The immunopeptidome from a genomic perspective

The immunopeptidome from a genomic perspective: Establishing the non-canonical landscape of MHC class I–associated peptides.

Georges Bedran[1], Hans-Christof Gasser[2], Kenneth Weke[1], Tongjie Wang[2], Dominika Bedran[1], Alexander Laird[3,4], Christophe Battail[5], Fabio Massimo Zanzotto[6], Catia Pesquita[7], Håkan Axelson[8], Ajitha Rajan[2], David J. Harrison[9], Aleksander Palkowski[1], Maciej Pawlik[10], Maciej Parys[11], Robert O'Neill[12], Paul M. Brennan[13], Stefan N. Symeonides[4], David R. Goodlett[1,14,15], Kevin Litchfield[16,17], Robin Fahraeus[1,18], Ted R. Hupp[1,4], Sachin Kote[1,*], Javier A. Alfaro[1,2,14,*]

1 International Centre for Cancer Vaccine Science, University of Gdansk, Gdansk, Poland
2 School of Informatics, University of Edinburgh, Edinburgh, UK
3 Urology Department, Western General Hospital, NHS Lothian, Edinburgh, UK
4 Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK
5 CEA, Grenoble Alpes University, INSERM, IRIG, Biosciences and bioengineering for health laboratory (BGE) - UA13 INSERM-CEA-UGA, Grenoble, France
6 Department of Enterprise Engineering, University of Rome "Tor Vergata", Rome, Italy
7 LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
8 Division of Translational Cancer Research, Department of Laboratory Medicine, Lund University, Lund, Sweden
9 School of Medicine, University of St Andrews, St Andrews, UK
10 Academic Computer Centre CYFRONET, AGH University of Science and Technology, Cracow, Poland
11 Royal (Dick) School of Veterinary Studies and The Roslin Institute, University of Edinburgh, Edinburgh, UK
12 Cambridge Oesophagogastric Centre, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
13 Translational Neurosurgery, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK
14 Department of Biochemistry and Microbiology, University of Victoria, Victoria, Canada
15 University of Victoria Genome BC Proteome Centre, Victoria, Canada
16 Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK
17 Tumour Immunogenomics and Immunosurveillance Laboratory, University College London Cancer Institute, London, UK
18 Inserm UMRS1131, Institut de Génétique Moléculaire, Université Paris 7, Paris, France

* **Correspondence** should be addressed to Javier A. Alfaro; mailing address: ul. Kładki 24 80-822 Gdańsk, Poland; e-mail address: javier.alfaro@proteogenomics.caand Sachin Kote; mailing address: ul. Kładki 24 80-822 Gdańsk, Poland; e-mail address: sachin.kote@ug.edu.pl

Running title: The MHC class I non-canonical landscape.

Conflicts of interest statement: The authors declare no potential conflicts of interest.

Keywords: Cancer, tumor antigens, non-canonical MHC class I–associated peptides, mass spectrometry, shared antigens.

Financial support: G.B., D.B., K.W., A.P., R.F., T.R.H., S.K., and J.A.A. received support from Fundacja na rzecz Nauki Polskiej (FNP) (grant ID: MAB/3/2017). D.R.G. received support from Genome Canada & Genome BC (grant ID: 264PRO). D.J.H. received support from NuCana plc (grant ID: SMD0-ZIUN05). H.A. received support from Swedish Cancer Foundation (grant ID: 211709). H.G. received support from United Kingdom Research and Innovation (UKRI) (grant ID: EP/S02431X/1). C.P. received support from Fundação para a Ciência e a Tecnologia (FCT) through LASIGE Research Unit (grant ID: UIDB/00408/2020 and UIDP/00408/2020). A.L. F.M.Z., C.P., A.R., A.P., and J.A.A. received support from European Union's Horizon 2020 research and innovation programme (grant ID: 101017453). C.B. received support from Agence Nationale de la Recherche (ANR) through GRAL LabEX (grant ID: ANR-10-LABX-49-01) and CBH-EUR-GS 32 (grant ID: ANR-17-EURE-0003). S.N.S. received support from Cancer Research UK (CRUK) and the Chief Scientist's Office of Scotland (CSO): Experimental Cancer Medicine Centre (ECMC) (grant ID: ECMCQQR-2022/100017). A.L. received support from Chief Scientist's Office of Scotland (CSO) NRS Career Researcher Fellowship. R.O.N. received support from CRUK Cambridge Centre Thoracic Cancer Programme (grant ID: CTRQQR-2021\100012).

Abstract

Tumor antigens can emerge through multiple mechanisms, including translation of non-coding genomic regions. This non-canonical category of tumor antigens has recently gained attention; however, our understanding of how they recur within and between cancer types is still in its infancy. Therefore, we developed a proteogenomic pipeline based on deep learning *de novo* mass spectrometry to enable the discovery of non-canonical MHC class I–associated peptides (ncMAPs) from non-coding regions. Considering that the emergence of tumor antigens can also involve post-translational modifications, we included an open search component in our pipeline. Leveraging the wealth of mass spectrometry–based immunopeptidomics, we analyzed data from 26 MHC class I immunopeptidomic studies across 11 different cancer types. We validated the *de novo* identified ncMAPs, along with the most abundant post-translational modifications, using spectral matching and controlled their false discovery rate (FDR) to 1%. The non-canonical presentation appeared to be 5 times enriched for the A03 HLA supertype, with a projected population coverage of 54.85%. The data reveal an atlas of 8,601 ncMAPs with varying levels of cancer selectivity and suggest 17 cancer-selective ncMAPs as attractive therapeutic targets according to a stringent cutoff. In summary, the combination of the open-source pipeline and the atlas of ncMAPs reported herein could facilitate the identification and screening of ncMAPs as targets for T-cell therapies or vaccine development.

Introduction

The accelerated adoption of mass spectrometry (MS) for high-throughput profiling of immunopeptidomes in cancer has led to several discoveries. Leveraging these studies to improve cancer immunotherapy involves connecting the wealth of immunopeptidomic data to immunogenomics, where the goal is to carefully choose effective targets for T-cell therapies or vaccine development.

The discovery of cancer antigens has mainly focused on mutated tumor-specific antigens (neoantigens) arising from patient-specific somatic mutations. It has been shown that only a small percentage of the numerous non-synonymous mutations in a tumor actually produce neoantigens (1,2). The challenging task of identifying those that can evoke a suitable tumor rejection was addressed by Ebrahimi-Nik et al. (3). Using a combination of genomics, shotgun MS immunopeptidomics, and targeted MS, they found that (I) MS-identified neoepitopes are a rich source of tumor rejection–mediating antigens, (II) neoantigens derive from passenger mutations, and (III) binding affinity and $CD8^+$ T-cell responses in tumor-bearing hosts are poor predictors of antitumor activity *in vivo*. Although neoantigens confer an advantage to patients undergoing immunotherapy (4), their patient-specific nature is a major bottleneck when producing off-the-shelf treatments for a large number of individuals. Alternatively, shared neoantigens (5) (*i.e.,* recurrent mutations in cancer) could offer a new line of population-level immunotherapy. However, high-throughput experimental profiling of such broadly presented neoantigens across the human population is a long-term goal with many milestones to be achieved.

Recently, tumor antigens that exceed the exome boundaries (*i.e.,* non-canonical) have attracted attention as potential targets as a result of their immunogenicity and recurrence among cancer patients (6). These antigens find their way to the cell surface through rapid degradation (7) of “non-coding” translation products stemming from novel open reading

frames (nORF) (8). In addition, “non-coding” translation products can originate from other sources (9), including intron retention (IR) (10), ribosomal slippage (11), and frameshift mutations (12). In 2016, Laumont *et al.* (13) demonstrated their association with MHC molecules using a reductionist approach based on 6-frame translation and subsequently their recurrence between patients (14). Ribo-Seq has proven to be an immensely valuable tool for identifying non-canonical MHC class I–associated peptides (ncMAPs) as it provides experimental evidence for their non-canonical translation and MHC class I presentation when combined with MS immunopeptidomics (6,15,16). Despite previous efforts to study non-canonical immunopeptidomes, the requirements of such multi-level experimental data (Ribo-seq and/or RNA-Seq) or computational struggles when dealing with large MS databases have hindered their large-scale profiling in a harmonized manner across multiple cancer types from hundreds of samples.

With these considerations in mind, we developed COD-dipp (Closed Open *De novo* – deep immunopeptidomics pipeline), a pipeline based on deep learning *de novo* MS to enable the discovery of ncMAPs. Owing to the potential involvement of post-translational modifications (PTMs) in this process (1), we added an open search component for their discovery. We applied COD-dipp to a large-scale dataset using immunopeptidome profiles of over 772 samples from 26 (1,13,14,17–40) published studies and 11 cancer types. We identified a range of PTMs of potential interest from a therapeutic standpoint and tackled the non-canonical immunopeptidome. We validated the *de novo* identified ncMAPs and controlled their false discovery rate (FDR) to 1% using a second-round search with tuned PTM parameters, in addition to a series of quality-control steps. Our large-scale analysis revealed 8,601 ncMAPs, accounting for 1.7% of immunopeptidomes. These peptides had varying levels of tumor selectivity, defined by their parent gene expression levels in normal tissues. We suggest 17 ncMAPs as attractive therapeutic targets using a stringent tumor-selectivity cutoff.

Materials and methods

Dataset selection

Twenty-four studies were selected based on a list of keywords related to immunopeptidomics (**Supplementary Method S1**). Low-resolution analyses were eliminated, and MHC class I–related datasets conducted with at least one of the following instruments were kept: Q Exactive, Q Exactive plus/HF/HFX, LTQ Orbitrap Velos, LTQ Orbitrap Elite, Orbitrap Fusion, and Orbitrap Fusion Lumos (**Supplementary Table S1**). An additional study was considered from the MassIVE (RRID:SCR_013665) database, as it incorporates 95 HLA-A, -B, -C, and -G mono-allelic cell lines (28,40). An auxiliary immunopeptidomic dataset (39) covering 30 healthy tissues from 21 healthy individuals was also used to partly assess cancer selectivity.

Proteogenomic database generation

Canonical protein database for MS database search

A protein database was downloaded using ENSEMBL r94 BioMart (RRID:SCR_002344); decoy sequences were appended by reversing the target sequences, and 116 contaminant proteins were added (41).

Non-canonical protein database for alignment using BLAST-like alignment tool (BLAT)

A pre-mRNA 3-frame translation (3FT) database was generated from genes with a protein-coding biotype based on ENSEMBL r94 (RRID:SCR_002344) using the AnnotationHub and Biostrings (RRID:SCR_016949) R packages.

COSMIC mutated protein database for BLAT alignment

COSMIC (RRID:SCR_002260) coding Mutants (42) VCF v95 was downloaded along with ENSEMBL v94 CDS and GTF files. An in-house Python (RRID:SCR_008394) package was used along with the previously mentioned inputs to generate a FASTA file containing the corresponding mutated protein sequences.

MS computational analysis

Algorithms representing three main philosophies of peptide-spectrum matching including open search, *de novo* sequencing, and closed search were used. The open search approach allowed the identification of distantly related peptides and could identify PTMs and single amino acid variations. The *de novo* sequencing approach derived sequences from first-principle analysis of the MS^2 spectra. The closed search approach, used as a validation step, assumed a specific set of reference protein sequences and allowed for limited post-translational modifications. Although each approach has its own limitations, our strategy addressed them by combining a closed search with a *de novo* sequencing approach and implementing multiple filtering steps for accuracy control and quality control checkpoints (see **Supplementary Figure S1**).

Data conversion

The proprietary RAW files acquired from the selected instruments were converted to mzML and MGF formats using msconvert (ProteoWizard version 3.0.19295. c8b8b470d, RRID:SCR_012056) with the peak-picking and TPP compatibility filters.

Open search analysis

The MSFragger (43) v2.2 search engine was used to conduct an open search analysis against the ENSEMBL r94 protein database in combination with PTMiner (44) v1.1.2, to apply a transfer FDR and a false localization rate of 1% (FLR, the rate of falsely localizing

the site of modification). Unspecific cleavage generating peptides 8 to 25 amino acids long with no fixed/variable PTMs was considered. Further analysis revealed that the frequent unexplained mass shifts observed during the open-search annotations were caused by non-specific cleavage. To address this issue, an open-search post-processing algorithm, PTMiner, was employed to effectively corrects for mass shifts introduced by in-source fragmentation, nonspecific digestion, or missed cleavage, by adding or deleting amino acids from the peptide N- or C-termini. For instance, a deviation of -128.1 to -128.08 Dalton on lysine residues was frequently detected on the first 2 or last 2 amino acids of peptides. The deviation was caused by non-specific cleavage during the open search and resulted in an incorrect assignment of a negative mass shift of a lysine due to the presence of an additional lysine in the sequence. As these cases are not biologically meaningful, unexplained mass shifts were removed from the final results of the study.

De novo analysis

DeepNovoV2 (45) is a neural-network-based *de novo* peptide sequencing model that integrates convolutional neural networks (CNNs) and long short-term memory (LSTM) architectures. This deep-learning design extracts features from both the spectrum and the language of the presented peptides. DeepNovo has demonstrated improved performance compared to the state-of-the-art *de novo* sequencing algorithms by large margins (45). The model can be tuned on a restricted peptide space to improve its performance. The training, testing, and validation sets were derived from MS-GF+ (v2019.04.18, RRID:SCR_015646) database search results for each sample. The search used the ENSEMBL v94 protein database and 8 to 25 amino acid peptides with unspecific cleavage, no fixed/variable PTMs and an FDR of 1% applied by Scavager (46). The trained models were used to perform *de novo* (prediction) on the remaining unmatched spectra of each sample (from MS-GF+ after 1% FDR control). Accuracy was calculated by comparing the *de novo* predicted sequences and MS-GF+ results on the validation set. *A de novo* score threshold that controlled the

accuracy at 90% within the validation set was applied to the predicted sequence in a sample-specific manner.

De novo peptide annotation

De novo peptides from canonical human proteins were identified using BLAT (47) (RRID:SCR_011919) alignment against the target-decoy protein database. Sequences perfectly matching any protein sequence were considered exonic (one mismatch allowed for the isobaric amino acids leucine and isoleucine). All remaining sequences unexplained by proteins were considered potential non-canonical peptides and were aligned against the pre-mRNA 3FT database. Stringently, peptides perfectly matching a 3FT sequence without any mismatch were required to have at least three mismatches with any known protein sequence before being considered non-canonical. Since peptide-spectrum matches (PSMs) can be assigned without complete sequencing accuracy, requiring a 3 amino acid difference alongside the 90% accuracy cutoff above increases the confidence that the peptides assigned fall far outside the standard human proteome. Remaining *de novo* peptides without any canonical or non-canonical annotation were labeled as 'unmapped peptides' and discarded.

Second-round search

A second-round search was performed using the FragPipe (41,43) headless pipeline, which includes MSFragger v3.4, MSBooster (bioRxiv 2022.10.19.512904), and Philosopher (41). Non-canonical peptides from all samples were concatenated with the ENSEMBL v94 protein into a custom database. Only four of the most abundant PTMs were considered to avoid a large search space complexity, inflated FDR, and decreased sensitivity. The following variable PTMs were included: methionine oxidation, N-terminal acetylation, cysteinylation, and cysteine carbamidomethylation (for samples treated with iodoacetamide). Unspecific cleavage generating peptides 7 to 15 amino acids long was considered. The ion, PSM, and peptide-level FDR were maintained at 1%.

Alignment of immunopeptides to the genome

Second-round search non-canonical peptide coordinates were retrieved from the 3FT database FASTA headers and stored in BED format.

Open reading frame analysis

Upstream genomic sequences of ncMAPs were scanned for start codons up to the first encounter with a stop codon. Sequences were centered around the detected start codons and stretches of 100 nucleotides from each side were extracted. Translation initiation site (TIS) scores were predicted for each sequence using TITER (48), a deep-learning-based framework for accurately predicting TIS on a genome-wide scale based on QTI-seq data. A TIS score greater than 0.5 was considered a positive prediction.

Intron retention analysis

For each intron in the UCSC hg38 KnownGene table (RRID:SCR_005780), the first codon coordinates of the corresponding upstream exon in-frame with the canonical translation were extracted and stored in BED format (see **Pseudocode 1**). Intronic coordinates from the generated BED file were intersected with the ncMAPs BED file using pybedtools (49) (RRID:SCR_021018). Intronic retention events were considered possible when ncMAPs within introns were in-frame with their upstream exons (see **Pseudocode 2**).

```
// Pseudo-code 1: extracts the start coordinate of the first in-frame codon for
each exon (inframeCoordinate variable)
for each transcript
  remainderValue = 0
  for each exon
    if strand is positive
      if downstream intron exists
  leftoverBases = remainder of (ExonEndCoordinate - remainderValue - ExonStart +
1) / 3
        if remainderValue is equal to 0
          inframeCoordinate = ExonStartCoordinate
        else
          inframeCoordinate = ExonStartCoordinate – remainderValue
        if leftoverBases is greater than 0
          remainderValue = 3 - leftoverBases
        addToTable(transcript, chromosome, ExonStart, ExonEnd, inframeCoordinate,
IntronStart, IntronEnd)
    if strand is negative
      if downstream intron exists
        leftoverBases = remainder of (ExonSart - ExonEndCoordinate +
remainderValue + 1) / 3
        if remainderValue is equal to 0
          inframeCoordinate = ExonEndCoordinate
        else
          inframeCoordinate = ExonEndCoordinate + remainderValue
        if leftoverBases is greater than 0
          remainderValue = 3 - leftoverBases
        addToTable(transcript, chromosome, ExonStart, ExonEnd, inframeCoordinate,
IntronStart, IntronEnd)

// Pseudo-code 2: checks if each intronic ncMAP is in-frame with its upstream
exon.
  ncMAPIsInFrame = False
  if strand is positive
    // firstCoordinate = start coordinate of ncMAP
    // secondCoordinate = start coordinate of the first inframe codon from
previous exon
    coordinateDifference = firstCoordinate – secondCoordinate
    if remainder of (coordinateDifference / 3) is equal to 0
      ncMAPIsInFrame = True
  else:
    // firstCoordinate = start coordinate of first inframe codon from previous
exon
    // secondCoordinate = end coordinate of ncMAP
    coordinateDifference = firstCoordinate – secondCoordinate
    if remainder of (coordinateDifference / 3) is equal to 0
      ncMAPIsInFrame = True
```

Frameshift mutation analysis

The COSMIC (42) v95 coding mutations (RRID:SCR_002260) VCF file was downloaded and converted into a protein FASTA file using a VCF-to-Proteogenomics toolkit (https://github.com/immuno-informatics/VCFtoProteogenomics) ncMAPs were then aligned

to the resulting 16 GB FASTA using BLAT v35 (47). Only hits with exact matches to sequences from frameshift mutations were considered.

Comparison of the identified non-canonical MHC class I–associated peptides between studies

ncMAPs from 4 different studies (6,13,16,50) were collected. First, sequences were aligned to the human proteome (ENSEMBL v94) using BLAT v35 (47). Sequences found in human proteins were discarded, and the remaining sequences were aligned to the 3FT database with one mismatch allowance for the isobaric amino acids leucine and isoleucine, as allowed for COD-dipp ncMAPs. Genomic coordinates of the sequences found in the 3FT database were extracted and overlapped between studies using the ChIPpeakAnno (51) R package (RRID:SCR_012828). A minimum overlap of 21 nucleotides (7 amino acids) between two sequences was required.

Cancer selectivity of the non-canonical MHC class I–associated peptides

Tumor specificity has been previously implied when peptide parent genes are either completely absent or present in trace amounts in healthy tissues (6,14,16) since MHC class I presentation is preferentially derived from highly abundant transcripts (28,30). While tumor specificity implies the expression of an antigen solely in tumor samples, the experimental design of this study cannot guarantee this constraint. Instead, cancer-selective ncMAPs were conservatively identified through three iterative steps:

Step 1: Panel of normal immunopeptidomes

In addition to the 88 healthy MS samples from the initial set of the 25 considered studies, the HLA Ligand Atlas (39) was used to extend the panel of normal immunopeptidomes and

partly assess the cancer selectivity of the 8,601 identified ncMAPs. The HLA Ligand Atlas is a pan-tissue immunopeptidomic reference for 30 healthy tissue types obtained from 21 human subjects. The resulting 334 healthy samples (see **Supplementary Table S1**) were analyzed in the same manner as in the second-round search (see *Second-round search* above). ncMAPs identified in the panel of normal immunopeptidomes were labeled as non-cancer selective.

Dimensionality reduction of the HLA-binding motif space

Binding affinity prediction was employed to identify similarities and differences in HLA-binding motifs among the 65 healthy and 51 tumor-only HLA alleles. NetMHCpan-4.1 was utilized to evaluate the binding of 1,000,000 random peptides to each allele, which resulted in a binding matrix (BM) of 116 alleles and 1,000,000 peptides. A value of 1 was assigned to strong binders (EL rank ≤ 0.5%) in the BM; otherwise, a value of 0 was assigned. A pairwise cosine distance matrix (DM) was then calculated to assess the similarity of binding between alleles. The DM was then reduced using t-SNE to visualize the data in 2D with a perplexity of 20 and 500 iterations.

Step 2: Parental gene expression levels in healthy tissue

The gene expression levels of the identified ncMAPs were retrieved from the GTEx v8 (52) dataset, consisting of 29 tissues from 948 healthy donors and 17,382 overall samples. Considering all individuals, the 90th percentile value of normalized expression was assigned to each gene per tissue as a strict step to guarantee the upper-end gene expression in healthy tissues. A stringent cutoff for cancer selectivity was used to shortlist ncMAPs whose parent genes fell below a 1 TPM expression cutoff (excluding the testis tissue given its immune-privileged status). It is worth noting that this stringent threshold removes 92% of protein-coding genes that show expression above 1 TPM in any tissue within the GTEx v8.

Step 3: Protein expression levels in healthy tissue

The protein expression levels of ncMAPs passing the 1 TPM cutoff were retrieved from the Human Protein Atlas V22.0 database (53). ncMAPs without parent protein expression in healthy tissues were labeled as cancer-selective (excluding the testis tissue given its immune-privileged status).

Code availability

The COD-dipp code, intended for high-performance computing (HPC), is available on the GitHub repository: https://github.com/immuno-informatics/COD-dipp.

Data availability

The data analyzed in this study were obtained from PRIDE at PXD004746, PXD014017, PXD012308, PXD011628, PXD012083, PXD011766, PXD013057, PXD011723, PXD007203, PXD004233, PXD003790, PXD001898, PXD007860, PXD011257, PXD007935, PXD009749, PXD009753, PXD009750, PXD009751, PXD009752, PXD009754, PXD009755, PXD004023, PXD007596, PXD009531, PXD010808, PXD008937, PXD009738, PXD006939, PXD005231, PXD000394, PXD004894, PXD019643 and from massIVE at MSV000080527, MSV000084172, MSV000084442. The results of this study are available within the article and its supplementary data files and are accessible on the following figshare repository: https://doi.org/10.6084/m9.figshare.16538097.

Results

Immunopeptidomic MS datasets

We selected 25 immunopeptidomic MS studies (see **Supplementary Table S1**) to create a cancer-centered dataset of MHC class I presentation. Data-dependent acquisition (DDA) studies covered eleven cancer types distributed across the brain (Glioblastoma and Meningioma), lung, skin, liver, blood (Leukemia and Lymphoma), colon, ovaries, kidneys, and breast. Moreover, tumor and healthy samples were derived from either cell lines or patient tissues (**Fig. 1a** and **Supplementary Method S1**). We selected publicly available studies with data generated using high-resolution MS instruments (LTQ Orbitrap, Q Exactive Plus/HF/HFX, and Fusion Lumos) to minimize the bias associated with older tandem MS instrumentation (**Fig. 1b**). Within our dataset, the most commonly used monoclonal antibody for HLA class I immunoprecipitation (IP) was W6/32 in comparison to the other antibodies (BB7.2 and G46-2.6) (**Fig. 1c**, see **Supplementary Table S1**). The selected studies covered five different HLA class I genes, with HLA-A, B, and C being the most studied compared to HLA-E and -G (**Fig. 1d**). Furthermore, the included MS samples covered 114 HLA alleles (**Fig. 1e**).

Closed Open *De novo* – deep immunopeptidomics pipeline (COD-dipp)

We present COD-dipp, an open-source high-throughput pipeline with novel post-processing steps, to deeply interrogate immunopeptidomic datasets (**Fig. 2**)**.** We used this pipeline to screen for ncMAPs in datasets utilizing DDA due to its widespread use. To identify post-translationally modified MHC class I–associated peptides (ptmMAPs), we performed an open-search analysis with MSFragger (43) and controlled both FDR and the FLR to 1% with

PTMiner (44). To identify ncMAPs, we used DeepNovoV2 (45) for *de novo* analysis. In combination with the PSM level information of MS-GF+ (54), DeepNovoV2 was trained to interpret the raw MS data in a sample-specific manner. The training step for such a deep learning approach is crucial for learning the features of tandem mass spectra, fragment ions, and leveraging sequence patterns in the immunopeptidome to impute missing MS^2 fragments. All high-quality *de novo* peptides (90% accuracy) were sequentially mapped (47) to (I) the human reference proteome to reveal the *de novo*-based canonical MHC class I–associated peptides, and (II) to a 3FT database to reveal the *de novo*-based ncMAPs. Finally, an orthogonal validation step was performed by a second-round search to control a 1% FDR for the *de novo* identified ncMAPs while considering the most abundant PTMs found by the open-search strategy. Applying the COD-dipp pipeline across the dataset revealed the breadth of (I) post-translationally modified MHC class I–associated peptides referred to as ptmMAPs, and (II) non-canonical MHC class I–associated peptides referred to as ncMAPs.

ptmMAPs

The open search analysis reported 4.03% of the MS spectra showing post-translational modifications (**Fig. 3a**). Some identified PTMs were confirmatory, representing chemical modifications from sample preparation methods (cysteine carbamidomethylation) or common chemical derivatives (methionine oxidation and di-oxidation). We also observed PTMs that are extremely common in proteins, such as protein N-terminal acetylation, affecting multiple properties such as half-life time, folding, and interaction. On the other hand, some of the identified PTMs have been reported previously to increase immunogenicity against diseases (55) and protect against degradation (tri-oxidation of cysteine (56), cysteinylation (57), and N-term serine acetylation, see **Fig. 3b and Supplementary Table S2**. Furthermore, 1.12% of spectra from open search showed unknown mass shifts, as illustrated in **Fig. 3a** (green and red). This category was partly populated by computational artifacts and was excluded

from the final results. To validate these findings, we performed an independent post-search by crosschecking the identifications from our open search with those of the original studies. The results showed 96.1% agreement in peptide-spectrum matches, which are detailed in **Supplementary Method S2**: Validation 1 and **Supplementary Figure S2**.

ncMAPs

We explored the ncMAP landscape in cancer using our workflow (**Fig. 2**) and identified 10,413 unique *de novo*-based ncMAPs from intragenic non-coding regions (before the second-round search validation), which accounted for 3.7% of the identified *de novo* sequences. We took two additional validation steps, including checking the identification scores as well as the correlation between the experimental and theoretical liquid chromatography retention times, to guarantee the correctness of these identifications (see **Supplementary Method S2**: Validation 2 and 3, and **Supplementary Figure S2**). The *de novo* non-canonical peptides showed strong evidence of high-quality identification (*i.e.,* correctly predicted complete peptide sequences). Even with this strong evidence, it was possible that chromatic behavior remained unchanged in certain instances where neighboring amino acids were in flipped positions, or that a 90% accuracy rate still led to an uncertain FDR percentage. Hence, we confirmed the identified 10,413 *de novo*-based ncMAPs by performing a second-round search for additional validation and controlling the FDR at 1%. Several PTMs were also considered in the parameters from the *a priori* knowledge provided by the open search strategy. Of the 516,382 uniquely identified peptides in the second-round search, 1.7% (8,601) were non-canonical (**Fig. 3c** and **Supplementary Table S3**). The PTM profiles (**Fig. 3d**) of canonical (dark gray) and non-canonical (light gray) peptides appeared to be similar, with M oxidation being the most prevalent modification. The identified ncMAPs showed comparable spectra from patients within the same studies and

from different studies (**Supplementary Figures S3**, **S4**, and **S5** provide examples of such similarities). The binding affinities of all 8,601 ncMAPs resulting from the second-round search were further investigated using NetMHCpan 4.1 (58). The binding prediction analysis showed a comparable binding rate for both the canonical (90%) and non-canonical (93%) MAPs, as depicted in **Fig. 3e**. We further took four additional independent post-search validation steps, including checking retention time shifts induced by PTMs, mass accuracy, and spectra comparison to those of the original studies, guaranteeing the correctness of the ncMAPs identified by the second-round search (see **Supplementary Method S2**: Validation 4, 5, 6, and 7, and **Supplementary Figure S2**).

Comparison of COD-dipp ncMAPs with the literature

To assess the performance of our COD-dipp method, we conducted a comparison with the results of peptide-PRISME by Erhard *et al.* 2020 (50). Our comparison was based on three common studies (1,14,34) and resulted in 3,453 at 1% FDR from COD-dipp along with 4,576 ncMAPs at 10% FDR from Erhard *et al.* We first aligned Erhard *et al.'s* ncMAPs to the human proteome and eliminated a small fraction (1.4%) that matched the canonical protein sequences (**Fig. 4a**, left-hand side). Since the COD-dipp ncMAPs were restricted to the 3FT of protein-coding genes, we aligned the remaining ncMAPs from Erhard *et al.* to the same 3FT database for comparison purposes. **Fig. 4a (left-hand side)** shows that 68.25% of ncMAPs were successfully mapped to the 3FT database. The rest (30.35%) that did not align to any of the human proteome or the 3FT database are shown in yellow on **Fig. 4a** left-hand side. This unmapped fraction consisted of ncMAPs from regions of the genome not studied herein, such as intergenic regions, anti-sense translation, etc. The successfully mapped fraction to the 3FT database (navy) of 3,123 ncMAPs along with 3,453 ncMAPs from COD-dipp were then compared, as shown in **Fig. 4a** right-hand side (see **Supplementary Table S4**). peptide-PRISME shared 38% (1,197) of its ncMAPs (intersection) with COD-dipp (**Fig. 4a** right-hand side) and showed 62% (1,926) of exclusive ones. Adjusting the higher FDR used by peptide-PRISME from 10% to 1% increased the

shared fraction to 48.9% (**Fig. 4b**), along with a ~ 2.4-fold decrease in total ncMAPs (from 4,576 to 1,916). At an FDR of 1%, COD-dipp identified 2.34 times more exclusive ncMAPs (2,298 vs. 979) from the 3FT of protein-coding genes.

To contextualize our findings from COD-dipp within the existing literature on ncMAPs, we compared our results with those of three previous studies: (I) Laumont *et al.* 2016 (13), (II) Chong *et al.* 2020 (6), and (III) Ouspenskaia *et al.* 2021 (16), as shown in Figure 4c. We used the same mapping procedure that was applied to peptide-PRISME results. We eliminated a fraction of sequences mapping to known proteins, which was 4%, 5%, and 3% of sequences for Chong *et al.* 2020, Laumont *et al.* 2016, and Ouspenskaia *et al.* 2021, respectively (see **Fig. 4c** left-hand side). **Fig. 4c** left-hand side shows in navy the fractions of ncMAPS that were successfully mapped to the 3FT database, which was 34.38% for Chong *et al.* 2020, 63.69% for Laumont *et al.* 2016, and 72.74% for Ouspenskaia *et al.* 2021. The remaining ncMAPs that did not align (**Fig. 4c** left-hand side in yellow) to any of the human proteome or the 3FT database originate from sources not studied herein. For instance, Laumont *et al.* 2016 included 6-frame translation in their MS search database, which accounts for intergenic regions, anti-sense translation, long non-coding RNA, and retroelement sources. Both Chong *et al.* 2020 and *Ouspenskaia et al.* 2021 added Ribo-Seq detected proteins to their MS database searches, accounting for all possible nORFs, even those outside of known genes. The fractions successfully mapped to the 3FT database (navy) from these three studies, along with the 8,601 ncMAPs from COD-dipp, were then compared, as shown in **Fig. 4c** right-hand side (**Supplementary Table S4**). Intersections with COD-dipp were 31.42% for Chong *et al.* 2020, 38.3% for Ouspenskaia *et al.* 2021, and 45.8% for Laumont *et al.* 2016, respectively. In contrast, intersections with all other studies were 40% for Chong *et al.* 2020, 38.66% for Ouspenskaia *et al.* 2021, and 65.93% for Laumont *et al.* 2016. Hence, COD-dipp ncMAPs alone accounted for 78.55% of Chong *et al.* 2020's intersection, 96.07% of Ouspenskaia *et al.* 2021's intersection, and 69.47% of Laumont *et al.* 2016's intersection. COD-dipp ncMAPs accounted, on average, for 81.36% of

the intersection when comparing three previously published ncMAP sets, thus validating our approach. With 2,168 ncMAPs (25%) shared with the literature and 6,433 new ncMAPs, we have revealed an atlas of non-canonical MHC class I presentation.

Properties and origins of ncMAPs

We compared the sequence lengths of canonical and non-canonical MAPs (**Fig. 5a**) and found them to be similar, with a slight skew of the non-canonical category toward longer lengths. This could be due either to an actual preference of ncMAPs toward longer sequence lengths or simply the consequence of requiring 3 amino acid differences from any known proteins favoring longer sequences. Next, we inspected ncMAPs according to their relative positions within protein-coding genes (**Fig. 5b**). Exonic regions translated in alternative frames were the main source of ncMAPs (19.2%). These events could arise from frameshift mutations, initiation codon readthrough (59), nORFs, or ribosomal slippage (11) during translation (i.e., ribosome frameshifting). Intronic regions were the second most abundant source of ncMAPs (12.2%). These events can arise from frameshift mutations, nORFs, or IR. Interestingly, 5'-UTRs contributed to 10.2% of ncMAPs and have been shown to produce translation products through upstream ORFs along with a non-AUG start codon (60). Lastly, 3'-UTRs contributed the least toward ncMAPs (3.2%), potentially through stop codon read-through (61). It is important to note that these categorizations are not mutually exclusive and that an ncMAP sequence may have multiple assignments due to the overlapping nature of transcripts. We conducted three analyses to estimate how well the nORFs (I), IR (II), and frameshift mutations (III) could explain the detected ncMAPs. (I) ncMAPs with upstream start codons (AUG, CUG, UUG, GUG, and ACG) accounted for 63.4%, and 41.5% were predicted to be TIS using TITER (48) (**Fig. 5c** left-hand side). The breakdown of the TIS start codon distribution (**Fig. 5c**, right-hand side) showed CUG (L) as the most abundant nORF start codon, and 70% of the predicted TIS showed non-AUG start codons, in line with previous findings (15). (II) Translation frames of ncMAPs from intronic regions were checked for compatibility with upstream exons, and 49.4% were found in-frame with upstream exons,

making IR events a possible source (**Fig. 5d**). (III) A total of 597 ncMAPs were found in aberrant proteins from frameshift mutations in cancer (42) (**Fig. 5e** and **Supplementary Table S5**). Eventually, 70.1% of ncMAPs were explicable by novel ORFs, IR, or frameshift mutations (**Fig. 5f**). ncMAPs were found to be presented by all 113 alleles in our dataset, except for the HLA-C*07:17 allele, mostly because of low sample coverage by MS for this allele (see **Supplementary Figure S6**). Furthermore, the average non-canonical presentation per HLA supertype (62) was 1%, except for A03, which was 5% (see **Supplementary Figure S6**).

Cancer selectivity of ncMAPs

Of the 8,601 identified ncMAPs, 2,758 were detected in the panel of normal healthy tissue by MS and were labeled as non-cancer-selective. The panel of normals originally consisted of healthy MS samples from all considered studies, which we extended by adding the HLA Ligand Atlas (39), a pan-tissue immunopeptidomic reference of 30 healthy tissue types obtained from 21 human subjects. **Fig. 6a** shows the ability of the extended panel of normals to capture several more ncMAPs (12.85%) in healthy tissues that were not observed in our original panel of normals (19.22%). We assessed the coverage of tumor-only HLA alleles in healthy samples using the panel of normal samples. The 334 healthy samples covered 53% of the HLA alleles expressed in the tumor samples. Analysis of a subset of ncMAPs represented by the 57 shared alleles (i.e., present in both healthy and tumor samples) revealed a substantial overlap in HLA-binding motifs between the panel of normal samples and other samples. This was demonstrated by (I) the majority of identified ncMAPs being retained (7,513 out of 8,601) and (II) a comparable percentage of ncMAPs being detected in healthy samples through MS (36.46% with shared alleles versus 32% with all alleles) (see **Supplementary Figure S7**). To better understand the similarity between the HLA-binding motifs of the alleles represented in tumor-only samples and those represented in healthy samples, we generated a matrix of cosine distances of binding affinities and used t-SNE to reduce the dimensionality and visualize the data. Our results indicated a high level of

similarity between the two, further supporting the notion that the 65 alleles in the panel of normal samples were representative of the tumor-only alleles (**Supplementary Figure S7**).

However, the lack of ncMAP detection in the panel of normals does not confirm cancer selectivity owing to the sensitivity limitations of MS. Proper cancer selectivity assessment should be performed at the gene expression level in healthy tissues. Hence, we retrieved the parent gene expression values (in TPM) of the remaining ncMAPs from the Genotype-Tissue Expression project (GTEx v8) (52). We first compared the gene expression levels of the following two groups: (I) ncMAPs detected in the panel of normals by MS (blue), and (II) remaining ncMAPs without detection in the panel of normals (red). **Fig. 6b** shows significantly higher gene expression for ncMAPs detected in healthy tissues (blue) than for those that were left undetected (red). Moreover, to ensure low toxicity levels in normal tissues, we filtered ncMAPs to retain those with parent genes expressed below 1 TPM and without evidence of protein expression in any healthy tissue except the testis (immune-privileged site) (**Fig. 6c**). By applying this stringent filter, we identified 24 ncMAPs derived from genes not expressed or expressed only in trace amounts in healthy tissues. Of these, 17 were associated with proteins not detected in healthy tissues. **Table 1** provides a summary of these 17 cancer-selective ncMAPs, which we suggest as promising targets for clinical applications (see **Supplementary Table S3** for more details).

Discussion

The cartography of non-canonical antigen presentation revealed in our study arose from a harmonized large-scale analysis of immunopeptidomic data mapped to the human genome. Our innovations over the most recent trends in computational MS identified a diversity of peptides mapping to canonical and non-canonical translation products. We mapped deviations away from the reference proteome as mass shifts (PTMs) and applied a sequential approach to tackle the non-canonical immunopeptidome. Our proteogenomic

pipeline allowed the identification of thousands of ncMAPs (8,601) derived from non-coding regions of protein-coding genes with an FDR of 1%. This was accomplished by analyzing a large collection of publicly available studies using COD-dipp, a highly modular large-scale pipeline that bypasses the challenge of multi-omics requirements and large MS databases when identifying ncMAPs.

Recent studies have suggested that the immunopeptidome is rich in PTMs (63), which can have profound effects on immune tolerance. T cells can discriminate between modified and non-modified epitopes, which has been demonstrated in the case of ubiquitination (64), glycosylation (65), phosphorylation (1,66). T-cell reactivity to PTMs is an effect of their central tolerance escape from the thymus (67). PTMs may also alter proteolytic activity, and consequently, peptide presentation by the MHC system (68). The open-search component sheds light on several PTMs implicated in immunogenicity (serine N-terminal acetylation, cysteinylation, and cysteine tri-oxidation) and could provide insights for future studies on PTM-based epitopes. For instance, tri-oxidation of cysteine has the potential to alter the immune response (56); however, its mechanism of interaction with HLA molecules and T cells is still in its infancy**.** Additionally, T cells can discriminate between cysteinylated and unmodified cysteine residues (57). Likewise, N-terminal serine acetylation is known for multifunctional regulation, acting as a protein degradation signal, inhibitor of endoplasmic reticulum (ER) translocation, and mediator of protein complex formation. Methionine sulfone (methionine dioxidation) has been found to occur *in vivo* in *Proteus mirabilis* (69), a Gram-negative bacterium present in malignant cancers (70), although it can result from the use of a strong oxidizing agent.

The validity of ncMAPs was rigorously tested using retention time correlation (experimental vs. theoretical), orthogonal second-round search, mass accuracy, PTM retention time shifts, HLA binding prediction, and PSM comparison with previously published results. Twenty-five percent of the identified ncMAPs accounted, on average, for 81.36% of intersections when

compared with three other high-profile studies (6,13,16). In addition, COD-dipp revealed 6,433 new ncMAPs from protein-coding genes. Considering the high-quality and rigorous computational validation, the identification rate discrepancy is partly due to the performance of COD-dipp and the size of our dataset collection, making it the most exhaustive non-canonical library of MHC class I–associated peptides.

Our survey of the possible sources of ncMAPs revealed that 70.1% could be attributed to nORFs, IR, or frameshift mutations. We identified 597 ncMAPs downstream of known frameshift mutations in COSMIC, an understudied source of antigens in immunopeptidomic studies. Certainly, other biological processes not accounted for in this study could generate ncMAPs. For instance, mechanisms such as ribosomal slippage (11) and stop codon readthrough could explain some of the remaining ncMAPs (29.9%).

This study focuses on peptides from non-coding regions of the genome, referred to as non-canonical peptides. Unlike neoantigens, which derive from patient-specific mutations in cancer, these non-canonical peptides are not mutated and are present in both cancer and healthy individuals. Although their presence in healthy samples makes their tumor specificity less clear, non-canonical peptides tend to be more abundant in cancer cells than in healthy cells. Over two decades ago, Ishii et al. (71) purified an octamer non-canonical antigen (IPGLPLSL or pRL1a) associated with heat shock proteins (HSPs) and validated their findings using MS. The isolated octamer non-canonical antigen pRL1a was derived from the 5'-untranslated region of the *AKT* gene in leukemia and induced tumor rejection. To the best of our knowledge, this was the first demonstration of a non-canonical antigen that confers immunity. Subsequent studies have shown that HSPs are beneficial for anticancer vaccines (72) because they bind canonical/non-canonical antigens with tumor rejection properties that end up being presented by MHC I and II molecules (73).

Numerous studies have suggested various possible candidates for cancer vaccines over the past 2–3 decades, and each has failed, at least partly, due to the issue of specificity. We used a conservative definition of cancer selectivity that follows three iterative steps. We searched for the identified ncMAPs over a panel of 334 normal MS samples and confirmed a fraction (32%) of non-cancer-selective ncMAPs. The remaining fraction (5,843, 68%) contained both cancer-selective ncMAPs and non-cancer-selective ncMAPs that were not detected by MS. We used the expression levels of the ncMAPs' parent genes across 29 healthy tissues as a means of prioritization (6,14,16). ncMAPs whose parent genes were expressed in any normal tissue above a threshold of 1 TPM were not considered cancer selective. However, we caution that this definition excludes the consideration of 92% of protein-coding genes. We revealed 17 rigidly defined candidates as cancer-selective ncMAPs, originating from genes and proteins that were completely absent or available in trace amounts in healthy tissues. We hope that this offers a sufficiently stringent approach to reducing toxicity in clinical applications. We provide a complete breakdown of all detected ncMAPs in **Supplementary Table S3.** We report the parent gene and protein expression values across healthy tissue types from the GTEx cohort and Human Protein Atlas, respectively. Moreover, we report their cancer-selectivity status conditioned on a gene expression cutoff (1 TPM) and lack of protein expression in healthy tissues. This will allow the research community to make decisions regarding the peptides that should be retained or removed from their analyses. It is particularly important that we do not filter all peptides, as aberrant intron-retention and frame-shift mutations that are certainly cancer-specific may lie within these results and would not need this stringent filtering if found in subsequent studies.

Here, we provide a free and open-source informatics pipeline to study non-canonical peptides, along with a reservoir of potential targets that could be used in combination with T-cell therapies or cancer vaccines. We anticipate that this will help pave the way for future research on antigens from non-canonical sources and engage further oncology research on alternative sources of antigens.

We acknowledge that our study presents several limitations. First, our approach relies on a DDA MS, which is known for its dynamic range limitations. Thus, only the most abundant ncMAPs were identified. Moreover, owing to the technical limitations of MS, we require that our ncMAPs be at least 3 amino acids different from any known human protein. Thus, a substantial fraction could be eliminated, leading to underestimation of the non-canonical fraction. Second, because immunogenicity prediction is still in its infancy, the identified ncMAPs require further validation to qualify as tumor rejection–mediating antigens for clinical applications. Despite our efforts to identify cancer-selective targets, the toxicity of these peptides in healthy tissues requires further investigation.

Acknowledgments

This work was supported by the International Centre for Cancer Vaccine Science (Fundacja na rzecz Nauki Polskiej: MAB/3/2017) project is carried out within the International Research Agendas programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017453. This work was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising. The authors would like to thank ‘CI-TASK, Gdansk’, and the ‘PLGrid Infrastructure, Poland’ for providing their hardware and software resources.

Author Contributions

J. A. and G. B. conceived and initiated the project. J. A. and S. K. coordinated and supervised the project. G. B. and J. A wrote the first draft of the manuscript. G. B. collected

the online studies, developed the computational approach and software, processed the data, and coordinated the manuscript. G.B, J.A and H. G. created and revised the figures. The manuscript was reviewed and approved by all the authors. A. L., C. B., F. M. Z., C. P., H. A., A. R., D. J. H., T. R. H., and S. S., part of the 'KATY' consortium, and T. W., M. P., R. O., P. B. and K. L. revised the manuscript. G. B., D. B., K. W., A. P., D. R. G., R. F., S. K., J. A. Part of the 'International Centre for Cancer Vaccine Science' revised the manuscript.

References

1. Bassani-Sternberg M, Bräunlein E, Klar R, Engleitner T, Sinitcyn P, Audehm S, et al. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. Nat Commun. 2016;7:1–16.

2. Newey A, Griffiths B, Michaux J, Pak HS, Stevenson BJ, Woolston A, et al. Immunopeptidomics of colorectal cancer organoids reveals a sparse HLA class I neoantigen landscape and no increase in neoantigens with interferon or. J Immunother Cancer. 2019;7:309.

3. Ebrahimi-Nik H, Michaux J, Corwin WL, Keller GLJ, Shcheglova T, Pak H, et al. Mass spectrometry–driven exploration reveals nuances of neoepitope-driven tumor rejection. JCI Insight. 2019;4:e129152.

4. Blass E, Ott PA. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. Nat Rev Clin Oncol. 2021;18:215–29.

5. Pearlman AH, Hwang MS, Konig MF, Hsiue EH-C, Douglass J, DiNapoli SR, et al. Targeting public neoantigens for cancer immunotherapy. Nat Cancer. 2021;2:487–97.

6. Chong C, Müller M, Pak H, Harnett D, Huber F, Grun D, et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. Nat Commun. 2020;11:1293.

7. Malabat C, Feuerbach F, Ma L, Saveanu C, Jacquier A. Quality control of transcription start site selection by nonsense-mediated-mRNA decay. eLife. 2015;4:e06722.

8. Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, et al. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. eLife. 2014;3:e03528.

9. Rivero-Hinojosa S, Grant M, Panigrahi A, Zhang H, Caisova V, Bollard CM, et al. Proteogenomic discovery of neoantigens facilitates personalized multi-antigen targeted T cell immunotherapy for brain tumors. Nat Commun. 2021;12:6689.

10. Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, et al. Intron retention is a source of neoepitopes in cancer. Nat Biotechnol. 2018;36:1056–8.

11. Zook MB, Howard MT, Sinnathamby G, Atkins JF, Eisenlohr LC. Epitopes Derived by Incidental Translational Frameshifting Give Rise to a Protective CTL Response. J Immunol. 2006;176:6928–34.

12. Fang W, Wu C-H, Sun Q-L, Gu Z-T, Zhu L, Mao T, et al. Novel Tumor-Specific Antigens for Immunotherapy Identified From Multi-omics Profiling in Thymic Carcinomas. Front Immunol. 2021;12:748820.

13. Laumont CM, Daouda T, Laverdure J-P, Bonneil E, Caron-Lizotte O, Hardy M-P, et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat Commun. 2016;7:10238.

14. Laumont CM, Vincent K, Hesnard L, Audemard E, Bonneil E, Laverdure J-P, et al. Noncoding regions are the main source of targetable tumor-specific antigens. Sci Transl Med. 2018;10.

15. Ruiz Cuevas MV, Hardy M-P, Hollý J, Bonneil É, Durette C, Courcelles M, et al. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. Cell Rep. 2021;34:108815.

16. Ouspenskaia T, Law T, Clauser KR, Klaeger S, Sarkizova S, Aguet F, et al. Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. Nat Biotechnol. 2022;40:209–17.

17. Olsson N, Schultz LM, Zhang L, Khodadoust MS, Narayan R, Czerwinski DK, et al. T-Cell Immunopeptidomes Reveal Cell Subtype Surface Markers Derived From Intracellular Proteins. PROTEOMICS. 2018;18:1700410.

18. Demmers LC, Heck AJR, Wu W. Pre-fractionation Extends but also Creates a Bias in the Detectable HLA Class I Ligandome. J Proteome Res. 2019;18:1634–43.

19. Khodadoust MS, Olsson N, Chen B, Sworder B, Shree T, Liu CL, et al. B-cell lymphomas present immunoglobulin neoantigens. Blood. 2019;133:878–81.

20. Komov L, Kadosh DM, Barnea E, Milner E, Hendler A, Admon A. Cell Surface MHC Class I Expression Is Limited by the Availability of Peptide-Receptive “Empty” Molecules Rather than by the Supply of Peptide Ligands. PROTEOMICS. 2018;18:1700248.

21. Zeiner PS, Zinke J, Kowalewski DJ, Bernatz S, Tichy J, Ronellenfitsch MW, et al. CD74 regulates complexity of tumor cell HLA class II peptidome in brain metastasis and is a positive prognostic marker for patient survival. Acta Neuropathol Commun. 2018;6:18.

22. Bichmann L, Nelde A, Ghosh M, Heumos L, Mohr C, Peltzer A, et al. MHCquant: Automated and Reproducible Data Analysis for Immunopeptidomics. J Proteome Res. 2019;18:3876–84.

23. Chong C, Marino F, Pak H, Racle J, Daniel RT, Müller M, et al. High-throughput and Sensitive Immunopeptidomics Platform Reveals Profound Interferonγ-Mediated Remodeling of the Human Leukocyte Antigen (HLA) Ligandome. Mol Cell Proteomics. 2018;17:533–48.

24. Koumantou D, Barnea E, Martin-Esteban A, Maben Z, Papakyriakou A, Mpakali A, et al. Editing the immunopeptidome of melanoma cells using a potent inhibitor of

endoplasmic reticulum aminopeptidase 1 (ERAP1). Cancer Immunol Immunother. 2019;68:1245–61.

25. Marino F, Mommen GPM, Jeko A, Meiring HD, van Gaans-van den Brink JAM, Scheltema RA, et al. Arginine (Di)methylated Human Leukocyte Antigen Class I Peptides Are Favorably Presented by HLA-B*07. J Proteome Res. 2017;16:34–44.

26. Narayan R, Olsson N, Wagar LE, Medeiros BC, Meyer E, Czerwinski D, et al. Acute myeloid leukemia immunopeptidome reveals HLA presentation of mutated nucleophosmin. PloS One. 2019;14:e0219547.

27. Shraibman B, Kadosh DM, Barnea E, Admon A. Human Leukocyte Antigen (HLA) Peptides Derived from Tumor Antigens Induced by Inhibition of DNA Methylation for Development of Drug-facilitated Immunotherapy. Mol Cell Proteomics. 2016;15:3058–70.

28. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. Immunity. 2017;46:315–26.

29. Di Marco M, Schuster H, Backert L, Ghosh M, Rammensee H-G, Stevanovic S. Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices. J Immunol Baltim Md 1950. 2017;199:2639–51.

30. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. Mol Cell Proteomics MCP. 2015;14:658–73.

31. Khodadoust MS, Olsson N, Wagar LE, Haabeth OAW, Chen B, Swaminathan K, et al. Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. Nature. 2017;543:723–7.

32. Andreatta M, Nicastri A, Peng X, Hancock G, Dorrell L, Ternette N, et al. MS-Rescue: A Computational Pipeline to Increase the Quality and Yield of Immunopeptidomics Experiments. Proteomics. 2019;19:e1800357.

33. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, et al. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. PLoS Comput Biol. 2017;13:e1005725–e1005725.

34. Ternette N, Olde Nordkamp MJM, Muller J, Anderson AP, Nicastri A, Hill AVS, et al. Immunopeptidomic Profiling of HLA-A2-Positive Triple Negative Breast Cancer Identifies Potential Immunotherapy Target Antigens. Proteomics. 2018;18:e1700465.

35. Pearson H, Daouda T, Granados DP, Durette C, Bonneil E, Courcelles M, et al. MHC class I–associated peptides derive from selective regions of the human genome. J Clin Invest. 2016;126:4690–701.

36. Erhard F, Halenius A, Zimmermann C, L'Hernault A, Kowalewski DJ, Weekes MP, et al. Improved Ribo-seq enables identification of cryptic translation events. Nat Methods. 2018;15:363–6.

37. Newey A, Griffiths B, Michaux J, Pak HS, Stevenson BJ, Woolston A, et al. Immunopeptidomics of colorectal cancer organoids reveals a sparse HLA class I neoantigen landscape and no increase in neoantigens with interferon or MEK-inhibitor treatment. J Immunother Cancer. 2019;7:309.

38. Loffler MW, Mohr C, Bichmann L, Freudenmann LK, Walzer M, Schroeder CM, et al. Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma. Genome Med. 2019;11:28.

39. Marcu A, Bichmann L, Kuchenbecker L, Kowalewski DJ, Freudenmann LK, Backert L, et al. HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. J Immunother Cancer. 2021;9:e002071.

40. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, et al. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat Biotechnol. 2020;38:199–209.

41. da Veiga Leprevost F, Haynes SE, Avtonomov DM, Chang H-Y, Shanmugam AK, Mellacheruvu D, et al. Philosopher: a versatile toolkit for shotgun proteomics data analysis. Nat Methods. 2020;17:869–70.

42. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47:D941–7.

43. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. Nat Methods. 2017;14:513.

44. An Z, Zhai L, Ying W, Qian X, Gong F, Tan M, et al. PTMiner: Localization and Quality Control of Protein Modifications Detected in an Open Search and Its Application to Comprehensive Post-translational Modification Characterization in Human Proteome*. Mol Cell Proteomics. 2019;18:391–405.

45. Qiao R, Tran NH, Xin L, Chen X, Li M, Shan B, et al. Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. Nat Mach Intell. 2021;3:420–5.

46. Ivanov MV, Levitsky LI, Bubis JA, Gorshkov MV. Scavager: A Versatile Postsearch Validation Algorithm for Shotgun Proteomics Based on Gradient Boosting. PROTEOMICS. 2019;19:1800280.

47. Kent WJ. BLAT---The BLAST-Like Alignment Tool. Genome Res. 2002;12:656–64.

48. Zhang S, Hu H, Jiang T, Zhang L, Zeng J. TITER: predicting translation initiation sites by deep learning. Bioinformatics. 2017;33:i234–42.

49. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics. 2011;27:3423–4.

50. Erhard F, Dölken L, Schilling B, Schlosser A. Identification of the Cryptic HLA-I Immunopeptidome. Cancer Immunol Res. 2020;8:1018–26.

51. Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics. 2010;11:237.

52. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreservation Biobanking. 2015;13:311–9.

53. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015;347:1260419.

54. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun. 2014;5:5277.

55. Ojha R, Prajapati VK. Cognizance of posttranslational modifications in vaccines: A way to enhanced immunogenicity. J Cell Physiol. 2021;jcp.30483.

56. Trujillo JA, Croft NP, Dudek NL, Channappanavar R, Theodossis A, Webb AI, et al. The Cellular Redox Environment Alters Antigen Presentation. J Biol Chem. 2014;289:27979–91.

57. Parker R, Partridge T, Wormald C, Kawahara R, Stalls V, Aggelakopoulou M, et al. Mapping the SARS-CoV-2 spike glycoprotein-derived peptidome presented by HLA class II on dendritic cells. Cell Rep. 2021;35:109179.

58. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. 2020;48:W449–54.

59. Bullock TN, Eisenlohr LC. Ribosomal scanning past the primary initiation codon as a mechanism for expression of CTL epitopes encoded in alternative reading frames. J Exp Med. 1996;184:1319–29.

60. Starck SR, Tsai JC, Chen K, Shodiya M, Wang L, Yahiro K, et al. Translation from the 5′ untranslated region shapes the integrated stress response. Science. 2016;351:aad3867.

61. Goodenough E, Robinson TM, Zook MB, Flanigan KM, Atkins JF, Howard MT, et al. Cryptic MHC class I-binding peptides are revealed by aminoglycoside-induced stop codon read-through into the 3′ UTR. Proc Natl Acad Sci. 2014;111:5670–5.

62. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a revised and updated classification. BMC Immunol. 2008;9:1.

63. Yi X, Liao Y, Wen B, Li K, Dou Y, Savage SR, et al. caAtlas: An immunopeptidome atlas of human cancer. iScience. 2021;24:103107.

64. Gavali S, Liu J, Li X, Paolino M. Ubiquitination in T-Cell Activation and Checkpoint Inhibition: New Avenues for Targeted Cancer Immunotherapy. Int J Mol Sci. 2021;22:10800.

65. Malaker SA, Ferracane MJ, Depontieu FR, Zarling AL, Shabanowitz J, Bai DL, et al. Identification and Characterization of Complex Glycosylated Peptides Presented by the MHC Class II Processing Pathway in Melanoma. J Proteome Res. 2017;16:228–37.

66. Penny SA, Abelin JG, Malaker SA, Myers PT, Saeed AZ, Steadman LG, et al. Tumor Infiltrating Lymphocytes Target HLA-I Phosphopeptides Derived From Cancer Signaling in Colorectal Cancer. Front Immunol. 2021;12:723566.

67. Raposo B, Merky P, Lundqvist C, Yamada H, Urbonaviciute V, Niaudet C, et al. T cells specific for post-translational modifications escape intrathymic tolerance induction. Nat Commun. 2018;9:353.

68. Kacen A, Javitt A, Kramer MP, Morgenstern D, Tsaban T, Shmueli MD, et al. Post-translational modifications reshape the antigenic landscape of the MHC I immunopeptidome in tumors. Nat Biotechnol. United States; 2022;

69. Buzy A, Bracchi V, Sterjiades R, Chroboczek J, Thibault P, Gagnon J, et al. Complete amino acid sequence ofProteus mirabilis PR catalase. Occurrence of a methionine sulfone in the close proximity of the active site. J Protein Chem. 1995;14:59–72.

70. Lin L, Jia L, Fu Y, Zhao R, Huang Y, Tang C, et al. A comparative analysis of infection in patients with malignant cancer: A clinical pharmacist consultation study. J Infect Public Health. 2019;12:789–93.

71. Ishii T, Udono H, Yamano T, Ohta H, Uenaka A, Ono T, et al. Isolation of MHC Class I-Restricted Tumor Antigen Peptide and Its Precursors Associated with Heat Shock Proteins hsp70, hsp90, and gp96. J Immunol. 1999;162:1303–9.

72. Bloch O, Lim M, Sughrue ME, Komotar RJ, Abrahams JM, O'Rourke DM, et al. Autologous Heat Shock Protein Peptide Vaccination for Newly Diagnosed Glioblastoma: Impact of Peripheral PD-L1 Expression on Response to Therapy. Clin Cancer Res. 2017;23:3575–84.

73. Binder RJ. Immunosurveillance of cancer and the heat shock protein-CD91 pathway. Cell Immunol. 2019;343:103814.

Tables

ID	Peptide	Gene name	Mean expression in healthy tissues (TPM)	Number of healthy tissues with protein expression	Annotation
1	AFAPFPTQF	CXorf49B	0.01	0 of 56	Cancer selective
1	AFAPFPTQF	CXorf49	0.01	0 of 56	Cancer selective
1	AFAPFPTQF	RP11-402P6.15	0.10	0 of 56	Cancer selective
2	DYIHFVHHF	RP11-325B23.2	0.00	0 of 56	Cancer selective
3	EALSASQALYTR	HIST1H4L	0.04	43 of 56	
4	ELIKAFSK	GNGT1	0.05	1 of 56	
5	ESAGLFQVPR	SUN3	0.13	3 of 56	
6	EVEKILIQY	KCNU1	0.05	0 of 56	Cancer selective
7	EVPGAQGQQGPR	CTAG2	0.15	0 of 56	Cancer selective
7	EVPGAQGQQGPR	CTAG1B	0.03	0 of 56	Cancer selective
7	EVPGAQGQQGPR	CTAG1A	0.06	0 of 56	Cancer selective
8	FPVDVDHAVL	CTAG2	0.15	0 of 56	Cancer selective
8	FPVDVDHAVL	CTAG1B	0.03	0 of 56	Cancer selective
8	FPVDVDHAVL	CTAG1A	0.06	0 of 56	Cancer selective
9	ILSDNIRNL	C1orf94	0.14	0 of 56	Cancer selective
10	IPKDKSKNK	C2orf83	0.02	0 of 56	Cancer selective
11	KLLELIKAFSK	GNGT1	0.05	1 of 56	
12	KNNIYAFKI	RP11-231I13.2	0.01	0 of 56	Cancer selective
13	KTLHLTIVK	C12orf50	0.07	0 of 56	Cancer selective
14	KYLSRFRPK	TRPC5	0.08	0 of 56	Cancer selective
15	MVRSPEEGSLR	TEX19	0.13	0 of 56	Cancer selective
16	MVRSVSAAAR	HIST1H2BB	0.26	44 of 56	
17	MVRSVSAAARR	HIST1H2BB	0.26	44 of 56	
18	REEAPRGVRM	CTAG2	0.15	0 of 56	Cancer selective
18	REEAPRGVRM	CTAG1B	0.03	0 of 56	Cancer selective
18	REEAPRGVRM	CTAG1A	0.06	0 of 56	Cancer selective
19	SAGLFQVPR	SUN3	0.13	3 of 56	
20	SQVHKFFLL	OR9Q1	0.04	0 of 56	Cancer selective
21	SYGIYIYTY	SLC15A5	0.06	0 of 56	Cancer selective
22	TVSHQIIFY	EXD1	0.06	0 of 56	Cancer selective
23	VIQKVILVV	MGAT4D	0.03	0 of 56	Cancer selective
24	YYFILEHAKY	SOX30	0.29	0 of 56	Cancer selective

Table 1: List of cancer-selective non-canonical MHC-associated peptides. The mean parent gene expression in TPM was derived from 29 healthy tissues from the GTEx v8 dataset. The number of healthy tissues with protein expression was obtained from the Human Protein Atlas v22.0.

Figure legends

Figure 1: Infographics of immunopeptidomic datasets included in this study. a) Different types of cancer considered in this study with the number of samples and sample types per cancer type. **b)** Proportions of different mass spectrometry instruments used in this study. **c)** Antibodies used for immunoprecipitation (IP) **d)** Overall count of HLA alleles per HLA gene. **e)** Overall count of mass spectrometry immunopeptidomic samples per HLA allele.

Figure 2: COD-dipp: A new high-throughput pipeline for a deep interrogation of immunopeptidomic datasets. Samples are first analyzed with an open search strategy to detect the landscape of post-translational modifications (PTMs). A false localization rate (FLR) for the PTMs and false discovery rate (FDR) of 1% are applied. Simultaneously, the samples are analyzed using a novel *de novo* approach to identify non-canonical peptides. The *de novo* strategy trains a model per sample using quality-controlled peptide-spectrum matches from the MS-GF+ search engine to learn the direct interpretation of sample-specific mass spectra. The MS-GF+ results are split into three groups: training and testing to tune the hyperparameters and account for overfitting, and a validation group to approximate the accuracy per sample. *De novo* predicted peptides with an accuracy of at least 90% are sequentially mapped against the Human proteome (HP) then a 3-frame translation (3FT) database of protein-coding genes (1 mismatch allowed between leucine/isoleucine, i.e., Xle). Predicted *de novo* peptides matching any known protein are labeled “canonical”. Peptides mapping to the 3FT database with at least 3 amino acids mismatches from any known protein sequence are labeled “non-canonical”. Lastly, a second-round search is performed as a validation approach. Four of the most abundantly identified PTMs and a custom database consisting of ENSEMBL proteins and non-canonical peptides are considered. The resulting canonical and non-canonical peptides are controlled to an FDR of 1% and aligned to the hg38 human genome.

Figure 3: Landscape of post-translationally modified and non-canonical MHC class I–associated peptides (ncMAPs). Open search: a) Overview of post-translational modifications (PTMs) identified by open search (blue: spectra without PTMs, orange: spectra with a known UNIMOD PTM localized on a specific amino acid on the peptide. Green: The mass shift is localized, however the known PTM options do not fit the modified residue. Red: Otherwise). **b)** Most abundant “annotated PTMs” grouped by type. **Second-round search: c)** Fraction of canonical (dark gray) and non-canonical (light gray) MAPs in the immunopeptidome. **d)** Proportion of canonical (dark gray) and non-canonical (light gray) MAPs with/without post-translational modifications. **e)** Fraction of binders versus non-binders for both canonical and non-canonical MAPs using NetMHCpan 4.1.

Figure 4: Comparison of COD-dipp non-canonical MHC class I–associated peptides (ncMAPs) with other studies. Since the COD-dipp ncMAPs are restricted to the 3-frame translation (3FT) of protein-coding genes, sequences from the literature were aligned to the same 3FT database for comparison purposes. The intersection is based on genomic coordinates to deal with sequences that partially match (i.e., longer, shorter, or partially overlapping). Since the Venn is generated by overlapping genomic coordinates of the ncMAPs, the original counts for each study are listed from left to right (i.e., on the right-hand side of panel c, the notation 29/41 refers to 29 instances for Chong *et al.* 2020 and 41 for COD-dipp). **a)** Comparison with peptide-PRISM published ncMAPs at a 10% FDR. COD-dipp ncMAPs were restricted to 3 studies in common with Erhard *et al.* 2020. **b)** Comparison with Peptide-PRISM published ncMAPs at a 1% FDR. COD-dipp ncMAPs were restricted to 3 studies in common with Erhard *et al.* 2020. **c)** Comparison of the atlas of ncMAPs revealed by COD-dipp to 3 previous studies.

Figure 5: Origins of non-canonical MHC class I–associated peptides (ncMAPs). a) Peptide length distribution of canonical (dark gray) and non-canonical (light gray) MAPs. **b)** Annotation of ncMAPs across gene features. **c)** Analysis of ncMAPs that could originate from novel open reading frames (ORF). Upstream start codons of non-canonical MAPs are analyzed for their potential to initiate translation and produce ORFs (left-hand side) as a source of ncMAPs. The frequencies of different start codons for positively predicted translation initiation sites (TIS) are shown on the right-hand side. **d)** Analysis of ncMAPs from intronic regions that may originate from intron retention (IR) events. Translation of MAPs from IR sources should be in-frame with the corresponding upstream exons. **e)** Analysis of ncMAPs that could originate from frameshift mutations in cancer. ncMAPs are aligned to an in-silico translated protein database of COSMIC somatic frameshift mutations. **f)** Summary indicating whether the ncMAPs can be accounted for by any of the analyses conducted in panels c, d, or e.

Figure 6: Cancer selectivity of non-canonical MHC class I–associated peptides (ncMAPs). (a) Percentage of ncMAPs that were solely in healthy and/or tumor samples by MS (blue) and ncMAPs undetected in healthy samples by MS (red). **b)** Parent gene expression of ncMAPs in TPM in 29 healthy tissues from 17,382 samples (GTEx v8 dataset). ncMAPs are distributed over two groups: (I) ncMAPs detected in healthy samples by MS in blue, (II) ncMAPs undetected in healthy samples by MS in red. **c)** Parent gene expression of ncMAPs in TPM in 29 healthy tissues from 17,382 samples (GTEx v8 dataset). A limit on the gene expression (y-axis) of 1.2 TPM was set to visualize cancer-selective ncMAPs in black.

Figure 1

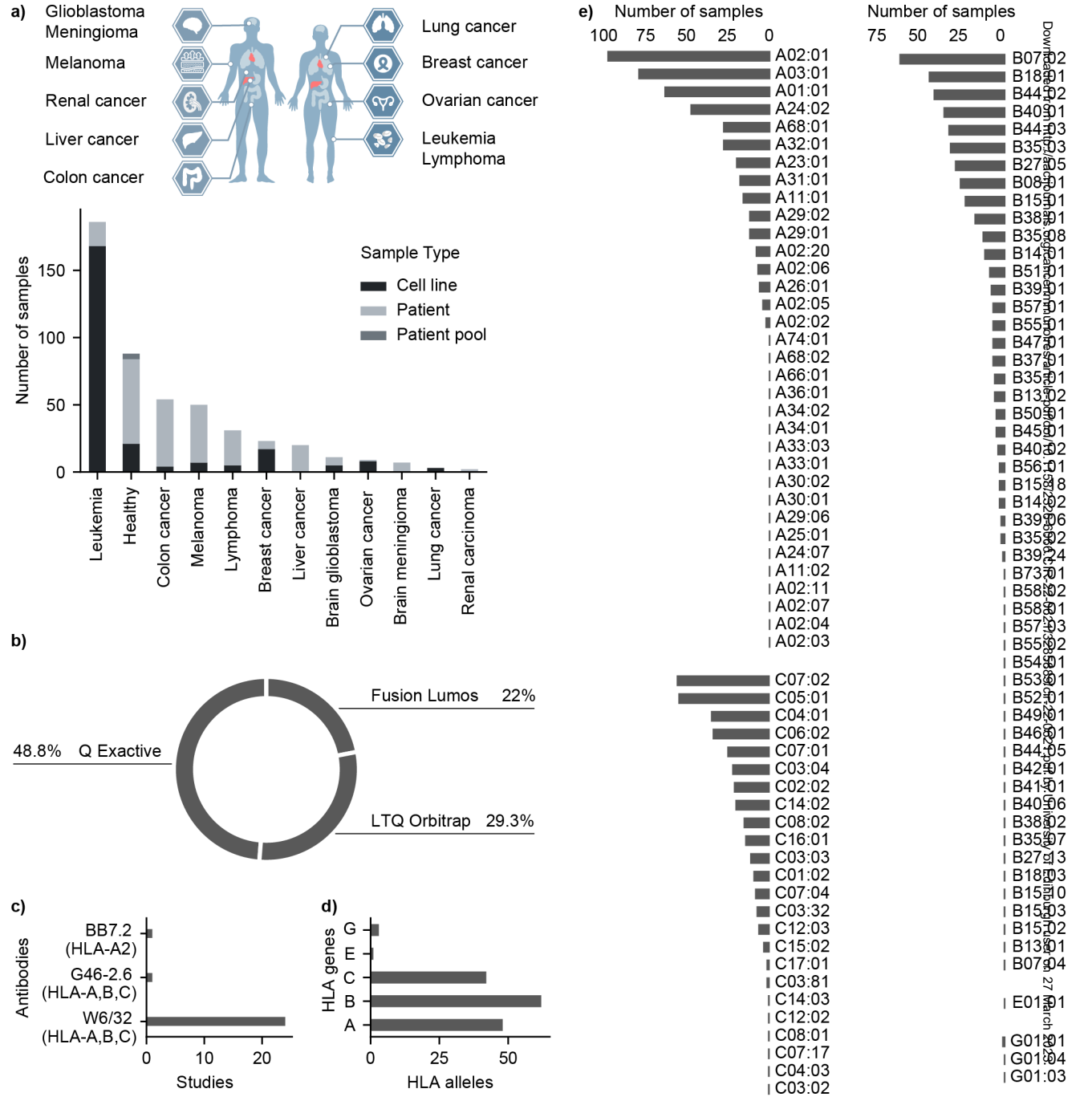


Figure 2

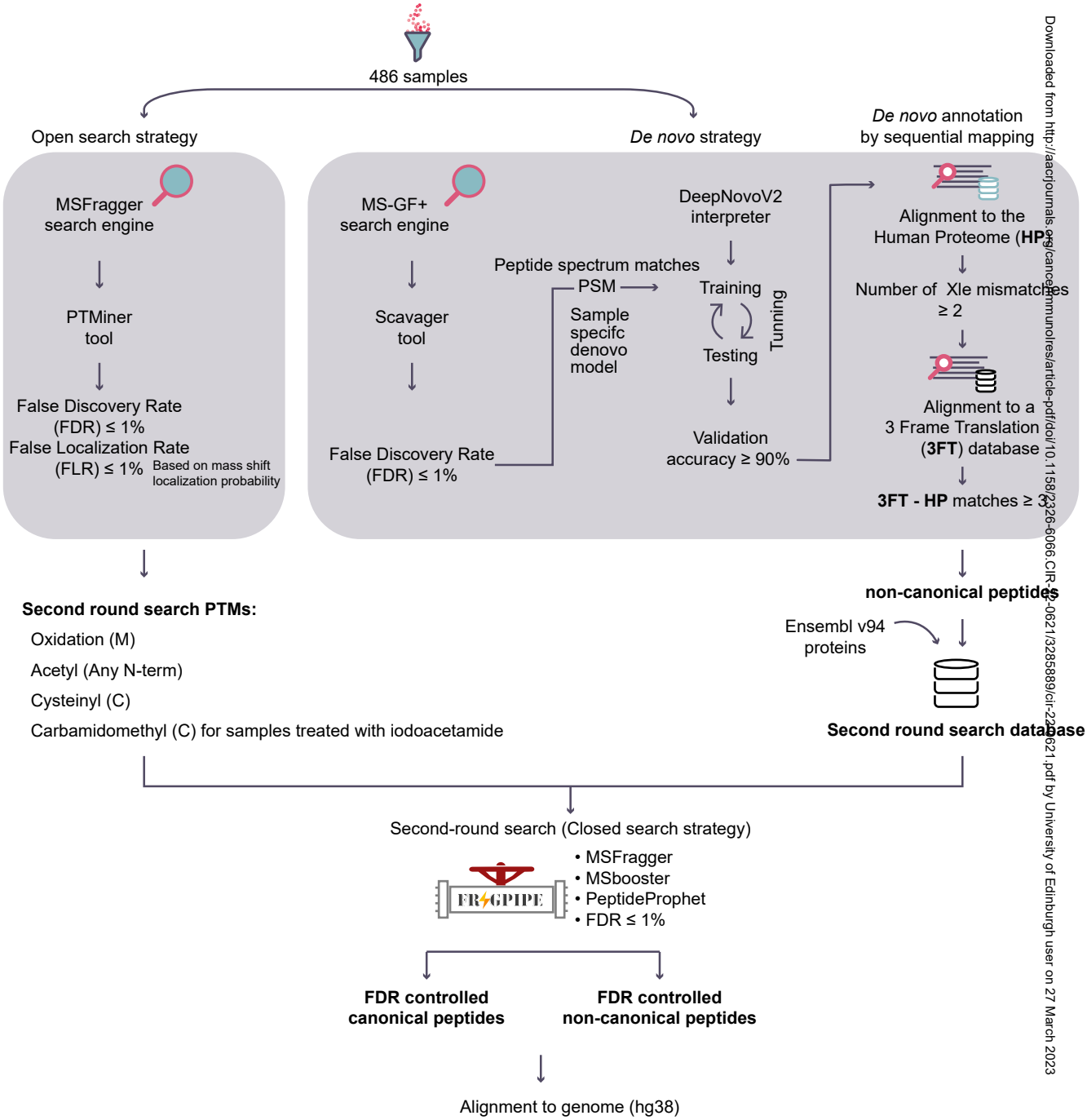
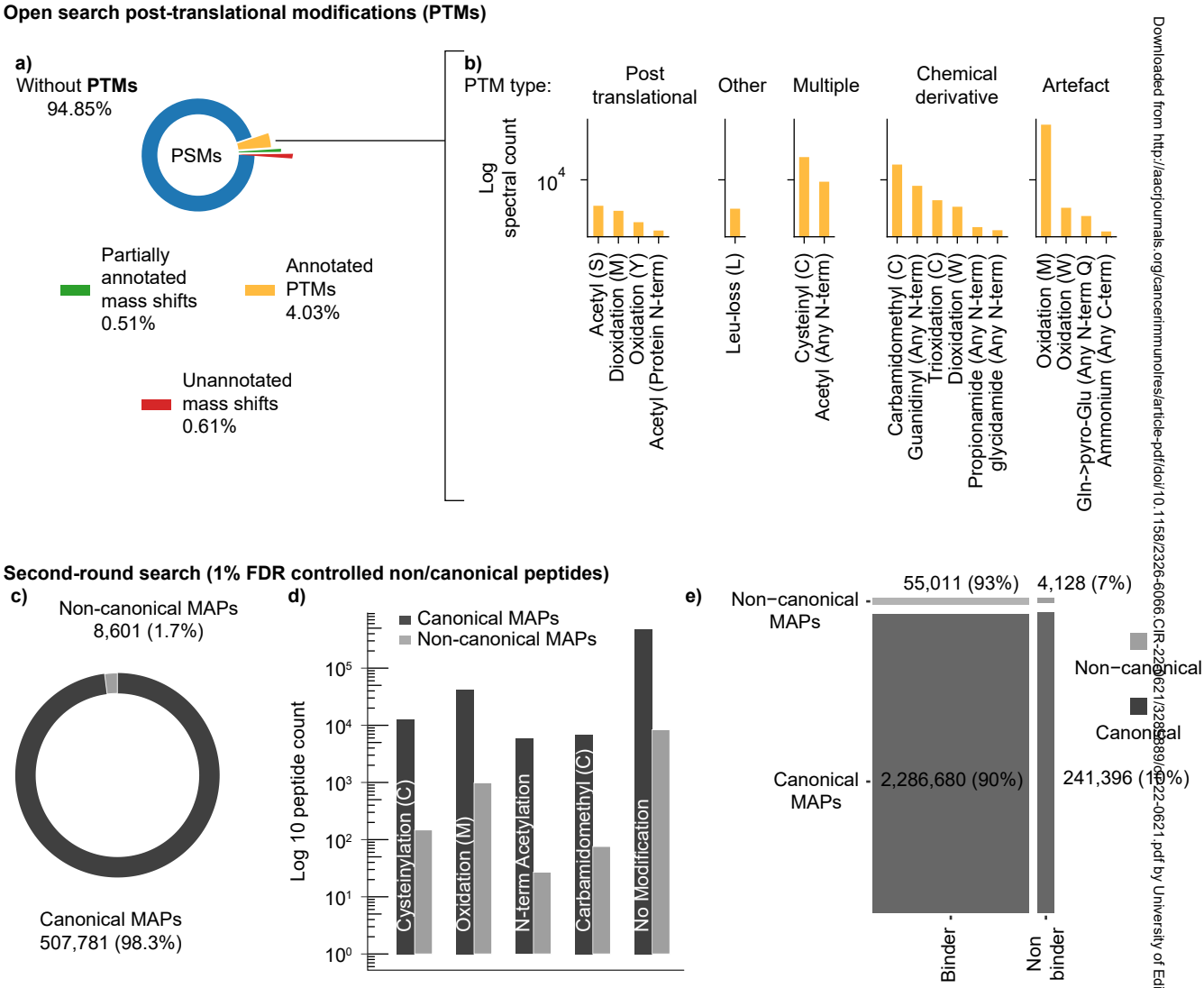
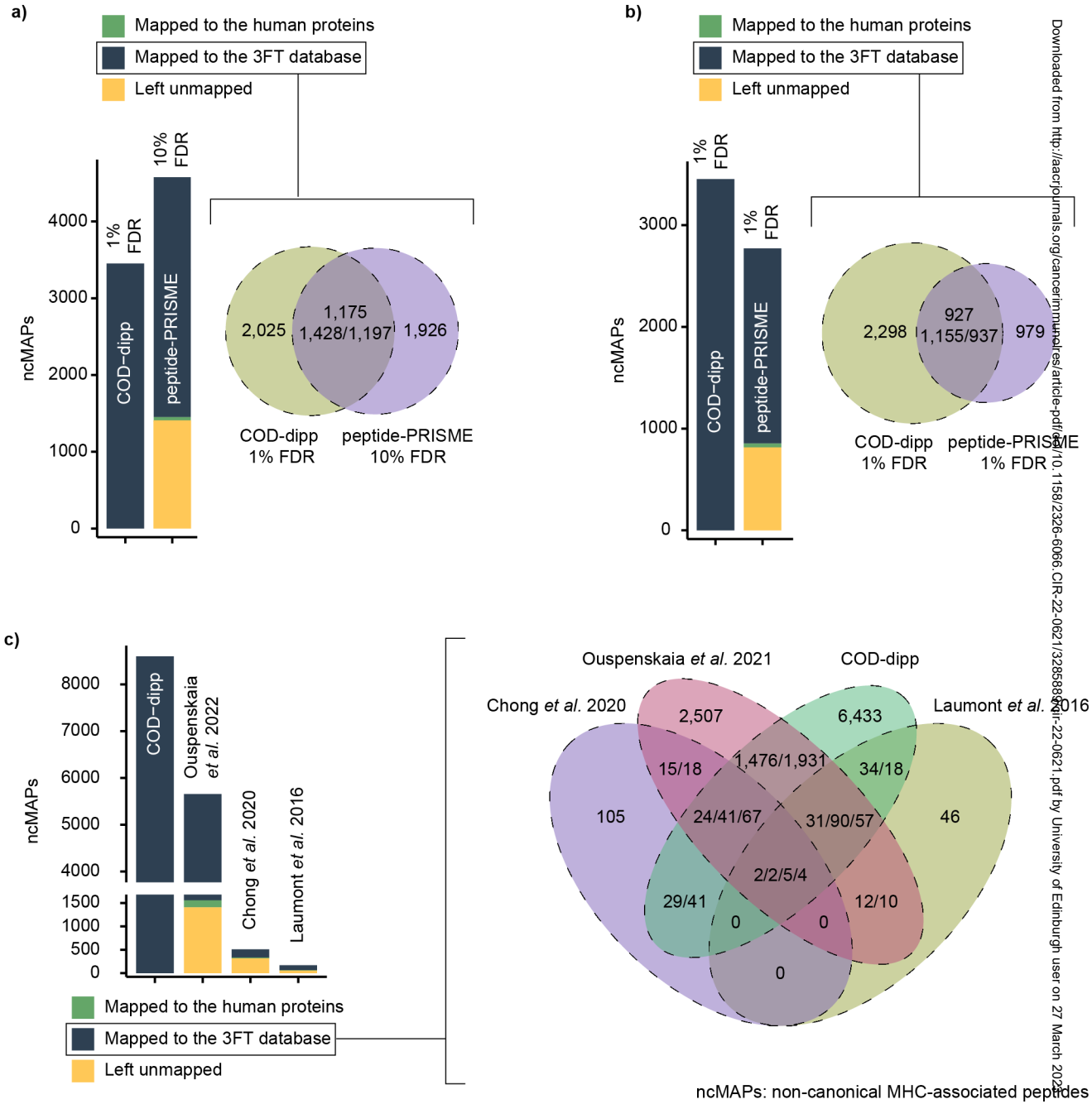


Figure 3





ncMAPs: non-canonical MHC-associated peptides
3FT database: 3-frame translation database

Figure 5

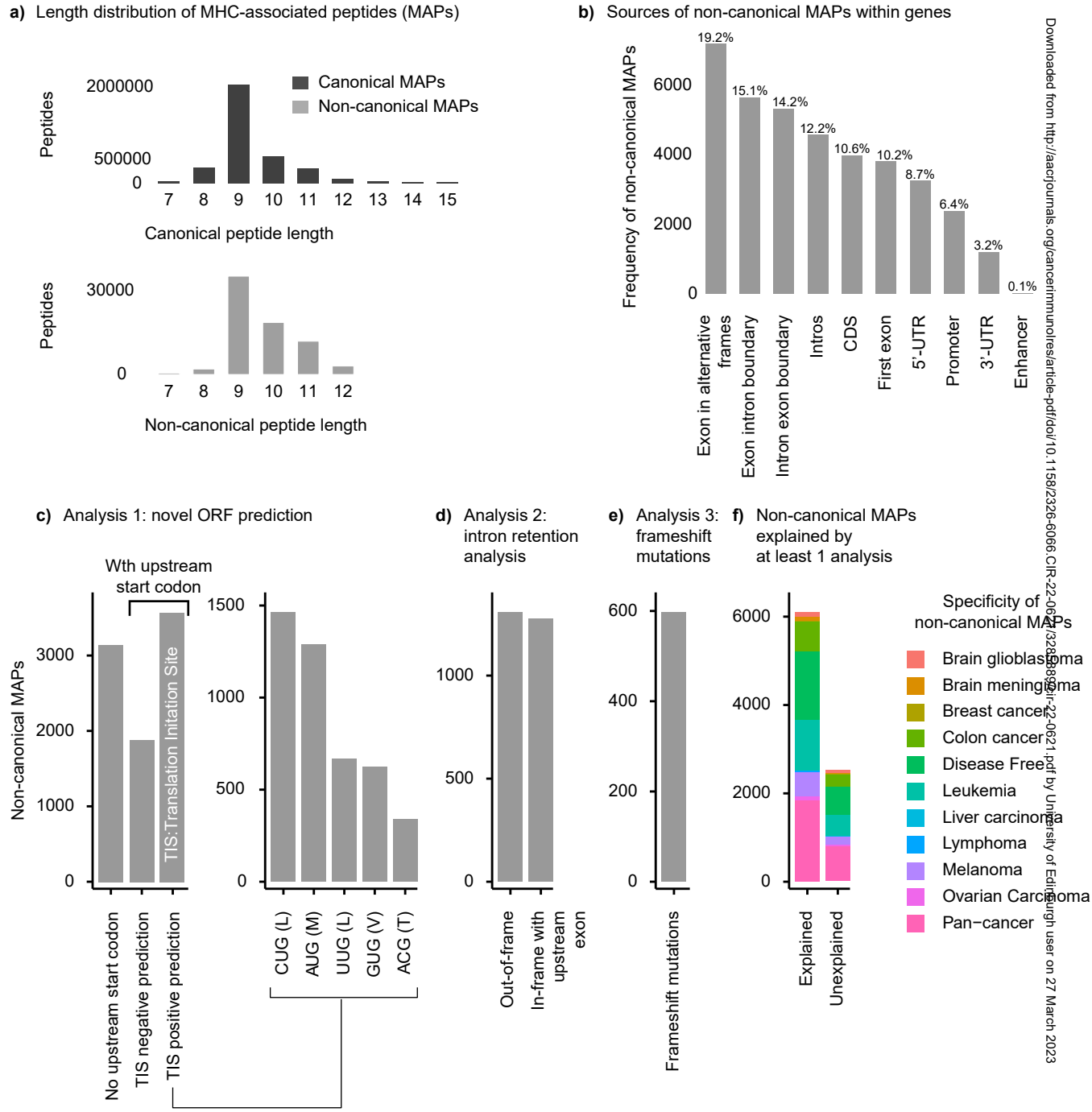


Figure 6