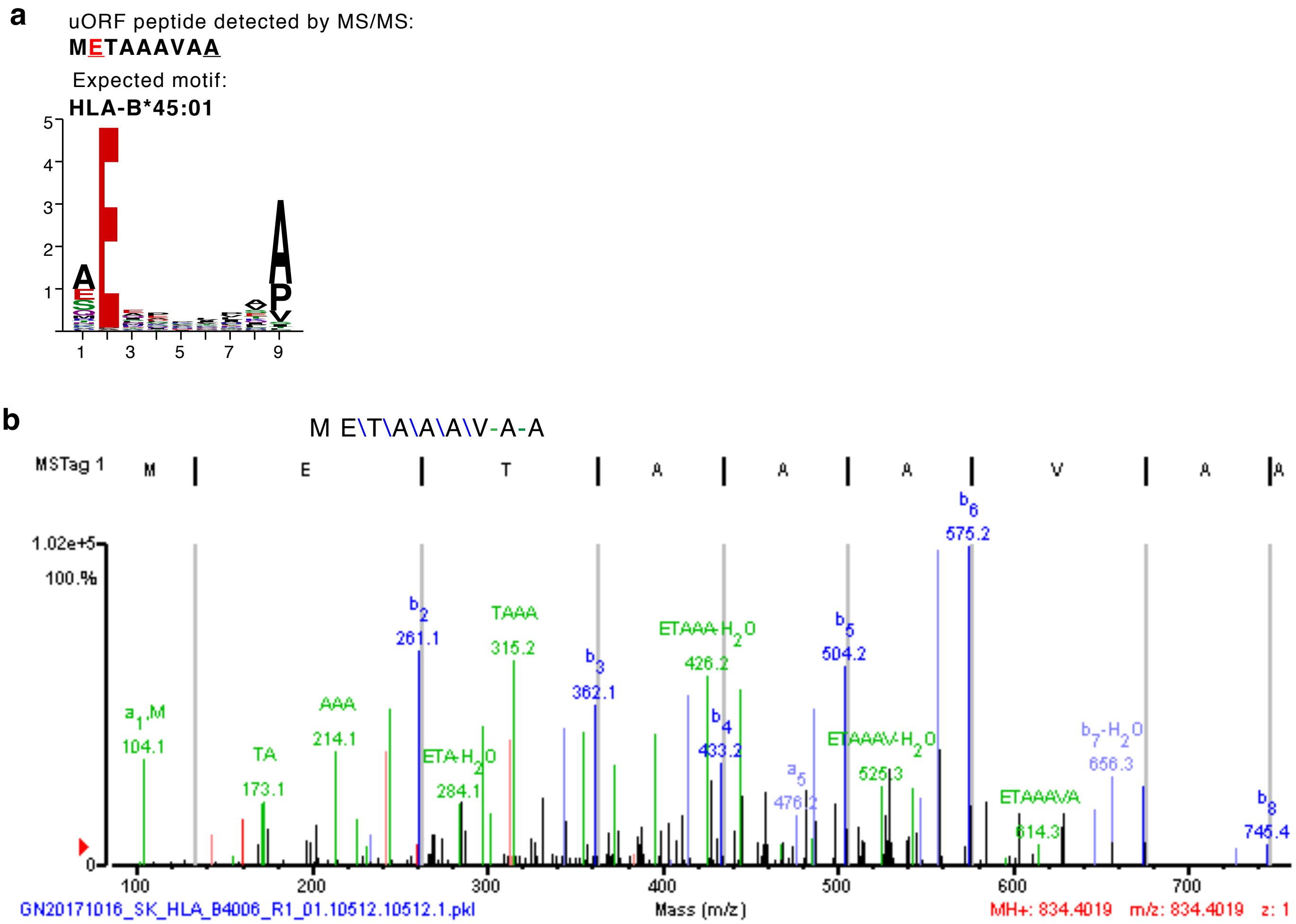
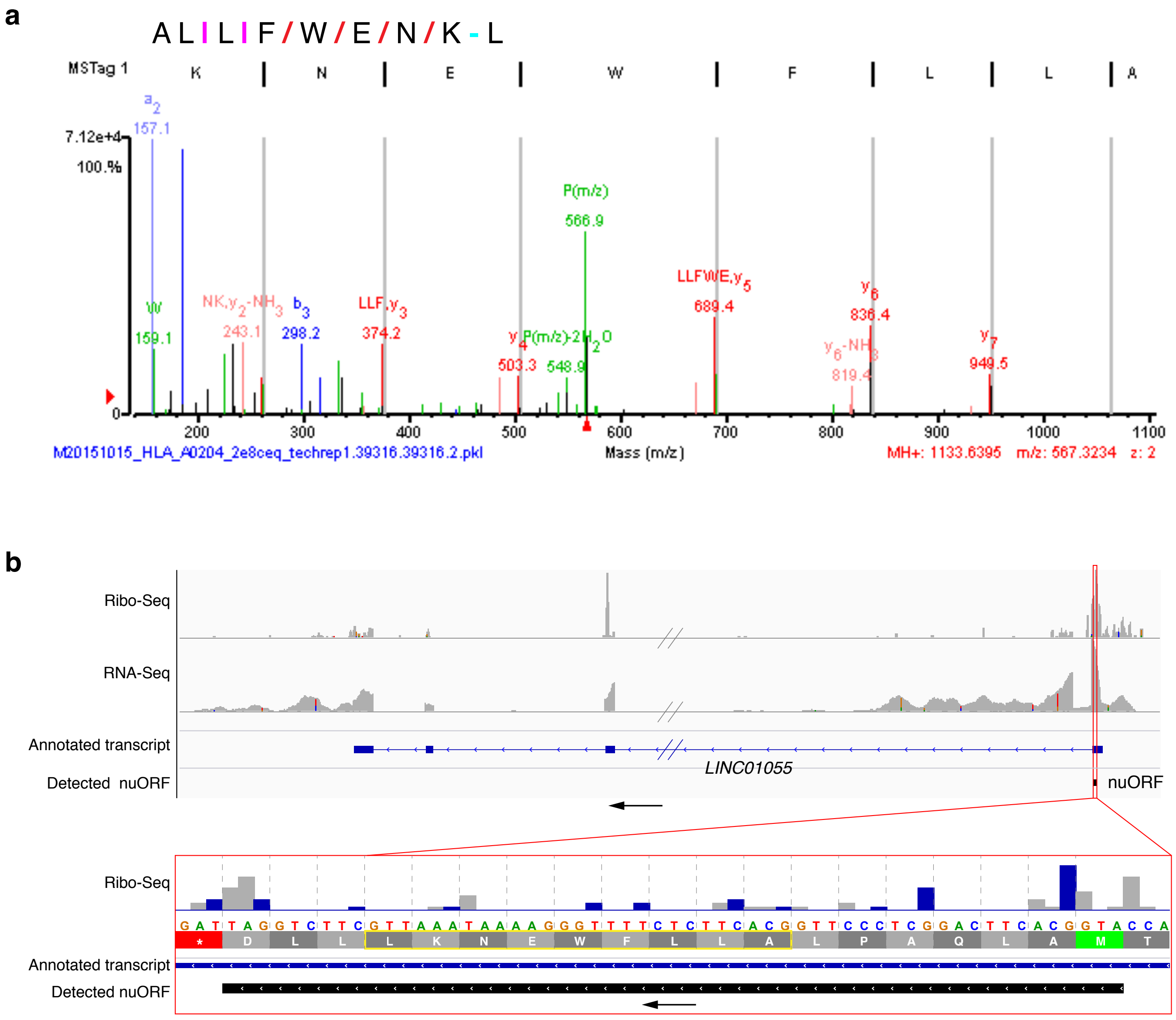
Supp. Fig. 1

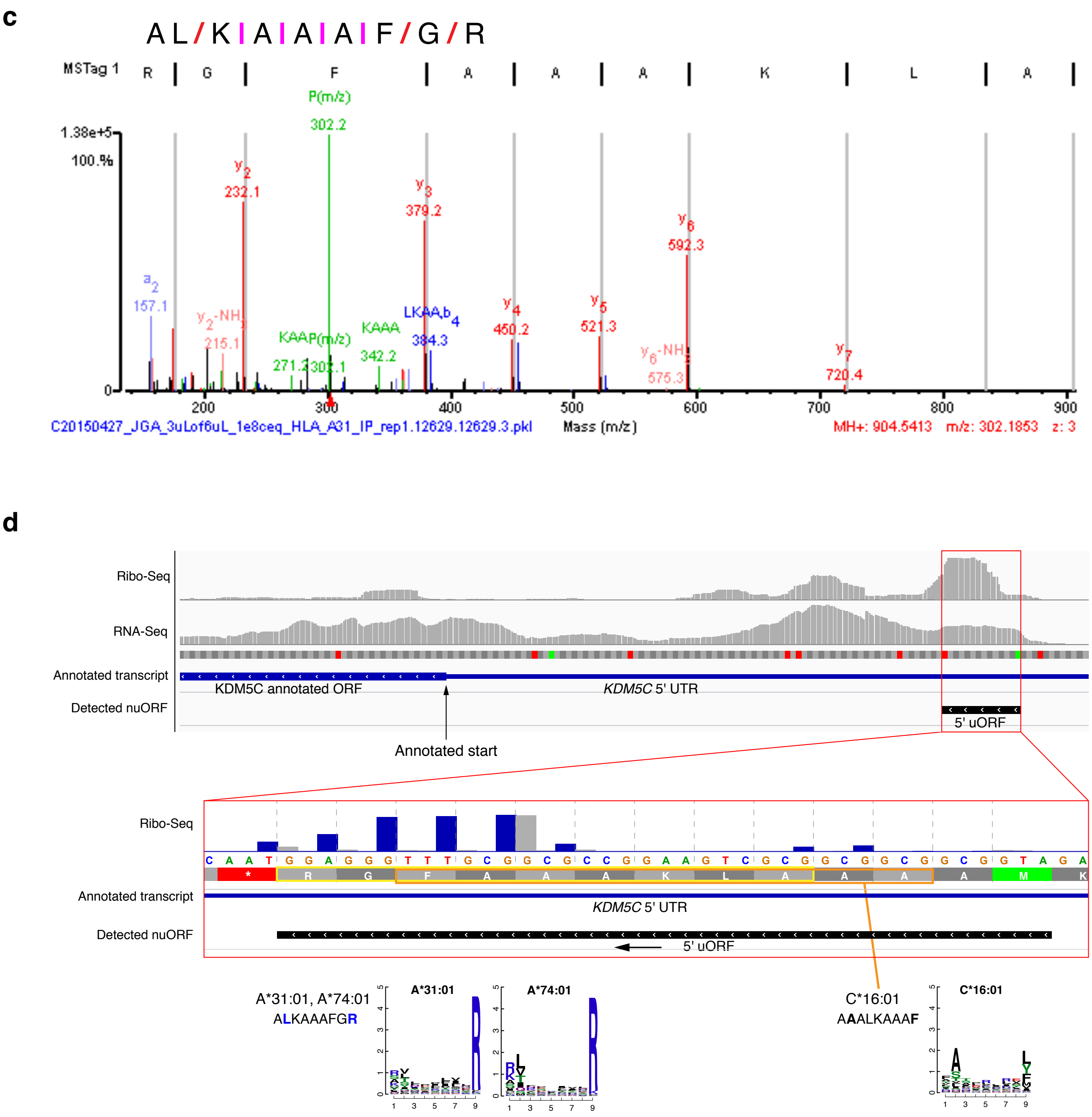


Sup. Fig. 1. Short nuORFs are presented on MHC I without post-translational protease processing. a. The peptide comprising the full-length sequence of the 5' uORF from the *ARAF* gene and the expected binding motif for the HLA allele B*45:01, from which it was presented and detected. **b.** LC-MS/MS spectrum of the *ARAF* 5' uORF.

Supp. Fig. 2


a
A L I L I F / W / E / N / K - L
MSTag 1
K N E W F L L A
a2 157.1
7.12e+4
100 %
W 159.1
NK,y2-NH3 243.1
b3 298.2
LLF,y3 374.2
y4 503.3
P(m/z)-2H2O 548.9
P(m/z) 566.9
LLFWE,y5 689.4
y6-NH3 819.4
y6 836.4
y7 949.5
0
200 300 400 500 600 700 800 900 1000 1100
M20151015_HLA_A0204_2e8ceq_techrep1.39316.39316.2.pkl
Mass (m/z)
MH+: 1133.6395 m/z: 567.3234 z: 2
b
Ribo-Seq
RNA-Seq
Annotated transcript
Detected nuORF
LINC01055
nuORF
Ribo-Seq
G A T T A G G T C T T C G T T A A A T A A A A G G G T T T T C T C T T C A C G G T T C C C T C G G A C T T C A C G G T A C C A
* D L L L K N E W F L L A L P A Q L A M T
Annotated transcript
Detected nuORF

Supp. Fig. 2


c
A L / K I A I A I A I F / G / R
MSTag 1
R G F A A A K L A
P(m/z)
302.2
1.38e+5
100.%
y2 232.1
y3 379.2
y6 592.3
a2 157.1
y2-NH3 215.1
KAAP(m/z)
271.2 303.1
KAAA 342.2
LKAA,b4 384.3
y4 450.2
y5 521.3
y6-NH3 575.3
y7 720.4
0
200 300 400 500 600 700 800 900
C20150427_JGA_3uLof6uL_1e8ceq_HLA_A31_IP_rep1.12629.12629.3.pkl
Mass (m/z)
MH+: 904.5413 m/z: 302.1853 z: 3
d
Ribo-Seq
RNA-Seq
Annotated transcript
KDM5C annotated ORF
KDM5C 5' UTR
Detected nuORF
5' uORF
Annotated start
Ribo-Seq
C A A T G G A G G G T T T G C G G C G C C G G A A G T C G C G G C G G C G G C G G T A G A
* R G F A A A K L A A A A M K
Annotated transcript
KDM5C 5' UTR
Detected nuORF
5' uORF
A*31:01, A*74:01
ALKAAAFGR
A*31:01
A*74:01
C*16:01
AAALKAAAF
C*16:01
0 1 2 3 4 5
1 3 5 7 9

Sup. Fig. 2. Some nuORF-derived peptides map to the same MS/MS spectra as peptides that have previously been proposed to be derived from proteasomal splicing in Faridi et al.

We found 308 nuORF-derived peptides, supported by Ribo-seq, that map to the same MS/MS spectra as 343 peptides that have previously been proposed to be derived from proteasomal splicing in Faridi et al. in either of two scenarios: (1) for 98 cases, the nuORF-derived peptide sequence is identical to a proposed spliced peptide; or (2) for 210 cases, the partial sequence present in the MS/MS spectrum matched to a nuORF-derived peptide is also consistent with one or more different, yet similar, spliced peptide sequences (Supplementary Table 5). **a.** MS/MS spectrum of the peptide ALLFWENKL presented by HLA allele A*02:04 that can be derived from the translated *LINC01055* lncRNA nuORF was previously proposed to be derived from proteasomal cis-splicing. The peptide contains L at both position 2 and the C-terminus, consistent with the anchor motif for allele A*02:04 ligands. **b.** RNA-seq and Ribo-seq reads aligned to the *LINC01055* lncRNA locus. Red box marks the MHC I IP LC-MS/MS detected nuORF. Bottom panel shows a magnified view of the reads supporting nuORF translation. Detected peptide is outlined in yellow. **c,d.** Partial sequence present in the MS/MS spectra assigned to spliced peptides are also consistent with different, yet similar, nuORF peptide sequences. **c.** MS/MS spectrum of a peptide presented by allele A*31:01 yields near complete fragmentation with explicit sequence evidence for the order of all residues except the first two: AL/KIAIAIAIF/G/R. Both peptides ALKAAAFGR (derived from *KDM5C* 5' uORF) and LAKAAAFGR (proposed to be derived from proteasomal cis-splicing) are consistent with the fragment ions present in the spectrum. Leucine in position 2 (nuORF) and arginine at the C-terminus are consistent with the anchor motif for allele A*31:01 ligands, while alanine in position 2 (spliced) is not. In addition, 3 more peptides from *KDM5C* 5' uORF were detected on HLA A*74:01 and C*16:01, further supporting the nuORF translation and presentation (Supplementary Table 6). **d.** RNA-seq and Ribo-seq reads aligned to the *KDM5C* locus. Red box marks the 5' uORF detected by MHC I IP LC-MS/MS. Detected peptides and the expected peptide sequence motifs are outlined in yellow and orange.

Supplementary Information

Supplementary Note 1

Improved MHC I immunopeptidomics mapping with NuORFdb compared to RNA-seq based references

We benchmarked the performance of nuORFdb as a reference for MHC I immunopeptidome MS/MS spectra mapping compared to UCSCdb, RNAdb and TransDb, using data from 6 representative HLA alleles expressed in B721.221 mono-allelic cells[10]. Compared to nuORFdb, searching spectra against RNAdb and TransDb took 2.3- and 3.1-fold longer (**Extended Data Fig. 2a**), resulted in increased detection of nuORF peptides at the expense of respectively 15% and 20% decrease in the number of spectra mapped to annotated peptides (**Extended Data Fig. 2b,c**), and had higher FDR rates of 8.1 and 8.7%, respectively *vs*. 5.5% for nuORFdb at a 1% global set FDR (**Extended Data Fig. 2d**). Among nuORF peptides identified on 6 alleles, 152 peptides were identified across all three nuORF-containing databases. TransDb contained the most unique nuORF peptides (154), not identified using the other databases (**Extended Data Fig. 2e**). Most RNAdb- and TransDb-specific nuORFs were derived from lncRNAs (**Extended Data Fig. 2f**), consistent with their high representation of lncRNA nuORFs (**Fig. 1f**). Among TransDb-specific peptides, only 12% were derived from nuORFs supported by RNA-seq or Ribo-seq; while only 25% of RNAdb-specific peptides were supported by Ribo-seq reads, suggesting that the spectra were misassigned (**Extended Data Fig. 2g**). Misassignment of spectra when using larger search databases was much more likely: the median number of candidate peptides per spectrum using TransDb was over 10^5, nearly 10-fold higher than using UCSCdb or nuORFdb (**Extended Data Fig. 2h**), and the median identification score difference between the top ranked peptide sequence and the next best was the smallest for the TransDb and RNAdb-specific peptides (**Extended Data Fig. 2i**). Moreover, only 73% and 62% of peptides identified from RNAdb and TransDb were predicted by the MHC I predictor *HLAthena*[10] to bind to MHC I alleles as compared to 96% and 85% for annotated and shared nuORF peptides, respectively (**Extended Data Fig. 2j**). Finally, while chromatographic retention times for shared nuORF peptides correlated as well with predicted hydrophobicity indices as they did for annotated peptides[24,25], the retention times for nuORF peptides detected uniquely in specific databases frequently did not agree with their predicted hydrophobicity, suggesting that the peptide amino acid composition is incorrectly identified (**Extended Data Fig. 2k**). As a result, when searching larger databases, the sensitivity of peptide identification decreases and maintaining confident peptide spectrum matching must rely on higher quality MS/MS spectra with more complete peptide sequence coverage with enough specificity to discriminate amongst the peptides in the larger search space.

Supplementary Note 2

HLA-peptide immunoprecipitation with cysteine alkylation

Because neither free sulfhydryl cysteine nor disulfide cross-linked cysteines tend to withstand sample handling and chromatographic separation steps to yield readily identifiable MS/MS spectra, it is common practice in proteomics experiments to employ reduction and alkylation at room temperature with reagents like dithiothreitol (DTT) and iodoacetamide (IAA). For detection of HLA-presented peptides, addition of a low amount of IAA (10 mM) not only inhibits cysteine proteases, but also increases (~5x) the LC-MS/MS detection of peptide sequences containing a cysteine residue[9,10,14]. Prior reduction by adding DTT during IP is impractical as it would dissociate the antibody. Adding DTT afterward to the eluted peptides would not diminish possible artefacts arising from off-site alkylation (peptide N-termini), by tying up residual IAA, because the residual IAA would have already been washed away before eluting peptides from the beads. However, the low temperature (4°C) of the incubation inhibits off-site reactions. While alkylation could decrease the binding affinity of certain peptides, we observed that cysteine residues were detected only in regions of the peptides not tucked away in the HLA binding pocket, e.g. residue 3-8[10]. In the absence of alkylation, some cysteine containing peptides could be recovered once the search accounted for cysteinylation[9]. We processed the allele C*08:02 both with and without IAA. While we recovered carbamidomethylated peptides upon addition of IAA, we also observed cysteinylated peptides in both cases[10], suggesting that cysteinylation is a posttranslational modification and occurs at some cysteines but not all.

Supplementary Methods

Cell cultures

A375 cells were cultured in DMEM media (Gibco), supplemented with 5% fetal bovine serum (FBS). HCT116 cells were cultured in McCoy's 5A Medium (Thermo Fisher Scientific), supplemented with 5% FBS.

Generation of HLA mono-allelic B721.221 cells

The HLA mono-allelic cell lines were generated as previously described[9,10]. Briefly, single HLA allele-expressing cDNA vectors in a pcDNA-3 backbone were ordered from GenScript™. The HLA class I deficient B721.221 cell line was transfected with the HLA allele expression vectors using lipofectamine, as described previously[9]. Cell lines with stable surface HLA expression were generated first through selection using 800μg/ml G418 (Thermo Fisher Scientific), followed by enrichment of HLA positive cells through up to 2 serial rounds of fluorescence-activated cell sorting (FACS) and isolation using a pan-HLA antibody (W6/32; Santa Cruz) on a FACSAria II instrument (BD Biosciences).

Primary human cells and generation of cancer cell lines

All human tissues were obtained following informed consent through DFCI or Partners Healthcare approved IRB protocols. Conditions for growth and *in vitro* propagation of melanoma and GBM tumor cell lines were described previously[3,4]. PBMCs from fresh healthy donor whole blood were isolated using Ficoll density gradient medium. $CD19^+$ B cells were isolated using EasySep Human CD19 Positive Selection Kit, obtaining between 25 and 54 million B cells per donor. For fresh CLL samples, PBMCs were isolated using Ficoll density gradient medium, enriched for CD19 positive CLL tumor cells and were used in IP/MS analysis and Ribo-seq. For cryopreserved CLL samples, live cells were isolated with an OptiPrep density gradient medium. Surgically resected clear cell renal cell carcinoma (ccRCC) tissue was mechanically dissociated with scalpels, and then enzymatically dissociated using a mixture of collagenase D (Roche), Dispase (STEMCELL Technologies), and DNase I (New England BioLabs) at room temperature, and filtered through a 100 micron cell strainer using the sterile plunger of a syringe. Red blood cells were lysed using ammonium-chloride-potassium buffer (Gibco). The cell suspension was stained for viability (Zombie Aqua; BioLegend), anti-CD45 (BV605; BD Biosciences), and anti-carbonic anhydrase IX (PE; R&D Systems). Viable, CD45$^-$, CAIX$^+$ tumor cells were isolated by FACS (BD FACSAria II cell sorter; BD Biosciences). Cells were cultured in a specialized growth medium consisting of OptiMEM GlutaMax media (Gibco), 5% fetal bovine serum, 1mM sodium pyruvate (Gibco), 100 units/mL penicillin and streptomycin, 50 micrograms/mL gentamicin, 5 micrograms/mL insulin (Sigma), and 5 ng/mL epidermal growth factor (Sigma). Following successive passages, CAIX expression was confirmed by flow cytometry (anti-CAIX, PE-conjugated; R&D Systems) and by immunohistochemical analysis of a cell pellet.

Ovarian cancer patient-derived cells were propagated within a xenograft model, which was generated by serial passaging of tumor cells from a patient with advanced ovarian cancer. These cells originated from solid tumor that were injected orthotopically in the abdominal cavity in NOD-SCID mice (8-week old, Jackson labs). Tumor growth was monitored weekly by observing mice for signs of abdominal distension. Cells were harvested 4 months after initial injection and cryopreserved.

Primary human epidermal melanocytes (Thermo C0025C) were cultured according to manufacturer's protocol.

Ribosome profiling library preparation

Ribosome profiling was performed according to the manufacturer's protocol (TruSeq Ribo Profile - RPHMR12126, Illumina, discontinued), with the following modifications. For adherent cell lines (melanoma, primary melanocytes, HCT116, A375), culture media was removed, cells were washed with ice-cold PBS containing cycloheximide (0.1mg/ml) and lysed in the Lysis Buffer according to the Illumina protocol. For suspension cell lines and primary blood samples, cells were spun 1,000rpm for 5 minutes, washed once with ice-cold PBS containing cycloheximide (0.1mg/ml) and lysed in the Lysis Buffer. To perform Ribo-seq on small samples, such as primary B cells and melanocytes, cells were lysed in 230 µl of lysis buffer, such that the entire lysate could be used in library preparation. Ribosomes containing ribosome-protected mRNA fragments (RPFs) were enriched using MicroSpin S-400 columns (GE Healthcare, catalog # 27-5140-01). Ribo-zero rRNA Removal Kit (Illumina, MRZH11124, discontinued) was used to deplete rRNA from RPFs. The entire RPF sample was loaded on a 15% urea-polyacrylamide gel. Samples were eluted from the gel overnight at 4°C. Subsequently, end repair, adapter ligation and reverse transcription were carried out according to the manufacturer's protocol. For the cDNA gel purification, the reverse transcription reaction was loaded on a 10% urea-polyacrylamide gel. The samples were eluted from the gel overnight at room temperature. Subsequently, RPFs were circularized and 5 µl of circDNA was used for library amplification. The number of amplification cycles was determined based on the observed sample quality and expected yield, but usually ranged between 8 and 10 cycles. Following amplification, the library was gel-purified using 4% E-Gel EX Agarose Gel (ThermoFisher G401004) and Zymoclean Gel DNA Recovery Kit (Zymo Research D4007), with 4 volumes of ADB buffer to accommodate 4% agarose gel.

RNA-seq library preparation

For CLL4 and CLL5 samples, 20 µl of the lysate pre-RNase I digestion, prepared for the Ribo-seq library preparation, was used to extract total RNA using the Zymo RNA Clean & Concentrator-25 (R1017). Stranded RNA-seq library was prepared using Illumina TruSeq® Stranded mRNA Library Prep kit (20020594) according to manufacturer's instructions. The libraries were sequenced on the Illumina NovaSeq platform, 150 cycles paired-end.

Newly generated and published RNA-seq data processing

Newly generated CLL4 and CLL5 paired-end RNA-seq data, as well as published GSE51424[37] and GSE100007[39] (single-end[37] and paired-end[39]) RNA-seq data was aligned using STAR (v2.7) with the following parameters: --alignIntronMin 20 --alignIntronMax 100000 --outFilterType

BySJout --quantMode TranscriptomeSAM --twopassMode Basic --outSAMtype BAM SortedByCoordinate --limitBAMsortRAM 30000000000. Expression was quantified using RSEM (v1.2.31).

HLA-peptide sequencing by tandem mass spectrometry

Peptides were resuspended in 3% ACN, 5% FA and loaded onto an analytical column (20-30 cm, 1.9 µm C18 Reprosil beads (Dr. Maisch HPLC GmbH), packed in-house PicoFrit 75 µm inner diameter, 10 µm emitter (New Objective)). Peptides were eluted with a linear gradient (EasyNanoLC 1000 or 1200, ThermoFisher Scientific) ranging from 6-30% Buffer B (either 0.1% FA or 0.5% AcOH and 80% or 90% ACN) over 84 min, 30-90% B over 9 min and held at 90% Buffer B for 5 min at 200 nl/min. During data dependent acquisition, peptides were analyzed on a QExactive Plus (QE+, Tune 2.1), QExactive HF (QE-HF, Tune 2.1) or Fusion Lumos (ThermoFisher Scientific, Tune 3.1). Full scan MS was acquired at a resolution of 70,000 (QE+) or 60,000 (QE-HF and Lumos) from 300-1,800 m/z or 300-1,700 m/z (Lumos). AGC target was set to 1e6 and 5 msec max injection time for QE type instruments and 4e5 and 50 ms for Lumos. The top 10 (Lumos, QE+), 12 (QE+), 15 (QE-HF) precursors per cycle were subjected to HCD fragmentation at resolution 17,500 (QE+) or 15,000 (QE-HF, Lumos). The isolation width was set to 1.7 m/z with a 0.3 m/z offset for QE and 1.0 m.z and no offset for Lumos, the collision energy was set to optimal for the instrument used ranging from 25 to 30 NCE, AGC target was 5E4 and max fill time 120 ms (QE+ and Lumos) or 100 ms (QE-HF). For Lumos measurements, precursors of 800-1700 m/z were also subjected to fragmentation if they were singly charged. Dynamic exclusion was enabled with a duration of 15 sec (QE+), 10 secs (QE-HF) or 5 sec (Lumos).

HLA peptide identification using Spectrum Mill

MS/MS spectra were excluded from searching if they did not have a precursor MH+ in the range of 600-4000, had a precursor charge >5, or had a minimum of <5 detected peaks. Merging of similar spectra with the same precursor m/z acquired in the same chromatographic peak was disabled. Prior to searches, all MS/MS spectra had to pass the spectral quality filter with a sequence tag length >2 (*i.e.*, minimum of 4 masses separated by the in-chain masses of 3 amino acids). MS/MS search parameters included: no-enzyme specificity; fixed modification: cysteinylation of cysteine; variable modifications: carbamidomethylation of cysteine, oxidation of methionine, and pyroglutamic acid at peptide N-terminal glutamine; precursor mass tolerance of ±10 ppm; product mass tolerance of ± 10 ppm, and a minimum matched peak intensity of 30%. Variable modification of carbamidomethylation of cysteine was only used for HLA alleles that included an alkylation step (Supplementary Table 7).

Peptide spectrum matches (PSMs) for individual spectra were automatically designated as confidently assigned using the Spectrum Mill auto-validation module to apply target-decoy based FDR estimation at the PSM level of <1% FDR. Peptide auto-validation was done separately for each HLA allele with an auto thresholds strategy to optimize score and delta Rank1 – Rank2 score thresholds separately for each precursor charge state (1 through 4) across all LC-MS/MS runs for an HLA allele. Score threshold determination also required that peptides had a minimum sequence length of 7, and PSMs had a minimum backbone cleavage score (BCS) of 5. BCS is a peptide sequence coverage metric and the BCS threshold enforces a uniformly higher minimum sequence coverage for each PSM, at least 4 or 5 residues of unambiguous sequence. The BCS score is a sum after assigning a 1 or 0 between each pair of adjacent AA's in the sequence (max score is peptide length-1). To receive a score, cleavage of the peptide backbone must be supported by the presence of a primary ion type for HCD: b, y, or internal ion C-terminus (*i.e.*, if the internal ion is for the sequence PWN then BCS is credited only for the backbone bond after the N). The BCS metric serves to decrease false-positives associated with spectra having fragmentation in a limited portion of the peptide that yields multiple ion types.

Benchmarking of search databases

We selected 6 representative HLA alleles expressed in B721.221 mono-allelic cells: A*02:01, A*24:02, B*15:01, B*44:02, C*05:01, and C*07:02. When benchmarking searches against the 4 databases (UCSCdb, nuORFdb, RNAdb and TransDb (GENCODE[22] and MiTranscriptome[23])), we sought to track the number of unique candidate sequences tested for each peptide spectrum match (**Extended Data Fig. 2h**) to characterize the size of the search space for each database. While the default Spectrum Mill search strategy does this, for a database as large as TransDb it would require an impractically large amount of memory. We therefore made several changes to Spectrum Mill operation to achieve that goal using a full-length N-mer strategy (FLmers). The 4 protein sequence databases were converted into 4 peptide databases, with each peptide entry containing a single unique peptide of length N. N-mer subsequences derived from each protein sequence ranged from 8-10 amino acids long with a single copy of each unique peptide sequence retained for each database. For Spectrum Mill searches, the digest parameter was changed from no enzyme to full length (FL), which eliminates the step of making peptide subsequences from each protein, when matching to each MS/MS spectrum's precursor m/z. All other search parameters and FDR filtering were as described above.

Peptide hydrophobicity index calculation

Hydrophobicity index was predicted using SSRCalc[56], http://hs2.proteome.ca/SSRCalc/SSRCalcQ.html). Modification of cysteine was checked for alleles B*56:01 and A*74:01. For A*02:01 and C*03:04 free cysteine was specified.

Whole proteome analysis and interpretation

Protein expression of the B721.221 and GBM H4152-BT145 cell lines was assessed as described previously[57]. Briefly, cell pellets of B721.221 cells expressing A*03:01, B*55:01 and C*07:01, as well as pellets of GBM6 with and without IFNγ treatment were lysed in 8M Urea and digested to peptides using LysC and Trypsin (Promega). B721 analysis was performed label free with a 1:1:1 mix using 100 µg each of the three monoallelic cell lines. For GBM, 100 µg peptides were labeled with TMT6 reagents (Thermo Fisher) 126 (untreated) and 127 (IFNγ) and then pooled for subsequent fractionation and analysis. Pooled peptides were separated into 24 fractions using offline high pH reversed phase fractionation. 1 µg per fraction was loaded onto an analytical column (20-30 cm, 1.9 µm C18 Reprosil beads (Dr. Maisch HPLC GmbH), packed in-house PicoFrit 75 µM inner diameter, 10 µM emitter (New Objective)). Peptides were eluted with a linear gradient (EasyNanoLC 1000 or 1200, Thermo Scientific) ranging from 6-30% Buffer B (either

0.1%FA or 0.5% AcOH and 80% or 90% ACN) over 84 min, 30-90% B over 9 min and held at 90% Buffer B for 5 min at 200 nl/min. During data dependent acquisition, peptides were analyzed on a Fusion Lumos (Thermo Scientific). Full scan MS was acquired at a 60,000 from 300 - 1,800 m/z. AGC target was set to 4e5 and 50 ms. The top 20 precursors per cycle were subjected to HCD fragmentation at 15,000 resolution with an isolation width of 0.7 m/z, 30 NCE, 3e4 AGC target and 50ms max injection time. For TMT experiments, resolution was set to 60,000 and 34 NCE. Dynamic exclusion was enabled with a duration of 45 sec.

Spectra were searched using Spectrum Mill against the same database as the one used for the MHC I IP/MS spectra analysis (described above), specifying Trypsin/allow P (allows K-P and R-P cleavage) as digestion enzyme, allowing 4 missed cleavages. Carbamidomethylation of cysteine was set as a fixed modification. For the GBM dataset TMT labeling was required at lysine, but peptide N-termini were allowed to be either labeled or unlabeled. Allowed variable modifications were acetylation at the protein N-terminus, oxidized methionine, pyroglutamic acid, deamidated asparagine and pyrocarbamidomethyl cysteine. Match tolerances were set to 20 ppm on MS1 and MS2 level. PSMs score thresholding used the Spectrum Mill auto-validation module to apply target-decoy based FDR in 2 steps: at the peptide spectrum match (PSM) level and the protein level. In step 1 PSM-level autovalidation was done first using an auto-thresholds strategy with a minimum sequence length of 8; automatic variable range precursor mass filtering; and score and delta Rank1 – Rank2 score thresholds optimized to yield a PSM-level FDR estimate for precursor charges 2 through 4 of <1.0% for each precursor charge state in each LC-MS/MS run. To achieve reasonable statistics for precursor charges 5-6, thresholds were optimized to yield a PSM-level FDR estimate of <0.5% across all LC runs per experiment (instead of per each run), since many fewer spectra are generated for the higher charge states. In step 2, protein-polishing autovalidation was applied to each experiment to further filter the PSMs using a target protein-level FDR threshold of zero, the protein grouping method expand subgroups, top uses shared (SGT) with an absolute minimum protein score of 9. After assembling protein groups from the autovalidated PSMs, protein polishing determined the maximum protein level score of a protein subgroup that consisted entirely of distinct peptides estimated to be false-positive identifications (PSMs with negative delta forward-reverse scores); B721: 11.6, GBM: 10.5. PSMs were removed from the set obtained in the initial peptide-level autovalidation step if they contributed to protein subgroups that had protein scores below the maximum false-positive protein score. In Spectrum Mill the protein score was the sum of the scores of distinct peptides. When a peptide sequence of >8 residues was shared by multiple protein entries in the sequence database, the proteins were grouped together. In some cases there were unshared peptides that uniquely represent a subgroup, i.e. lower scoring member of the group, typically isoforms, family members, or different species. As a consequence of these two peptide and protein level steps, each identified protein subgroup was comprised of multiple peptides, unless a single excellent scoring peptide was the sole match.

In the cases where a spectrum could be matched to multiple peptide sequences from different ORFs, the same decision tree was followed for the whole proteome analysis as for the MHC I described above.

Peptide sequence correlation, clustering and visualization

Peptide distance computation and visualization were performed as before[9]. Briefly, peptide distances were defined as:

$$D(A, s_1, s_2) = \frac{1}{L} \sum_{i=1}^{L} distPMBEC(s_{1i}, s_{2i}) * (1 - H_{Ai})$$

where A is the allele; $s_1$and $s_2$are peptide sequences; L is the length of the peptide sequences, $n \in \{8,9,10,11\}$; H is the entropy of the amino acid residues at each position in the peptide, $distPMBEC = maxPMBEC - PMBEC$ is a 20x20 matrix of residue dissimilarities derived from a pre-computed matrix of residue similarities biased by their HLA binding properties[58]. For each allele, peptide distances between every pair of peptides in the MS datasets was computed and the pairwise distance matrices were reduced to two dimensions with non-metric multidimensional scaling (NMDS) (nmds() function from ecodist R package).

Peptide sequence motif correlation (for **Figure 2F**) was calculated per allele using all detected 9AA peptides. For each peptide, the frequency of each amino acid at each position was calculated to generate a vector of 180 features long. Using these vectors, the position entropy weighted correlation was found between nuORF peptides and all annotated peptides, or between 10,000 random subsets of annotated peptides the same size as the nuORF set and all annotated peptides (minus the subset). Correlations were calculated for all 92 measured alleles independently.

Whole genome sequencing and analysis

PCR-free Whole Genome Sequencing (WGS) was performed on cultured melanoma patient 11 cells and matched healthy PBMCs at the Broad Genomics Platform. Libraries were prepared using the Kapa Biosciences HyperPrep library construction kit, and sequenced to 60x coverage (Illumina 2x150bp reads, NovaSeq). Cancer-specific variants were identified using GATK Best Practices (GATK v3.x nightly-2017-09-30)[59] and Strelka2 v2.8.4[60]. In particular, we first aligned sequenced reads to human genome reference assembly GRCh37 using BWA-MEM[61] v0.7.15-r1140 with default parameters. We then sort aligned reads by coordinates and removed PCR duplicates using Picard tool v2.12.1 [http://broadinstitute.github.io/picard/]. Next, we applied base quality score recalibration to the de-duplicated BAM files using GATK. The recalibrated BAM files were used as inputs for both GATK and Strelka2 for calling somatic variants. For GATK, we followed best practices and used MuTect2 with --dbsnp set to dbSNP build 138 [https://www.ncbi.nlm.nih.gov/snp/] and --cosmic set to Cosmic v82 [https://cosmic-blog.sanger.ac.uk/cosmic-release-v82/]. For Strelka2, we first ran Manta[62] v1.2.1 to detect structural variants and indels as recommended by Strelka2 user guide [https://github.com/Illumina/strelka/blob/v2.9.x/docs/userGuide/README.md]. We then ran Strelka2 with --indelCandidates option set to Manta outputs and other options set to default values. We merged variants called using GATK and Strelka2 together.

Supplementary References

56. Krokhin, O. V. *et al.* b. New sequence-specific correction factors for prediction of peptide retention in RP--HPLC: application to protein identification by off-line HPLC-MALDI-MS. in *Proceedings of the 52-th ASMS Conference on Mass Spectrometry and Allied Topics. Nashville, TN, USA* (2004).

57. Mertins, P. et al. Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography-mass spectrometry. Nat. Protoc. 13, 1632–1661 (2018).

58. Kim, Y., Sidney, J., Pinilla, C., Sette, A. & Peters, B. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. BMC Bioinformatics 10, 394 (2009).

59. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. in Current Protocols in Bioinformatics (eds. Bateman, A., Pearson, W. R., Stein, L. D., Stormo, G. D. & Yates, J. R., III) vol. 467 11.10.1–11.10.33 (John Wiley & Sons, Inc., 2002).

60. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods 15, 591–594 (2018).

61. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN] (2013).

62. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32, 1220–1222 (2016).