Supporting Information

Supplementary Tables

Supplementary Table S1:List of experiments in the non-HLA build and the protease used.

Supplementary Table S2: Listing of peptides that are mapped to Ribo-seq ncORFs from the Human non-HLA PeptideAtlas 2023-06 build

Supplementary Table S3: Listing of Ribo-seq ncORFs annotated in the Human non-HLA PeptideAtlas 2023-06 build

Supplementary Table S4: Listing of peptides that are mapped to Ribo-seq ncORFs from the Human HLA PeptideAtlas 2023-11 build

Supplementary Table S5: Listing of detected Ribo-seq ncORFs from the Human HLA PeptideAtlas 2023-11 build

Supplementary Table S6: List of HLA build MS runs, including the HLA type of each MS run.

Supplementary Table S7: List of 677 HLA-I peptides, including their sequence, best allele, the 22 features that the model used for training, and the output probabilities from the model.

Supplementary Table S8: List of 7,264 ncORFs along with the features that were used to train machine learning models and output probabilities of the model.

Supplementary Table S9: FDR metrics for the non-HLA build analysis

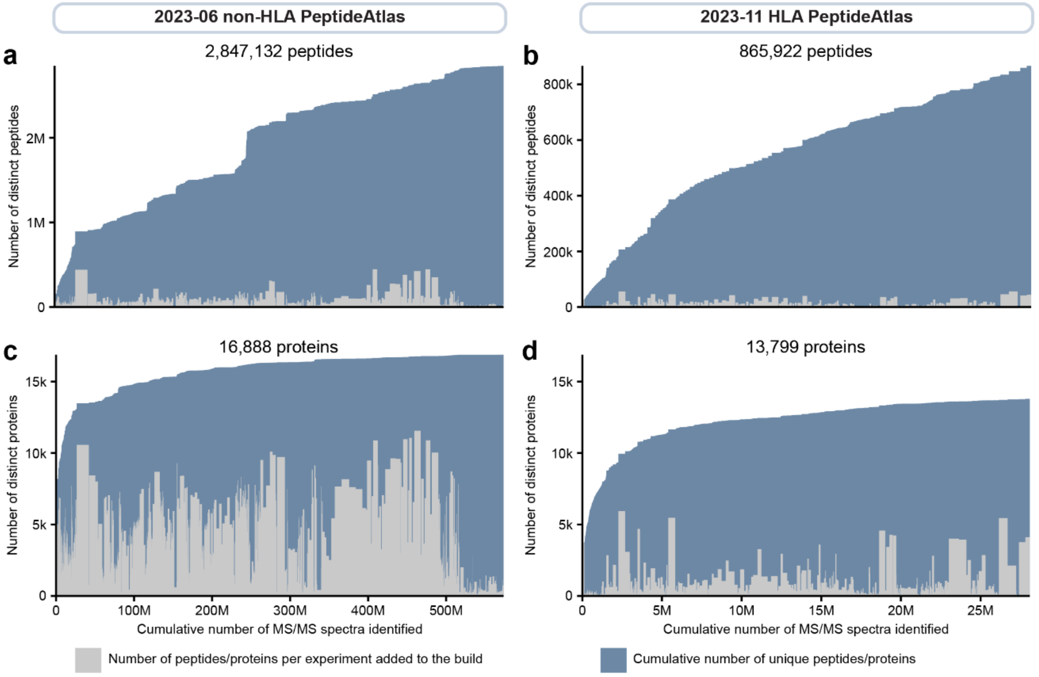
Supplementary Table S10: FDR metrics for the HLA build analysis

Supplementary Documents

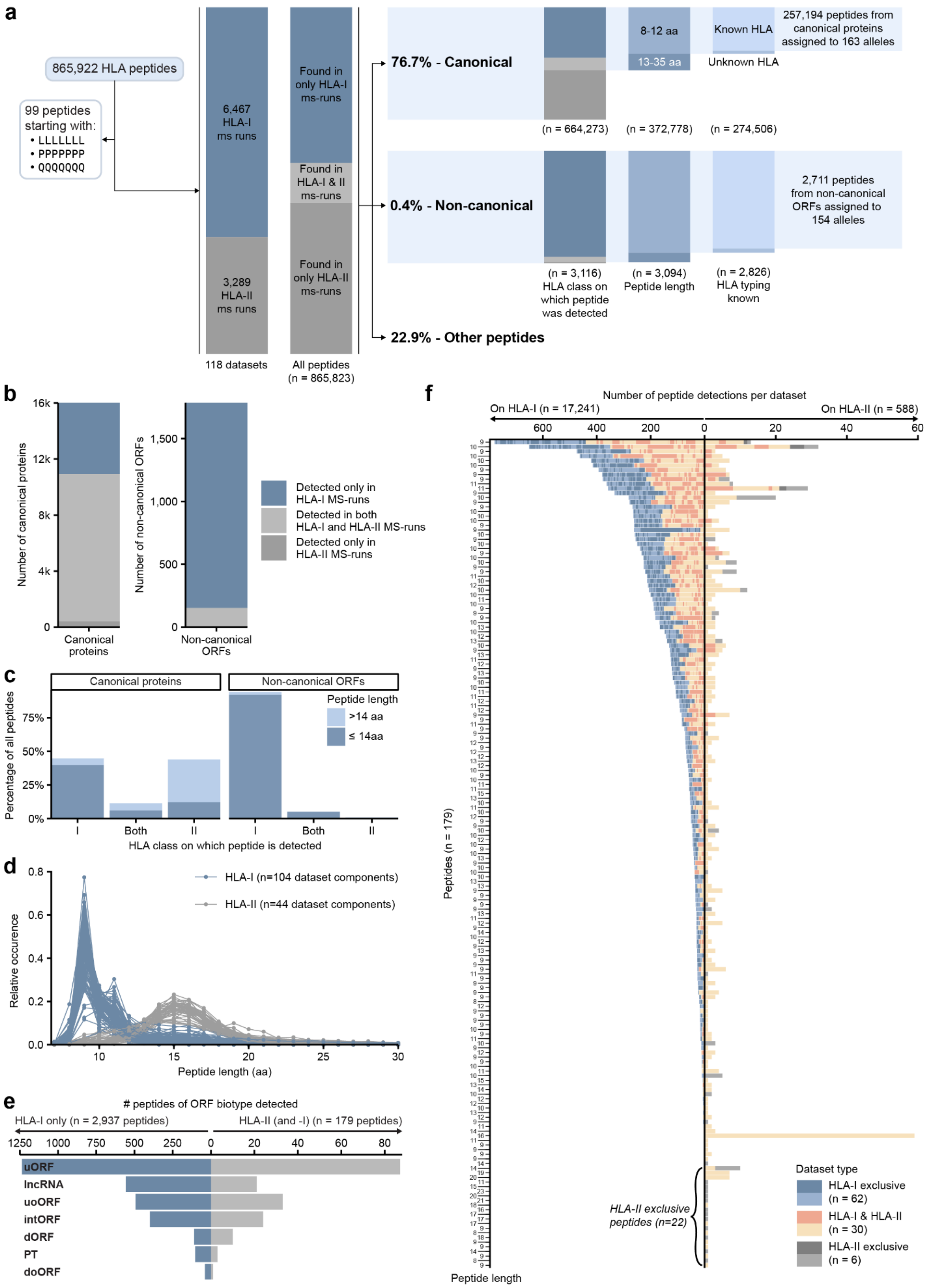
Supplementary Document S1 - Discussion of ncORF detections in non-HLA data

Supplementary Document S2 - Discussion of machine learning results predicting detectability of ncORFs.

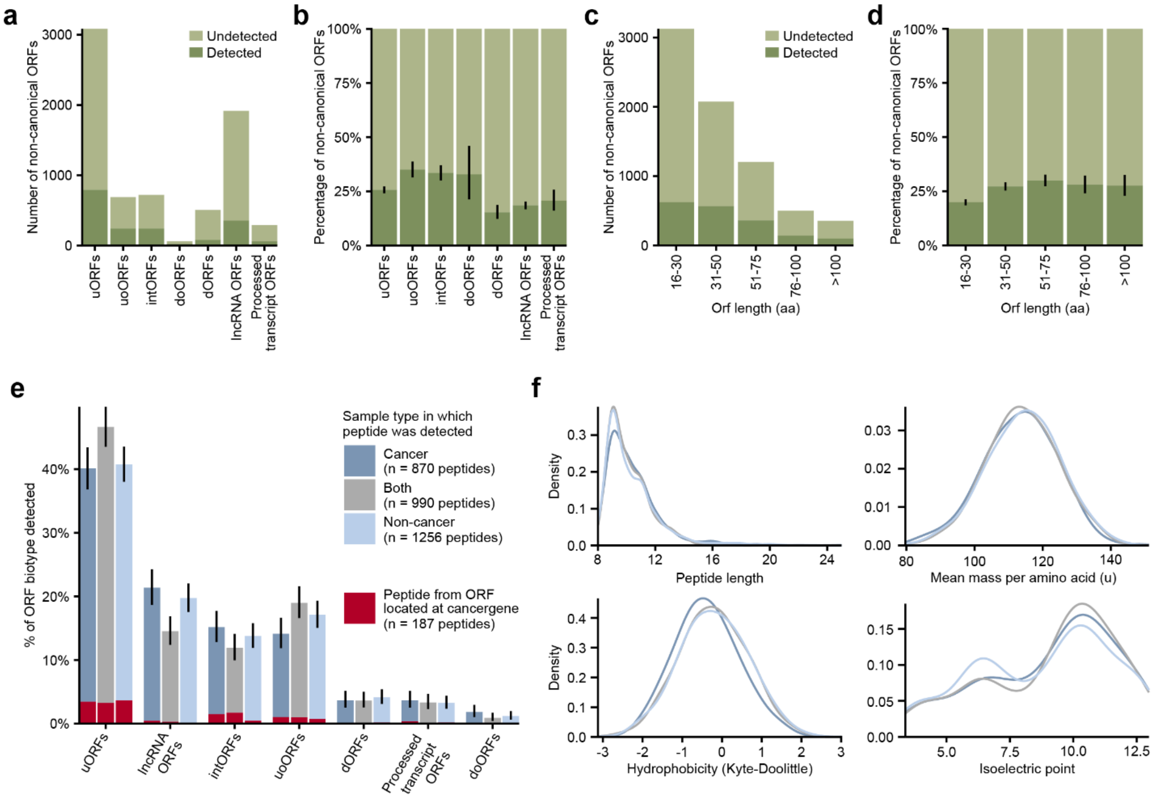
Supplementary Figures



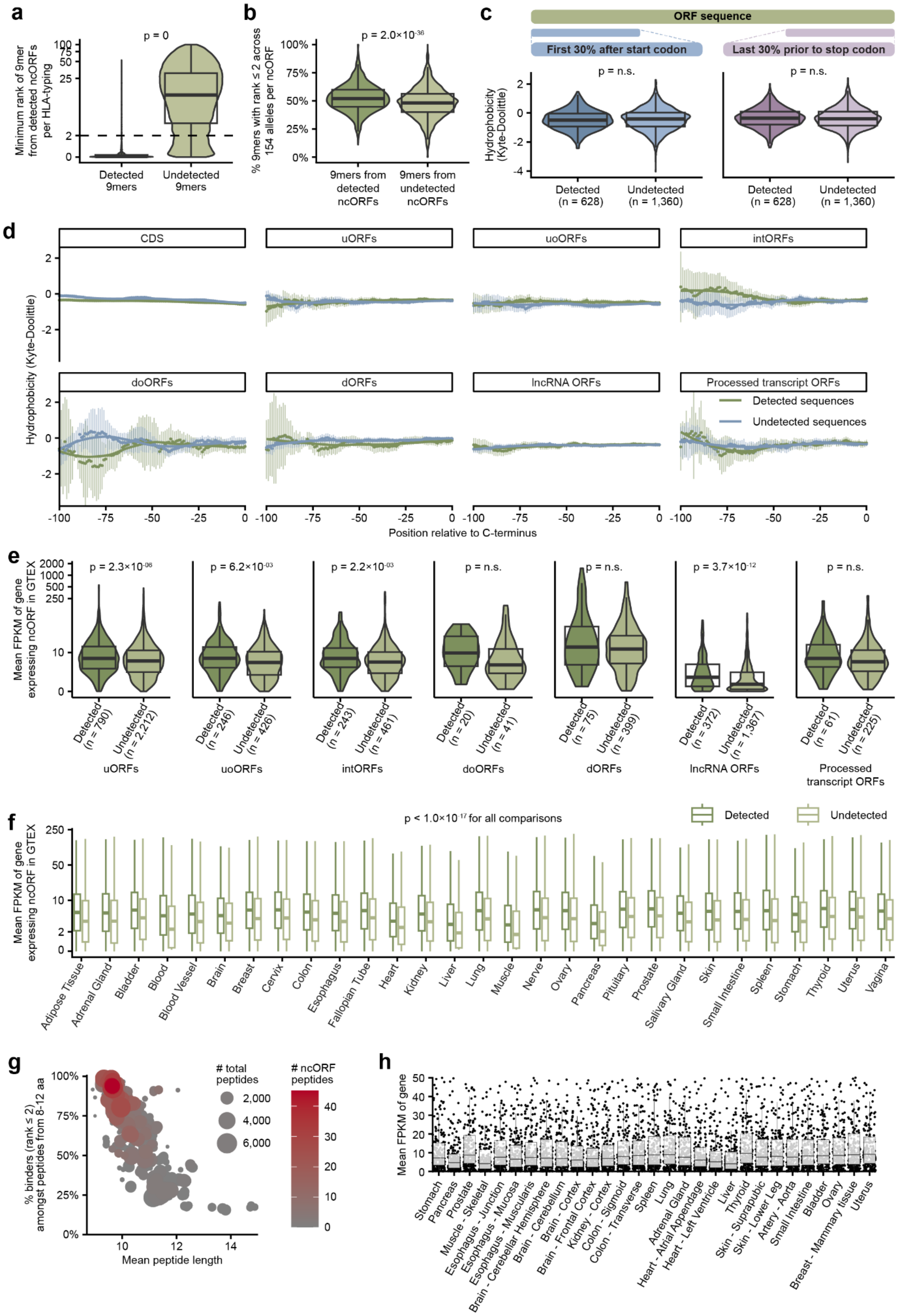
Supplementary Figure S1. The number of distinct peptides and proteins as datasets were added to the Human non-HLA (left) and HLA (right) PeptideAtlas. (**a**) Over 2.8 million distinct peptides have been observed in the 573 million PSMs in the non-HLA build. Each rectangle is one of the 1,172 experiments. Blue rectangles represent the cumulative number of distinct peptides in the build, while the gray rectangles depict the total number of distinct peptides within each experiment. **(b**) Over 0.86 million distinct peptides have been observed in the 28 million PSMs in the HLA build. Each rectangle is one of the 592 experiments. Blue rectangles represent the cumulative number of distinct peptides in the build, while the gray rectangles depict the total number of distinct peptides within each experiment. (**c**) The blue rectangles depict the cumulative 16,888 canonical proteins that have been cataloged in the 2023-06 Human non-HLA PeptideAtlas, whereas the gray rectangles show the total number of proteins present in each of the 1,172 experiments. (**d**) The blue rectangles depict the cumulative 13,799 canonical proteins that have been cataloged in the 2023-11 Human HLA PeptideAtlas, whereas the gray rectangles show the total number of proteins present in each of the 592 experiments. Although the total number of peptides continues to increase steadily, progress in the number of proteins is now very slow. Over the last 100 million PSMs, the cumulative counts are increasing by ~2,000 peptides per million PSMs and ~1 newly identified protein per million PSMs.


a
865,922 HLA peptides
99 peptides starting with:
• LLLLLLL
• PPPPPPP
• QQQQQQQ
6,467 HLA-I ms runs
3,289 HLA-II ms runs
118 datasets
Found in only HLA-I ms-runs
Found in HLA-I & II ms-runs
Found in only HLA-II ms-runs
All peptides (n = 865,823)
76.7% - Canonical
8-12 aa
13-35 aa
Known HLA
Unknown HLA
257,194 peptides from canonical proteins assigned to 163 alleles
(n = 664,273)
(n = 372,778)
(n = 274,506)
0.4% - Non-canonical
2,711 peptides from non-canonical ORFs assigned to 154 alleles
(n = 3,116) HLA class on which peptide was detected
(n = 3,094) Peptide length
(n = 2,826) HLA typing known
22.9% - Other peptides
b
Number of canonical proteins
16k
12k
8k
4k
0
Canonical proteins
Number of non-canonical ORFs
1,500
1,000
500
0
Non-canonical ORFs
Detected only in HLA-I MS-runs
Detected in both HLA-I and HLA-II MS-runs
Detected only in HLA-II MS-runs
c
Canonical proteins
Non-canonical ORFs
Percentage of all peptides
75%
50%
25%
0%
Peptide length
>14 aa
≤ 14aa
I
Both
II
I
Both
II
HLA class on which peptide is detected
d
Relative occurence
0.8
0.6
0.4
0.2
0.0
HLA-I (n=104 dataset components)
HLA-II (n=44 dataset components)
10
15
20
25
30
Peptide length (aa)
e
# peptides of ORF biotype detected
HLA-I only (n = 2,937 peptides)
HLA-II (and -I) (n = 179 peptides)
1250
1000
750
500
250
0
20
40
60
80
uORF
lncRNA
uoORF
intORF
dORF
PT
doORF
f
Number of peptide detections per dataset
On HLA-I (n = 17,241)
On HLA-II (n = 588)
600
400
200
0
20
40
60
Peptides (n = 179)
HLA-II exclusive peptides (n=22)
Peptide length
Dataset type
HLA-I exclusive (n = 62)
HLA-I & HLA-II (n = 30)
HLA-II exclusive (n = 6)

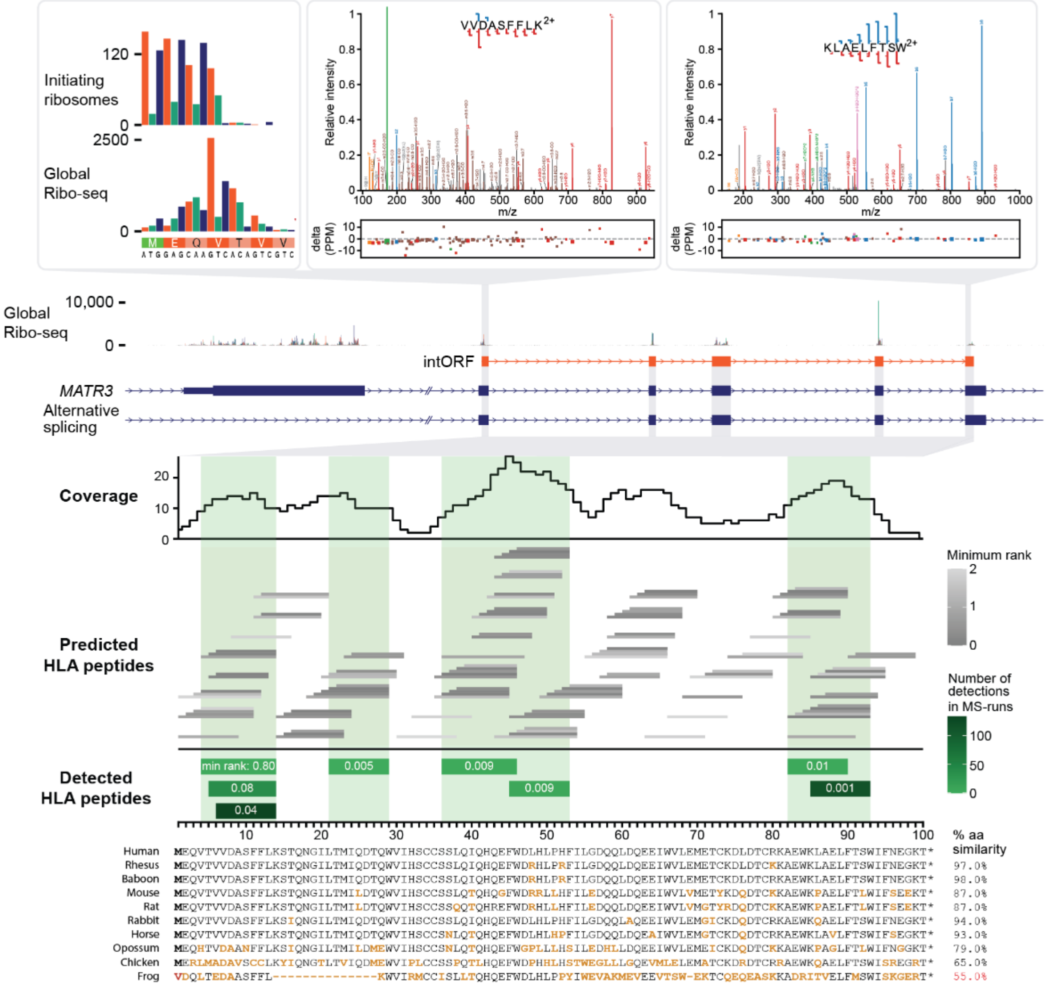
Supplementary Figure S2. Detection of ncORF peptides in HLA-I and HLA-II, and in cancer and non-cancer samples. (**a**) Schematic illustrating the total numbers of peptides (from both normal proteins and ncORFs) extracted from the total set of peptides. Depending on the analysis, peptides were further selected to those detected on HLA-I, had a length from 8-12 amino acids, and originated from an MS run with a known HLA-typing. The counts below each bar denote the number of distinct peptides. The distinction between “canonical”, “non-canonical”, and “other peptides” is defined in the methods. (**b**) Barplots showing for the detected canonical proteins (left) and detected ncORFs (right) whether their detected peptides were exclusively detected in HLA-I or HLA-II MS-runs, or in both. **(c)** Barplots showing for canonical proteins and ncORFs the percentage of peptides found exclusively in HLA-I and HLA-II MS-runs, or in both. Bars are colored by the peptide length being ≤14aa, or >14aa in length. **(d)** Line graph showing the peptide length distribution per dataset component split by HLA-class. **(e)** Bar plot showing the number of peptides detected per ORF exclusive to HLA-I samples (left) and those present in HLA-II samples, possibly in addition to HLA-I samples (right). Please note the x-axis scales differ by an order of magnitude between the left and right part of this panel. HLA-I and HLA-II peptide detection is not mutually exclusive as HLA-I peptides might be accidentally recovered from HLA-II pulldown experiments. **(f)** The frequency of peptide detection in HLA-I MS-runs (left) and HLA-II MS-runs (right) per peptide. Each alternating shade corresponds to a different dataset, with shades grouped by dataset type. The left axis denotes peptide lengths. Please note x-axis scales differ by an order of magnitude between the left and right part of this panel. Only peptides detected in at least one HLA-II sample are included. 22 of the 179 distinct peptides were exclusively detected in HLA-II samples. Fourteen of these peptides have a length 14 amino acids or greater, suggesting a potential presentation by HLA-II. This is still a minority in contrast to the total amount of 3,116 non-canonical ORF derived HLA peptides.



Supplementary Figure S3. (**a-d**) Comparisons of the detected and undetected non-canonical ORFs. (**a**) The total number of non-canonical ORF per ORF type and the number of ORFs for which a peptide was observed. (**b**) As in (**a**), but now shown in percentages. (**c**) The total number of non-canonical ORF grouped by length and the number of ORFs for which a peptide was observed. (**d**) As in (**c**), but now shown in percentages. Black lines on bar graphs indicate 95% confidence intervals. (**e**) Bar plot showing the proportion of Ribo-seq ORF-derived HLA peptides detected per biotype, categorized by whether the peptide was exclusively identified in immunopeptidomics analyses of cancer tissues or cell lines, non-cancer samples, or both. No significant changes in ORF biotype recovery are observed between these sample types. Peptides originating from ncORFs located on a known cancer gene are colored red. Black lines on bar graphs indicate 95% confidence intervals. (**f**) Density plots comparing the Ribo-seq ORF-derived HLA peptides differentiated on sample type (as depicted in (**e**)): cancer, non-cancer, or both. The plots compare peptides by their length, mass, hydrophobicity (Kyte-Doolittle scale), and isoelectric point. No significant changes between the distributions of these density plots can be observed.


a
p = 0
Minimum rank of 9mer from detected ncORFs per HLA-typing
100
75
50
25
2
0
Detected 9mers
Undetected 9mers
b
p = 2.0×10⁻³⁶
% 9mers with rank ≤ 2 across 154 alleles per ncORF
100%
75%
50%
25%
0%
9mers from detected ncORFs
9mers from undetected ncORFs
c
ORF sequence
First 30% after start codon
Last 30% prior to stop codon
p = n.s.
p = n.s.
Hydrophobicity (Kyte-Doolittle)
Detected (n = 628)
Undetected (n = 1,360)
d
CDS
uORFs
uoORFs
intORFs
doORFs
dORFs
lncRNA ORFs
Processed transcript ORFs
Detected sequences
Undetected sequences
Hydrophobicity (Kyte-Doolittle)
Position relative to C-terminus
e
p = 2.3×10⁻⁰⁶
p = 6.2×10⁻⁰³
p = 2.2×10⁻⁰³
p = n.s.
p = n.s.
p = 3.7×10⁻¹²
p = n.s.
Mean FPKM of gene expressing ncORF in GTEX
2000
1000
500
250
10
0
Detected (n = 790)
Undetected (n = 2,212)
uORFs
Detected (n = 246)
Undetected (n = 426)
uoORFs
Detected (n = 243)
Undetected (n = 461)
intORFs
Detected (n = 20)
Undetected (n = 41)
doORFs
Detected (n = 75)
Undetected (n = 399)
dORFs
Detected (n = 372)
Undetected (n = 1,367)
lncRNA ORFs
Detected (n = 61)
Undetected (n = 225)
Processed transcript ORFs
f
p < 1.0×10⁻¹⁷ for all comparisons
Detected
Undetected
Mean FPKM of gene expressing ncORF in GTEX
250
50
10
2
0
Adipose Tissue
Adrenal Gland
Bladder
Blood
Blood Vessel
Brain
Breast
Cervix
Colon
Esophagus
Fallopian Tube
Heart
Kidney
Liver
Lung
Muscle
Nerve
Ovary
Pancreas
Pituitary
Prostate
Salivary Gland
Skin
Small Intestine
Spleen
Stomach
Thyroid
Uterus
Vagina
g
% binders (rank ≤ 2) amongst peptides from 8-12 aa
100%
75%
50%
25%
0%
10
12
14
Mean peptide length
# total peptides
2,000
4,000
6,000
# ncORF peptides
40
30
20
10
0
h
Mean FPKM of gene
50
40
30
20
10
0
Stomach
Pancreas
Prostate
Muscle - Skeletal
Esophagus - Junction
Esophagus - Mucosa
Esophagus - Muscularis
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex
Kidney - Cortex
Colon - Sigmoid
Colon - Transverse
Spleen
Lung
Adrenal Gland
Heart - Atrial Appendage
Heart - Left Ventricle
Liver
Thyroid
Skin - Suprapubic
Skin - Lower Leg
Artery - Aorta
Small Intestine
Bladder
Ovary
Breast - Mammary tissue
Uterus

Supplementary Figure S4. Potential determinants of ncORF detection. **(a)** Violin plot comparing for all MS-runs grouped by HLA-typing the minimum binding prediction rank for detected and undetected 9mers. **(b)** Violin plot comparing all 9mers from detected and undetected non-canonical ORFs. The y-axis shows per ORF the percentage of 9mers with a NetMHCpan rank ≤ 2 across all 154 alleles associated with ncORF peptides. **(c)** Violin plots similar to **(4a)** comparing the hydrophobicity by the Kyte-Doolittle scale between detected and undetected ncORFs for the first 30% of the ncORF sequence after the start codon, or the last 30% of the ncORF sequence. Statistical tests were performed with the two-sided Wilcoxon test, reported p-values were adjusted for multiple testing with Bonferroni correction. **(d)** Comparison of the hydrophobicity similar to **(4b)** between detected and undetected ncORFs/CDS per ncORF biotype. Each dot represents the average hydrophobicity of the amino acids at that position and the 14 amino acids before that position per ncORF biotype or CDS grouped by whether these were detected or not in the immunopeptidomics data. The lines were fitted using Local Polynomial Regression Fitting. Vertical bars represent 95% confidence intervals. Note that because ncORFs are mostly smaller than 100 aa, confidence intervals get larger with increasing C-terminus offset. **(e)** Comparison of the expression levels of detected and undetected ncORFs similar to **(4c)**, but split per biotype. On the y-axis, the mean FPKM in GTEX of genes expressing an ncORF is shown on a pseudo-log scale. 326 ncORFs for which the gene id was not present in GTEX are not shown. Significance was determined using two-sided Wilcoxon tests, reported p-values were adjusted for multiple testing with Bonferroni correction. **(f)** Comparison of the expression levels of detected and undetected ncORFs similar to **(e)**, but split per tissue. Outliers are not shown in the graph. Significance was determined using two-sided Wilcoxon tests, and p-values were adjusted for multiple testing with Bonferroni correction. All comparisons were found to be significant. **(g)** Dot plot similar to **(3e)**, for MS-runs originating from the HLA-ligand-atlas. The plot visualizes the correlation between mean peptide length and the percentage of predicted binders amongst peptides with a length between 8 and 12 amino acids (NetMHCpan rank ≤ 2) per MS run. Dot size corresponds to the total number of peptides per MS-run. Dot color corresponds with the percentage of non-canonical ORF-derived peptides per MS-run. Statistical tests were performed with the two-sided Wilcoxon test, reported p-values were adjusted for multiple testing with Bonferroni correction. **(h)** Comparison of the GTEx expression of 224/277 genes from which ncORFs in the HLA ligand atlas originate (53 genes with ncORFs in the HLA ligand atlas were not present in GTEx). GTEx tissues comparable to those from the HLA ligand atlas were selected, and sorted in the same way as in **(4e)**. Each represents the mean FPKM of a gene across these tissue samples in GTEx. Only genes with a mean FPKM lower than 50 are plotted for clarity, but all 224 genes were included for the boxplots.



Supplementary Figure S5. Overview of data available for c5norep142, an intORF in the *MATR3* gene. Ribo-seq data shows the initiation of translation at the methionine translation initiation codon (green), as determined by enrichment of ribosomes at the TIS. Two peptide spectral matches for HLA-I peptides VVDASFFLK and KLAELFTSW are shown having nearly complete sequence coverage (USIs are mzspec:PXD037270:Liv32_1176935F:scan:33690:VVDASFFLK/2 and mzspec:PXD011628:PBMC009_msms37:scan:16281:KLAELFTSW/2, respectively). The lowest panel shows the position of all 8 peptides that were observed in the immunopeptidomics data. The color shading indicates the number of MS runs in which each peptide was observed. The middle panel shows all peptides that are predicted with NetMHCpan to be observable in the MS runs (i.e. they are predicted to bind with NetMHCpan score <2 to at least one allele in one of the samples in which peptides were observed). The top part shows the number of predicted binding peptides in which each amino acid was located. Green shadings indicate which part of the ORF sequence was observed. Except for the region near the offset of 62, detected peptides occurred in the regions with the highest numbers of predicted binders.