Student: Sarah Chen
High school: Phillips Academy, Andover, MA
State: Massachusetts
Country: USA

Mentors: Tamara Ouspenskaia and Travis Law
Institute: Broad Institute of MIT and Harvard,
Cambridge, Massachusetts, USA

Title: Seeking Candidate Neoantigens from Retained Introns

Seeking Candidate Neoantigens from Retained Introns

Sarah Chen[1], Tamara Ouspenskaia[2], Travis Law[2], Karl Clauser[2]
[1] Phillips Academy, Andover, MA
[2] Broad Institute of MIT and Harvard, Cambridge, MA

Abstract

Cancer-specific peptides produced by somatic mutations in tumor cells can be presented by MHC-I molecules on the surfaces of cells. The identification of neoantigens enables neoantigen-based immunotherapies such as personalized cancer vaccines. While the current approach is to search for neoantigens derived from cancer-specific somatic variants, it often falls short for cancers with few somatic mutations. One potential source of neoantigens is intron retention in tumor cells as a result of splicing errors. Here we identify retained intron candidates from RNA-seq data, generate features from Ribo-seq support, and validate candidates by mass spectrometry as a step toward the identification of neoantigens from retained introns.

Keywords

Ribo-seq, RNA-seq, neoantigens, HLA, cancer, retained introns, mass spectrometry

Table of Contents

Background

The major histocompatibility complex class I (MHC I) enables the immune system to distinguish self and non-self molecules. The MHC I complex in humans is encoded by the human leukocyte antigen (HLA) genes. MHC I molecules present peptides from cytosolic proteins on the surface of cells. Cytotoxic T cells can recognize the presented antigens, and infected or cancerous cells that present non-self antigens can elicit an immune response (Swain, 1983). Neoantigens are tumor-specific antigens that result from somatic mutations in cancer cells. Neoantigens have been targeted in patient-specific immunotherapies, in order to treat patients with melanoma and glioblastoma (Keskin et al., 2019; Ott et al., 2017; Sahin et al., 2017). Currently, neoantigens are predicted from cancer-specific somatic mutations in protein-coding regions of the genome (Gubin et al., 2015). Yet, this approach falls short for patients with low somatic mutation burden (Rajasagi et al., 2014).

Retained introns derived from splicing errors in cancer cells are another potential source of neoantigens. Neoantigens predicted from the tumor-specific retention and translation of introns have been computationally identified using RNA-seq data (Smart et al., 2018). In order to determine if retained introns are bona fide sources of neoantigens in cancer cells, the MHC I complex can be biochemically isolated and MHC I-bound peptides subjected to analysis by mass spectrometry (Abelin et al., 2017; Hunt et al., 1992). Despite a large number of predicted retained introns, only a handful was confirmed by mass spectrometry to be presented by MHC I in cancer cell lines (Smart et al., 2018), suggesting that there is still much we do not understand about intron retention, translation, processing, and MHC I presentation.

Ribosome profiling (Ribo-seq) has emerged as a powerful approach to investigate the translated portion of the transcriptome in cells and tissues (Ingolia et al., 2009). Here, I propose to use a combination of RNA-seq, Ribo-seq and mass spectrometry to determine the extent of retained intron contribution to the MHC I immunopeptidome in healthy and cancer cells.

Methods and Results

Retained intron (RI) prediction and analysis were performed using RNA-seq, Ribo-seq, and mass spectrometry data from B721.221 cells engineered to express a single class I HLA-allele (HLA-A*01:01, HLA-A*33:03, HLA-B*15:01, HLA-B*44:02). These cells were used for the analysis because of the vast amount of the MHC I immunopeptidome MS data previously acquired from 95 HLA alleles individually expressed in these cells (Abelin et al., 2017).

Data Preprocessing

RNA-seq reads were trimmed of adapter sequences and aligned to the genome. Adapters were removed with Cutadapt 1.15 (Martin, 2011). Reads below the chosen length threshold of 80 nt or with too many unknown nucleotides were discarded, leaving 99.29% of the original 150 million read pairs. Reads were aligned to the genome with STAR 2.5.3a, using reference gene annotations (Dobin et al., 2013). The reference transcriptome consisted of GENCODE gene annotations as well as transcripts annotated in MiTranscriptome, which was generated by de novo transcriptome assembly of RNA-seq data from over four thousand cancer and healthy samples (Harrow et al., 2012; Iyer et al., 2015).

Ribo-seq reads were trimmed of primers, barcodes, and unique molecular identifiers (UMIs) with Cutadapt, stripped of contaminants such as rRNA with BowTie (Langmead et al., 2009), and aligned to the genome with STAR, using reference annotations (Figure 1). Read alignments were offset-corrected with RibORF (Ji, 2018). Offset-correction is performed in order to truncate each read to 1 nt and place it at the predicted position of the ribosomal A-site.

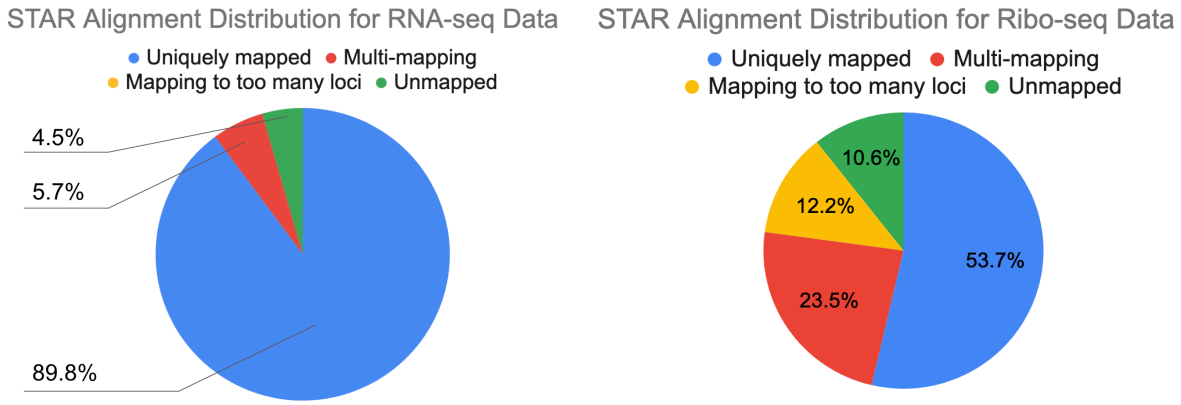


Figure 1: STAR alignment summary

RNA-seq data (paired, 150 nt long reads) had much higher rates of alignment and much higher rates of unique alignment to the genome compared to Ribo-seq data (single, 28 nt long reads).

Retained Intron Analysis

In order to identify RIs, *de novo* transcripts were assembled from aligned RNA-seq data using StringTie (Pertea et al., 2015). Transcripts containing RIs were identified by comparing my *de novo* transcripts to the reference transcriptome using GffCompare (v0.11.2, https://ccb.jhu.edu/software/stringtie/gffcompare.shtml). RI candidates that were contained within the coding sequence of any annotated transcript were discarded.

The *de novo* assembly and RI identification were performed on RNA-seq data from individual alleles and also on RNA-seq data combined across all alleles. A superset of RI candidates was constructed from the predictions from each allele and from the combined alleles for further processing (Figure 2). 1801 RI candidates were identified overall.

Analyzing alleles both individually and collectively preserves sample-specific differences but also captures overall trends to a greater extent. The alleles are technical replicates, and they are also biologically identical apart from their HLA alleles. Combining alignments across alleles amplifies the presence of lowly-expressed transcripts and enables their identification in *de novo* transcript assembly. These lowly-expressed transcripts are potential sources of RI candidates. 493 RI candidates were predicted only in the combined analysis. The transcripts containing those candidates trended toward lower expression levels compared to transcripts containing RI candidates predicted in the individual analysis (Figure 3). Here, transcript expression levels are quantified using TPM (transcripts per million), which essentially measures the number of reads aligning to each transcript normalized by transcript length and total sequencing depth. StringTie calculates the TPM for each transcript during transcript assembly (Pertea et al., 2015).

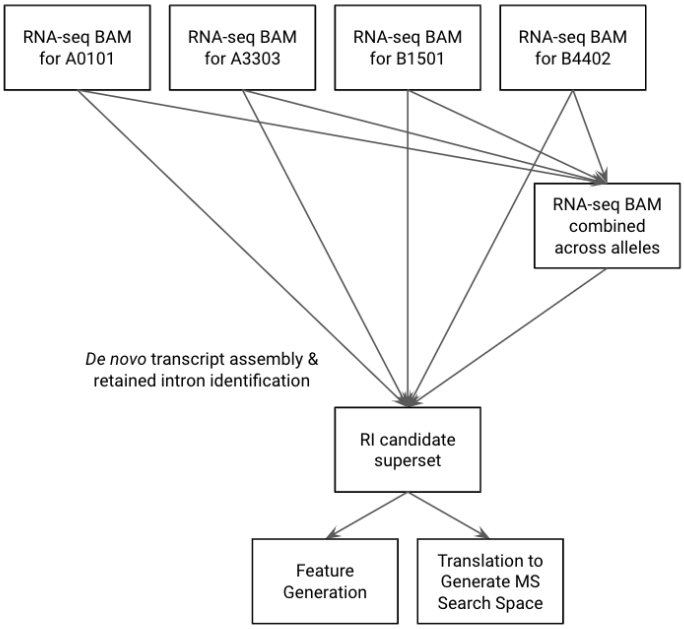


Figure 2: RI analysis schematic, after adapter trimming and genome alignment
BAM files are generated after RNA-seq reads are trimmed and aligned to the genome. Aligned reads are assembled into de novo *transcripts, and RIs are identified. Each allele is processed individually, and RI candidates are also predicted from RNA-seq alignments combined across all alleles. For the superset of candidates, features are generated from RNA-seq and Ribo-seq data. Candidates are also translated into proteins so that they can be searched in the MS data.*

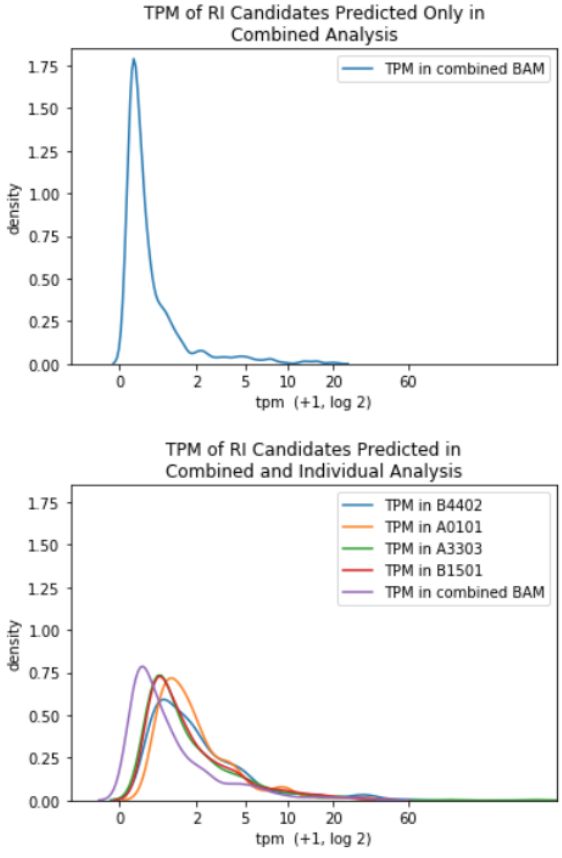


Figure 3: *RI candidates identified exclusively by the combined analysis have lower expression than those identified in individual analysis. The median TPM of the transcripts containing candidates unique to the combined analysis (top) was 0.41, whereas the median TPM of the remaining transcripts (bottom) was 0.77.*

Following MHC-I immunoprecipitation, peptides were sequenced with LC-MS/MS data for all alleles. In order to determine if the RIs generate antigens for MHC I presentation, I constructed a database of RI candidates amenable to searching the MHC I immunopeptidome mass spectrometry data. For the database, each RI and its flanking 45 nt in the neighboring exons was translated in 3 frames, and potential open reading frames (ORFs) that ended with a stop codon and were at least 8 AA (amino acids) long were added to the search space (Figure 4). Reads shorter than 8 AA were discarded because MHC I-presented antigens are typically 9-11 AA or, less frequently, 8 or 12 AA.

Entire RI candidates are considered rather than just their exon-adjacent regions due to the diverse variations of intron retention (Figure 5).

To determine the extent of the contribution of RIs to the overall MHC I immunopeptidome MS search space, I identified all possible 9 amino acid long peptides that could be generated from the RI candidates as well as from the GENCODE and MiTranscriptome references. Adding RI candidates to the search space yields a 0.94% increase in the number of unique 9-mers compared to the GENCODE and MiTranscriptome reference alone (Figure 6).

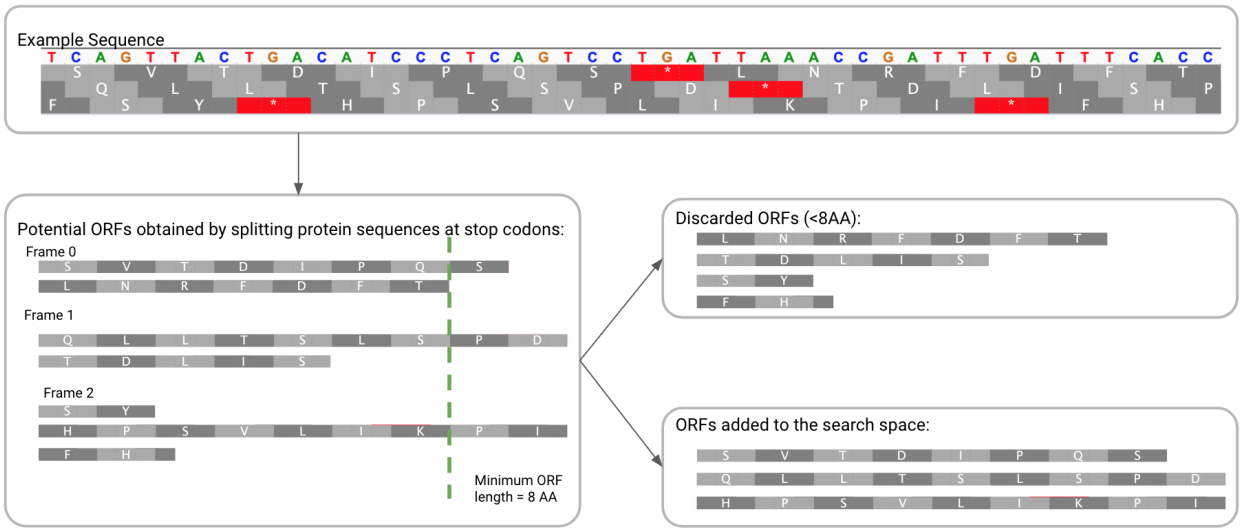


Figure 4: Generating ORFs from a RI Candidate
RI candidates and the 45 nt on the 5' and 3' flanks are translated in 3 frames. The sequence here is arbitrarily selected to illustrate the ORF identification process, and it is shortened for clarity.

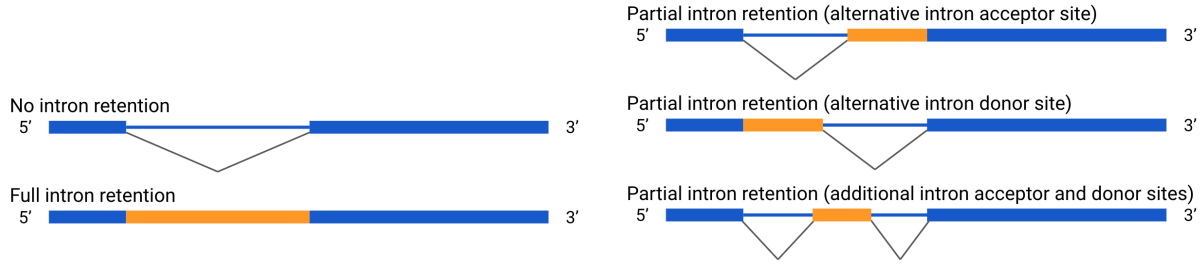


Figure 5: Intron Retention Variations
Exonic regions are signified by thick blue lines. Intronic regions are signified by thin blue lines, and intronic regions that are retained are signified by thick orange lines.

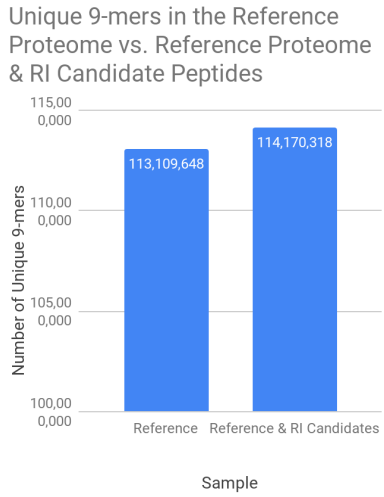


Figure 6: *The RI candidate peptides introduce 1,060,670 unique 9-mers as potential presented peptides.*

Figure 7A presents an example of an ORF within a RI candidate supported by RNA-seq, Ribo-seq, and MS data. The ORF is contained within a RI candidate on the forward strand in the DNAH17-AS1 gene. It has Ribo-seq support and RNA-seq support, and the peptide EHQKEGSRLLL (highlighted) has also been found in the mass spectra.

Given the number of identified RI candidates and the previously reported results, where few RIs were validated by the MHC I immunopeptidome MS analysis, I am planning to use additional RNA-seq and Ribo-seq features to narrow down the RI candidates (Smart et al., 2018) (Table 1). Many RI candidates with RNA-seq support lack Ribo-seq alignments supporting their translation. The latter half of gene CCNT2 is visualized in IGV in Figure 7B. The exons demonstrate RNA-seq and Ribo-seq support. The third intron also demonstrates RNA-seq support, to a smaller extent, but lacks of Ribo-seq support and therefore does not appear to be translated. The high rate of such false positives emphasizes the need to use Ribo-seq information to supplement RNA-seq data. Features generated from Ribo-seq information may be able to better distinguish true positives from false positives and increase the precision of predictions.

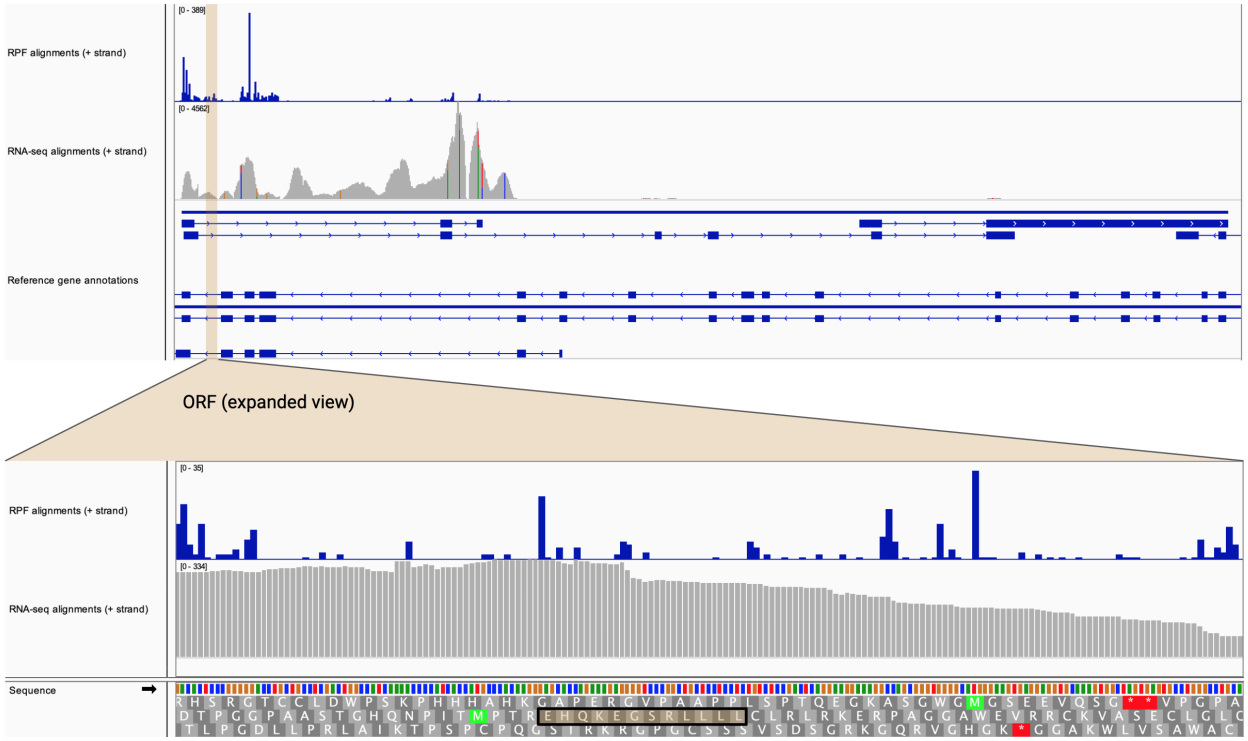
Table 1: RNA-seq and Ribo-seq features for RI candidates

Feature for each RI candidate	Definition/Calculation	Purpose
RNA-seq TPM of the candidate's transcript	StringTie reports the TPM of all assembled transcripts. RI candidates are mapped to StringTie transcripts, and the TPM is extracted.	Quantify candidate's transcript expression level
RNA-seq TPM	Standard TPM formula (Wagner et al., 2012)	Identify candidate expression level
In-frame Ribo-seq TPM	TPM calculated from Ribo-seq data, considering only reads in the translational frame for each candidate	Quantify amount of Ribo-seq support of the candidate's translation
Percentage of maximum entropy (PME) of aligned RPFs (ribosome protected fragments)	Entropy of RPF distribution out of the entropy of a uniform distribution (Ji, 2018)	Evaluate the distribution of RPF alignments across a RI candidate. For example, a concentration of reads in a single base (which would have low PME) does not provide very strong evidence of translation.
Mode RPF length	Most frequent length of RPFs aligned to the RI candidate	Distinguish Ribo-seq support stemming from true translation events from RPF alignments that are artifacts of the protocol. True ribosomal footprints should be ~28 nt.

In order to determine the extent of intron contribution to the MHC I immunopeptidome, I have taken advantage of the vast MHC I immunopeptidome MS data that has been previously generated in the lab. However, in order to find cancer-specific RIs that could be used for targeted immunotherapy, I have also applied my pipeline to RNA-seq data acquired from patient-derived melanoma cultures for which MHC I immunopeptidome MS data is also available. I have generated a patient-specific RI database that will be used to search MS spectra. Ultimately, I will compare the RI candidates as well as MS-identified RI antigens in tumor samples to their equivalents in healthy samples in order to find truly cancer-specific RIs.

A.

Gene DNAH17-AS1



B.

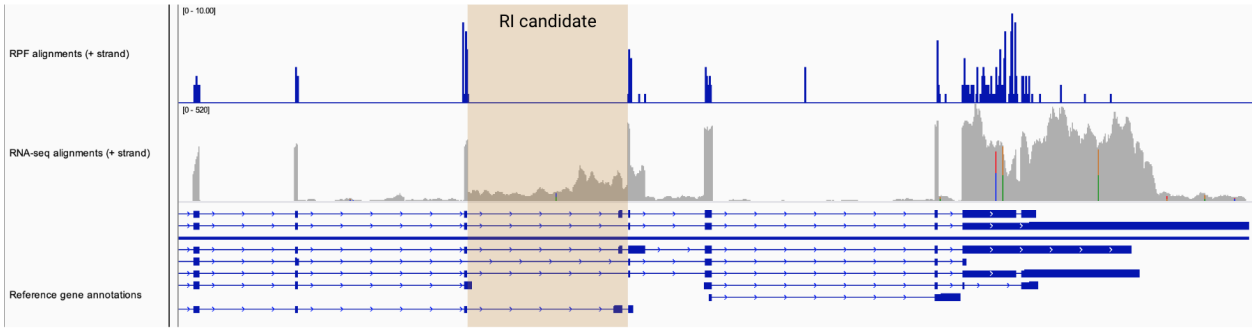


Figure 7: RI candidate ORF examples with RNA-seq and RPF data

A. Example of RI candidate supported by RNA-seq, Ribo-seq, and MS data

The ORF at chr17:76,481,371-76,481,557(+) is shown in IGV. The ORF is within a RI candidate in the DNAH17-AS1 gene. The peptide EHQKEGSRLLL (highlighted in the ORF's translation table) has been found in the mass spectra.

B. Example of RI candidate that is supported by RNA-seq but does not appear to be translated

The latter half of gene CCNT2 is visualized in IGV. The third intron (highlighted) appears to be a false positive RI candidate that is transcribed but not translated.

Discussion

More accurate prediction of intron retention is an important step toward improved identification of neoantigens derived from intron retention. RNA-seq data supports the prediction of almost 2000 RIs in this data, but few of them are likely to be true positives. Considering Ribo-seq data in addition to RNA-seq data when predicting RIs has the potential to lower the false positive rate by providing information about the translation of RI candidates and by distinguishing translated candidates from non-translated candidates.

References

Abelin, J.G., Keskin, D.B., Sarkizova, S., Hartigan, C.R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G.L., Eisenhaure, T.M., et al. (2017). Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. Immunity *46*, 315–326.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Gubin, M.M., Artyomov, M.N., Mardis, E.R., and Schreiber, R.D. (2015). Tumor neoantigens: building a framework for personalized cancer immunotherapy. J. Clin. Invest. *125*, 3413–3421.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

Hunt, D.F., Henderson, R.A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A.L., Appella, E., and Engelhard, V.H. (1992). Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. Science *255*, 1261–1263.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. *47*, 199–208.

Ji, Z. (2018). RibORF: Identifying Genome-Wide Translated Open Reading Frames Using Ribosome Profiling. Curr. Protoc. Mol. Biol. *124*, e67.

Keskin, D.B., Anandappa, A.J., Sun, J., Tirosh, I., Mathewson, N.D., Li, S., Oliveira, G., Giobbie-Hurder, A., Felt, K., Gjini, E., et al. (2019). Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. Nature *565*, 234–239.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.

Ott, P.A., Hu, Z., Keskin, D.B., Shukla, S.A., Sun, J., Bozym, D.J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., et al. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. Nature.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

Rajasagi, M., Shukla, S.A., Fritsch, E.F., Keskin, D.B., DeLuca, D., Carmona, E., Zhang, W., Sougnez, C., Cibulskis, K., Sidney, J., et al. (2014). Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. Blood *124*, 453–462.

Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.P., Simon, P., Lower, M., Bukur, V., Tadmor, A.D., Luxemburger, U., Schrors, B., et al. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature *547*, 222–226.

Smart, A.C., Margolis, C.A., Pimentel, H., He, M.X., Miao, D., Adeegbe, D., Fugmann, T., Wong, K.-K., and Van Allen, E.M. (2018). Intron retention is a source of neoepitopes in cancer. Nat. Biotechnol.

Swain, S.L. (1983). T cell subsets and the recognition of MHC class. Immunol. Rev. *74*, 129–142.

Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. *131*, 281–285.

Acknowledgments

Sarah Chen performed computational analysis on the data. Tamara Ouspenskaia and Travis Law, the mentors, suggested the research topic, provided computational resources, and provided general guidance. Karl Clauser performed the mass spectrometry analysis.

Declaration of Academic Integrity

本参赛团队声明所提交的论文是在指导老师指导下进行的研究工作和取得的研究成果。尽本团队所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果。若有不实之处，本人愿意承担一切相关责任。

参赛队员：Sarah Chen　　指导老师：Tamara Ouspenskaia, Travis Law

2019 年　9 月　7 日