Published in final edited form as:

Nat Biotechnol. 2018 December ; 36(11): 1056–1058. doi:10.1038/nbt.4239.

Intron retention is a source of neoepitopes in cancer

Alicia C. Smart[#1,2], **Claire A. Margolis**[#1,2], **Harold Pimentel**[3], **Meng Xiao He**[1,2], **Diana Miao**[1,2], **Dennis Adeegbe**[1,4], **Tim Fugmann**[5], **Kwok-Kin Wong**[1,4], and **Eliezer M. Van Allen**[*, 1,2]

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA 02179, USA. [3]Department of Genetics and Biology, Stanford University, Stanford, CA 94305, USA. [4]Perlmutter Cancer Center at NYU Langone Medical Center, New York, NY 10016, USA. [5]Philochem AG, Otelfingen, Switzerland.

[#] These authors contributed equally to this work.

Abstract

We present an *in silico* approach to identify neoepitopes derived from intron retention events in tumor transcriptomes. Using mass spectrometry immunopeptidome analysis, we show that retained intron (RI) neoepitopes are processed and presented on MHC-I on the surface of cancer cell lines. RNA-derived neoepitopes should be considered for prospective personalized cancer vaccine development.

Personalized cancer vaccines comprising neoepitope peptides generated from somatic mutations have shown potential as targeted immunotherapies[1–3]. Other types of aberrant peptides, including cancer germline antigens generated from genes that are transcriptionally silent in adult tissues, have been shown to act as tumor neoepitopes in immune rejection[4, 5]. Dysregulation of RNA splicing through intron retention, which is common in tumor transcriptomes[6, 7], represents another potential source of tumor neoepitopes, but has not

[*]Correspondence to eliezerm_vanallen@dfci.harvard.edu (EMV).

Authors' Contributions

Conception and design: A. C. Smart, C. A. Margolis, E. M. Van Allen

Development of methodology: C. A. Margolis, A. C. Smart, H. Pimentel, M. X. He, T. Fugmann, D. Miao, K. Wong, E. M. Van Allen

Analysis and interpretation of data (e.g., pipeline development, statistical analysis, computational analysis): C. A. Margolis, A. C. Smart, D. Adeegbe

Writing, review, and/or revision of the manuscript: C. A. Margolis, A. C. Smart, H. Pimentel, M. X. He, D. Miao, D. Adeegbe, T. Fugmann, K. Wong, E. M. Van Allen

Study supervision: E. M. Van Allen

Disclosure of Potential Conflicts of Interest

EMV holds consulting roles with Tango Therapeutics, Invitae, and Genome Medical and receives research support from Bristol-Myers Squibb and Novartis.

Data availability

Raw RNA-Seq data for the Snyder et al. 2014 patient cohort are available on dbGaP under accession code phs001038.v1.p1 and for the Hugo et al. 2016 cohort on the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under the accession number SRA: SRP070710.

Code availability

Pipeline code is publicly accessible on GitHub at https://github.com/vanallenlab/retained-intronneoantigen-pipeline.

been previously explored. Intron retention is caused by splicing errors that lead to inclusion of an intron in the final mRNA transcript. RI transcripts are translated and degraded by the nonsense-mediated decay (NMD) pathway, which generates peptides for endogenous processing, proteolytic cleavage, and presentation on MHC-I[8–10].

We developed a computational approach for detecting intron retention events from tumor RNA-seq data (Fig. 1A, Methods). Intron fragments likely to be translated based on their position downstream from a translated exon and upstream from an in-frame stop codon were identified. Predicted binding affinities between retained intron (RI) peptide sequences and sample-specific HLA class I alleles were calculated to identify candidate RI neoepitopes. We filtered and thresholded preliminary results to exclude artifacts. This process (**Methods**) generated a robust list of putative RI neoepitopes for each sample.

We applied this method to tumor sequencing data from two cohorts of melanoma patients treated with checkpoint inhibitors[11, 12] to identify putative RI neoepitopes (n = 48 melanomas; Supplementary Tables S1 and S2). Apart from one outlier, both cohorts had comparable levels of intron retention and predicted RI neoepitopes (Fig. 1B). Slight variation in RI neoepitope load between cohorts was expected given differences in RNA sequencing run, depth, and quality[13]. The total predicted neoepitope load included RI neoepitopes, and somatic mutation neoepitopes derived computationally using published methods (Supplementary Fig. S1, Supplementary Table S1, Methods). Most patients showed substantially augmented total neoepitope loads with the additional consideration of RI neoepitopes. Mean somatic neoepitope load was 2,218 and mean RI neoepitope load was 1,515, yielding a ~0.7-fold increase in mean total neoepitope load with the addition of RI neoepitopes (Fig. 1C). Excluding one outlier sample with a vastly higher level of somatic neoepitopes than the rest, incorporation of RI neoepitopes roughly doubled the total neoepitope load. There was not a significant correlation between somatic neoepitope load and RI neoepitope load (Ordinary Linear Regression p = 0.63) (Supplementary Fig. S2).

To demonstrate that RI neoepitopes are processed and presented on MHC-I, we predicted RI neoepitopes from six human tumor cell lines and detected neoepitopes that were complexed to MHC-I by mass spectrometry (Supplementary Table S3). In melanoma cell line MeWo, the predicted RI neoepitopes *EVYAAGKYV* and *YAAGKYVSF* from *KCNAB2* (chr1:6142308–6145287) were experimentally discovered in complex with MHC-I via mass spectrometry with high confidence (Fig. 2A). We identified RI neoepitopes in another melanoma cell line, SK-MEL-5 (*AMSDVSHPK* and *LAMSDVSHPK* from *SMARCD1*), in B cell lymphoma cell lines CA46 (*FRYVAQAGL* from *LRSAM1*) and DOHH-2 (*TLFLLSLPL* and *FLLSLPLPV* from *CYB561A3*), and in leukemia cell lines HL-60 (*SVLDDVRGW* from *TAF1*) and THP-1 (*LTSQGKSAF* from *ZCCHC6*) (**Fig. 2B,** Supplementary Fig. S3). Applying this method to somatic mutation-derived neoepitopes, a comparable percentage of predicted neoepitopes were detected by mass spectrometry (Supplementary Table S4). The discovery of peptides in complex with MHC-I in cell lines using mass spectrometry with RI neoepitope sequences predicted computationally with our pipeline provides direct evidence of the processing and presentation of RI neoepitopes through the MHC-I pathway.

Given that somatic neoepitope burden is a known correlate of checkpoint inhibitor response in melanoma[14], we next examined whether RI neoepitope load might be similarly associated with response. However, there was no association between RI neoepitope load and clinical benefit from checkpoint inhibitor therapy, nor was there correlation with expression of canonical markers of immune cytolytic activity, CD8A, GZMA, PRF1[15], or clinical covariates (Pearson correlation $p > 0.05$ for all, and Supplementary Fig. S4-S6). Rather, there was a non-significant trend of association between high RI neoepitope load and lack of benefit (Two-sided Mann-Whitney U $p = 0.29$ Snyder cohort, 0.61 Hugo cohort). High RI neoepitope load tumors and checkpoint inhibitor nonresponder tumors, with only 38% overlap, shared common transcriptional programs consistent with cell cycle and DNA damage repair activity (Supplementary Fig. S7 and Supplementary Table S5).

Here, we demonstrate that tumor-specific RI neoepitopes can be identified computationally in both patient- and cell line-derived samples and a subset can be validated as presented in complex with MHC-I. These data support the hypothesis that aberrant splicing results in intron retention, which generates abnormal transcripts that are translated into immunogenic peptides, loaded on MHC-I and presented to the immune system, underscoring their relevance in patients receiving immunotherapy. Further studies will be necessary to clinically validate the immunogenicity of specific RI neoepitopes in patients, including identification of T cells specific to predicted RI neoepitopes.

Furthermore, we found that RI neoepitope load is not associated with checkpoint inhibitor response and discovered that patients with high RI neoepitope load are transcriptionally similar to immunotherapy nonresponders; both patient groups have enrichment of cell cycle and DNA damage repair-related gene sets. Intron retention has been shown to regulate the cell cycle in both non-malignant[16] and malignant cells[17]. These findings warrant further investigation and experimental validation, given the emerging synergistic relationship between cell cycle inhibition and immune checkpoint blockade therapies[18–20].

Identification of a wider array of tumor neoepitopes, including those derived from somatic mutation, aberrant gene expression, and splicing dysregulation, will contribute to a more complete understanding of the tumor immune landscape. Additional work dissecting the relationship between the prediction, processing and presentation, and ultimate immunogenicity of neoepitopes derived from different sources will be required to ensure clinical relevance of this approach. It has been shown that melanoma in particular may feature certain shared epitopes across patients that are derived from incomplete splicing processes, which may render these cancers more susceptible to RI-derived neoepitopes[21, 22]. Similar approaches across different tissues will provide further clarity on the role of RI neoepitopes in tumor immunity across cancer contexts. Currently, our findings are limited by the availability of clinically annotated cohorts with high quality RNA sequencing and matched normal tissue. Incorporation of matched normal tissue will improve exclusion of retained introns that represent normal gene expression and may help increase precision of our filtering approach. Prediction of patient-specific RI-neoepitopes has the potential to contribute to the development of personalized cancer vaccines.

Online Methods

Clinical cohorts

Analysis was conducted on published cohorts of melanoma patients treated with immune checkpoint inhibitors. The Hugo et al. cohort included samples from 27 melanoma patients (26 pretreatment, 1 on-treatment) treated with the PD-1 inhibitor pembrolizumab[11]. Patient outcomes were classified as responding to therapy (R) (n=14) or not responding to therapy (NR) (n=13), as described in the original publication. These samples were sequenced from fresh frozen tissue using a standard, poly(A) selected protocol. The Snyder cohort included post-treatment samples for 21 melanoma patients treated with ipilimumab (anti-CTLA-4 therapy)[12, 23]. Outcomes were classified as receiving long-term clinical benefit (LB) (n=8) or not receiving clinical benefit (NB) (n=13), as described in the original publication. RNA sequencing of the Snyder cohort was performed on fresh frozen tissue using a standard, poly(A) selected protocol.

RI neoepitope pipeline

Raw RNA-Seq FASTQ files were pseudoaligned to an augmented hg19 (GENCODE Release 19, GRCh37.p13)[24] transcriptome index containing both exonic and intronic transcript sequences, and transcript expression was quantified via kallisto[25]. The KMA algorithm[26], implemented as a suite of Python scripts within an R package, was used to identify the genomic loci of expressed intron retention events with limited false positives. Using these RI loci, the UCSC Table Browser[27] database was queried via public MySQL server to obtain the nucleotide sequences corresponding to the intronic regions and fragments of the previous exonic sequences, as well as the open reading frame orientation at the start of the intron. RI peptide sequences of 9–10 amino acids, with at least one intronic amino acid, were generated by translating open reading frames into intronic sequences until hitting an in-frame stop codon. These peptides, along with sample HLA Class I alleles identified via the POLYSOLVER algorithm[28], were assessed for putative peptide-MHC I binding affinity via NetMHCpan v3.1[29]. A threshold of rank < 0.5% was used to identify putative RI neoepitopes.

Several filters were applied at various steps throughout the pipeline to eliminate likely false positive RIs and RI neoepitopes. After expression quantification, RIs expressed at a level ≤ 1 transcript per million, likely artifactual, were eliminated from the analysis. Additional expression-based filters were applied within the KMA algorithm: RIs that did not reach a level of at least five unique counts in at least 25% of samples in a cohort and whose neighboring exons did not reach a level of at least one transcript per million in at least 25% of samples in a cohort were eliminated as false positives[26]. Due to the absence of matched normal RNA-Seq data for our melanoma clinical cohorts, a ‘panel of normals’ approach was taken in an attempt to filter out introns commonly retained in normal skin tissue, which would not produce immunogenic peptides due to likely host immune tolerance. RIs were identified in six normal skin samples (three individuals, two samples per individual: Individual ERS326932 with samples ERR315339 and ERR315376, Individual ERS326943 with samples ERR315372 and ERR315460, and Individual ERS327007 with samples ERR315401 and ERR315464) from the Human Protein Atlas. RNA-Seq paired-end FASTQ

files for each sample were downloaded from the following open-access link: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1733/samples/. All normal sample retention profiles were highly concordant, both within and across individuals (Supplementary Fig. S8A). The final filter set of 7,050 normal RIs was obtained by intersecting the sets of RIs shared by each unique combination of one sample per individual —eight groups total (Supplementary Fig. S8B, Supplementary Table S6). These RIs were eliminated from downstream tumor sample analyses. In addition, RI peptides with amino acid sequences present in the normal proteome, derived from the UniProt human reference proteome version 2017_03, downloaded on 07/05/2017, were filtered due to likely host immune tolerance[30]. Finally, a set of RIs that were flagged due to abnormally high expression values and discovered upon manual review via Integrative Genomics Viewer[31] to be erroneously-annotated in either the reference transcriptome or the Table Browser database were eliminated from the analysis (Supplementary Fig. S9A-D, Supplementary Table S6).

Pipeline code is publicly accessible on GitHub at https://github.com/vanallenlab/retained-intron-neoepitope-pipeline.

Clinical cohort somatic neoepitope analysis

Putative somatic neoepitopes were identified *in silico* for each sample as described in Van Allen et al. 2015[14]. Briefly, BAM files from each cohort underwent sequencing quality control to ensure concordance between tumor and matched normal sequences and adequate depth of sequencing coverage. Single nucleotide variants were called using MuTect[32] and insertions and deletions were called using Strelka[33]. Annotation of identified variants was done using Oncotator (http://www.broadinstitute.org/cancer/cga/oncotator). Sequences of 9–10 amino acid peptides with at least one mutant amino acid were generated. These peptides, along with HLA Class I alleles called with POLYSOLVER were analyzed using NetMHCpan v3.0 to identify HLA-peptide binding interactions[28, 29]. For each patient, all peptides with predicted binding rank ≤ 2.0% for at least one patient HLA Class I allele were called somatic neoepitopes.

Cell line analyses

Raw RNA-Seq data from the following published[34] cell lines: CA46, DOHH-2, HL-60, THP-1, MeWo, SK-MEL-5 were obtained from the Cancer Cell Line Encyclopedia[35] via the NCI Genomic Data Commons and run through our computational pipeline as previously described, with minor adaptations as described henceforth. HLA Class I alleles were used for each cell line as enumerated in publication. A threshold of predicted binding rank ≤ 2.0% for at least one HLA Class I allele was used to distinguish cell line RI neoepitopes. All pipeline filters applied to patient data described above were implemented on the cell line data *except* RI neoepitopes expected to be retained in normal tissue were not filtered due to the fact that these experiments were focused on presentation of RI neoepitopes rather than immune system stimulation once presented.

Mass spectrometric data from Ritz et al.[34], as well as previously unpublished data for cell lines MeWo, DOHH-2, and SK-MEL-5, was searched against a database consisting of 93,250 sequences of the human reference proteome downloaded from UniProt on July 7,

2017 concatenated with putative retained intron sequences (TPM > 1), or concatenated with 133,811 intron sequences with TPM < 1 (not retained) as negative control. Fragment mass spectra were searched with SEQUEST and filtered to a 1% false discovery rate with percolator to identify high confidence events.

Gene set enrichment analysis

Gene expression was quantified in patient samples using kallisto[25]. Gene set enrichment analysis (GSEA) was run to compare both top quartile vs. bottom quartile RI load patients and immunotherapy responders vs. nonresponders. Initially, 50 Hallmark gene sets were tested[36]. GSEA analyses of the Founders gene sets underlying the Hallmark gene sets that were significantly enriched in both top quartile vs. bottom quartile RI load patients and immunotherapy responders vs. nonresponders were subsequently performed. All statistical values reported are Benjamini-Hochberg FDR q values corrected for multiple hypothesis testing.

Statistical analyses

Assessment of difference in means or medians for a continuous variable between two clinical response groups (i.e., clinical benefit vs. no clinical benefit) was performed using the two-sided nonparametric Mann-Whitney U test for non-normally-distributed variables (e.g., RI neoepitope burden). All statistical analyses were conducted in the R statistical software environment (v.3.3.1).

Life Sciences Reporting Summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful to Dario Neri for fruitful discussions, Danilo Ritz for the purification of HLA peptides from cell lines, and Mahmoud Ghandi for assistance in coordinating access to cell line transcriptome data.

Grant support

This work was supported by the BroadNext10, NIH K08 CA188615, NIH R01 CA227388, and Prostate Cancer Foundation-V Foundation Challenge Award.

References and Notes:

1. Ott PA et al. Nature 547, 217–221 (2017). [PubMed: 28678778]
2. Sahin U et al. Nature 547, 222–226 (2017). [PubMed: 28678784]
3. Carreno BM et al. Science 348, 803–808 (2015). [PubMed: 25837513]
4. Hunder NN et al. N Engl J Med 358, 2698–2703 (2008). [PubMed: 18565862]
5. Robbins PF et al. Clin Cancer Res 21, 1019–1027 (2015). [PubMed: 25538264]
6. Dvinge H & Bradley RK Genome Med 7, 45 (2015). [PubMed: 26113877]

7. Jung H et al. Nat Genet 47, 1242–1248 (2015). [PubMed: 26437032]
8. Apcher S et al. Proc Natl Acad Sci U S A 108, 11572–11577 (2011). [PubMed: 21709220]
9. Rock KL, Farfan-Arribas DJ & Shen L J Immunol 184, 9–15 (2010). [PubMed: 20028659]
10. Pearson H et al. J Clin Invest 126, 4690–4701 (2016). [PubMed: 27841757]
11. Hugo W et al. Cell 165, 35–44 (2016). [PubMed: 26997480]
12. Snyder A et al. N Engl J Med 371, 2189–2199 (2014). [PubMed: 25409260]
13. Li S et al. Nat Biotechnol 32, 888–895 (2014). [PubMed: 25150837]
14. Van Allen EM et al. Science 350, 207–211 (2015). [PubMed: 26359337]
15. Rooney MS, Shukla SA, Wu CJ, Getz G & Hacohen N Cell 160, 48–61 (2015). [PubMed: 25594174]
16. Middleton R et al. Genome Biol 18, 51 (2017). [PubMed: 28298237]
17. Dominguez D et al. Elife 5 (2016).
18. Deng J et al. Cancer Discov 8, 216–233 (2018). [PubMed: 29101163]
19. Schaer DA et al. Cell Rep 22, 2978–2994 (2018). [PubMed: 29539425]
20. Goel S et al. Nature 548, 471–475 (2017). [PubMed: 28813415]
21. Lupetti R et al. J Exp Med 188, 1005–1016 (1998). [PubMed: 9743519]
22. Andersen RS et al. Oncoimmunology 2, e25374 (2013). [PubMed: 24073381]
23. Nathanson T et al. Cancer Immunol Res 5, 84–91 (2017). [PubMed: 27956380]
24. Harrow J et al. Genome Res 22, 1760–1774 (2012). [PubMed: 22955987]
25. Bray NL, Pimentel H, Melsted P & Pachter L Nat Biotechnol 34, 525–527 (2016). [PubMed: 27043002]
26. Pimentel H et al. Nucleic Acids Res 44, 838–851 (2016). [PubMed: 26531823]
27. Karolchik D et al. Nucleic Acids Res 32, D493–496 (2004). [PubMed: 14681465]
28. Shukla SA et al. Nat Biotechnol 33, 1152–1158 (2015). [PubMed: 26372948]
29. Nielsen M & Andreatta M Genome Med 8, 33 (2016). [PubMed: 27029192]
30. The UniProt C Nucleic Acids Res 45, D158–D169 (2017). [PubMed: 27899622]
31. Robinson JT et al. Nat Biotechnol 29, 24–26 (2011). [PubMed: 21221095]
32. Cibulskis K et al. Nat Biotechnol 31, 213–219 (2013). [PubMed: 23396013]
33. Saunders CT et al. Bioinformatics 28, 1811–1817 (2012). [PubMed: 22581179]
34. Ritz D et al. Proteomics 16, 1570–1580 (2016). [PubMed: 26992070]
35. Barretina J et al. Nature 483, 603–607 (2012). [PubMed: 22460905]
36. Subramanian A et al. Proc Natl Acad Sci U S A 102, 15545–15550 (2005). [PubMed: 16199517]

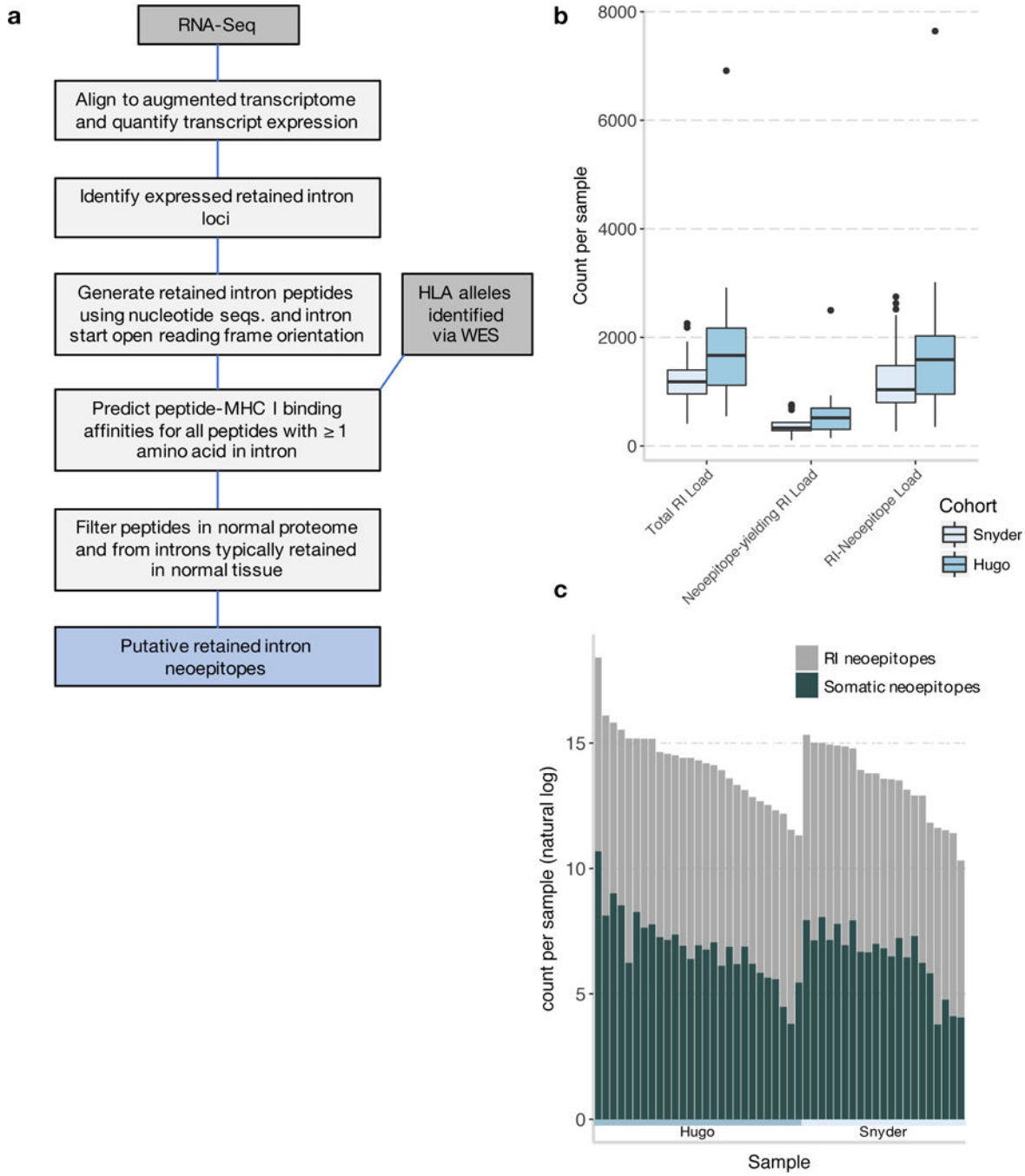

Figure 1.

A, *In silico* pipeline detects intron retention events from transcriptome sequencing, determines open reading frames extending into introns, and identifies putative HLA-specific neoepitopes. **B,** Distribution of total RI load, neoepitope-yielding RI load, and RI neoepitope load in patient cohorts (n = 27 Hugo samples, n = 21 Snyder samples). Boxplots show the median, first, and third quartiles, whiskers extend to 1.5 x the interquartile range, and outlying points are plotted individually. **C,** Somatic and RI neoepitope load by patient.

Within each cohort, patients are sorted by total neoepitope load. Neoepitope counts (y-axis values) are represented in natural log format.

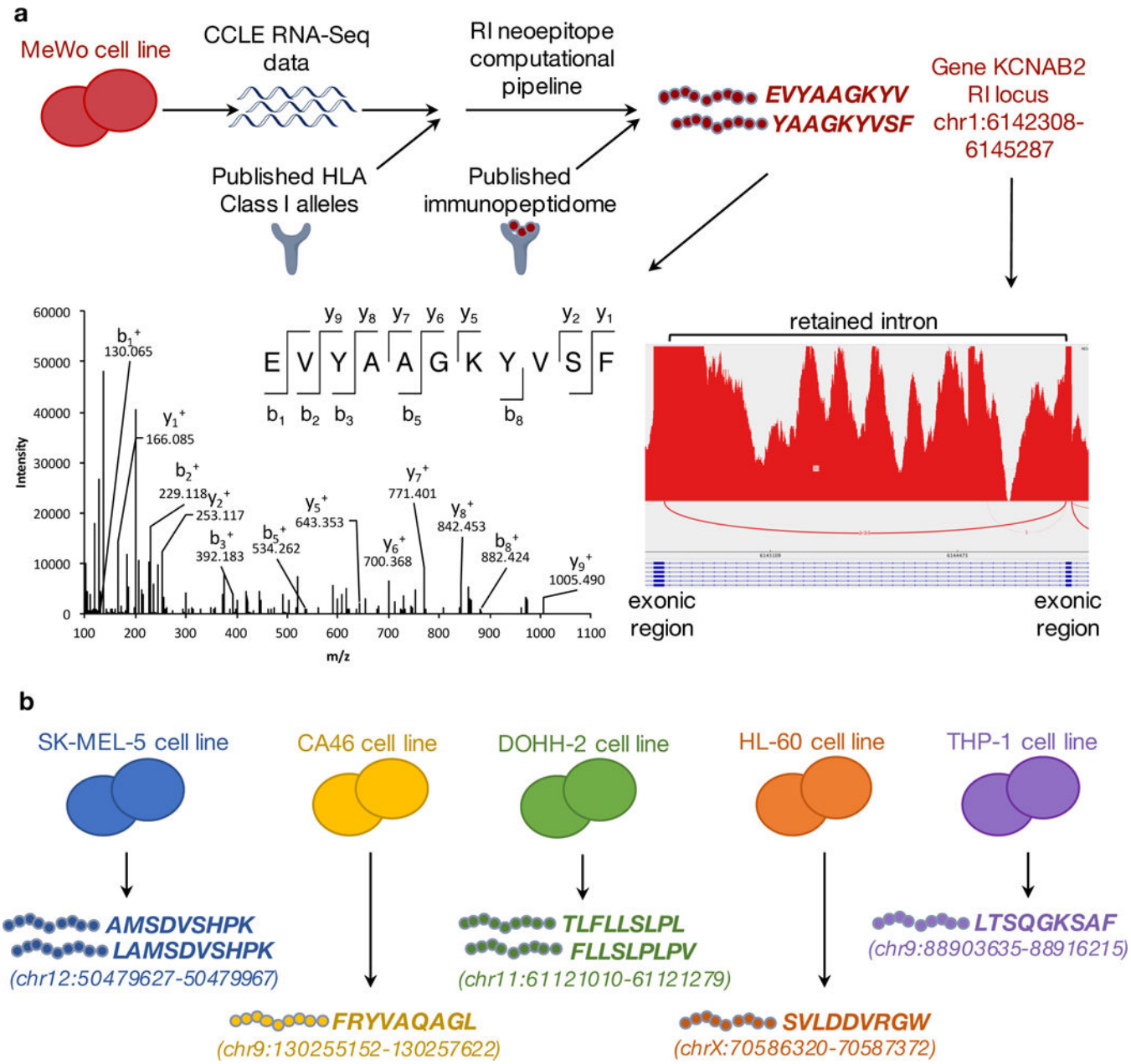

Figure 2.

A, Two RI neoepitopes identified in the MeWo cell line originating from gene *KCNAB2* were both predicted *in silico* and found by mass spectrometry in the MeWo immunopeptidome. Integrative Genomics Viewer (IGV) sashimi plot indicating RNA-Seq read depth (RI expression in TPM=5.13, percent-spliced-in [PSI] value=1.07%) and mass spectra. Experiments were repeated five times with independent measurements for cell line MeWo. Neoepitopes shown had one peptide-to-spectrum match (PSM) and were identified in one replicate within 1% false discovery rate (FDR). **B,** Predicted RI neoepitopes were found to have mass spectrometric evidence supporting their presentation in complex with

MHC I using the same methodology in additional tumor cell lines: SK-MEL-5, CA46, DOHH-2, HL-60, and THP-1.