In silico prediction of retained intron-derived neoantigens in leukemia

Sarah Chen[1]

Under the direction of Tamara Ouspenskaia[2], Nicoletta Cieri[2,3], Travis Law[2], Kari Stromhaug[2,3], Aviv Regev[2], Cathy Wu[2,3]

[1] Phillips Academy, Andover, MA
[2] Broad Institute of MIT and Harvard, Cambridge, MA
[3] Dana-Farber Cancer Institute, Cambridge, MA

MIT PRIMES 2020-2021 Paper Submission

Abstract

Alternative splicing is critical for the regulation and diversification of gene expression. Conversely, splicing dysregulation, caused by mutations in splicing machinery or splice junctions, is a hallmark of cancer. Tumor-specific isoforms are a potential source of neoantigens, cancer-specific peptides presented by human leukocyte antigen (HLA) class I molecules and potentially recognized by T cells. For cancers such as acute myeloid leukemia (AML) with a low mutation burden but widespread splicing aberrations, splice variants and retained introns (RIs) in particular, may broaden the number of suitable targets for immunotherapy. I developed a computational pipeline to predict AS-derived neoepitopes from tumor RNA-Seq. I first used the B721.221 B cell line as a model system, for which RNA-Seq, Ribo-Seq, and immunoproteome data from >90 HLA class I monoallelic lines were available. I performed *de novo* transcriptome assembly with StringTie, identifying on average 694±73 AS isoforms across 4 technical replicates. Using HLAthena, I identified 1,087 AS-derived neoepitopes predicted to bind across 4 frequent HLA alleles. Of them, 192 (18%) also displayed evidence of mRNA translation, measured as the alignment of ≥1 Ribo-Seq. To further increase prediction accuracy, I am currently analyzing the HLA I immunopeptidome to define the features of predicted AS isoforms more likely to be not only translated but also HLA presented. Finally, I applied my prediction pipeline to AML cell lines (n=8) and primary samples (n=7). I identified 682±113 AS isoforms in AML cell lines, similar to the 694 in B721, but the proportion of isoforms containing RIs (as opposed to alternative 5' and 3' splice sites or cassette exons) was 3.5x higher than in B721, in line with the biological relevance of RIs in particular in this disease setting. Primary AML samples yielded 1496±294 AS isoforms, more than twofold the number in B721 or AML cell lines, thus reinforcing the significant contribution of AS to the cancer immunopeptidome. Accurate prediction of AS-derived neoantigens through this pipeline will contribute to the design of novel cancer immunotherapies.

Introduction

Alternative splicing (AS) is a process essential for the regulation and diversification of gene expression. Eukaryotic genes are composed of a variable number of exons interrupted by intervening sequences known as introns, which are removed in a process termed RNA splicing to yield mature mRNA transcripts (**Figure 1A**). Alternative splicing, or the use of alternative combinations of exons, enables a single gene to increase its coding capacity, allowing the synthesis of structurally and functionally distinct protein isoforms (**Figure 1B**). The regulation of splicing can vary in cancer in particular due to mutations in splicing machinery or splice junctions, and TCGA analysis has found that cancer samples include up to 30% more alternative splicing events than normal samples (Kahles et al., 2018). Resultant aberrant AS diversifies the cancer transcriptome by introducing tumor-specific transcript isoforms. The protein-products of those isoforms can be translated and yield tumor-specific peptides, which in turn may be presented on HLA class I and elicit an anti-tumor immune response (**Figure 1C**).

Types of alternative splicing include exon skipping, mutually exclusive exons, cassette exons, alternative 3′ splice sites, alternative 5′ splice site, and intron retention (**Figure 1B**). Exon skipping and mutually exclusive exons involve alternative combinations of canonical exons while cassette exons, alternative 3′ splice sites, alternative 5′ splice sites, and intron retention involve the inclusion of intronic regions in spliced mRNA transcripts. I considered the latter 4 types in this analysis and referred to them using ‘AS’ in this paper.

Among these types, intron retention is particularly relevant in acute myeloid leukemia (AML). 20% of AML cases have a recurrent somatic mutation of splicing factor 3b subunit 1 (SF3B1), and other common mutated splicing modulators in AML include SRSF2, SF3B1, U2AF1, and ZRSR2 (Zhou and Chng, 2017). In a TCGA analysis evaluating the role of intron retention across different solid and hematological tumors, almost all cancer types demonstrated elevated levels of intron retention relative to normal tissues, but AML demonstrated the highest level of differential upregulation of retained introns (RIs) (**Figure 2**) (Dvinge and Bradley, 2015).

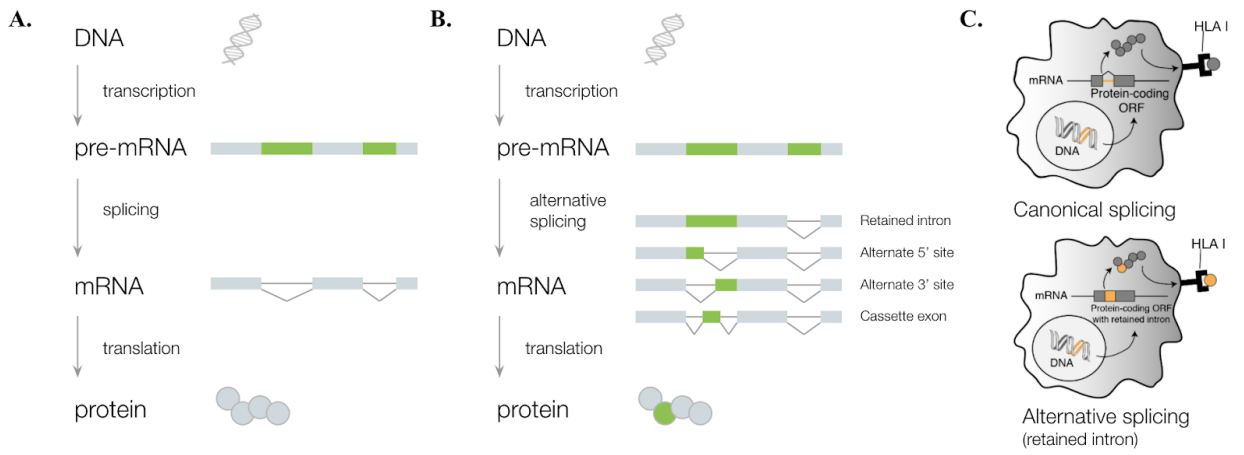
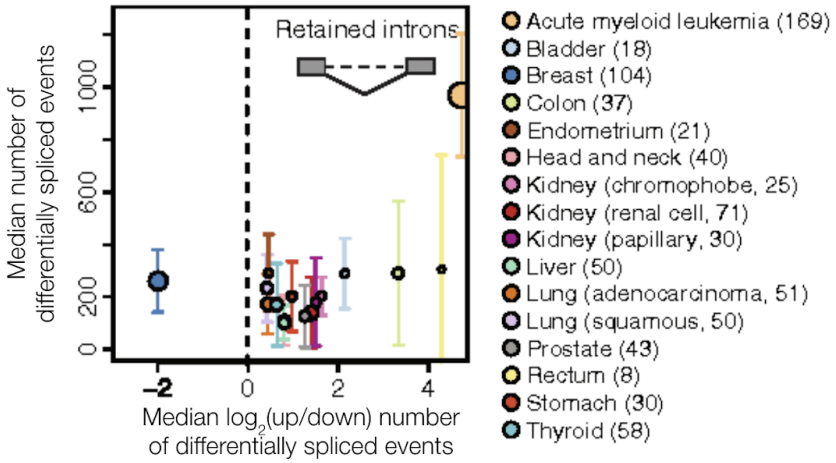


Figure 1

A. *DNA is transcribed into pre-mRNA which is spliced, so introns are removed and exons retained, then translated into protein.* ***B***. *In the case of alternative splicing, a single pre-mRNA transcript can be spliced in different patterns to yield multiple mRNA isoforms and downstream proteins. Types of AS events include retained introns, alternative 3′ splice sites, alternative 5′ splice sites, and cassette exons (top to bottom). Each of these types of AS preserves intronic regions that would otherwise have been spliced out, and those regions (green) are reflected in the protein product. Note: other types of AS such as exon skipping and mutually exclusive exons exist as well but are not reported in this figure because they are not considered in this analysis.* ***C***. *Peptides derived from aberrant alternative splicing may be presented on HLA class I.*



(Dvinge and Bradley. *Genome Medicine* 2015)

Figure 2

A plot (adapted from Dvinge and Bradley, Genome Medicine 2015) depicting intron retention in 16 cancer types. The X-axis indicates the direction of change in AS (upregulation vs. downregulation when comparing tumor and normal samples), and the Y-axis indicates the magnitude of change (number of differential splicing events). Intron retention is upregulated across a variety of cancer types, AML (top left) is an extreme outlier with the highest degree of upregulation by far.

As cancer-specific AS generates novel transcripts not present in normal tissues, it has recently been explored as a source of neoantigens, which are peptides arising from tumor-specific protein sequences presented on HLA class I that can be selectively recognized by T cells and hence represent promising immunotherapeutic targets. Indeed, cancer vaccines as well as T cell-based therapies can target neoantigens to foster the immune system to recognize the malignant cells and generate an anti-tumor response. Neoantigens have been targeted in personalized cancer vaccines across different disease settings such as melanoma and glioblastoma (Keskin et al., 2019; Ott et al., 2017; Sahin et al., 2017). However, in current neoantigen-based immunotherapies, neoantigens are predicted only from cancer-specific somatic mutations in protein-coding regions of the genome (Gubin et al., 2015). As a result, this approach falls short for tumors with low somatic mutation burden, such as AML, where AS events become an important additional source of neoantigens (Rajasagi et al., 2014).

RIs in mammalian cells were originally overlooked due to the difficulty of detecting them but have since been recognized as an important means of diversifying as well as regulating gene expression (Vanichkina et al., 2018). Advances in next-generation sequencing and the introduction of computational tools including SplAdder (Kahles et al., 2016) and rMats (Shen et al., 2014) have enabled the identification of AS events from RNA-seq data in cancer and normal samples, and translation of the resultant protein products has been validated via mass spectrometry (MS). Application of these tools to large cancer datasets has suggested that tumor-specific AS-derived peptides might significantly expand the pool of potential neoantigens. In a seminal analysis of TCGA breast cancer and ovarian serous cystadenocarcinoma patients, considering peptides derived from alternatively spliced noncanonical junctions in addition to SNV-derived peptides increased the percentage of samples with at least one putative neoantigen (validated via MS and NetMHC) from 30% to 75% (Kahles et al., 2018). However, it should be noted that translation at novel junctions is only a small subset of all translation introduced by tumor-specific AS. Peptides generated from the sequence of retained introns (RIs) and regions downstream of an AS event causing a frameshift were not included in this analysis but have the potential to broaden the number of potential neoantigens even further (**Figure 3**).

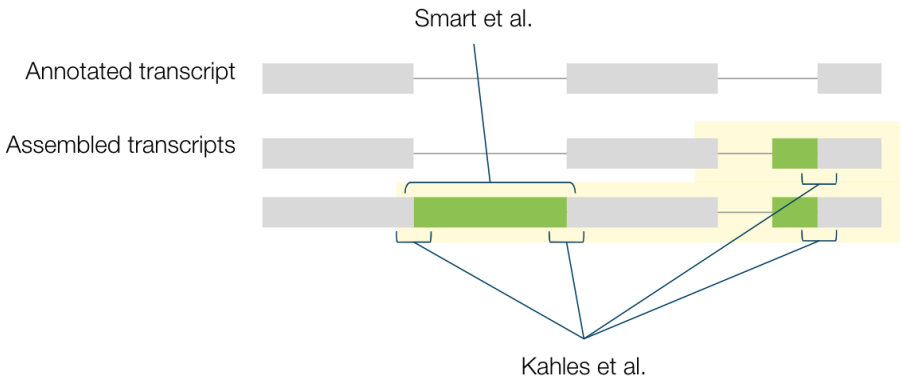


Figure 3
Smart et al. and Kahles et al. considered only a subset of potential AS-derived peptides. Smart considered peptides from retained introns, ignoring, for example, the intronic region introduced by the alternative 3' splice site in this diagram. Kahles considered novel splice junctions, fully overlooking intronic regions introduced by RIs or alternate splice sites. Neither study considered novel downstream peptides translated due to a frameshift introduced by an upstream AS event. My pipeline considers the full scope of potential AS-derived peptides (highlighted).

In another study from Smart and colleagues (Smart et al., 2018), potential neoantigens derived from tumor-specific RIs have been computationally identified using RNA-seq data and validated using HLA class I immunopeptidome MS data from tumor cell lines (including one AML cell line). However, this work has only considered the contributions of fully retained introns to the proteomic diversity and overlooked intronic regions retained in the transcriptome due to alternative 5' or 3' splice sites as well as downstream regions that may be translated in a different frame due to upstream RIs (**Figure 3**).

Accurately detecting the full spectrum of RIs poses unique challenges due to sources of spurious transcriptional signal. Noise from DNA contaminants or unprocessed mRNA transcripts generates spurious intronic RNA-seq reads, and since intronic regions are also rich in repetitive sequences, abundant multi-mapping reads exacerbate the problem (Vanichkina et al., 2018). RI prediction tools addressing those difficulties have recently been published, including IRFinder, KMA, and IntEREst (Middleton et al., 2017; Oghabian et al., 2018; Pimentel et al., 2015).

Motivated by the possibility of expanding this novel, promising class of neoantigens, here I present a pipeline that considers the full scope of potential neoantigens introduced by AS events. By working on the AS isoform rather than event level, I consider a superset of peptides with respect to previous pipelines for prediction of neoantigens from aberrant splicing: peptides from

RIs, peptides from intronic regions retained due to alternative 5' and 3' splice sites, peptides spanning novel junctions, and peptides resulting from frame-shift inducing upstream AS events (**Figure 3**). Finally, I apply that pipeline to the clinically relevant setting of AML, to identify novel AS-derived peptides in both AML cell lines and primary samples.

Results

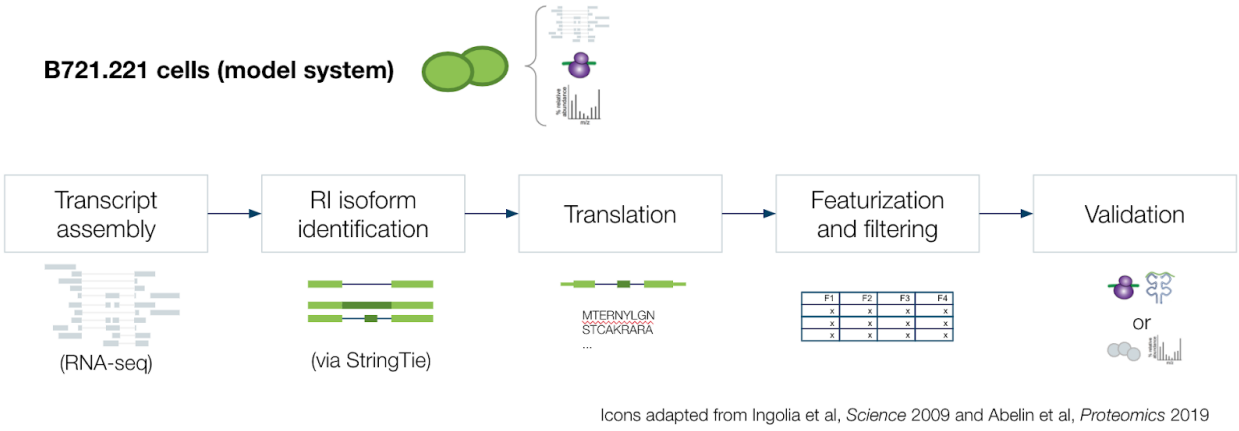


Figure 4
Pipeline overview. Schematic outlining the steps of transcript assembly from RNA seq data using StringTie, AS isoform identification, coding sequence identification and translation, featurization and potential filtering, and AS-derived peptide validation via Ribo-seq and HLAthena. See Methods for more detailed discussions of each step.

I present a pipeline to predict potential AS-derived neoantigens by characterizing AS events at the isoform level as summarized in **Figure 4**. I constructed this pipeline using the B721.221 B cell line as a model system, leveraging the available RNA-seq, Ribo-seq (which provides a readout of mRNA translation), and monoallelic immunopeptidome mass spectrometry data (Abelin et al., 2017; Ouspenskaia et al., 2020; Sarkizova et al., 2020). The pipeline leverages StringTie to assemble transcripts from RNA-seq data and predict AS isoforms and then featurizes, filters, and in silico translates predicted isoforms to identify novel AS-derived peptides (Kovaka et al., 2019). In the B721.221 model system, to validate predicted AS-derived peptides, I determined the presence of Ribo-seq support, which acted as evidence of translation; additionally, I required peptides to be predicted as HLA binders by HLAthena, a recently published algorithm for HLA binding prediction which outperforms the current gold standard netMHCpan (Abelin et al., 2017; Sarkizova et al., 2020). To further increase the prediction accuracy, I am in the process of validating the resulting AS-derived peptides using HLA class I immunopeptidome LC-MS/MS data available for >90 HLA class I monoallelic B721 cell lines.

In the B721.221 cell line, I predicted a mean of 694±73 AS isoforms in each of 4 monoallelic B721 cell lines (A*01:01, A*33:03, B*15:01, B*44:02), which for the purpose of this analysis

can be considered as technical replicates (**Figure 5A**). These isoforms yielded 164,742 potential novel peptides of length 8-11 (which represent the standard length of peptides presented on HLA class I) that were not found in the canonical proteome. 70,000 of those peptides had evidence of translation (because a Ribo-seq read mapped to the peptide sequence) and approximately 1,000 peptides were predicted to be HLA binders by HLAthena, using a threshold of 0.5% rank. 192 peptides fulfilled both criteria, and I considered those as validated AS-derived peptides (**Figure 5B**).

I also quantified the RNA-seq read support of predicted AS-isoforms and AS-derived peptides to calculate RNA-seq features possibly associated with Ribo-Seq and HLAthena validated AS-isoforms. I thoroughly characterized the profile of these peptides, providing downstream users of the pipeline with a rich set of 10 primary features (**Table 1**). I explored the relationship of those RNA-seq features with the likelihood of a predicted peptide's validation and found that the likelihood of validation was strongly linked to an AS isoform's transcript expression, with more highly expressed transcripts exhibiting increased rates of validation (**Figure 6**). Observations such as this can be exploited to define thresholds to generate a higher confidence prediction set, which would be essential when applying the pipeline in the absence of Ribo-Seq or immunoproteomic support. By complementing my current validation scheme with the analysis of mass spectrometry data to obtain a more complete portrait of AS-derived peptides presented in the immunopeptidome, I expect to have gathered all the necessary data in B721 to devise a formal model to distinguish AS-derived peptides with the highest probability of existence and presentation on HLA class I. Importantly, this feature-based filtering of the prediction set would be applied to enrich for validated peptides in settings where mass spectrometry and Ribo-seq (less common and/or more expensive protocols than RNA-seq) are not available, enabling the application of this pipeline to a broader array of samples.

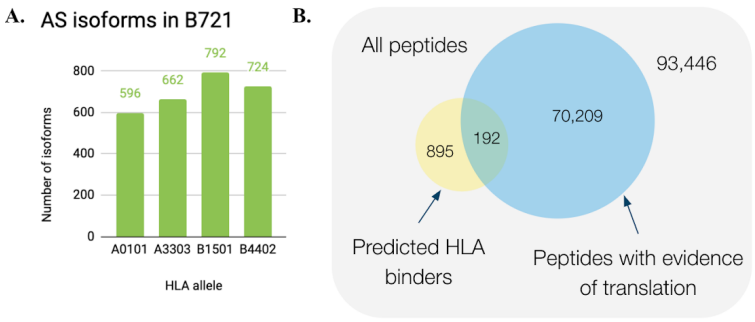


Figure 5

A. *Number of alternatively spliced isoforms predicted in each B721.221 allele.* ***B***. *AS-derived novel peptides validated via Ribo-seq and HLAthena analysis. 192 peptides have evidence of translation (supported by ≥1 Ribo-seq read) and were predicted HLA binders (using 0.5% rank threshold).*

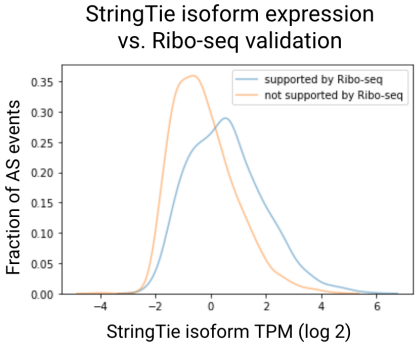


Figure 6

Density plot of transcript TPM (log 2 normalized) for AS events in B721 supported (blue) and not supported (orange) by Ribo-seq reads in the intronic segment introduced. StringTie isoform TPM varies significantly between AS events supported by Ribo-seq and not supported by Ribo-seq (p=4.3e-29 by independent t-test with unequal variable).

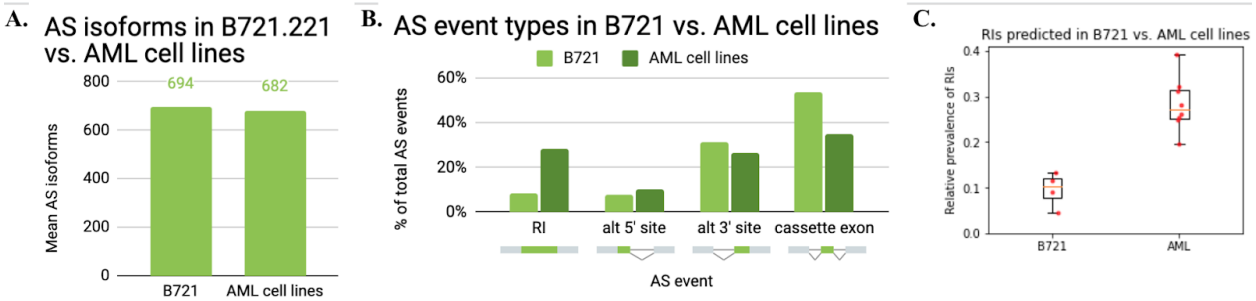


Figure 7

A. *Mean number of alternatively spliced isoforms in B721.221 alleles and AML cell lines.* ***B***. *Distribution of mean AS event counts across B721 alleles and AML cell lines.* ***C***. *Relative proportion of AS events that were RIs in B721 vs. AML cell lines, for each B721 allele and AML cell lines. I consider RIs, alternate 5' and 3' splice sites, and novel cassette exons in this analysis.*

Feature
AS isoform transcripts per million (TPM)
Canonical isoform TPM and ratio of canonical isoform TPM relative to AS isoform TPM
AS isoform base coverage
RNA-seq reads mapped to the intronic segment introduced by AS
RNA-seq reads mapped to adjacent canonical exons and ratio of those reads to reads mapped to the intronic segment
RNA-seq reads supporting the 5' and 3' boundaries of an intronic segment in an AS isoform (e.g. the number of reads spanning an intron-exon boundary for a retained intron event)
Fraction of multi-mapping reads and or reads with indels, clipping, or mismatches in intronic segments and adjacent canonical exons
Intronic segment GC content
Presence of a premature termination codon introduced by an upstream AS event
Splice motifs of the 5' and 3' splice junctions adjacent to an intronic segment, if applicable

Table 1

Features calculated by the AS prediction pipeline, involving RNA-seq expression or isoform structure. The AS events considered in this analysis (RIs, alternative 5' and 3' splice sites, and cassette exons) introduce a canonically intronic sequence to AS isoforms, here referred to as an 'intronic segment.' See methods for more detailed descriptions of each feature.

I then applied the pipeline to 8 AML cell lines (CMK, KASUMI-1, MUTZ-3, OCI-AML-3, OCI-M2, SET-2, TF-1, and THP-1) (Quentmeier et al., 2019). The analysis yielded a mean of 70,672 non-canonical AS-derived peptides of length 8-11, derived from a mean of 682 AS isoforms, comparable to the AS isoforms detected in B721 (**Figure 7A**). Of interest, AML cell lines greatly differed from B721 in the distribution of AS types, with a much higher proportion of RIs, compared to other types of AS events, roughly 30% vs. 10% (**Figure 7B and C**). While filtering peptide predictions and enriching for true positives using RNA-seq feature thresholds learned in B721 will enable the identification of a smaller, higher confidence prediction set, this current analysis already serves to characterize the landscape of potential AS-derived neoantigens in AML cell lines.

Finally, I applied my pipeline to 7 AML primary samples to yield a mean of 128,267 AS-derived peptides of length 8-11. The distribution of AS types was slightly different from the distribution in the AML cell lines and was dominated by novel cassette exons, but primary samples still maintained a higher proportion of predicted RIs than B721 (**Figure 8A and B**). These peptides

derived from a mean of 1,496 AS isoforms, more than double the number of isoforms assembled in B721 or AML cell lines. Furthermore, I examined the distribution of predicted AS events across the primary samples (**Figure 8C**). While most events were patient-specific (an expected result given the high heterogeneity of the disease), 1147 out of 7876 total events (14.56%) were predicted in at least 2 samples. Notably, one advantage of AS-derived neoantigens over SNV-derived neoantigens is that they may be shared across different patients with any given cancer type, a promising characteristic for the development of universal immunotherapeutic products.

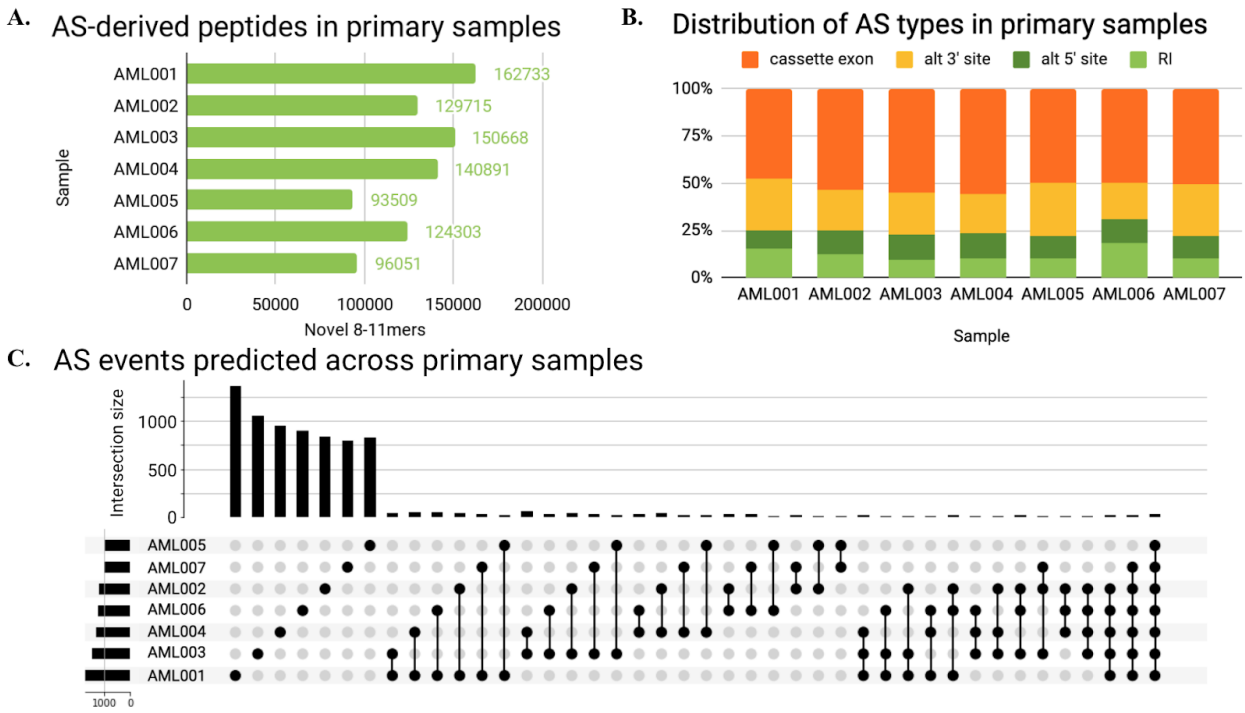


Figure 8:
***A**. Analysis of AML primary samples yielded a mean of 128,267 AS-derived peptides of length 8-11 amino acids. **B**. Distribution of AS event types within primary AML samples. **C**. Upset plot of AS events shared by primary samples. While the majority of AS events were patient-specific, 1147 out of 7876 total events (14.56%) were common to ≥2 samples. Sets with less than 9 AS events were omitted for brevity.*

Discussion

More accurate prediction of AS is an important step toward improved identification of AS-derived neoantigens, a still poorly investigated and yet highly promising category of tumor neoantigens, especially in cancer types characterized by a low mutation burden such as AML. Here, I present a novel pipeline I developed to predict those potential neoantigens that has been optimized in the B721 model system and then applied to 8 AML cell lines and 7 primary AML samples, thus generating a patient-specific AS database of potential AML-specific neoantigens. Through the integrated analysis of RNA-Seq and Ribo-Seq in the B721 model system, I also generated a thorough set of features characterizing the RNA-seq support of predicted peptides and isoforms and begin exploring which features have an impact on an isoform being translated and presented by HLA.

While past studies have also presented pipelines to predict potential AS-derived neoantigens, the pipeline I designed is the first to consider the full scope and potential of AS events (i.e. intron retention, alternative 5' and 3' splice sites, and novel cassette exons). Additionally, I worked on the isoform level to capture novel peptides introduced by combinations of AS events with other AS events or downstream canonical exons. I have predicted and validated novel AS-derived peptides in B721.221 cells and successfully characterized AS and its potential as a source of neoantigens in AML cell lines and primary samples, uncovering tens and hundreds of thousands of novel peptides derived from canonically intronic sequences. I have also identified 15% of predicted AS events in the cohort of primary AML samples I analyzed as shared between $\geq$2 patients, demonstrating the potential to identify neoantigens with therapeutic potential across patients.

Continuation of this work will complete mass spectrometry validation efforts in B721 and subsequently develop an RNA-seq-feature-based filtering and validation methodology for predictions in AML settings where Ribo-seq and MS data are not available. The pipeline could then be applied to further cell lines and tumor samples to characterize the potential AS-derived neoantigens in AML or other cancer types using only RNA-seq data.

In terms of technical refinements, my current and future efforts will be dedicated to the comparison of predicted AS-derived peptides in tumor samples vs. healthy samples, a process necessary to further verify the cancer-specific nature of predicted peptides. Additionally, the advancement of nanopore sequencing technology promises to enable more accurate identification of AS transcripts than short RNA-seq reads by more definitively determining what combination of alternative splicing events appear in a single transcript, and efforts to characterize AS using nanopore reads have already begun (Tang et al., 2020).

Finally, to realize this work's potential impact on cancer immunotherapies, the most promising AML-specific neoantigens predicted through this pipeline, once subjected to HLA binding prediction, will need to be validated through the detection of antigen-specific T cell responses, and my group is invested in realizing this fundamental step of validation.

More accurately and comprehensively predicting AS-derived peptides explores the fascinating biology of splicing as well as its therapeutic applications as a potential source of cancer neoantigens. My work contributes to shedding light on aberrant splicing, and intron retention in particular, in the relevant setting of acute myeloid leukemia. Identifying novel potential AML-specific peptides is a step toward the design of long-awaited AML-specific immunotherapies.

Methods

Data

I used RNA-seq and Ribo-seq data from the B721.221 B cell line. As >90 HLA class I monoallelic variants were generated for this cell line that served as my model system, I focused on the monoallelic cell lines carrying the following HLA alleles: HLA-A*01:01, HLA-A*33:03, HLA-B*15:01, HLA-B*44:02 (Abelin et al., 2017; Ouspenskaia et al., 2020; Sarkizova et al., 2020). I used RNA-seq data from 8 AML cell lines from the LL-100 panel: CMK, KASUMI-1, MUTZ-3, OCI-AML-3, OCI-M2, SET-2, TF-1, and THP-1 (Quentmeier et al., 2019). I also used RNA-seq data generated in the Wu lab from 7 AML primary samples not yet published but from patients part of a DFCI clinical trial (Ho et al., 2017).

Preprocessing

I trimmed RNA-seq reads of adapter sequences using Cutadapt 1.15, discarding reads below the chosen length threshold of 80 nt or with any unknown nucleotides (Martin, 2011). I then aligned reads to the genome with STAR, using GENCODE gene annotations (Dobin et al., 2013).

I trimmed Ribo-seq reads of primers and barcodes with Cutadapt, stripped them of contaminants such as ribosomal RNA with BowTie (Langmead et al., 2009), and aligned them to the genome with STAR using GENCODE annotations. I then offset-corrected Ribo-seq read alignments with RibORF (Ji et al., 2015). Offset-correction truncates each read to 1 nt and places 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 9**).

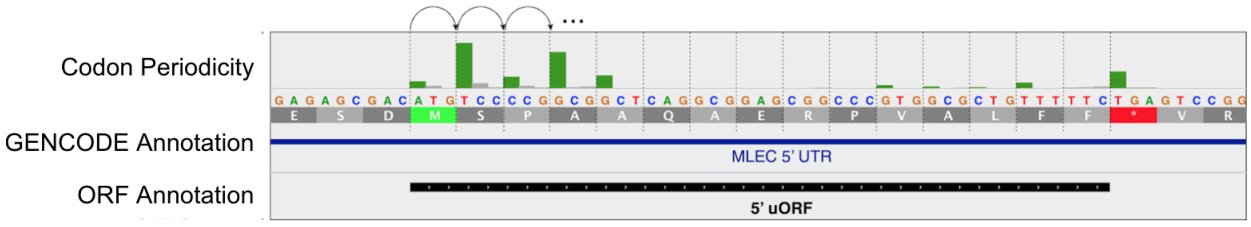


Figure 9

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.

AS-Derived Peptide Prediction Pipeline

To identify RIs, I assembled *de novo* transcripts from aligned RNA-seq data using StringTie (Kovaka et al., 2019). I ran StringTie in its conservative mode and used GENCODE hg38 annotations as a reference for B721 and AML cell line data and hg19 annotations for AML primary samples, which had been previously aligned to that annotation version (Harrow et al., 2012).

I used gffcompare (Pertea and Pertea, 2020) to compare assembled transcripts to annotated transcripts. I only considered assembled transcripts that were multi-exonic and share at least one splice junction with an annotated transcript isoform, as determined by gffcompare class codes "m", "n", and "j". I then identified the assembled transcripts that contain at least one non-canonical retained intron, alternative 5' splice junction, alternative 3' splice junction, or cassette exon and therefore included a portion of a canonically constitutive intron to a transcript. If that intronic region was fully included in any canonical transcript isoform annotation in GENCODE or RefSeq, I discarded the AS event. I considered all transcripts including any listed AS events as AS isoforms.

I translated AS isoforms that preserve a start codon from a canonical transcript isoform, beginning at that start codon and ending at the first in-frame stop codon (**Figure 10**). If the resulting protein sequence contained sequence from a noncanonical intronic region identified in prior steps, I added it to my protein search database. I worked conservatively, translating only in one frame and without seeking out noncanonical start codons, in favor of blunter approaches such as translating the entire transcript in three frames, to limit the size of the search space. This approach accounts for both intronic sequence, novel junctions, and downstream peptides, offering a more comprehensive approach than existing work.

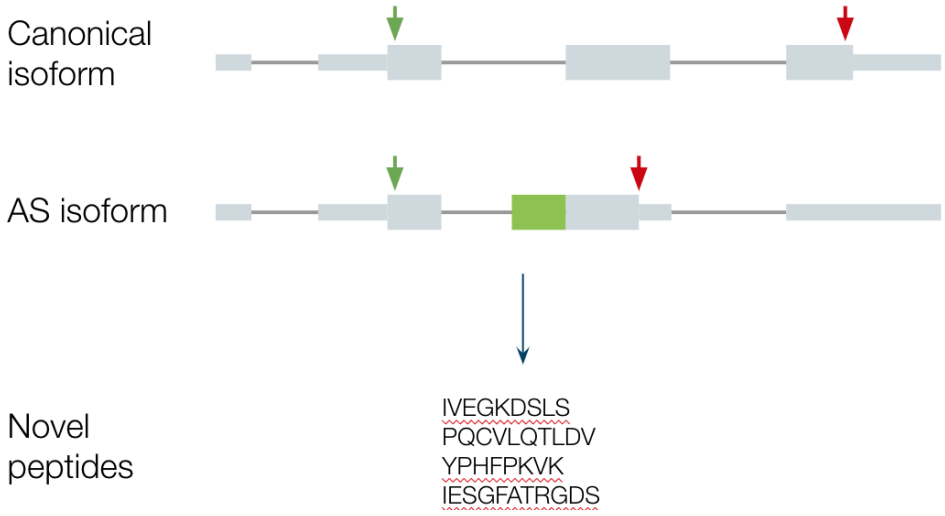


Figure 10

AS translation methodology. AS isoforms were translated from canonical start codons to the first downstream in-frame stop codon. I determine the canonical transcript(s) that the AS transcript most closely matches, use the canonical start codon(s) from those transcript(s), and I translate up to the first downstream, in-frame stop codon (which in this example is introduced in the second exon by a frameshift). I then find all 8-11mers from the resulting proteins and discard any peptides present in the canonical proteome to obtain a set of novel peptide predictions.

I then generated features quantifying the RNA-seq read support for intronic segments predicted to be retained in isoforms due to AS, their adjacent canonically exonic regions, and their boundaries (e.g. reads spanning a novel junctions introduced by an alternate 5' splice site, reads spanning the intron-exon boundary of a retained intron). I also noted AS isoforms' TPM, AS isoforms' expression level relative to corresponding canonical transcripts, and the percent of bases in the isoform supported by RNA-seq reads (base coverage). I also inspected the proportion of reads with mismatches, indels, and clipping and of reads that mapped to multiple genomic loci in intronic segments and their adjacent exons, which sheds light on the quality of reads being used to call a given AS event. Additionally, I calculate intronic segments' GC content, as higher GC content may characterize a subset of RIs (Jacob and Smith, 2017). I checked whether AS events introduce a premature termination codon (PTC) into transcripts, which may lead to nonsense-mediated decay (Monteuuis et al., 2019). In cases of alternative splice sites or cassette exons, I determine the splicing motifs introduced by novel splice junctions to determine whether they adhered to canonical expectations. Finally, when Ribo-seq data was available, I inspected Ribo-seq read support for predicted AS events as well. These features may

allow the filtering of AS isoforms and peptides to obtain a smaller but higher confidence set of AS-derived peptides.

Validation in B721

I validated predicted AS-derived peptides as translated by determining the number of Ribo-seq reads aligned to the corresponding genomic sequences. Ribo-seq, or ribosome profiling, 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). I require ≥1 Ribo-seq read mapped to a given peptide to validate its translation.

I validated predictions as likely to be HLA presented by using HLAthena to predict which peptides correspond to expected binding motifs, using a percent rank threshold of 0.5%. Efforts to search for predicted peptides in HLA class I immunopeptidome mass spectrometry data from monoallelic B721 lines to validate their translation and HLA presentation are also underway.

Acknowledgements

Aviv Regev, Tamara Ouspenskaia, and Travis Law served as my mentors in my first year of this project, and Cathy Wu, Nicoletta Cieri, and Kari Stromhaug are serving as my mentors in my second year.

Karl Clauser (affiliated with the Broad Institute) performed mass spectrometry analysis that was incorporated in a previous version of this paper. The results are not directly included in this version of the paper, but they shaped the course of the project. (As discussed, further MS analysis efforts are underway and will be included in a future version of this paper.) Sisi Sarkizova and Brian Haas have provided useful advice regarding computational methods for alternative splicing isoform identification and HLAthena analysis.

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.

Dvinge, H., and Bradley, R.K. (2015). Widespread intron retention diversifies most cancer transcriptomes. Genome Med. *7*, 45.

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.

Ho, V.T., Kim, H.T., Bavli, N., Mihm, M., Pozdnyakova, O., Piesche, M., Daley, H., Reynolds, C., Souders, N.C., Cutler, C., et al. (2017). Vaccination with autologous myeloblasts admixed with GM-K562 cells in patients with advanced MDS or AML after allogeneic HSCT. Blood Adv *1*, 2269–2279.

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.

Jacob, A.G., and Smith, C.W.J. (2017). Intron retention as a component of regulated gene expression programs. Hum. Genet. *136*, 1043–1057.

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*.

Kahles, A., Ong, C.S., Zhong, Y., and Rätsch, G. (2016). SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. Bioinformatics *32*, 1840–1847.

Kahles, A., Lehmann, K.-V., Toussaint, N.C., Hüser, M., Stark, S.G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Cancer Genome Atlas Research Network, et al. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell *34*, 211–224.e6.

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.

Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L., and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. *20*, 278.

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.

Middleton, R., Gao, D., Thomas, A., Singh, B., Au, A., Wong, J.J.-L., Bomane, A., Cosson, B., Eyras, E., Rasko, J.E.J., et al. (2017). IRFinder: assessing the impact of intron retention on mammalian gene expression. Genome Biol. *18*, 51.

Monteuuis, G., Wong, J.J.L., Bailey, C.G., Schmitz, U., and Rasko, J.E.J. (2019). The changing paradigm of intron retention: regulation, ramifications and recipes. Nucleic Acids Res. *47*, 11497–11513.

Oghabian, A., Greco, D., and Frilander, M.J. (2018). IntEREst: intron-exon retention estimator. BMC Bioinformatics *19*, 130.

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.

Ouspenskaia, T., Law, T., Clauser, K.R., Klaeger, S., Sarkizova, S., Aguet, F., Li, B., Christian, E., Knisbacher, B.A., Le, P.M., et al. (2020). Thousands of novel unannotated proteins expand the MHC I immunopeptidome in cancer.

Pertea, G., and Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. F1000Res. *9*, 304.

Pimentel, H., Conboy, J.G., and Pachter, L. (2015). Keep Me Around: Intron Retention Detection and Analysis.

Quentmeier, H., Pommerenke, C., Dirks, W.G., Eberth, S., Koeppel, M., MacLeod, R.A.F., Nagel, S., Steube, K., Uphoff, C.C., and Drexler, H.G. (2019). The LL-100 panel: 100 cell lines for blood cancer studies. Sci. Rep. *9*, 8218.

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.

Sarkizova, S., Klaeger, S., Le, P.M., Li, L.W., Oliveira, G., Keshishian, H., Hartigan, C.R., Zhang, W., Braun, D.A., Ligon, K.L., et al. (2020). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat. Biotechnol. *38*, 199–209.

Shen, S., Park, J.W., Lu, Z.-X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U. S. A. *111*, E5593–E5601.

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. *36*, 1056–1058.

Tang, A.D., Soulette, C.M., van Baren, M.J., Hart, K., Hrabeta-Robinson, E., Wu, C.J., and Brooks, A.N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat. Commun. *11*, 1438.

Vanichkina, D.P., Schmitz, U., Wong, J.J.-L., and Rasko, J.E.J. (2018). Challenges in defining the role of intron retention in normal biology and disease. Semin. Cell Dev. Biol. *75*, 40–49.

Zhou, J., and Chng, W.-J. (2017). Aberrant RNA splicing and mutations in spliceosome complex in acute myeloid leukemia. Stem Cell Investig *4*, 6.