plot, including a regression line. For all HLA-A01:01 alleles datasets, the predicted peptides were slightly higher than the expected ones (1.2 to

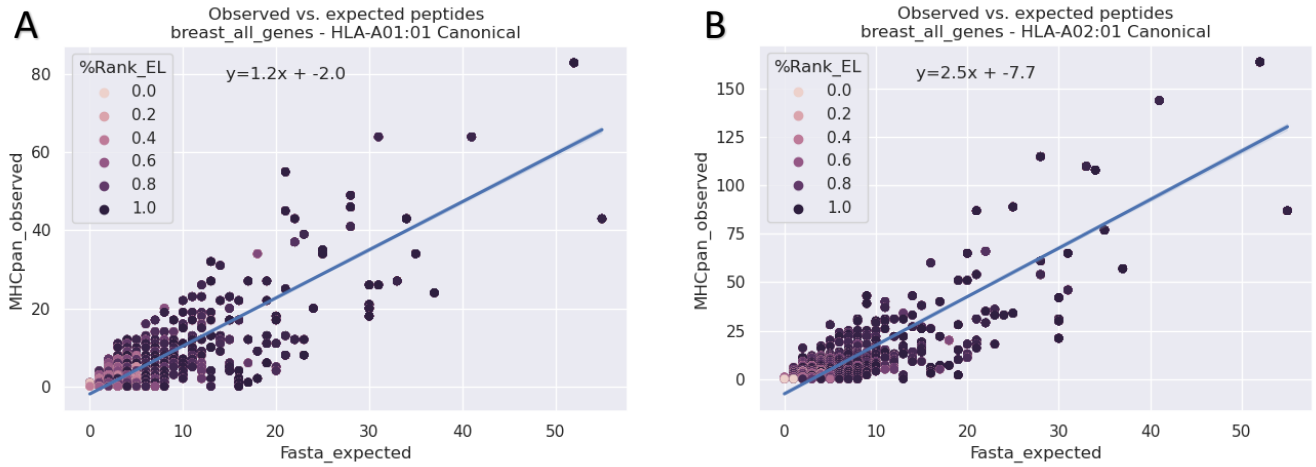


Figure 2, Scatterplots of observed vs. expected number of peptides in all genes breast dataset. Each dot represents a single read. A, dataset containing HLA-A01:01 allele predictions, regression analysis shows a 1.2 fold increase in the amount of observed peptides. B, dataset containing HLA-A02:02 allele predictions, regression analysis shows a 2.5 fold increase in the amount of observed peptides. Other HLA-A01:01 regression analysis show an increase, ranging from 0.8 to 1.4 times. For HLA-A02:01, this increases ranges from 1.6, to 2.5 (Supp. Fig. 2).

$$P_{Exp} = T_{Kmers} * C_{WB}$$

$$T_{Kmers} = L - k + 1$$

Equation 1, Equation to calculate number of expected peptides (P_{Exp}). Where,
T_{Kmers} is the total number of Kmers.
C_{WB} is the weak binder percentage cut-off,
L is the length of the read and
K is the length of the Kmers, (which has been fixed at 9 amino acids).

1.4 range Fig. 2A, Supp. Fig. 2A). The HLA-A02:01 alleles datasets however, had 2.0 to 3.0 fold more binders than expected (Fig. 2B, Supp. Fig. 2B). This difference is hypothesized to arise from the fact that HLA-A02:01 is more abundant in the population than HLA-A01:01 (33), meaning that NetMHCpan has been trained more on HLA-A02:01 data (32,34), resulting in more candidate peptides being found for HLA-A02:01 than HLA-A01:01.

The high number of predicted binders force us to focus on the top 1% as possible biomarker candidates. In this subset there is a clear difference in the number of HLA-A01:01 and HLA-02:01 binders (Fig. 3.). Among the total list of binders, the alleles showed a similar distribution in the Elution score ranking; HLA-A01:01 has a median of 0.52 (skin) and 0.51 (breast). HLA-A02:01 on the other hand, had the same median (0.44) for both the skin and breast dataset.

Due to HLA-A02:01 having significantly more peptides with a lower rank in the distribution plot than HLA-A01:01, we looked at the difference of allele distribution in the top 1% highest ranking peptides. (Fig. 3a). As expected, HLA-A02:01 had more of the peptides (61%), than HLA-A01:01 (39%) in the breast dataset (Fig. 2b). These results were similar for skin (HLA-A01:01 38%, HLA-A02:01 62%) (Supp. Fig. 3).

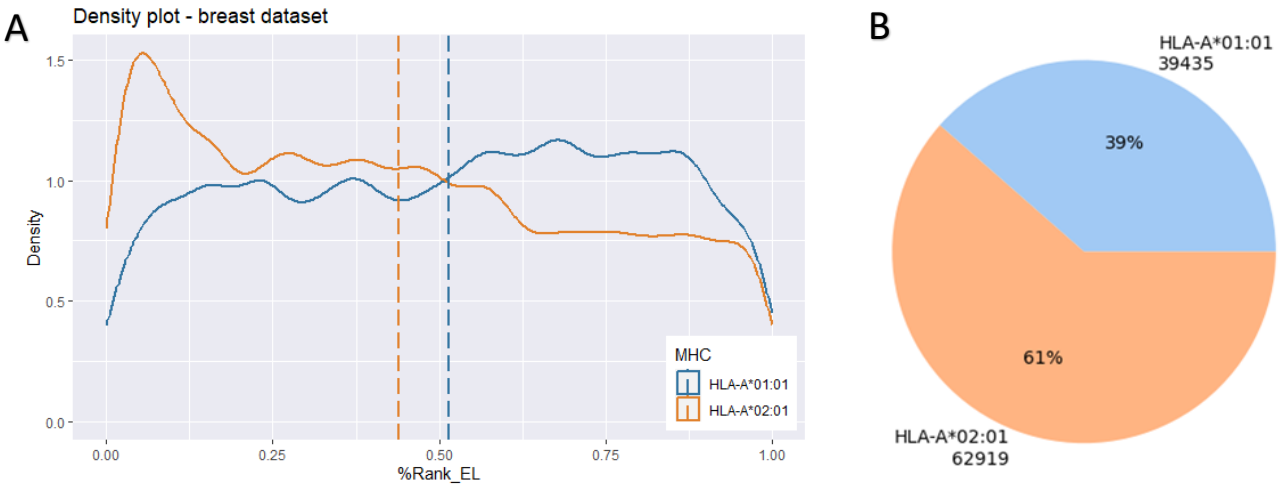


Figure 3, A, Density plot of the distribution of %Rank_EL for both HLA-A01:01 and HLA-A02:01 in the RNA-related, breast dataset. Dashed lines represent the median for both alleles. B, Pie plot with distribution for number of candidate peptides for both alleles in the top 1% highly expressed genes in the breast dataset. Figures for the skin dataset have similar results, shown in (Supp. Fig. 2).

POSSIBLE BIOMARKERS FOR EXPERIMENTAL TESTING

In our study, we identified cryptic peptide candidates (peptides with an elution rank of 1 or lower) in RNA-related genes and compared these with known, cryptic, Mass spectrometry peptides from other papers (2,3,16,32,35–42). Our final results included a total of 75 (5 HLA-A01:01 and 70 HLA-A02:01) peptides from the skin dataset, and 63 (1 HLA-A01:01 and 62 HLA-A02:01) peptides from the breast dataset (Supp Table 1). The difference in overlapping candidate peptides per allele could be caused by the fact that the amount of known peptides per allele differ. For HLA-A01:01, we gathered 2551 cryptic peptides, and 197 known canonical peptides. Whilst for HLA-A02:01, we accumulated 8842 cryptic peptides, and 599 known canonical peptides.

In Table 3 we assembled the top 5 overlapping peptides (based on elution score), including the gene function of the originating genes. Interestingly enough, one peptide was found originating from multiple transcripts in the breast dataset and was also found in the skin dataset. These peptides have been analyzed with Blast (43) to identify possible homologues.

We found strong evidence of peptide 'YLLEKFVAV' originating from ATP-dependent DNA helicase DDX11. Multiple studies (44,45) suggest that this lncRNA strand is a highly conserved oncogene. Furthermore, other research found this peptide in bladder cancer (46) and bone cancer (47), suggesting that it may be a potential pan-cancer biomarker. A perfect alignment was observed between peptide 'FLIPKFFEL' and the protein phosphatase 4 regulatory subunit (PPP4R1L) gene. Previous research (48) has shown that PPP4R1L is involved in protein dephosphorylation and has high expression in 27 different tissues, indicating that the peptide's origin cannot be traced back to a specific cell when cancer occurs.

In our study of the peptide 'ALAEVFHQL', we identified a novel homologue gene, which has been referenced by the synonyms 'KIAA0196' and 'WASHC5'. This gene has been primarily associated with the Ritscher-Schinzel (48) and hereditary spastic paraplegia syndrome (49).

In order to validate the identified peptide candidates, further research and experimental validation will be required.

Table 2, Shortlist of peptide candidates that have also been found by other papers, from both the skin and breast dataset, including both alleles.

Dataset	Pos	Peptide	Identity	%Rank_EL	Protein	Gene Description
TCGA-skin	498	YLLEKFVAV	ENST00000539757	0.001	ENSG00000111788	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide like pseudogene
TCGA-breast	449	YLLEKFVAV	ENST00000432996	0.001	DDX12P	DEAD/H-box helicase 12, pseudogene
TCGA-breast	498	YLLEKFVAV	ENST00000539757	0.001	ENSG00000111788	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide like pseudogene
TCGA-breast	196	FLIPKFFEL	ENST00000334187	0.002	PPP4R1L	protein phosphatase 4 regulatory subunit 1
TCGA-breast	64	ALAEVFHQL	ENST00000605862	0.002	ENSG00000242588	novel transcript

METHODS

DATA COLLECTION

Main data collection was performed using two databases; the National Cancer Institute's GDC Data Portal (https://portal.gdc.cancer.gov/) (50) and the Ensemble vertebrate genome browser (http://www.ensembl.org/) (51). The biggest challenge was gathering data that is publicly available. Overall in this project, only open, RNA-seq count data from multiple projects was tested and used.

The TCGA project (52) captures the most open-source patients RNA-count cancer data. To avoid biased data, other projects of the same cancer types were used for initial method testing. When no statistical differences were found, TCGA became the go-to project to use. Even though there are 55 tissue groups to choose from, we only looked at skin and breast cancer samples, due to the direct interest of our colleagues in the Molecular Targeted Therapies group.

DATA NORMALIZATION

Using R package DESEQ2, variance stabilization transformation was performed. This gene expression transformation showed an expression peak in the lower region of the distribution. To identify the top expressing genes in the dataset, the median for each gene in all patients has been calculated. The top percentile was selected as highest expressing genes, henceforth referred to as candidate genes.

SEQUENCE GATHERING

Gene sequences were downloaded from the Ensemble (http://www.ensembl.org/) database (53) at 11/24/2022. To this end, all the transcripts of a candidate gene are gathered, and filtered into coding & non-coding transcripts, depicted by the presence of a (known) coding region. When multiple non-coding transcripts overlap, they get elongated and conflated, containing the sequences of all overlapping transcripts. These non-coding transcripts will be referred to as cryptic transcripts. (Supp. Fig 4).

When all transcripts from the candidate genes have been gathered, the next step is to predict the translation. First, the transcripts with a known coding region are translated using the Bio-python package (54). The cryptic transcripts are translated by our in-house algorithm, which copies the alternative open reading frames method, developed by (31). This method implements new open reading frame (ORF) definitions, adopted by 37 research companies and academia around the world. This method (31) and other research (29,30) describes alternative open reading frames (aORF) being the direct cause of alternative translation. aORFs are reading frames that start and stop on different locations than the traditional start and stop codons in the CDS of a mRNA sequence. Internal out-of-frame ORFs are described as ORFs that have a reading frame shift inside the CDS, meaning that midway, other codons are being translated. These frame-shift ORFs were not used in our in-house algorithm, due to the fact that there is currently is no method known to accurately predict the locations of these internal frameshifts.

Finally, canonical sequences have been discarded from the main pipeline to ensure that all candidate peptides would be cryptic. These steps have also been summarized in Figure 1.

When the amino acid sequences have been determined, the actual HLA peptide binding prediction is done with NetMHCpan (32). The new version of this method uses elution data and data from *in vitro* binding essays for training. Because of this, NetMHCpan has shown to be one, if not the best tool for predicting the binding of MHC peptides (38). Furthermore, NetMHCpan is still used in state-of-the-art research (16,42,55–58).

The results from NetMHCpan have been filtered with user-set parameters;
--primary_site Skin OR Breast,
--cancer_project TCGA-SKCM or TCGA-BRCM,
--strong_binding_threshold 0.1,
--weak_binding_threshold 1,
--to_stop True,
--peptide_inclusive True.

Candidate peptides from NetMHCpan have been compared with known peptides from other papers (2,3,16,32,35–42). These peptides were downloaded from IEDB database (59) (https://www.iedb.org/) and are listed in the code provided.

Discussion

The main goal of this research was to develop a pipeline to detect cryptic neo-antigen peptides presented on MHC molecules. With this pipeline we showed that we are able to extract cryptic neo-antigens from open source, RNA count data. Performance of this pipeline was estimated with the use of known cryptic peptides from other research. However, experimental validation will be necessary to accurately test the performance of the pipeline.

One of the requirements in developing this pipeline was the strict use of open-source data. This made available data limited, resulting in only using RNA count data. Even though we were not able to obtain the sequence data, we were able to generate a workflow that predicts the sequence data. Obviously, we are not able to perform patient-specific mutation analysis. The mutational information could be promising for developing personalized treatments. but, patient-specific analysis was not in the scope of this project.

Further research could deliver insights into other cancer types. If significant changes are found, it could, in theory, mean that specific cancer/cell types deploy different cryptic translational pathways.

Cryptic peptide detection is a relatively new research subject, which will most likely improve when more data becomes available. Considering this, we made sure the pipeline is developed to be as flexible as possible, meaning the user should be able to tweak as much as they desire for their own research. This was achieved by adding plenteous parameters for the user to tweak.

During the development of this pipeline we found some interesting features and possible improvements that were not added due to our timeframe or project scope. Other research (57) has shown that aligning the candidate peptides to the genome can improve predictive accuracy. This could also be an added feature in future research.

Supplementary information

Supplementary figures and data can be found at the bottom of the paper. Supplementary data can also be downloaded from https://github.com/Sjonnie404/NeoPipeline/tree/Deployment/Data/Candidate%20peptides

Code availability

Source code for this project can be found at https://github.com/Sjonnie404/NeoPipeline.

ACKNOWLEDGEMENTS

The author would like to acknowledge the following people for helping with this research:
Dr. C. (Can) Kesmir
Dr. S.S. (Sabrina) Oliveira
Drs. J.K. (Jan Kees) van Amerongen
Dr. S.D. (Shreya) Dharadhar

REFERENCES

1. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries | Enhanced Reader.
2. Bassani-Sternberg M, Bräunlein E, Klar R, Engleitner T, Sinitcyn P, Audehm S, et al. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. Nat Commun. 2016;7(May).
3. Antwi K, Hanavan PD, Myers CE, Ruiz YW, Thompson EJ, Lake DF. Proteomic identification of an MHC-binding peptidome from pancreas and breast cancer cell lines. Mol Immunol. 2009;46(15):2931–7.
4. Bukowski K, Kciuk M, Kontek R. Mechanisms of multidrug resistance in cancer chemotherapy. Int J Mol Sci. 2020;21(9).
5. Unger FT, Witte I, David KA. Prediction of individual response to anticancer therapy: Historical and future perspectives. Cell Mol Life Sci. 2015;72(4):729–57.
6. El-Deiry WS, Goldberg RM, Lenz H, Shields AF, Gibney GT, Tan AR, et al. The current state of molecular testing in the treatment of patients with solid tumors, 2019. CA Cancer J Clin. 2019;69(4):305–43.
7. Spinney L. Caught in time. Nature. 2006;442(7104):736–8.
8. Roncarolo MG, Battaglia M. Regulatory T-cell immunotherapy for tolerance to self antigens and alloantigens in humans. Nat Rev Immunol. 2007;7(8):585–98.
9. Jiang T, Shi T, Zhang H, Hu J, Song Y, Wei J, et al. Tumor neoantigens: From basic research to clinical applications. J Hematol Oncol. 2019;12(1):1–13.
10. Zhang Z, Lu M, Qin Y, Gao W, Tao L, Su W, et al. Neoantigen: A New Breakthrough in Tumor Immunotherapy. Front Immunol. 2021;12(April):1–9.
11. Zhao X, Pan X, Wang Y, Zhang Y. Targeting neoantigens for cancer immunotherapy. Biomark Res. 2021;9(1):1–12.
12. Aebersold R, Mann M. Mass spectrometry-based proteomics: Abstract: Nature. Nature. 2003;422(6928):198–207.
13. Shen Y, Tolic N, Hixson KK, Purvine SO, Paš a-Tolic L, Qian W-J, et al. Proteome-Wide Identification of Proteins and Their Modifications with Decreased Ambiguities and Improved False Discovery Rates Using Unique Sequence Tags. Rapid Commun Mass Spectrom [Internet]. 1986;83(1):1871–82. Available from: http://www.unimod.org/andhttp://www.abrf.org/index.cfm/dm.home.
14. Boon T, Pel A Van. T cell-recognized antigenic peptides derived from the cellular genome are not protein degradation products but can be generated directly by transcription and translation of short subgenic regions. A hypothesis. Vol. 29. 1989.
15. Starck SR, Ow Y, Jiang V, Tokuyama M, Rivera M. A Distinct Translation Initiation Mechanism Generates Cryptic Peptides for Immune Surveillance. PLoS One [Internet]. 2008;3(10):3460. Available from: www.plosone.org
16. Ruiz Cuevas MV, Hardy MP, Hollý J, Bonneil É, Durette C, Courcelles M, et

al. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. Cell Rep. 2021;34(10).

17. Day S, Ramsland P, Apostolopoulos V. Non-Canonical Peptides Bound to MHC. Curr Pharm Des. 2009;15(28):3274–82.

18. Laumont CM, Perreault C. Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy. Cell Mol Life Sci. 2018;75(4):607–21.

19. Gladue DP, Calis J, Prusty BK, Lodha M, Erhard F, Dölken L. The Hidden Enemy Within: Non-canonical Peptides in Virus-Induced Autoimmunity. 2022; Available from: www.frontiersin.org

20. Dančík V, Addona TA, Clauser KR, Vath JE, Pevzner PA. De novo peptide sequencing via tandem mass spectrometry. J Comput Biol. 1999;6(3–4):327–42.

21. Matthiesen R, Jensen ON. Analysis of mass spectrometry data in proteomics. Vol. 453, Methods in Molecular Biology. 2008. 105–122 p.

22. Azari S, Xue B, Zhang M, Peng L. GA-novo: De novo peptide sequencing via tandem mass spectrometry using genetic algorithm [Internet]. Vol. 11454 LNCS, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer International Publishing; 2019. 72–89 p. Available from: http://dx.doi.org/10.1007/978-3-030-16692-2_6

23. Yilmaz M, Fondrie WE, Bittremieux W, Oh S, Noble WS. De novo mass spectrometry peptide sequencing with a transformer model. bioRxiv [Internet]. 2022;2022.02.07.479481. Available from: https://www.biorxiv.org/content/10.1101/2022.02.07.479481v2%0Ahttps://www.biorxiv.org/content/10.1101/2022.02.07.479481v2.abstract

24. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov G V, lo christopher, McDonald iii robert, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Available from: https://doi.org/10.1038/s41586-019-1186-3

25. Wang H, Wang Y, Xie S, Liu Y, Xie Z. Global and cell-type specific properties of lincRNAs with ribosome occupancy. Nucleic Acids Res. 2017;45(5):2786–96.

26. Choi SW, Kim HW, Nam JW. The small peptide world in long noncoding RNAs. Brief Bioinform. 2019;20(5):1853–64.

27. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Ridwan Amode M, et al. Ensembl 2021. Nucleic Acids Res. 2021;49(D1):D884–91.

28. Ensemble. Transcript flags [Internet]. 2022. Available from: https://www.ensembl.org/info/genome/genebuild/transcript_quality_tags.html

29. Erhard F, Halenius A, Zimmermann C, L'Hernault A, Kowalewski DJ, Weekes MP, et al. Improved Ribo-seq enables identification of cryptic translation events. Nat Methods. 2018;15(5):363–6.

30. Reuter K, Biehl A, Koch L, Helms V. PreTIS: A Tool to Predict Non-canonical 5' UTR Translational Initiation Sites in Human and Mouse. 2016; Available from: http://service.bioinformatik.uni-saarland.de/

31. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, et al. Standardized annotation of translated open reading frames [Internet]. 2022. Available from: www.nature.com/naturebiotechnology

32. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and

NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res [Internet]. 2020;48:449–54. Available from: http://www.cbs.dtu.dk/services/NetMHCIIpan-4.0/.

33. Que TN, Khanh NB, Khanh BQ, Van Son C, Van Anh NT, Anh TTT, et al. Allele and Haplotype Frequencies of HLA-A, -B, -C, and -DRB1 Genes in 3,750 Cord Blood Units From a Kinh Vietnamese Population. Front Immunol. 2022;13(June):1–11.

34. 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. J Immunol. 2017 Nov 1;199(9):3360–8.

35. Kumar P, Boyne C, Brown S, Qureshi A, Thorpe P, Synowsky SA, et al. Tumour-associated antigenic peptides are present in the HLA class I ligandome of cancer cell line derived extracellular vesicles. Immunology. 2022;166(2):249–64.

36. Garcia-Boronat M, Diez-Rivero CM, Reinherz EL, Reche PA. PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery. Nucleic Acids Res. 2008;36(Web Server issue):35–41.

37. Reche PA, Zhang H, Glutting JP, Reinherz EL. EPIMHC: A curated database of MHC-binding peptides for customized computational vaccinology. Bioinformatics. 2005;21(9):2140–1.

38. Boehm KM, Bhinder B, Raja VJ, Dephoure N, Elemento O. Predicting peptide presentation by major histocompatibility complex class I using one million peptides. bioRxiv. 2018;1–11.

39. Pezze PD, Ruf S, Sonntag AG, Langelaar-Makkinje M, Hall P, Heberle AM, et al. A systems study reveals concurrent activation of AMPK and mTOR by amino acids. Nat Commun. 2016;7:1–19.

40. Pritchard AL, Hastie ML, Neller M, Gorman JJ, Schmidt CW, Hayward NK. Exploration of peptides bound to MHC class I molecules in melanoma. Pigment Cell Melanoma Res. 2015;28(3):281–94.

41. Gloger A, Ritz D, Fugmann T, Neri D. Mass spectrometric analysis of the HLA class I peptidome of melanoma cell lines as a promising tool for the identification of putative tumor-associated HLA epitopes. Cancer Immunol Immunother. 2016;65(11):1377–93.

42. Erhard F, Dölken L, Schilling B, Schlosser A. Identification of the cryptic HLA-I immunopeptidome. Cancer Immunol Res. 2020 Aug 1;8(8):1018–26.

43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

44. Feng Y, Wu M, Hu S, Peng X, Chen F. LncRNA DDX11-AS1: a novel oncogene in human cancer. Hum Cell [Internet]. 2020;33(4):946–53. Available from: https://doi.org/10.1007/s13577-020-00409-8

45. Bhattacharya C, Wang X, Becker D. The DEAD/DEAH box helicase, DDX11, is essential for the survival of advanced melanomas. Mol Cancer. 2012;11:1–10.

46. Li Q, Wang S, Wu Z, Liu Y. DDX11-AS1exacerbates bladder cancer progression by enhancing CDK6 expression via suppressing miR-499b-5p. Biomed Pharmacother [Internet]. 2020;127(February):110164. Available from:

https://doi.org/10.1016/j.biopha.2020.110164

47. Zhang H, Lin J, Chen J, Gu W, Mao Y, Wang H, et al. DDX11-AS1 contributes to osteosarcoma progression via stabilizing DDX11. Life Sci [Internet]. 2020;254(November 2019):117392. Available from: https://doi.org/10.1016/j.lfs.2020.117392

48. Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics. 2014;13(2):397–406.

49. Ichinose Y, Koh K, Fukumoto M, Yamashiro N, Kobayashi F, Miwa M, et al. Exome sequencing reveals a novel missense mutation in the KIAA0196 gene in a Japanese patient with SPG8. Clin Neurol Neurosurg [Internet]. 2016;144:36–8. Available from: http://dx.doi.org/10.1016/j.clineuro.2016.02.031

50. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. N Engl J Med. 2016;375(12):1109–12.

51. Lykourentzou I, Giannoukos I, Nikolopoulos V, Mpardis G, Loumos V. Dropout prediction in e-learning courses through the combination of machine learning techniques. Comput Educ. 2009;53(3):950–65.

52. Guan J, Gupta R, Filipp F V. Cancer systems biology of TCGA SKCM: Efficient detection of genomic drivers in melanoma. Available from: www.nature.com/scientificreports

53. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Ridwan Amode M, Armean IM, et al. Ensembl 2022. Database issue Nucleic Acids Res [Internet]. 2022 [cited 2022 Sep 20];50:989. Available from: https://doi.org/10.1093/nar/gkab1049

54. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. 2009;25(11):1422–3. Available from: www.python.org

55. Tarke A, Coelho CH, Zhang Z, Dan JM, Yu ED, Methot N, et al. SARS-CoV-2 vaccination induces immunological T cell memory able to cross-recognize variants from Alpha to Omicron. Cell. 2022;185(5):847-859.e11.

56. Koşaloğlu-Yalçın Z, Lee J, Greenbaum J, Schoenberger SP, Miller A, Kim YJ, et al. Combined assessment of MHC binding and antigen abundance improves T cell epitope predictions. iScience. 2022;25(2).

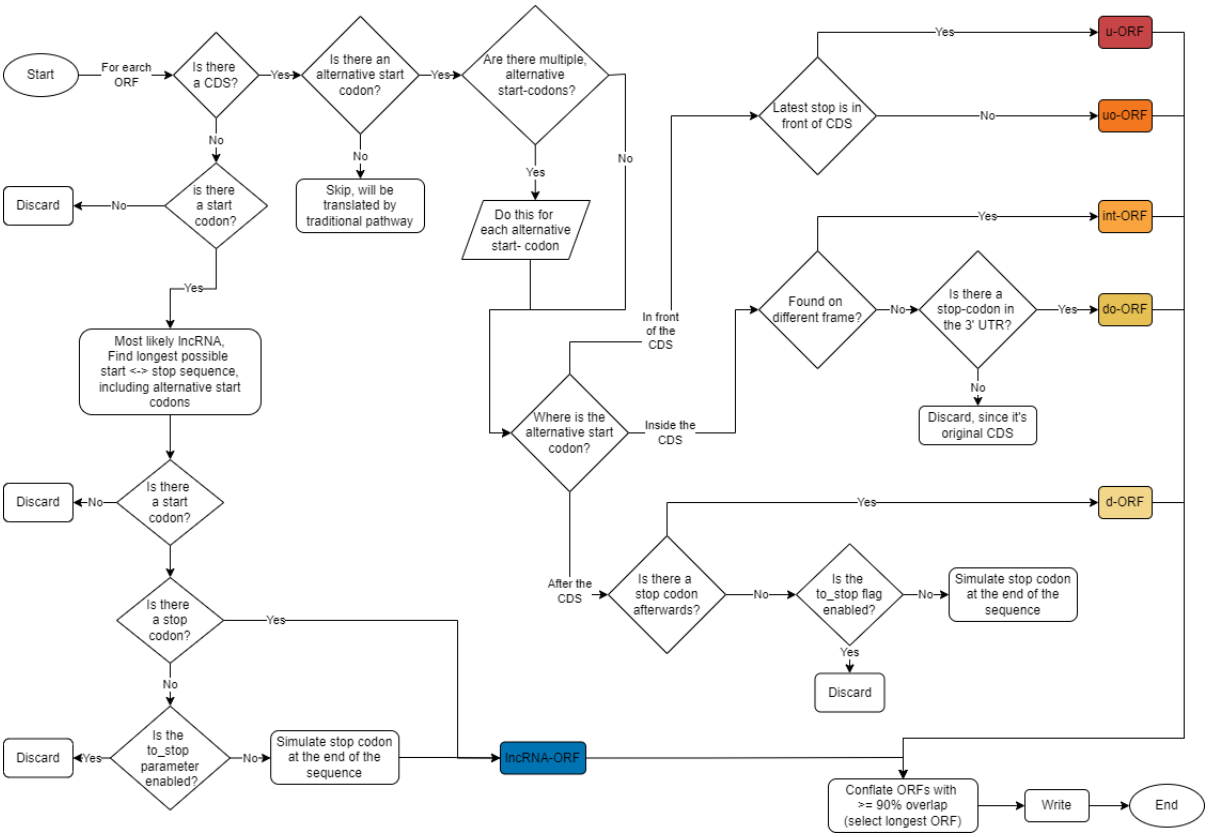
57. Bedran G, Wang T, Pankanin D, Weke K, Laird A, Battail C, et al. The Immunopeptidome From a Genomic Perspective: Establishing Immune-Relevant Regions for Cancer Vaccine Design. SSRN Electron J. 2022;1–26.

58. Minervina AA, Pogorelyy M V., Kirk AM, Crawford JC, Allen EK, Chou C-H, et al. SARS-CoV-2 antigen exposure history shapes phenotypes and specificity of memory CD8+ T cells. Nat Immunol. 2022;23(5):781–90.

59. Fleri W, Paul S, Dhanda SK, Mahajan S, Xu X, Peters B, et al. The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. Front Immunol. 2017;8(MAR):1–16.

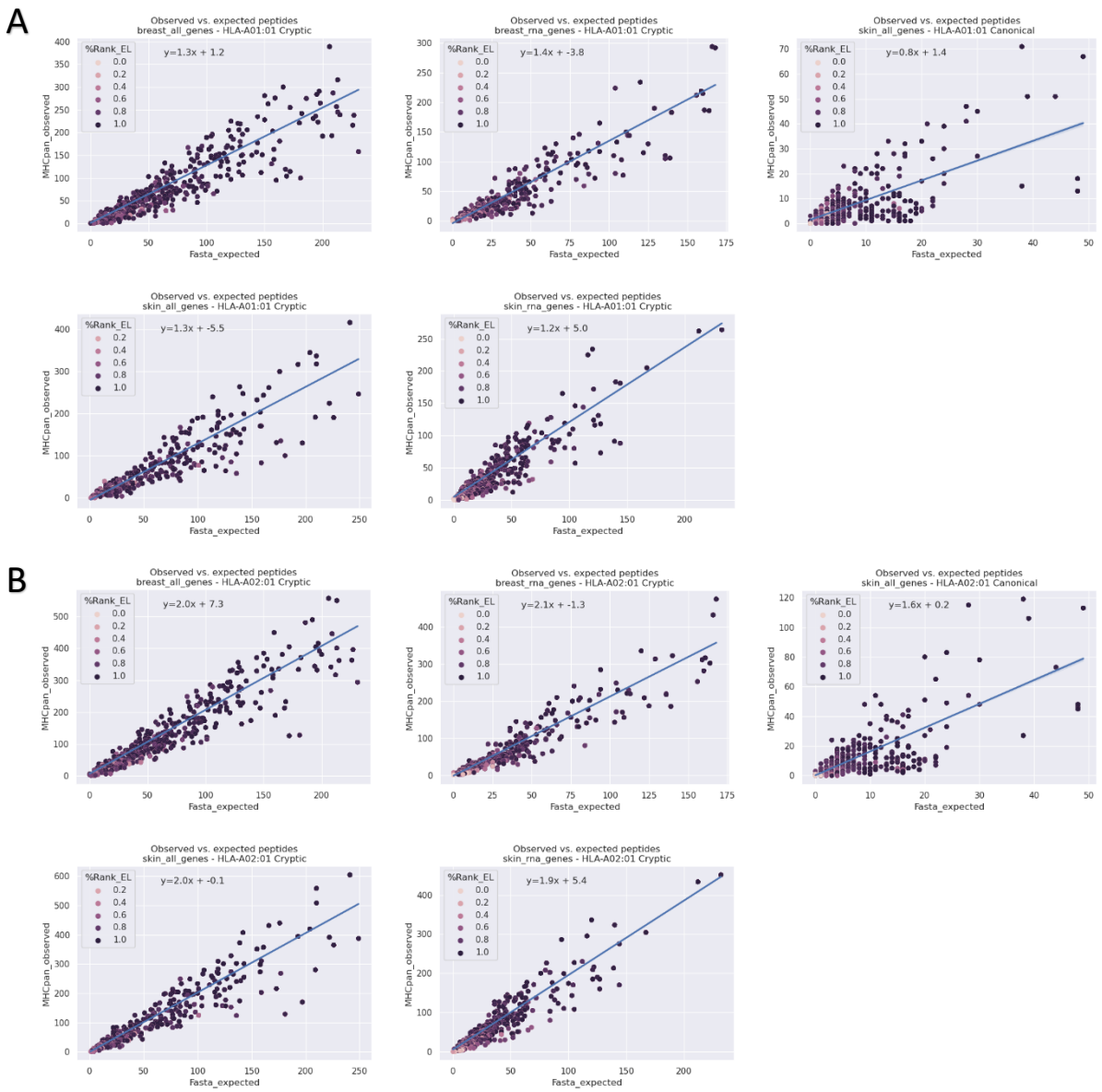
SUPPLEMENTARY INFORMATION

Supplementary figure 1



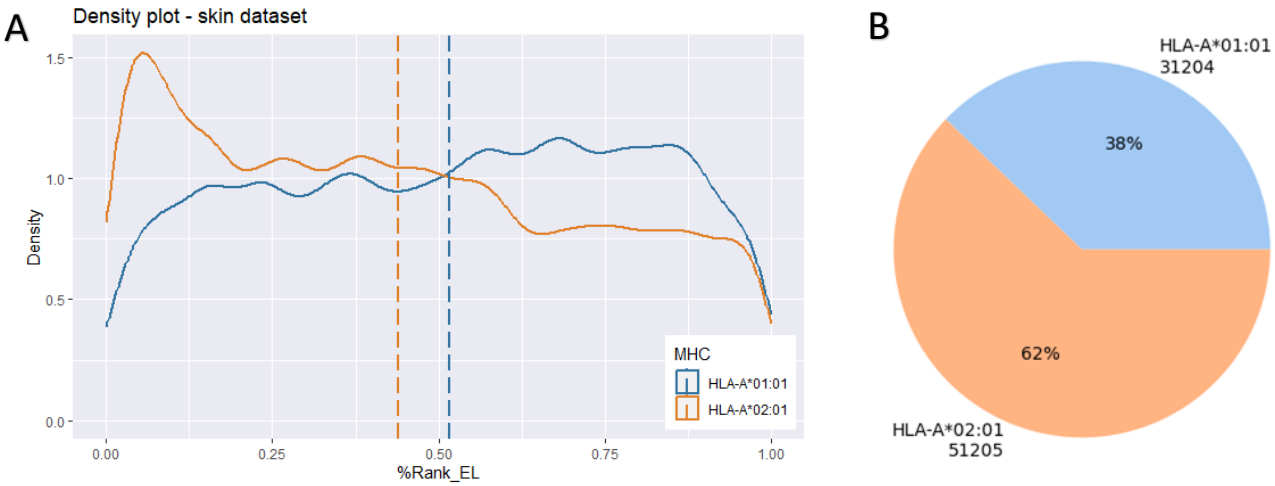
Supplementary Figure 1, Flowchart of in-house algorithm to detect alternative, Open Reading Frames (aORFs). Code methods and colors are inspired by the figures of other research (31).

Supplementary figure 2



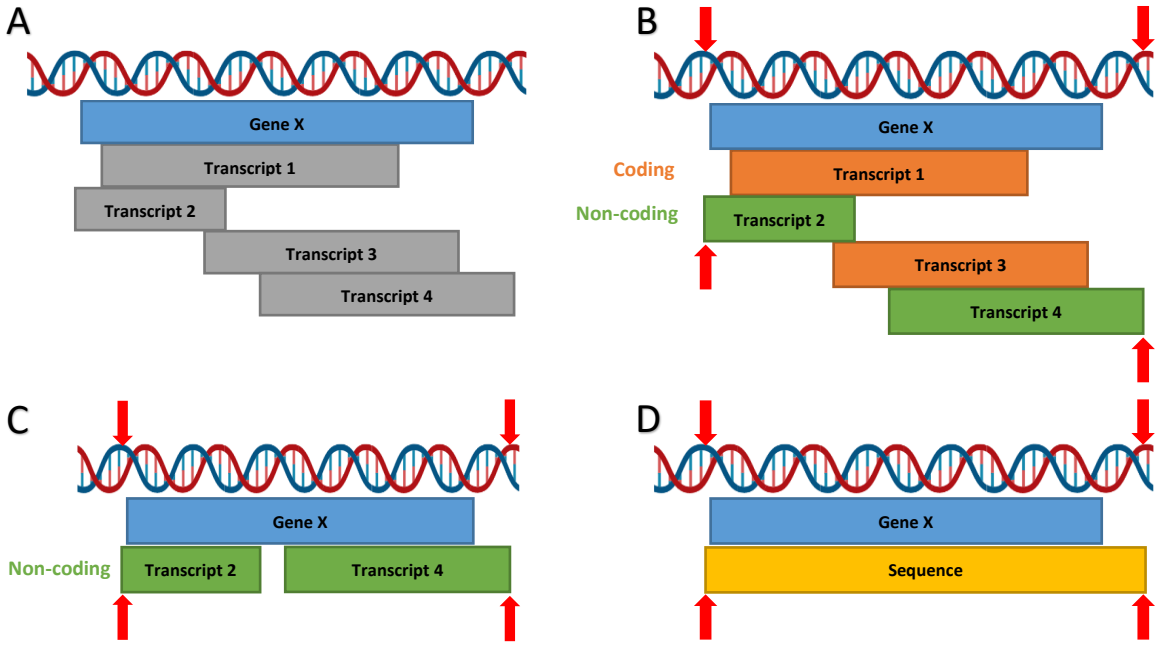
Supplementary Figure 2, Extra scatterplots (observed vs. expected) for candidate peptides for other parameters. A, Scatterplots regarding HLA-A01:01 allele predictions, regression analysis shows a 0.8 to 1.4 fold increase in the amount of observed peptides. B, Scatterplots regarding HLA-A02:01 allele predictions, regression analysis shows a 1.6 to 2.5 fold increase in the amount of observed peptides. (regression folds from Figure 1 are also taken into account.)

Supplementary figure 3



Supplementary Figure 3, A, Density plot of the distribution of %Rank_EL for both HLA-A01:01 and HLA-A02:01 in the RNA-related, skin dataset. Dashed lines represent the median for both alleles. B, Pie plot with distribution for number of candidate peptides for both alleles in the top 1% highly expressed genes in the skin dataset.

Supplementary figure 4



Supplementary Figure 4, Global overview of non-coding sequence gathering. A, first step is to identify all transcripts that are associated with the selected gene. B, second step is to identify and label which transcripts have a known coding region and which do not. Then the earliest start- and last stop codon coordinates are saved (indicated with the red arrows). C, Transcript sequences with a known CDS are saved and discarded from the list. D, The whole sequence is gathered from the earlier saved coordinates. Later in the pipeline, peptides that are both found in coding and non-coding sequences are discarded, due to possible contamination.

Supplementary Table 1, NetMHCpan output of candidate peptides for both the skin and breast dataset, including both alleles. (1 / 5)

Dataset	Pos	MHC HLA-A*	Peptide	Core	Of	Gp	Gl	Ip	Il	Icore	Identity	Score EL	%Rank EL	Score BA	%Rank BA	Aff(nM)	Bind Level
TCGA-Breast	72	01:01	HTEPLDELY	HTEPLDELY	0	0	0	0	0	HTEPLDELY	ENST00000523992	0.992338	0.004	0.73937	0.015	16.78	SB
TCGA-Breast	449	02:01	YLLEKFVAV	YLLEKFVAV	0	0	0	0	0	YLLEKFVAV	ENST00000432996	0.99692	0.001	0.939803	0.005	1.92	SB
TCGA-Breast	498	02:01	YLLEKFVAV	YLLEKFVAV	0	0	0	0	0	YLLEKFVAV	ENST00000539757	0.99692	0.001	0.939803	0.005	1.92	SB
TCGA-Breast	196	02:01	FLIPKFFEL	FLIPKFFEL	0	0	0	0	0	FLIPKFFEL	ENST00000334187	0.994211	0.002	0.927875	0.006	2.18	SB
TCGA-Breast	64	02:01	ALAEVFHQL	ALAEVFHQL	0	0	0	0	0	ALAEVFHQL	ENST00000605862	0.994985	0.002	0.860882	0.026	4.51	SB
TCGA-Breast	24	02:01	SLIEHLQGL	SLIEHLQGL	0	0	0	0	0	SLIEHLQGL	ENST00000439564	0.989222	0.005	0.851983	0.032	4.96	SB
TCGA-Breast	241	02:01	QLAQFVHEV	QLAQFVHEV	0	0	0	0	0	QLAQFVHEV	ENST00000432996	0.985561	0.006	0.837955	0.053	5.77	SB
TCGA-Breast	290	02:01	QLAQFVHEV	QLAQFVHEV	0	0	0	0	0	QLAQFVHEV	ENST00000539757	0.985561	0.006	0.837955	0.053	5.77	SB
TCGA-Breast	515	02:01	GLWGPVHEL	GLWGPVHEL	0	0	0	0	0	GLWGPVHEL	ENST00000414990	0.986793	0.006	0.824723	0.067	6.66	SB
TCGA-Breast	492	02:01	VLAKELVEV	VLAKELVEV	0	0	0	0	0	VLAKELVEV	ENST00000412962	0.983159	0.007	0.830255	0.059	6.28	SB
TCGA-Breast	453	02:01	RLWDEVMQA	RLWDEVMQA	0	0	0	0	0	RLWDEVMQA	ENST00000414990	0.980028	0.008	0.825399	0.066	6.61	SB
TCGA-Breast	6	02:01	VLGPIINKV	VLGPIINKV	0	0	0	0	0	VLGPIINKV	ENST00000424092	0.97678	0.01	0.732191	0.263	18.13	SB
TCGA-Breast	44	02:01	FLNDIFERI	FLNDIFERI	0	0	0	0	0	FLNDIFERI	ENST00000369385	0.971245	0.013	0.863228	0.025	4.39	SB
TCGA-Breast	30	02:01	ALDNGLFTL	ALDNGLFTL	0	0	0	0	0	ALDNGLFTL	ENST00000278882	0.966495	0.016	0.802336	0.098	8.49	SB
TCGA-Breast	387	02:01	KLGSVPVTV	KLGSVPVTV	0	0	0	0	0	KLGSVPVTV	ENST00000414990	0.959472	0.019	0.734274	0.256	17.73	SB
TCGA-Breast	607	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000550135	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Breast	1228	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000602436	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Breast	16	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000451424	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Breast	310	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000654763	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Breast	577	02:01	ALNNLLHSL	ALNNLLHSL	0	0	0	0	0	ALNNLLHSL	ENST00000522480	0.952066	0.023	0.751848	0.199	14.66	SB
TCGA-Breast	702	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000602458	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Breast	1056	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000659614	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Breast	1377	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000608477	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Breast	430	02:01	YVTDVLYRV	YVTDVLYRV	0	0	0	0	0	YVTDVLYRV	ENST00000522480	0.949292	0.024	0.837463	0.053	5.8	SB
TCGA-Breast	478	02:01	IVADVQISV	IVADVQISV	0	0	0	0	0	IVADVQISV	ENST00000414990	0.9292	0.032	0.715975	0.314	21.61	SB
TCGA-Breast	40	02:01	GLVNYQISV	GLVNYQISV	0	0	0	0	0	GLVNYQISV	ENST00000449061	0.929385	0.032	0.819201	0.074	7.07	SB
TCGA-Breast	424	02:01	YLGHLQQYV	YLGHLQQYV	0	0	0	0	0	YLGHLQQYV	ENST00000266746	0.927329	0.034	0.856079	0.029	4.75	SB

Supplementary Table 1, continued (2 / 5)

Dataset	Pos	MHC HLA-A*	Peptide	Core	Of	Gp	Gl	Ip	Il	Icore	Identity	Score EL	%Rank EL	Score BA	%Rank BA	Aff(nM)	Bind Level
TCGA-Breast	46	02:01	QLISIILRL	QLISIILRL	0	0	0	0	0	QLISIILRL	ENST00000551901	0.925103	0.035	0.701469	0.375	25.28	SB
TCGA-Breast	209	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000381800	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	346	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000701321	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	416	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000623673	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	57	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000635600	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	30400	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000597346	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	528	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000499521	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	3099	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000605862	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	210	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000456273	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	782	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000656196	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	1406	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000647856	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	1271	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000608477	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	1080	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000647856	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	309	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000562760	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	578	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000647856	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	6599	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000597346	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	488	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000551271	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	361	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000671580	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	663	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000551271	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	2	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000686891	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	565	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000621919	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	689	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000621919	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	1090	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000641433	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	1537	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000641433	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	1023	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000641433	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Breast	923	02:01	YLDQTLPRA	YLDQTLPRA	0	0	0	0	0	YLDQTLPRA	ENST00000539757	0.915058	0.041	0.721541	0.294	20.35	SB
TCGA-Breast	222	02:01	YQNQEIHNL	YQNQEIHNL	0	0	0	0	0	YQNQEIHNL	ENST00000374336	0.851158	0.074	0.582194	1.025	91.89	SB
TCGA-Breast	153	02:01	KVADLVLML	KVADLVLML	0	0	0	0	0	KVADLVLML	ENST00000374336	0.832821	0.085	0.654882	0.569	41.85	SB

Supplementary Table 1, continued (3 / 5)

Dataset	Pos	MHC HLA-A*	Peptide	Core	Of	Gp	Gl	Ip	Il	Icore	Identity	Score EL	%Rank EL	Score BA	%Rank BA	Aff(nM)	Bind Level
TCGA-Breast	195	02:01	RLQQLQHRV	RLQQLQHRV	0	0	0	0	0	RLQQLQHRV	ENST00000539757	0.800719	0.107	0.58412	1.01	89.99	WB
TCGA-Breast	121	02:01	RLQQLQHRV	RLQQLQHRV	0	0	0	0	0	RLQQLQHRV	ENST00000432996	0.800719	0.107	0.58412	1.01	89.99	WB
TCGA-Breast	287	02:01	LLWQSLILL	LLWQSLILL	0	0	0	0	0	LLWQSLILL	ENST00000301665	0.748795	0.148	0.751919	0.199	14.65	WB
TCGA-Breast	119	02:01	QLDWDVATV	QLDWDVATV	0	0	0	0	0	QLDWDVATV	ENST00000477247	0.692592	0.19	0.703558	0.366	24.72	WB
TCGA-Breast	170	02:01	VLITAVLLL	VLITAVLLL	0	0	0	0	0	VLITAVLLL	ENST00000414273	0.602202	0.278	0.684808	0.444	30.27	WB
TCGA-Breast	181	02:01	TLAEFQVIM	TLAEFQVIM	0	0	0	0	0	TLAEFQVIM	ENST00000414990	0.570685	0.311	0.589795	0.965	84.63	WB
TCGA-Breast	589	02:01	RQAEQEATV	RQAEQEATV	0	0	0	0	0	RQAEQEATV	ENST00000412962	0.487702	0.408	0.487586	1.919	255.75	WB
TCGA-Breast	90	02:01	YLIPIVVRY	YLIPIVVRY	0	0	0	0	0	YLIPIVVRY	ENST00000334187	0.281158	0.819	0.27039	6.765	2681.73	WB
TCGA-Skin	72	01:01	HTEPLDELY	HTEPLDELY	0	0	0	0	0	HTEPLDELY	ENST00000510506	0.992338	0.004	0.73937	0.015	16.78	SB
TCGA-Skin	325	01:01	SSDRKGGSY	SSDRKGGSY	0	0	0	0	0	SSDRKGGSY	ENST00000383620	0.974135	0.013	0.662035	0.037	38.73	SB
TCGA-Skin	166	01:01	EVDTFMEAY	EVDTFMEAY	0	0	0	0	0	EVDTFMEAY	ENST00000566851	0.905706	0.051	0.612493	0.061	66.2	SB
TCGA-Skin	145	01:01	FTATRPGVY	FTATRPGVY	0	0	0	0	0	FTATRPGVY	ENST00000427426	0.797859	0.101	0.610635	0.063	67.55	WB
TCGA-Skin	25	01:01	YLDPAQQNL	YLDPAQQNL	0	0	0	0	0	YLDPAQQNL	ENST00000421406	0.41373	0.378	0.25185	1.066	3277.44	WB
TCGA-Skin	498	02:01	YLLEKFVAV	YLLEKFVAV	0	0	0	0	0	YLLEKFVAV	ENST00000539757	0.99692	0.001	0.939803	0.005	1.92	SB
TCGA-Skin	196	02:01	FLIPKFFEL	FLIPKFFEL	0	0	0	0	0	FLIPKFFEL	ENST00000334187	0.994211	0.002	0.927875	0.006	2.18	SB
TCGA-Skin	64	02:01	ALAEVFHQL	ALAEVFHQL	0	0	0	0	0	ALAEVFHQL	ENST00000605862	0.994985	0.002	0.860882	0.026	4.51	SB
TCGA-Skin	25	02:01	YLDPAQQNL	YLDPAQQNL	0	0	0	0	0	YLDPAQQNL	ENST00000421406	0.992484	0.003	0.786521	0.131	10.07	SB
TCGA-Skin	290	02:01	QLAQFVHEV	QLAQFVHEV	0	0	0	0	0	QLAQFVHEV	ENST00000539757	0.985561	0.006	0.837955	0.053	5.77	SB
TCGA-Skin	492	02:01	VLAKELVEV	VLAKELVEV	0	0	0	0	0	VLAKELVEV	ENST00000412962	0.983159	0.007	0.830255	0.059	6.28	SB
TCGA-Skin	71	02:01	SLVELLVQL	SLVELLVQL	0	0	0	0	0	SLVELLVQL	ENST00000454465	0.980079	0.008	0.791773	0.121	9.52	SB
TCGA-Skin	240	02:01	ALLSGIVSI	ALLSGIVSI	0	0	0	0	0	ALLSGIVSI	ENST00000440904	0.970534	0.013	0.826899	0.064	6.51	SB
TCGA-Skin	503	02:01	GLYPQSPLL	GLYPQSPLL	0	0	0	0	0	GLYPQSPLL	ENST00000529624	0.961821	0.018	0.75225	0.198	14.59	SB
TCGA-Skin	16	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000451424	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Skin	718	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000660864	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Skin	252	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000668238	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Skin	301	02:01	GMNDMNHEV	GMNDMNHEV	0	0	0	0	0	GMNDMNHEV	ENST00000454465	0.950998	0.023	0.820898	0.072	6.94	SB
TCGA-Skin	702	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000602458	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Skin	634	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000564629	0.950274	0.023	0.915931	0.007	2.48	SB

Supplementary Table 1, continued (4 / 5)

Dataset	Pos	MHC HLA-A*	Peptide	Core	Of	Gp	Gl	Ip	Il	Icore	Identity	Score EL	%Rank EL	Score BA	%Rank BA	Aff(nM)	Bind Level
TCGA-Skin	3717	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000623726	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Skin	3370	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000623726	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Skin	1056	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000659614	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Skin	1014	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000623726	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Skin	4856	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000499624	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Skin	310	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000654763	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Skin	607	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000550135	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Skin	310	02:01	WLMPVIPAL	WLMPVIPAL	0	0	0	0	0	WLMPVIPAL	ENST00000437833	0.950274	0.023	0.915931	0.007	2.48	SB
TCGA-Skin	40	02:01	GLVNYQISV	GLVNYQISV	0	0	0	0	0	GLVNYQISV	ENST00000449061	0.929385	0.032	0.819201	0.074	7.07	SB
TCGA-Skin	424	02:01	YLGHLQQYV	YLGHLQQYV	0	0	0	0	0	YLGHLQQYV	ENST00000266746	0.927329	0.034	0.856079	0.029	4.75	SB
TCGA-Skin	2793	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000392097	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	452	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000392097	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	3099	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000605862	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	951	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000598377	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	707	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000609755	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	614	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000609755	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	782	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000656196	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	4138	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000499624	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	159	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000442526	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	663	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000551271	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	795	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000660864	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	1790	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000665286	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	987	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000513358	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	488	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000551271	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	4054	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000499624	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	626	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000499624	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	417	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000506172	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	557	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000413987	0.917518	0.04	0.835139	0.055	5.95	SB

Supplementary Table 1, continued (5 / 5)

Dataset	Pos	MHC HLA-A*	Peptide	Core	Of	Gp	Gl	Ip	Il	Icore	Identity	Score EL	%Rank EL	Score BA	%Rank BA	Aff(nM)	Bind Level
TCGA-Skin	557	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000413987	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	1127	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000416860	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	1952	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000623726	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	4550	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000623726	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	4445	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000499624	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	66	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000562082	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	57	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000635600	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	940	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000660724	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	1753	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000665286	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	1569	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000624350	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	292	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000654838	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	361	02:01	WLTPVIPAL	WLTPVIPAL	0	0	0	0	0	WLTPVIPAL	ENST00000671580	0.917518	0.04	0.835139	0.055	5.95	SB
TCGA-Skin	923	02:01	YLDQTLPRA	YLDQTLPRA	0	0	0	0	0	YLDQTLPRA	ENST00000539757	0.915058	0.041	0.721541	0.294	20.35	SB
TCGA-Skin	27	02:01	TILPAIILV	TILPAIILV	0	0	0	0	0	TILPAIILV	ENST00000427426	0.846628	0.077	0.7073	0.351	23.73	SB
TCGA-Skin	669	02:01	KVPEIEVTV	KVPEIEVTV	0	0	0	0	0	KVPEIEVTV	ENST00000638517	0.815565	0.097	0.547083	1.305	134.35	SB
TCGA-Skin	11	02:01	SLIAKVATA	SLIAKVATA	0	0	0	0	0	SLIAKVATA	ENST00000434500	0.81275	0.099	0.667022	0.517	36.7	SB
TCGA-Skin	195	02:01	RLQQLQHRV	RLQQLQHRV	0	0	0	0	0	RLQQLQHRV	ENST00000539757	0.800719	0.107	0.58412	1.01	89.99	WB
TCGA-Skin	287	02:01	LLWQSLILL	LLWQSLILL	0	0	0	0	0	LLWQSLILL	ENST00000301665	0.748795	0.148	0.751919	0.199	14.65	WB
TCGA-Skin	206	02:01	QLDWDVATV	QLDWDVATV	0	0	0	0	0	QLDWDVATV	ENST00000513466	0.692592	0.19	0.703558	0.366	24.72	WB
TCGA-Skin	211	02:01	QLDWDVATV	QLDWDVATV	0	0	0	0	0	QLDWDVATV	ENST00000510506	0.692592	0.19	0.703558	0.366	24.72	WB
TCGA-Skin	119	02:01	QLDWDVATV	QLDWDVATV	0	0	0	0	0	QLDWDVATV	ENST00000477247	0.692592	0.19	0.703558	0.366	24.72	WB
TCGA-Skin	80	02:01	ALSDPPALA	ALSDPPALA	0	0	0	0	0	ALSDPPALA	ENST00000650759	0.654447	0.229	0.499542	1.788	224.72	WB
TCGA-Skin	175	02:01	YLENGKETL	YLENGKETL	0	0	0	0	0	YLENGKETL	ENST00000383620	0.569159	0.312	0.461117	2.265	340.56	WB
TCGA-Skin	589	02:01	RQAEQEATV	RQAEQEATV	0	0	0	0	0	RQAEQEATV	ENST00000412962	0.487702	0.408	0.487586	1.919	255.75	WB
TCGA-Skin	139	02:01	RLNQTTFTA	RLNQTTFTA	0	0	0	0	0	RLNQTTFTA	ENST00000427426	0.40356	0.528	0.548145	1.295	132.82	WB
TCGA-Skin	124	02:01	VTWDAALYL	VTWDAALYL	0	0	0	0	0	VTWDAALYL	ENST00000510506	0.385044	0.559	0.599793	0.887	75.96	WB
TCGA-Skin	90	02:01	YLIPIVVRY	YLIPIVVRY	0	0	0	0	0	YLIPIVVRY	ENST00000334187	0.281158	0.819	0.27039	6.765	2681.73	WB
TCGA-Skin	103	02:01	FAYDGKDYI	FAYDGKDYI	0	0	0	0	0	FAYDGKDYI	ENST00000420110	0.250175	0.924	0.520666	1.563	178.8	WB