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 Neoantigen Candidates within Retained Introns

Seeking Neoantigen Candidates within Retained Introns

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

Abstract

Major histocompatibility complex class I (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. The identification of cancer-specific peptides (neoantigens) produced by somatic mutations in tumor cells and presented by MHC I molecules enables immunotherapies such as personalized cancer vaccines and adoptive T cell transfer. The state of the art approach searches for neoantigens derived from cancer-specific somatic variants and often falls short for cancers with few somatic mutations. Retained introns (RIs) resulting from splicing errors in cancer are an additional source of neoantigens. In this study, we identify RIs which are transcribed, translated, and contribute peptides to MHC I presentation. Using *de novo* transcriptome assembly of RNA-seq data,we identified 1799 RIs in B721.221 cells. Additionally, we detected 87 peptides from 83 RIs by liquid chromatography-tandem mass spectrometry of the MHC I immunopeptidome (LC-MS/MS). Finally, we use ribosome profiling (Ribo-seq), which provides a readout of mRNA translation, to identify RIs that are translated, a prerequisite for MHC I presentation. Previous studies have predicted thousands of RIs but have been able to validate only a handful through mass spectrometry. By distinguishing transcribed but untranslated versus translated candidates, Ribo-seq has the potential to improve RI predictions. We propose the use of a combination of RNA-seq and Ribo-seq, paired with mass spectrometry validation, to more accurately predict the contribution of RIs to the MHC I immunopeptidome, enabling the use of RI derived neoantigens in future immunotherapies.

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

Table of Contents

Background

The major histocompatibility complex class I (MHC I) complex in humans is encoded by the human leukocyte antigen (HLA) genes. MHC I molecules present peptides from proteins within the cell on the surface of cells, and cytotoxic T cells can recognize presented antigens and distinguish between self and non-self molecules. Infected or cancerous cells that present non-self antigens can elicit an immune response (Swain, 1983). Neoantigens are the tumor-specific antigens which result from somatic mutations in cancer cells and enable their immune identification. Neoantigens have been targeted in patient-specific immunotherapies targeting 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). However, this approach falls short for patients with low somatic mutation burden (Rajasagi et al., 2014).

Retained introns (RIs) can result from splicing errors in cancer cells and are another source of potential neoantigens. In order to determine if RIs 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). Neoantigens predicted from tumor-specific RIs have been computationally identified using RNA-seq data and validated using LC-MS/MS. Despite the thousands of predicted RIs, only a handful was confirmed by mass spectrometry to be presented by MHC I in cancer cell lines (Smart et al., 2018), suggesting that, due to limitations in the number of neoantigens used in immunotherapy vaccines, there are still necessary improvements to RI prediction prior to therapeutic applications.

Ribosome profiling (Ribo-seq) has emerged as a powerful approach to investigate the translated transcriptome in cells and tissues (Ingolia et al., 2009). It is based on enriching ribosome-protected mRNA footprints (RPFs) and enables the identification of translated open reading frames (Ji et al., 2015). Here, I present a combination of RNA-seq, Ribo-seq and mass spectrometry analysis to validate the contribution of RIs to the MHC I immunopeptidome in healthy and cancer cells, and I show progress towards improving the accuracy of translated and presented RI predictions.

Methods and Results

RI prediction and analysis were performed using RNA-seq, Ribo-seq, and mass spectrometry of the MHC I immunopeptidome 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 due to the large amount of the MHC I immunopeptidome MS data previously acquired from 92 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 any 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 annotations and transcripts annotated in MiTranscriptome, which were generated by de novo transcriptome assembly of RNA-seq data from over 4,000 cancer and healthy samples (Harrow et al., 2012; Iyer et al., 2015).

Ribo-seq reads were trimmed of primers and barcodes with Cutadapt, stripped of contaminants such as ribosomal RNA with BowTie (Langmead et al., 2009), and aligned to the genome with STAR, using annotations generated through *de novo* transcript assembly described in the following section. In contrast to RNA-seq, where long fragments of mRNA were converted to cDNA and sequenced, generating paired-end reads ~150 nt long, a Ribo-seq sequencing library mainly consists of 28-29 nt single reads, due to the size of the RPFs. Compared to RNA-seq, a greater proportion of Ribo-seq reads map to multiple genomic loci or remained unmapped (Figure 1).

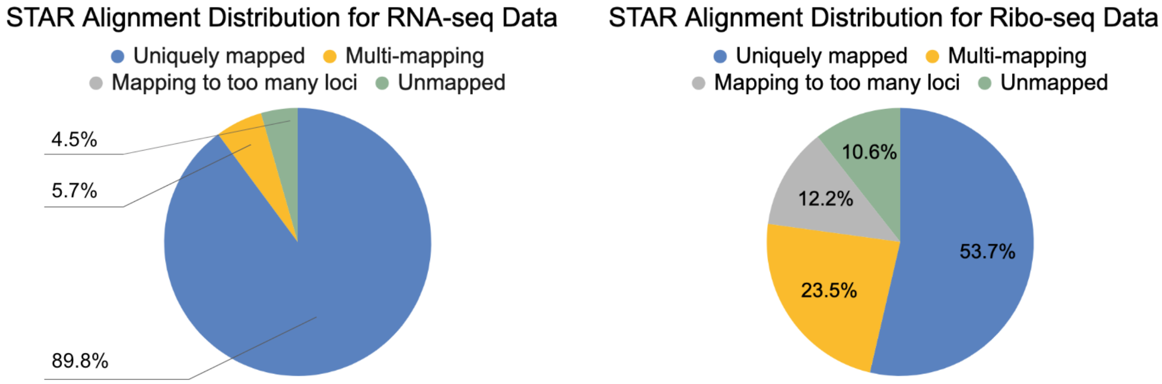

Figure 1: STAR alignment summary

Alignment metrics for RNA-seq (left) and Ribo-seq (right), showing reads aligned to a single locus (blue), multiple loci (red), too many loci (>20 loci for RNA-seq or >10 loci for Ribo-seq) (yellow), or that were unmapped (green).

Ribo-seq read alignments were then offset-corrected with RibORF (Ji et al., 2015). 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. Reads should exhibit trinucleotide periodicity supporting the translation of a given open reading frame (ORF) (Figure 2).

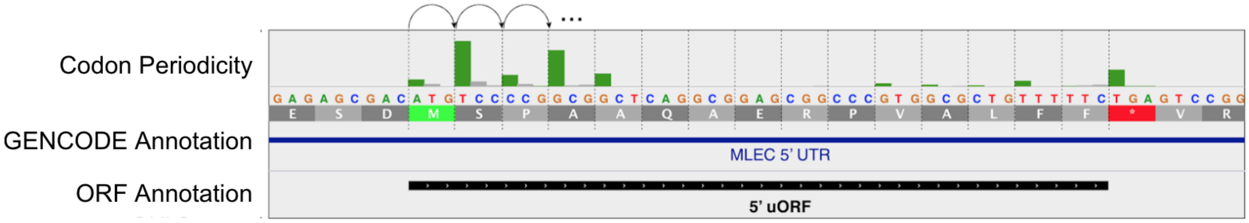

Figure 2: Offset Correction

An example of a translated ORF in the 5' UTR of MLEC supported by Ribo-seq. Offset-corrected reads shown in green are in-frame reads, supporting the translation of the ORF, while the reads shown in grey are out of frame. The start codon (M) is light green, the stop codon (*) is red.

Retained Intron Candidate Identification

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 *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 other annotated transcript were discarded.

The *de novo* assembly and RI identification were performed on RNA-seq data from 4 B721.221 cell lines individually and also on RNA-seq data combined across samples. A superset of 1799 RI candidates was constructed from all predictions and processed downstream (Figure 3).

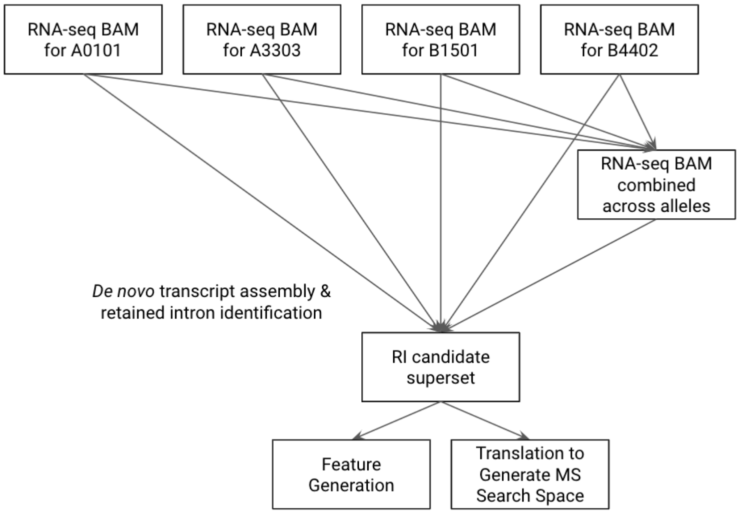
RNA-seq BAM for A0101
RNA-seq BAM for A3303
RNA-seq BAM for B1501
RNA-seq BAM for B4402
RNA-seq BAM combined across alleles
De novo transcript assembly & retained intron identification
RI candidate superset
Feature Generation
Translation to Generate MS Search Space

Figure 3: RI analysis schematic, after adapter trimming and genome alignment
BAM files are generated after RNA-seq reads are trimmed and aligned to the genome, and used for de novo *transcript and RI identification. RI candidates are predicted from each individual BAM file, as well as from a composite BAM file containing the reads from all 4 samples. For the superset of candidates, features are generated from RNA-seq and Ribo-seq data. Candidates are translated into proteins so that they can be searched in the MS data.*

Carrying out RI predictions at the sample level as well as across samples preserves sample-specific differences but also better captures overall trends. The 4 B721.221 cell lines are technical replicates, and they are biologically identical apart from their HLA alleles. Combining alignments across samples increases sensitivity to 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 (Figure 4.A). The transcripts containing those candidates trended toward lower expression levels compared to transcripts containing RI candidates predicted in the individual analysis.

The median TPM of transcripts containing RIs predicted only in the combined analysis was 0.41, whereas the median TPM of the remaining RI transcripts was 0.77 (Figure 4.B). Here, transcript expression levels are quantified using TPM (transcripts per million), which measures the number of reads aligning to each transcript normalized by transcript length and sequencing depth. StringTie calculates the TPM for each transcript during transcript assembly (Pertea et al., 2015).

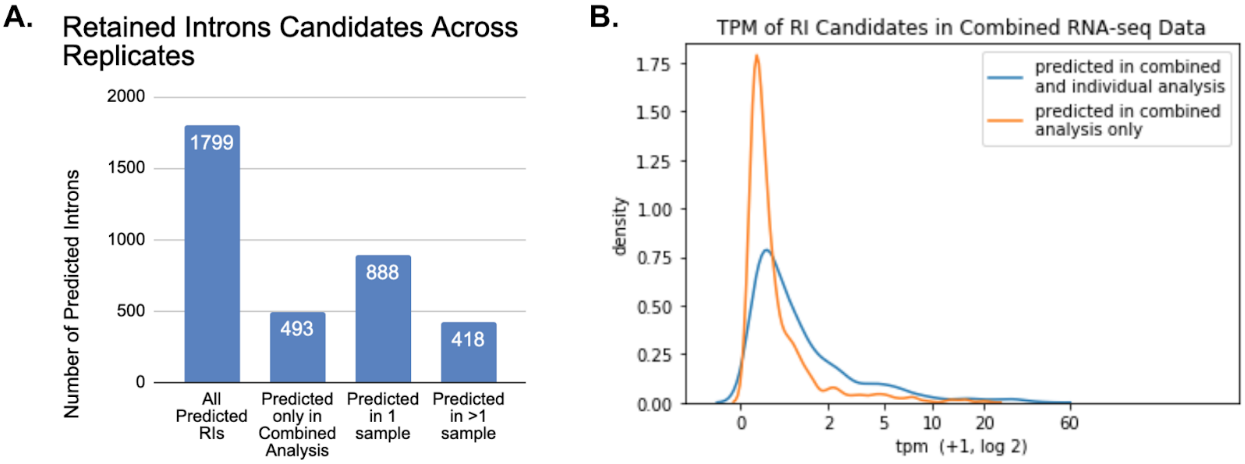

Figure 4:

A. The number of RI candidates predicted only in the combined analysis, in only 1 sample, and in >1 sample.

B. TPM of the transcripts containing RIs unique to the combined analysis (orange), or the transcripts identified in both the combined and individual searches (blue). The rank sum test p-value of the two distributions is 1.36e-33.

Mass Spectrometry Analysis

Following MHC-I immunoprecipitation, MHC I-bound peptides were analyzed by LC-MS/MS across 17 HLA alleles (Abelin et al., 2017), Sarkizova et al, 2019). In order to determine if the RIs generate antigens for MHC I presentation, I constructed a protein sequence database of RI candidates compatible with searching the MHC I immunopeptidome mass spectrometry data. For the database, each RI and its flanking 45 nt in neighboring exons were 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, such that each RI contributed multiple ORFs to the database (Figure 5.A). ORFs shorter than 8 AA were discarded because MHC I-presented antigens are typically 9-11 AA or, less frequently, 8 or 12 AA. Entire RIs were considered rather than just their exon-adjacent regions due to the diverse variations of intron retention (Figure 5.B).

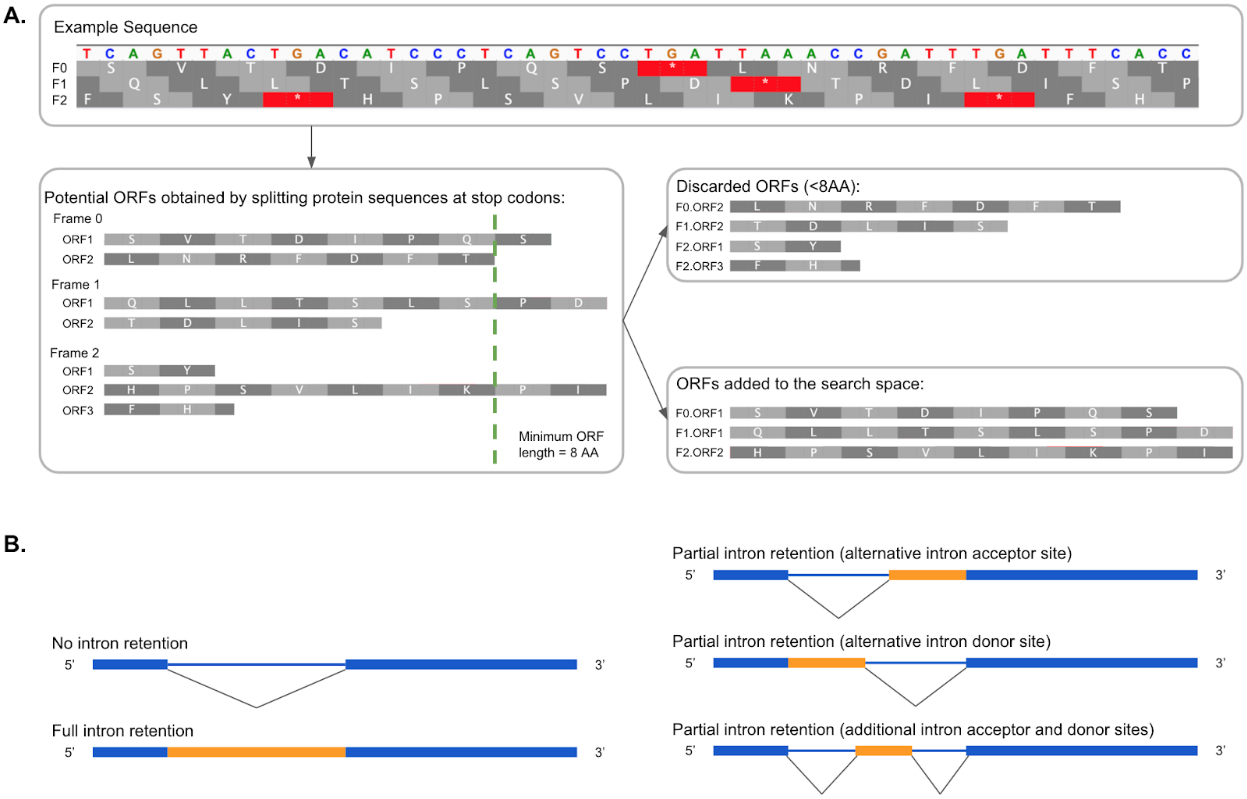
A.
Example Sequence
T C A G T T A C T G A C A T C C C T C A G T C C T G A T T A A A C C G A T T T G A T T T C A C C
F0 S V T D I P Q S * L N R F D F T
F1 Q L L T S L S P D * T D L I S P
F2 F S Y * H P S V L I K P I * F H
Potential ORFs obtained by splitting protein sequences at stop codons:
Frame 0
ORF1 S V T D I P Q S
ORF2 L N R F D F T
Frame 1
ORF1 Q L L T S L S P D
ORF2 T D L I S
Frame 2
ORF1 S Y
ORF2 H P S V L I K P I
ORF3 F H
Minimum ORF length = 8 AA
Discarded ORFs (<8AA):
F0.ORF2 L N R F D F T
F1.ORF2 T D L I S
F2.ORF1 S Y
F2.ORF3 F H
ORFs added to the search space:
F0.ORF1 S V T D I P Q S
F1.ORF1 Q L L T S L S P D
F2.ORF2 H P S V L I K P I
B.
No intron retention
Full intron retention
Partial intron retention (alternative intron acceptor site)
Partial intron retention (alternative intron donor site)
Partial intron retention (additional intron acceptor and donor sites)
5'
3'

Figure 5: Generating ORFs from a RI Candidate

A. A representative example of how an RI and its flanking regions contribute ORFs to the protein sequence database. Nucleotide sequence (top) is translated in 3 frames (F0, F1, F2). Stop codons (red), are marked with asterisks. Potential ORFs can be derived from the three frames (bottom). The green line marks the minimum required length of 8 AA for ORFs to be added to the database.

B. Schematic of types of intron retention. Exonic regions of a transcript 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.

MS spectra were mapped to a protein sequence database including translated RIs and proteins annotated in the UCSC Genome Browser. Mapped peptides were remapped to a protein sequence database also including GENCODE annotations to ensure that intron-mapping peptides did not map to any annotated exons. Of the 44,678 peptides found in the mass spectra, 87 were found exclusively in introns (Figure 6.A). 82 of 87 peptides mapped completely within RIs, while 5 peptides spanned an intron-exon boundary (Figure 6.B).

Out of the 493 RI candidates predicted only in the combined analysis, 23 were validated by MS, with the remaining 64 peptides mapped to RI candidates predicted in at least 1 individual sample. The distribution of RIs across the 3 prediction categories was proportional to the total number of candidates in each category (Figure 7). Thus, RI candidates with varying strengths of RNA-seq signal are translated and contribute antigens to MHC I presentation. Additionally, identifying RIs both at the sample level as well as across all samples enabled the identification MS-detected RIs that would otherwise not be found.

84 (96.6%) RI-assigned peptides mapped to just 1 RI, while the remaining 3 peptides mapped to 2 RIs (Figure 6.C). The 84 uniquely mapping peptides supported 77 unique RIs, while the 3 multi-mapping peptides supported 6 unique RIs. Of the 3 multi-mapping peptides, 1 mapped at the same genomic locus but was assigned to 2 overlapping but distinct introns from different isoforms on the gene AC093110.3. The other 2 peptides each mapped to 2 RIs at distinct genomic loci. 1 peptide mapped in loci that had identical flanking sequences, while the other peptide mapped in loci that had identical sequence only at the peptide region. The 3 out of 87 (3.4%) peptides that mapped to multiple loci reflected the percent of unique 9-mers in the RI-derived protein sequences that appear multiple times. Translated RI-derived ORFs introduced 1,069,750 9-mers to the MS-search space (Figure 9). 2.1% of those 9-mers appeared more than once in RI protein sequences, so such a proportion of multi-mapping peptides is expected.

In total, 87 RI-assigned peptides supported 85 unique ORFs, 5 of which were supported by 2 peptides when counting multi-mapping peptides multiple times (Figure 8.A). 4 of those 5 ORFs were supported by 2 peptides of identical sequence detected on different HLA alleles.
RI-assigned peptides supported 83 unique RIs out of the 1799 predicted, and 7 of those RIs were supported by more than 1 peptide, counting multi-mapping peptides multiple times (Figure 8.B). For 3 of those 7 RIs, mapping peptides had distinct sequences and supported more than 1 intronic ORF, and for the other 4 RIs, mapping peptides had identical sequence and were derived from distinct alleles.

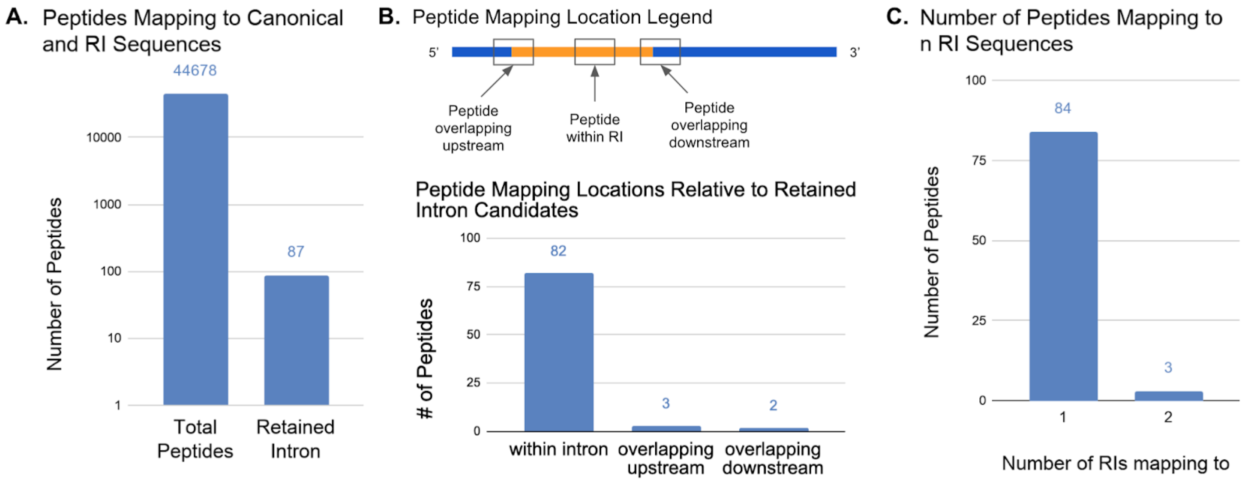

Figure 6: ***Peptide Mapping***

A. Number of peptides (y axis) found in total, mapped to canonical proteins, and to RIs (x axis).

B. Number of peptides (y axis), mapped entirely within a RI or overlapping a 3' or 5' exon-intron junction (x axis). Multimapping peptides were counted once.

C. Number of peptides (y axis) mapping to a unique or multiple RIs (x axis).

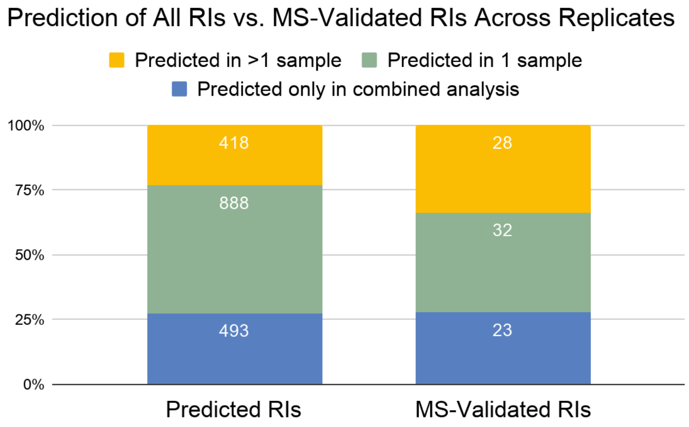

Figure 7: ***Prediction of RI Candidates Across Samples for All Predicted RIs vs. MS-Validated RIs***

The number of RI candidates predicted in >1 sample, 1 sample, and only in the combined analysis for all predicted RIs and for RIs to which a peptide has been assigned.

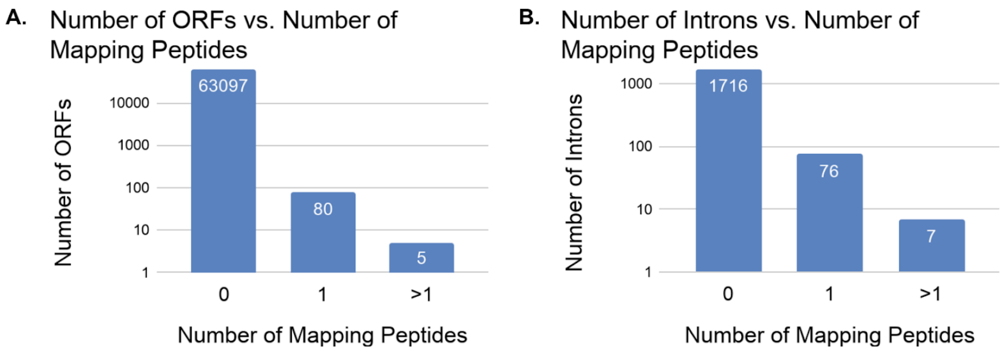

Figure 8: ***Peptide Support Across ORFs and Introns***

The number of ORFs or RIs (y-axis) with 0, 1, or >1 peptides mapping (x-axis). Peptides that map multiple times are multiply counted.

We set a global peptide-spectrum match FDR of 1%, which resulted in a detected peptide FDR of 1.33% globally. Applying the same aggregate false discovery rate (FDR) threshold to detected peptides mapped to canonical proteins and RIs resulted in a much higher FDR among RIs (29.89%) than among canonical proteins (1.27%) (Figure 9.A). The high FDR likely results from the size of the search space, and likely large number of spurious predictions. To quantify the relative contribution of RIs to the overall MHC I immunopeptidome MS search space, I identified all possible 9 AA long peptides that could be generated from the RI candidates as well as from the GENCODE references. Adding RI candidates to the search space yielded a 9.73% increase in the number of unique 9-mers compared to the UCSC reference alone (Figure 9.B). A larger search space increases the FDR as the same number of spectra are matched against a greater number of sequences, many of which are spurious. Reducing the number of RI candidates by filtering those that are less likely to be translated would decrease the size of the search space and improve the FDR.

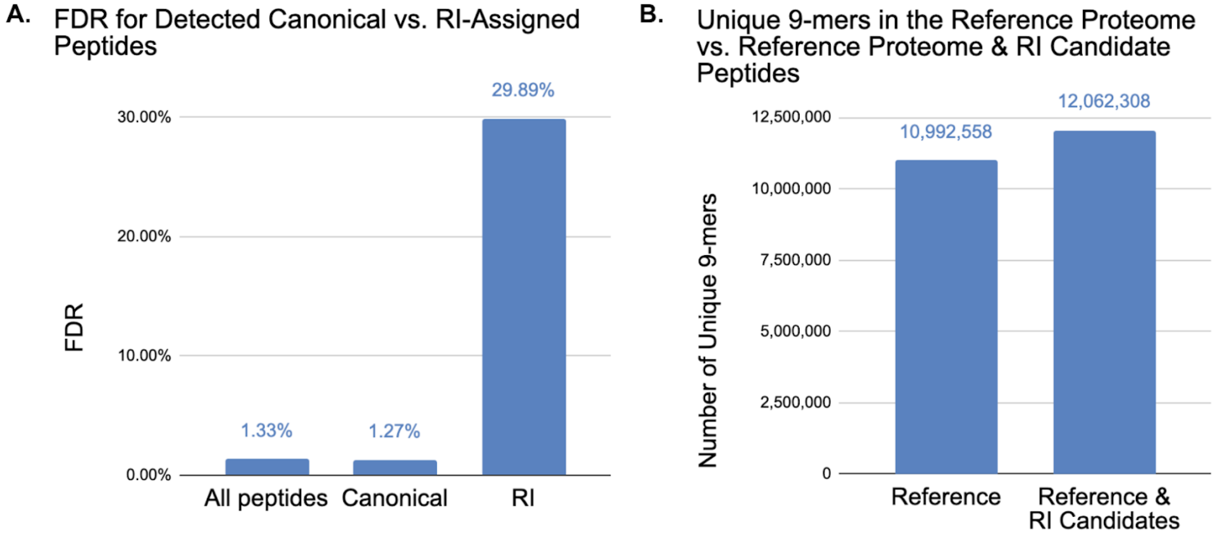

Figure 9: ***MS False Discovery Rate and Search Space***
A. The FDR (y axis) of peptides assigned to canonical proteins or RIs (x axis).
B. Number of unique 9 amino acid peptides (y axis) in the MS search database with and without RIs.

Ribo-seq Support and MS-Validation of Peptides, ORFs, and Introns

Given that peptides from just 83 out of 1799 RIs (4.6%) predicted by RNA-seq were detected in the MHC I immunopeptidome, in agreement with previous reports (Smart et al., 2018), I hypothesized that there are additional features that determine which RIs are translated and contribute antigens to the MHC I immunopeptidome. Because Ribo-seq provides a readout of the transcriptome that is actually translated by the ribosomes, I investigated whether Ribo-seq can improve RI prediction by comparing RPF support of MS-detected and undetected ORFs and introns.

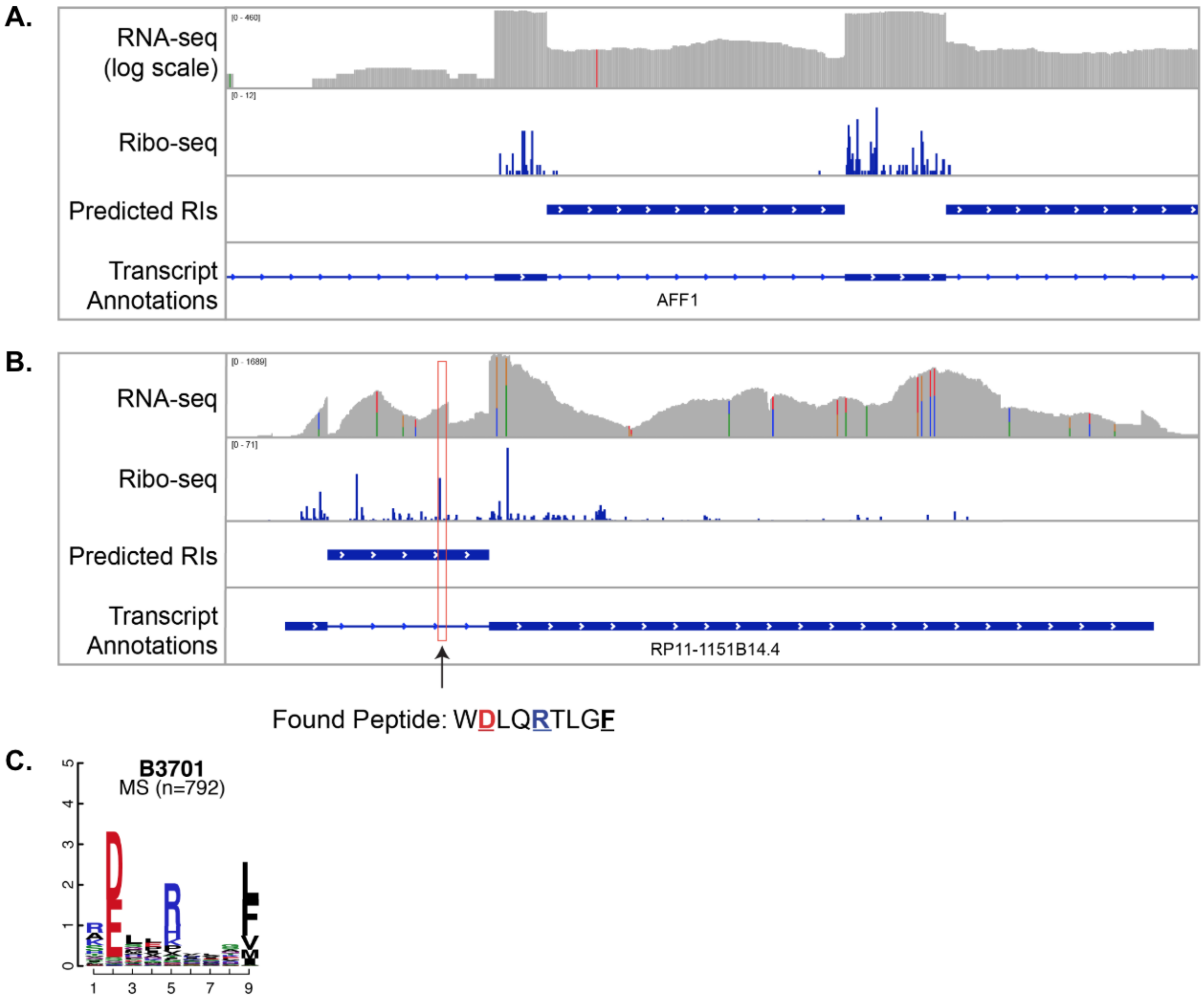

Figure 10: RI candidates with RNA-seq, Ribo-seq, and MS data

A. Example of RI candidate that is supported by RNA-seq but does not appear to be translated and presented based on a lack of RPF and MS support. The RI is from the AFF1 gene at chr4:88053016-88053422(+).

B. Example of RI candidate supported by RNA-seq, Ribo-seq, and MS data. The RI is from the RP11-1151B14.4 gene at chr4:88053016-88053422(+). The peptide highlighted has been found in the mass spectra for the allele B3701.

C. The peptide shown in B appears to match the expected binding motifs for its allele.

For example, the RI candidate on the gene AFF1 presents has RNA-seq support but lacks Ribo-seq and MS support (Figure 10.A), while the RI candidate on the gene RP11-1151B14.4 is supported by RNA-seq and Ribo-seq, and a peptide has been found in the mass spectra that maps to the RI and matches the expected binding motif (Figure 10.B). While the RI candidates are both supported by RNA-seq, the second RI candidate has much stronger evidence of translation. The first RI exemplifies the class of RI candidates I seek to filter out from RI predictions, while the second example embodies the class of RI candidates that I seek to enrich for.

Overall, MS-validated RIs and the ORFs derived from each RI (Figure 5) had higher rates of RPF support compared to non-MS-validated RIs and ORFs. While 14.93% of MS-validated ORFs have any in-frame RPFs, 6.38% of all predicted ORFs and 6.37% of non-MS-validated ORFs have any in-frame RPFs. The rate of RPF support in the MS-validated set was 2.3-fold greater than the rate in the non-MS-validated set (Figure 11.A). Mirroring the trend seen among ORFs, 80.72% of MS-validated introns had any RPFs, enriched in comparison to the 67.83% of all introns with any RPFs and 67.31% of non-MS-validated introns with RPF support (Figure 11.B). While the majority of predicted RIs had RPF support, a greater proportion of RIs validated by MS had RPF support than RIs not validated by MS. Thus, the number of both ORFs and introns with any RPFs was enriched in the MS-validated subsets.

In addition to examining how many MS-validated ORFs are RPF supported, I also looked at how many RPF-supported ORFs are MS-validated. ORFs and introns supported by Ribo-seq had higher levels of MS validation than ORFs and introns not supported by Ribo-seq. 0.21% of RPF-supported ORFs are detected by MS while 0.11% of non-RPF-supported ORFs are validated by MS, so MS-validation levels were nearly double for RPF-supported ORFs (Figure 12.A). Additionally, while 5.48% of RPF-supported introns were validated by MS, just 2.77% of non-RPF-supported introns were MS-validated (Figure 12.B). The 0.11% of ORFs and 2.77% of introns not supported by RPFs but validated by MS, present at low levels but still present, may be accounted for by a need for a greater sequencing depth.

Thus, RPF support is a factor positively related to MS-validation that provides further information about the translation of an RI candidate.

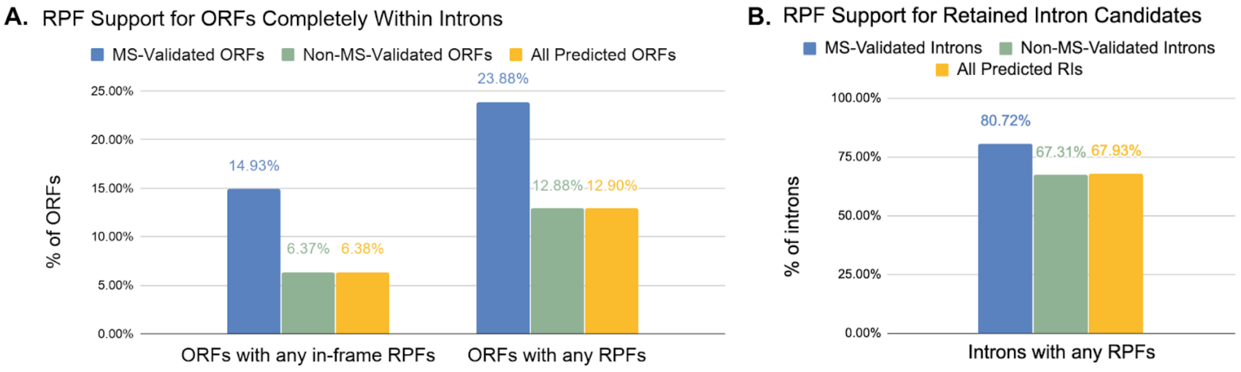

Figure 11: ***RPF Support for RI Candidates and RI Candidate ORFs***

A. The proportion of RI candidates that had any RPF reads for introns that were validated or not validated by MS. B. The proportion of RI ORFs with any supporting RPFs. Only ORFs fully contained within introns were considered in this figure so as not to artificially inflate RPF support levels by including ORFs in exonic flanking regions.

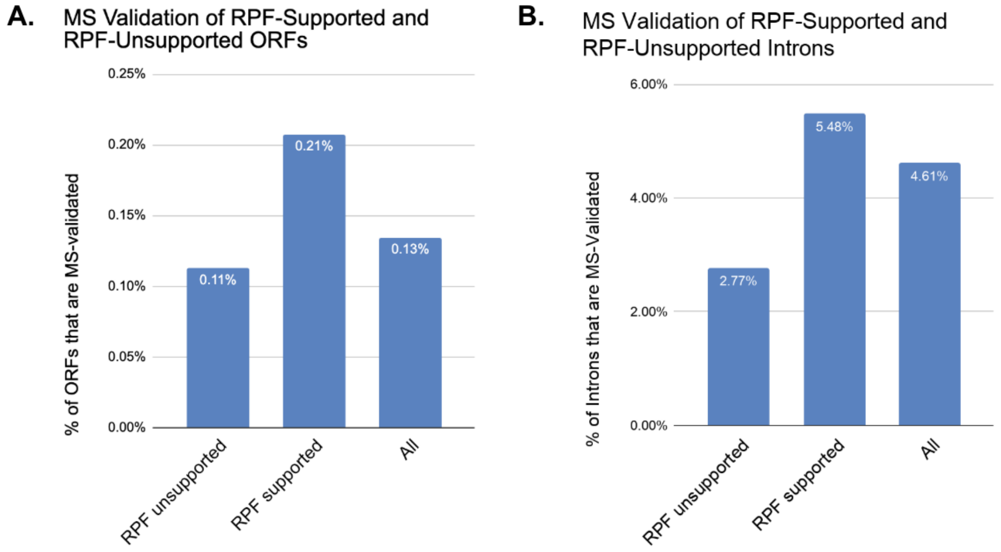

Figure 12: ***MS Validation of ORFs, partitioned by RPF support***

Percent of ORFs or introns that are supported by peptides for all ORFs or introns, ORFs or introns with at least 1 in-frame RPF, and ORFs or introns without any in-frame RPFs.

Ribo-seq Featurization

Additional features generated from Ribo-seq data may be able to better distinguish true positives from false positives and, through such filtering, increase the precision of RI predictions. I have generated additional Ribo-seq features for each RI-derived ORF, beyond just checking for the presence of any RPFs, and seek to use those features to filter RI candidates in this way (Table 1).

More highly translated ORFs should have higher numbers of aligned RPFs. In-frame Ribo-seq TPM provides a measure of the abundance of an ORF's translation normalized by length and sequencing depth. In addition, the majority of RPFs should be in-frame with a translated ORF, a feature that I quantify as purity, calculated as the ratio of reads in-frame vs. out-frame (Figure 13). Translated ORFs are also expected to have uniform coverage by RPFs, which I define as entropy (Figure 13). Furthermore, true ribosomal footprints should be approximately 28 nt long, so translated ORFs should have mapping RPFs with a mode length of approximately 28 nt. These additional features will help enable the differentiation of truly translated ORFs from those supported by spurious Ribo-seq reads resulting from RPF multimapping or from mRNA protection by non-ribosomal proteins (Ji et al., 2016).

Table 1: Ribo-seq features for RI candidates

Feature for each RI candidate	Definition/Calculation	Purpose
Presence of Any RPFs	Binary variable to indicate whether a peptide, ORF, or intron has 0 or >0 RPFs aligned	Indicate the lowest threshold of Ribo-seq support
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
Purity	Ratio of in-frame RPFs to total RPFs	Measure trinucleotide periodicity of aligned RPFs. Truly translated transcripts should demonstrate strong periodicity (Figure 13).
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 (Figure 13).
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.

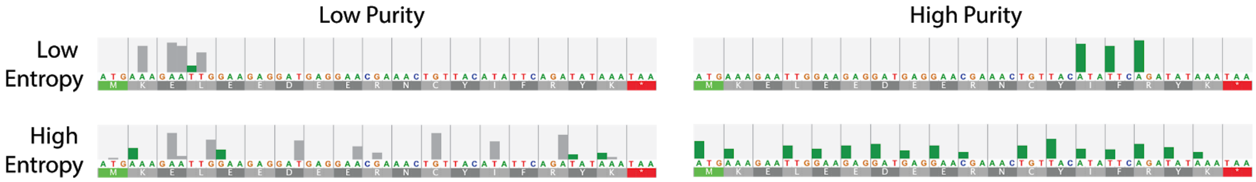
Low Purity
High Purity
Low Entropy
High Entropy
ATGAAAGAATTGGAAGAGGATGAGGAACGAAACTGTTACATATTCAGATATAAATAA
M K E L E E D E E R N C Y I F R Y K *

Figure 13: ***RPF Entropy and Purity***

Schematics of an ORF with varying purity and varying entropy. The start codon (M) is light green, the stop codon (*) is red. In-frame reads are shown in green, and out-of-frame reads are shown in gray. While in ORFs with high purity, nearly all reads are in-frame, in ORFs with low purity, few reads are in-frame. ORFs with high entropy have reads distributed throughout their lengths while ORFs with low entropy have RPFs aligning in only small subsets of their length.

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 1799 RIs in this data, but only 83 are validated by MS, reflecting previous findings (Smart et al., 2018). 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 to distinguish translated candidates from non-translated candidates.

6.38% of RI-derived ORFs were supported by any in-frame RPFs and were 2 times as likely to be MS validated compared to ORFs not supported by any in-frame RPFs (0.21% vs 0.11%). The rates of MS-validation of RPF-supported ORFs and introns were both enriched compared to rates of MS-validation of non-RPF-supported ORFs and introns (Figure 12). Accordingly, features derived from Ribo-seq data have the potential to improve the accuracy of RI prediction, in conjunction with RNA-seq data and features such as RNA-seq TPM of the transcripts containing RI candidates and the RNA-seq expression levels of the RI itself, which may also be useful in determining the likelihood of an RI candidate's translation and presentation.

Although the current FDR within intron-assigned peptides is 29.89%, filtering RI candidates using RPF features will decrease the contribution of RI-derived protein sequences to the search space, mitigate the number of spurious matches, and improve the FDR. Here, the small number of intron-assigned peptides and even smaller number of RPF-supported intron-assigned peptides in this data restricts meaningful quantitative analysis of the use of RPF features in improving FDR.

In addition, MHC I molecules derived from different HLA alleles bind a different repertoire of peptide ligands, presenting peptides with specific anchor motifs (Sidney et al., 2008). Therefore, not all RI-derived ORFs have peptides that are presentable by MHC I molecules. I hope to more closely examine the 6.37% of ORFs with in-frame RPF(s) supporting translation but not lacking MS support to determine whether they contain presentable peptides compatible with the MHC I binding motifs of the searched alleles or if they should not be expected to be validated by MS.

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.

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.

Ji, Z., Song, R., Regev, A., and Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife *4*.

Ji, Z., Song, R., Huang, H., Regev, A., and Struhl, K. (2016). Transcriptome-scale RNase-footprinting of RNA-protein complexes. Nat. Biotechnol. *34*, 410–413.

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.

S. Sarkizova, S. Klaeger, P. Le, L. Li, G. Oliveira, H. Keshishian, C. Hartigan, W. Zhang, D. Braun, P. Bachireddy, K. Ligon, I. Zervantonakis, J. Rosenbluth, T. Ouspenskaia, T. Law, S. Justesen, J. Stevens, W. Lane, T. Eisenhaure, G. L. Zhang, K. Clauser, N. Hacohen, S. Carr, D. Keskin, C. Wu. (2019). Improved prediction of endogenously presented HLA class I epitopes in human tumors based on 95 mono-allelic peptidomes. Nat. Biotechnol., In Press.

Sidney, J., Peters, B., Frahm, N., Brander, C., and Sette, A. (2008). HLA class I supertypes: a revised and updated classification. BMC Immunol. *9*, 1.

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.

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. The MIT PRIMES program and the Broad Institute of MIT and Harvard created this research opportunity.