nature biotechnology

Intron retention is a source of neoepitopes in cancer

Alicia C Smart[1,2,6], Claire A Margolis[1,2,6], Harold Pimentel[3], Meng Xiao He[1,2], Diana Miao[1,2], Dennis Adeegbe[1,4], Tim Fugmann[5], Kwok-Kin Wong[1,4] & Eliezer M Van Allen[1,2]

We present an *in silico* approach to identifying neoepitopes derived from intron retention events in tumor transcriptomes. Using mass spectrometry immunopeptidome analysis, we show that retained intron 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 been previously explored. Intron retention is caused by splicing errors that lead to inclusion of an intron in the final mRNA transcript. Retained intron (RI) transcripts are translated and degraded by the nonsense-mediated decay pathway, which generates peptides for endogenous processing, proteolytic cleavage and presentation on MHC type I[8–10].

We developed a computational approach to detecting intron retention events from tumor RNA-seq data (**Fig. 1a** and Online Methods). Intron fragments likely to be translated on the basis of their position downstream of a translated exon and upstream of an in-frame stop codon were identified. Predicted binding affinities between RI peptide sequences and the products of sample-specific HLA class I alleles were calculated to identify candidate RI neoepitopes. We filtered and thresholded preliminary results to exclude artifacts. This process (Online 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 1** and **2**). 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, as well as somatic mutation neoepitopes derived computationally using published methods (**Supplementary Fig. 1**, **Supplementary Table 1** and Online 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 no significant correlation between somatic neoepitope load and RI neoepitope load (ordinary linear regression P = 0.63; **Supplementary Fig. 2**).

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 3**). 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** and **Supplementary Fig. 3**). Applying this method to somatic mutation–derived neoepitopes, a comparable percentage of predicted neoepitopes were detected by mass spectrometry (**Supplementary Table 4**). 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 the canonical markers of immune cytolytic activity CD8A, GZMA or PRF1[15], or clinical covariates (Pearson correlation $P > 0.05$ for all; **Supplementary Figs. 4–6**). Rather, there was a nonsignificant trend toward association between high RI neoepitope load and lack of benefit (two-sided Mann–Whitney U, P = 0.29 Snyder[12] cohort, 0.61 Hugo[11] cohort). Tumors with high RI neoepitope load and tumors unresponsive to checkpoint inhibitors, with only 38% overlap, shared common transcriptional programs consistent with cell cycle and DNA damage repair activity (**Supplementary Fig. 7** and **Supplementary Table 5**).

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Department of Genetics and Biology, Stanford University, Stanford, California, USA. [4]Perlmutter Cancer Center at NYU Langone Medical Center, New York, New York, USA. [5]Philochem AG, Otelfingen, Switzerland. [6]These authors contributed equally to this work. Correspondence should be addressed to E.M.V. (eliezerm_vanallen@dfci.harvard.edu).

Received 10 October 2017; accepted 6 August 2018; published online 16 August 2018; doi:10.1038/nbt.4239

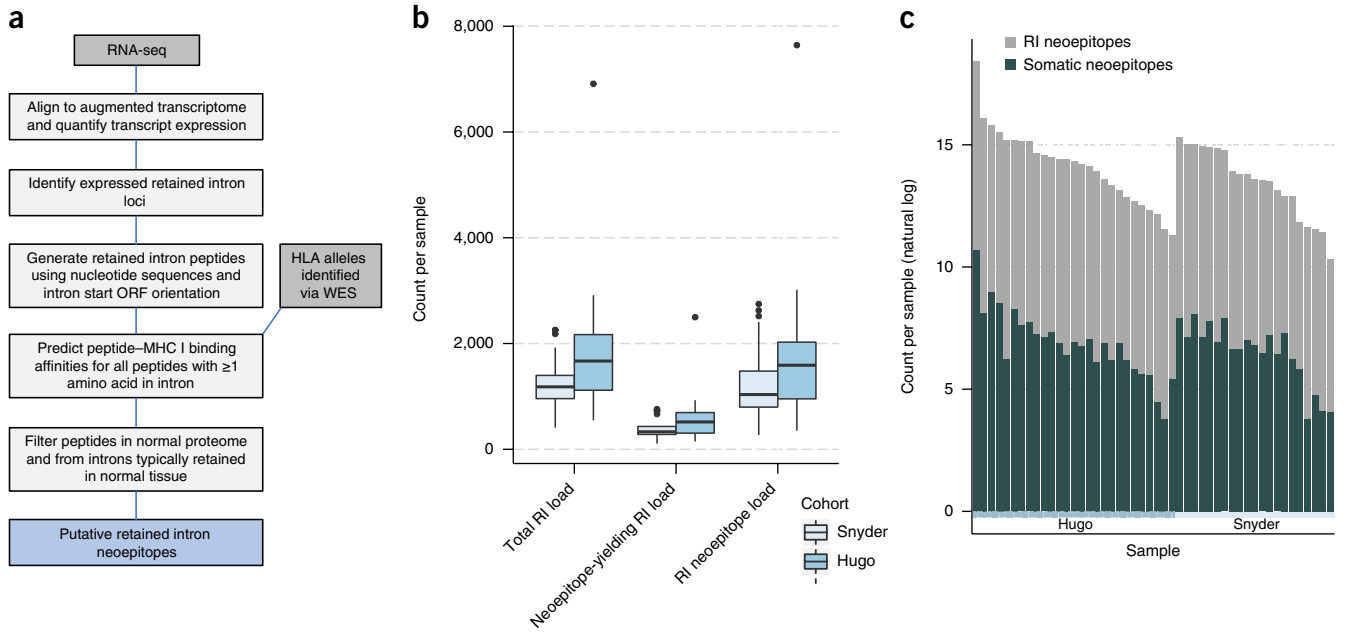

Figure 1 Computationally predicted RI neoepitopes detected in clinical patient cohorts. (**a**) An *in silico* pipeline detects intron retention events from transcriptome sequencing, determines open reading frames extending into introns, and identifies putative HLA-specific neoepitopes. ORF, open reading frame; WES, whole exome sequencing. (**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). Box plots show the median, first and third quartiles, whiskers extend to 1.5 × 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.

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 was not associated with checkpoint inhibitor response and discovered that samples from patients with high RI neoepitope load are transcriptionally similar to those whose tumors did not respond to immunotherapy: 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 nonmalignant[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 RIs 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.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to D. Neri for fruitful discussions, D. Ritz for the purification of HLA peptides from cell lines, and M. Ghandi for assistance in coordinating access to cell line transcriptome data. This work was supported by the BroadNext10, NIH K08 CA188615, NIH R01 CA227388 and a Prostate Cancer Foundation–V Foundation Challenge Award.

AUTHOR CONTRIBUTIONS

Conception and design: A.C.S., C.A.M., E.M.V.A. Development of methodology: C.A.M., A.C.S., H.P., M.X.H., T.F., D.M., K.-K.W., E.M.V.A. Analysis and interpretation of data (for example, pipeline development, statistical analysis, computational analysis): C.A.M., A.C.S., D.A. Writing, review and/or revision of the manuscript: C.A.M., A.C.S., H.P., M.X.H., D.M., D.A., T.F., K.-K.W., E.M.V.A. Study supervision: E.M.V.A.

COMPETING INTERESTS

E.M.V.A. holds consulting roles with Tango Therapeutics, Invitae and Genome Medical and receives research support from Bristol-Myers Squibb and Novartis. T.F. is an employee of Philochem AG.

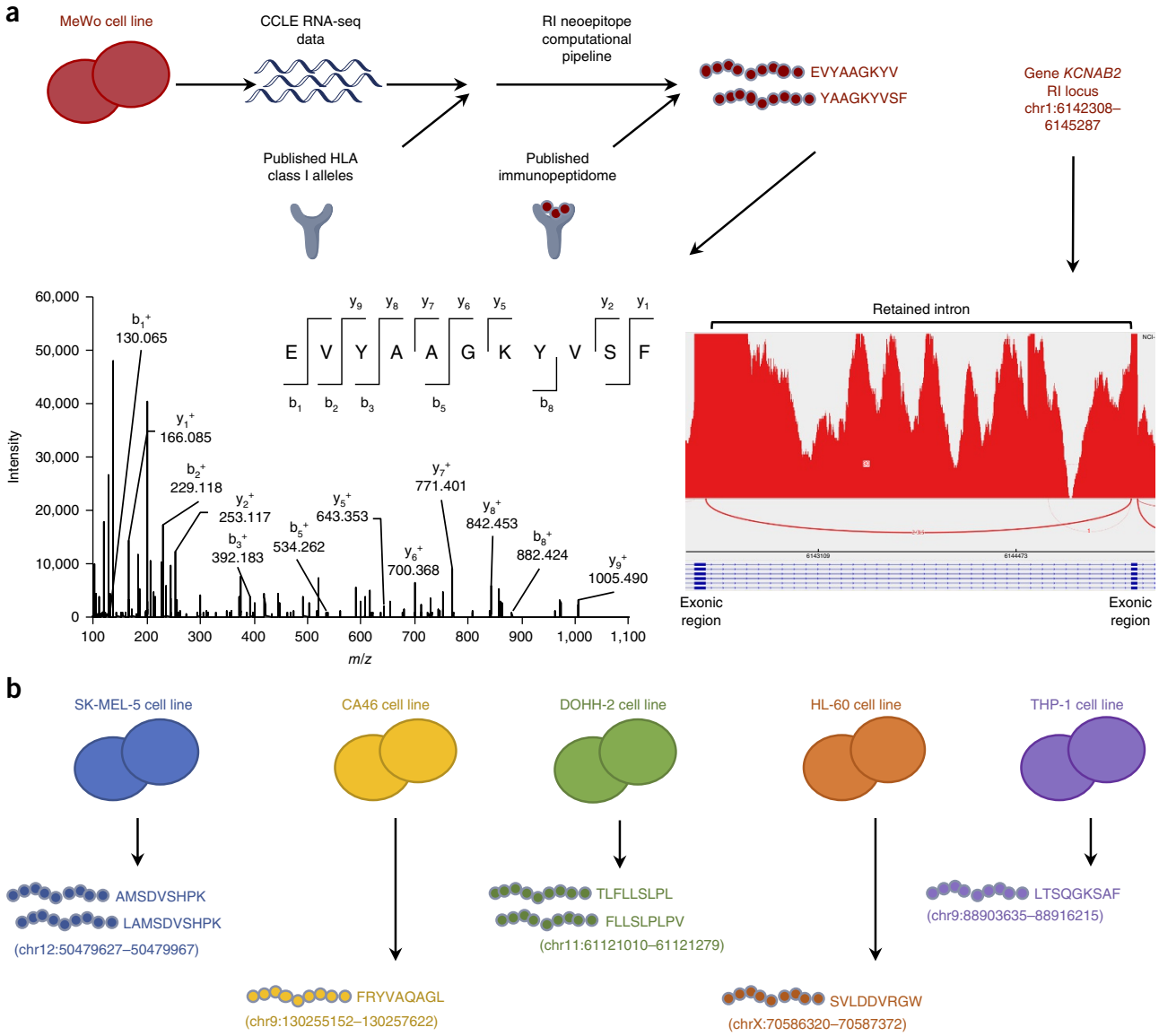

Figure 2 Predicted RI neoepitopes from human cancer cell lines are identified by mass spectrometry bound to MHC class I. (**a**) Two RI neoepitopes identified in the MeWo cell line originating from gene *KCNAB2* were both predicted *in silico* and found by mass spectrometry in the MeWo immunopeptidome. 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 a 1% false discovery rate. CCLE, Cancer Cell Line Encyclopedia. (**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.

1. Ott, P.A. *et al. Nature* **547**, 217–221 (2017).
2. Sahin, U. *et al. Nature* **547**, 222–226 (2017).
3. Carreno, B.M. *et al. Science* **348**, 803–808 (2015).
4. Hunder, N.N. *et al. N. Engl. J. Med.* **358**, 2698–2703 (2008).
5. Robbins, P.F. *et al. Clin. Cancer Res.* **21**, 1019–1027 (2015).
6. Dvinge, H. & Bradley, R.K. *Genome Med.* **7**, 45 (2015).
7. Jung, H. *et al. Nat. Genet.* **47**, 1242–1248 (2015).
8. Apcher, S. *et al. Proc. Natl. Acad. Sci. USA* **108**, 11572–11577 (2011).
9. Rock, K.L., Farfán-Arribas, D.J. & Shen, L. *J. Immunol.* **184**, 9–15 (2010).
10. Pearson, H. *et al. J. Clin. Invest.* **126**, 4690–4701 (2016).
11. Hugo, W. *et al. Cell* **165**, 35–44 (2016).
12. Snyder, A. *et al. N. Engl. J. Med.* **371**, 2189–2199 (2014).
13. Li, S. *et al. Nat. Biotechnol.* **32**, 888–895 (2014).
14. Van Allen, E.M. *et al. Science* **350**, 207–211 (2015).
15. Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G. & Hacohen, N. *Cell* **160**, 48–61 (2015).
16. Middleton, R. *et al. Genome Biol.* **18**, 51 (2017).
17. Dominguez, D. *et al. Elife* **5**, e10288 (2016).
18. Deng, J. *et al. Cancer Discov.* **8**, 216–233 (2018).
19. Schaer, D.A. *et al. Cell Rep.* **22**, 2978–2994 (2018).
20. Goel, S. *et al. Nature* **548**, 471–475 (2017).
21. Lupetti, R. *et al. J. Exp. Med.* **188**, 1005–1016 (1998).
22. Andersen, R.S. *et al. Oncoimmunology* **2**, e25374 (2013).

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 before treatment, 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)-selecting protocol. The Snyder *et al.* 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)-selecting 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 or 10 amino acids, with at least 1 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 5 unique counts in at least 25% of samples in a cohort and whose neighboring exons did not reach a level of at least 1 transcript per million in at least 25% of samples in a cohort were eliminated as false positives[26]. Owing 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 as a result of likely host immune tolerance. RIs were identified in six normal skin samples (three individuals, two samples per individual: subject ERS326932 with samples ERR315339 and ERR315376, subject ERS326943 with samples ERR315372 and ERR315460, and subject 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. 8a**). 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. 8b** and **Supplementary Table 6**). 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 5 July 2017, were filtered because of 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. 9a–d** and **Supplementary Table 6**).

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- or 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 published[34] cell lines CA46, DOHH-2, HL-60, THP-1, MeWo and 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 follows. 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 that RI neoepitopes expected to be retained in normal tissue were not filtered because 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, were searched against a database consisting of 93,250 sequences of the human reference proteome downloaded from UniProt on 7 July 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 patients in the top quartile vs. bottom quartile of RI load and patients whose tumors responded to immunotherapy vs. those whose did not. 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 of the above comparisons were subsequently performed. All statistical values reported are Benjamini–Hochberg false discovery rate 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 (for example, 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 research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. Pipeline code is publicly accessible on GitHub at https://github.com/vanallenlab/retained-intron-neoantigen-pipeline and as **Supplementary Software**.

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[11] cohort on the Sequence Read Archive under accession code SRP070710.

23. Nathanson, T. *et al. Cancer Immunol. Res.* **5**, 84–91 (2017).
24. Harrow, J. *et al. Genome Res.* **22**, 1760–1774 (2012).
25. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. *Nat. Biotechnol.* **34**, 525–527 (2016).
26. Pimentel, H. *et al. Nucleic Acids Res.* **44**, 838–851 (2016).
27. Karolchik, D. *et al. Nucleic Acids Res.* **32**, D493–D496 (2004).
28. Shukla, S.A. *et al. Nat. Biotechnol.* **33**, 1152–1158 (2015).
29. Nielsen, M. & Andreatta, M. *Genome Med.* **8**, 33 (2016).
30. The UniProt Consortium. *Nucleic Acids Res.* **45**, D158–D169 (2017).

doi:10.1038/nbt.4239

31. Robinson, J.T. *et al. Nat. Biotechnol.* **29**, 24–26 (2011).
32. Cibulskis, K. *et al. Nat. Biotechnol.* **31**, 213–219 (2013).
33. Saunders, C.T. *et al. Bioinformatics* **28**, 1811–1817 (2012).
34. Ritz, D. *et al. Proteomics* **16**, 1570–1580 (2016).
35. Barretina, J. *et al. Nature* **483**, 603–607 (2012).
36. Subramanian, A. *et al. Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).

nature research

Corresponding author(s): Eliezer Van Allen

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	All patient data used in this manuscript was downloaded from dbGaP or SRA at the following accession codes: phs001038.v1.p1 (dbGaP) and SRP070710 (SRA). RNA-seq data from the cell lines analyzed is available online via the Cancer Cell Line Encyclopedia, and immunopeptidome data was published.
Data analysis	All custom code from our retained intron neoepitope prediction pipeline has been made publicly available on GitHub at https://github.com/vanallenlab/neoantigen_calling_pipeline as well as in the manuscript Supplementary Information. Commercial code was additionally used from the following published methods: kallisto (v.0.43.1), KMA (v.0.1.0), POLYSOLVER, NetMHCpan (v.3.1). kallisto is a program for quantifying abundances of RNA-seq transcripts, based on the idea of pseudoalignment for repadly determining the compatibility of reads with targets, without the need for exact alignment. kallisto can be downloaded at https://pachterlab.github.io/kallisto/download. KMA is an R package that performs intron retention estimation and detection using biological replicates and resampling. Updated code can be found at https://github.com/pachterlab/kma. POLYSOLVER is a tool for HLA typing of MHC class I based on whole exome sequencing data and can be found at https://software.broadinstitute.org/cancer/cga/polysolver. NetMHCpan-3.0 is a neural network-based machine learning algorithm that predicts peptide binding and identifies MHC ligands. The method is available at www.cbs.dtu.dk/services/NetMHCpan-3.0. All statistical analyses were conducted in the R statistical software environment (v.3.3.1). Additional information about statistical tests performed in R can be found in the Methods section of the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Nature Biotechnology: doi:10.1038/nbt.4239

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For this analysis, data from two published cohorts of melanoma samples treated with immune checkpoint blockade therapy were analyzed. As clinically-annotated RNA-Seq immunooncology data is sparsely available, we used as many publically-available cohorts as we could find, without compromising our method (i.e., we did not include the Van Allen et al. Science 2015 cohort because these samples were FFPE and our methods are optimized for fresh frozen tissue). In addition, we restricted ourselves to melanoma only (as opposed to looking across cancer types) for this analysis, as we did not want tissue-specific expression profiles to confound our analysis.
Data exclusions	No data from either of the patient cohorts were excluded from this analysis. All samples were of adequate quality to be included in the final analysis.
Replication	Mass spectrometry immunopeptidome experiments were repeated five times with independent measurements for cell line MeWo. Neoepitope EVYAAGKYVSF had one peptide-to-spectrum match (PSM) and was identified in one replicate within 1% false discovery rate (FDR). Experiments were repeated four times with independent measurements for cell line SK-MEL-5. Neoepitope AMSDVSHPK had five peptide-to-spectrum matches (PSMs) and was identified in all four replicates within 1% false discovery rate (FDR). Experiments were repeated four times with independent measurements for CA46. Neoepitope FRYVAQAGL had two PSMs and was identified in two replicates within 1% FDR. Experiments were repeated three times with independent measurements for DOHH-2. Neoepitope ELTLFLLSL had one PSM and was identified in one replicate within 1% FDR. Experiments were repeated four times with independent measurements for HL-60. Neoepitope SVLDDVRGW had one PSM and was identified in one replicate within 1% FDR. Experiments were repeated three times with independent measurements for THP-1. Neoepitope FALTSQGKSAF had five PSMs and was identified in all three replicates within 1% FDR. Lack of peptide identification in all independent replicate experiments is most likely due to low sensitivity of mass spectrometry technology for identification of HLA-bound peptides.
Randomization	Our patient cohorts were designated based on the two publications from which we obtained our samples. These publications describe specific patient stratification criteria, which are based on clinical characteristics that qualify patients to receive immune checkpoint blockade therapy for treatment. Patients were not selected or excluded from our analysis based on their therapy response status.
Blinding	For our study, blinding is not relevant.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Nature Biotechnology: doi:10.1038/nbt.4239

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	RNA-Seq data for all cell lines used was obtained from the Broad Institute's Cancer Cell Line Encyclopedia. Immunopeptidome data was obtained from Ritz et al Proteomics 2016.
Authentication	Cancer Cell Line Encyclopedia (CCLE) cell line authentication procedures, including SNP fingerprint matching and transcriptional profiling, are described in Barretina et al Nature 2012.
Mycoplasma contamination	Mycoplasma contamination was performed globally across the CCLE via the following procedure: 1. Try to re-acquire the cell line, and if not possible, try to decontaminate 2. If 1. successful, use to generate data, and if previous data generated, discard. 3. If 1. not successful, discard cell line. If data already generated, do not publish.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this analysis.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Clinical information about the published patient cohorts used in this analysis is available in Snyder et al. NEJM 2014 (n = 21 patients used herein) and Hugo et al. Cell 2016 (n = 27 patients). All patients had melanoma and were treated with an immune checkpoint blockade agent (anti-CTLA-4 in the case of the Snyder cohort, anti-PD-1 for Hugo cohort). Patients were classified as either deriving clinical benefit from immunotherapy (n = 8 Snyder cohort patients; n = 14 Hugo cohort patients) or no clinical benefit (n = 13 Snyder; n = 13 Hugo).
Recruitment	Participants were not recruited for this study.

Nature Biotechnology: doi:10.1038/nbt.4239