moPepGen: Rapid and Comprehensive Proteoform Identification

Chenghao Zhu[a,b,c,d,†,#], Lydia Y. Liu[a,b,e,f,g,†], Takafumi N. Yamaguchi[a,b,c], Helen Zhu[e,f,g], Rupert Hugh-White[a,b,c], Julie Livingstone[a,b,c], Yash Patel[a,b,c], Thomas Kislinger[e,f,#], Paul C. Boutros[a,b,c,d,e,#]

[a] Department of Human Genetics, University of California, Los Angeles, CA, USA
[b] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA, USA
[c] Institute for Precision Health, University of California, Los Angeles, CA, USA
[d] Department of Urology, University of California, Los Angeles, CA, USA
[e] Department of Medical Biophysics, University of Toronto, Toronto, Canada
[f] Princess Margaret Cancer Centre, University Health Network, Toronto, Canada
[g] Vector Institute for Artificial Intelligence, Toronto, Canada
[†] These authors contributed equally to this work
[#] Correspondence: chenghaozhu@mednet.ucla.edu, thomas.kislinger@utoronto.ca, pboutros@mednet.ucla.edu

Abstract

Gene expression is a multi-step transformation of biological information from its storage form (DNA) into functional forms (protein and some RNAs). Regulatory activities at each step of this transformation multiply a single gene into a myriad of proteoforms. Proteogenomics is the study of how genomic and transcriptomic variation creates this proteoform diversity, and is limited by the challenges of modeling the complexities of gene-expression. We therefore created moPepGen, a graph-based algorithm that comprehensively enumerates proteoforms in linear time. moPepGen works with multiple technologies, in multiple species and on all types of genetic and transcriptomic data. In human cancer proteomes, it detects and quantifies previously unobserved noncanonical peptides arising from germline and somatic genomic variants, noncoding open reading frames, RNA fusions and RNA circularization. By enabling efficient identification and quantitation of previously hidden proteins in both existing and new proteomic data, moPepGen facilitates all proteogenomics applications. It is available at: https://github.com/uclahs-cds/package-moPepGen.

Main Text

A single stretch of DNA can give rise to multiple protein products through genetic variation and through transcriptional, post-transcriptional and post-translational processes (*e.g.* RNA editing, alternative splicing and RNA circularization)[1–4]. The number of proteoforms rises combinatorically with the number of possibilities at each of these levels, so despite advances in proteomics technologies[8,9] much of the proteome is undetected in high-throughput studies[7].

One strategy to detect peptide sequences not present in canonical reference databases[5–7] is *de novo* sequencing and open search. Despite continued algorithmic improvements, this strategy is computationally expensive, has elevated false-negative rates, and leads to variant identification issues and difficult data interpretation[10,11]. As a result the vast majority of proteogenomics studies use non-canonical peptide databases incorporating DNA and RNA alterations[7]. These databases are often generated using DNA and RNA sequencing of the same sample, and this improves error rates relative to community-based databases (*e.g.*, UniProt[12], neXtProt[13] and the Protein Mutant Database[14]) by focusing the search space[7,15].

This type of sample-specific proteogenomics relies on the ability to predict all potential protein products generated by the complexity of gene expression. Modeling transcription, translation and peptide cleavage to fully enumerate the combinatorial diversity of peptides is computationally demanding. To simplify the search-space, existing methods have focused on enumerating only a single peptide caused by one variant or variant type[16–32], greatly increasing error rates (**Extended Data Table 1**). To fill this gap, we created a graph-based algorithm for exhaustive elucidation of protein sequence variations and subsequent *in silico* non-canonical peptide generation. This method is moPepGen (multi-omics Peptide Generator; **Figure 1A**).

moPepGen captures peptides that harbour any combination of small variants (*e.g.*, SNPs, indels, RNA editing sites) occurring on canonical coding transcripts, as well as on non-canonical transcript backbones resulting from novel open reading frames (ORFs), transcript fusion, alternative splicing, and RNA circularization (**Supplementary Figure 1**). It performs variant integration, *in silico* translation and peptide cleavage in a series of three graphs for every transcript, enabling systematic traversal across every variant combination (**Online Methods**; **Extended Data Figure 1A-D**). All three reading frames are explicitly modeled to efficiently capture frameshift variants and facilitate three-frame ORF search (**Extended Data**

Figure 2A). Alternative splicing events (*e.g.*, retained introns, *etc.*) and transcript fusions are modeled as subgraphs with additional small variants (**Extended Data Figure 2B**). Graphs are replicated four times to fully cover peptides of back-splicing junction read-through in circular RNAs (circRNAs; **Extended Data Figure 2C**). moPepGen outputs non-canonical peptides that cannot be produced by the chosen canonical proteome database. It documents all possible sources of each peptide to eliminate redundancy – for example where different combinations of genetic and transcriptomic events can produce the same peptide.

We first validated moPepGen using 1,000,000 iterations of fuzz testing (**Supplementary Figure 2**). For each iteration, a transcript model, sequences and set of variants of all supported types was simulated. Then non-canonical peptides generated by moPepGen were compared to those from a ground-truth brute-force algorithm. moPepGen demonstrated perfect accuracy and linear runtime complexity (4.7×10^{-3} seconds per variant) compared to exponential runtime complexity for the brute-force method (**Figure 1B-C**). A comprehensive non-canonical peptide database of germline polymorphisms was generated with 15 GB of memory in 3.2 hours on a 16-core compute node; the brute-force method was unable to complete this task.

Having established the accuracy of moPepGen, we next compared it to two popular custom database generators, customProDBJ[18] and pyQUILTS[22]. We tested all three methods on five prostate tumours with extensive multi-omics characterization[33–35]. We first evaluated the simple case of germline and somatic point mutations and indels. Most peptides (84.0 ± 0.9%) were detected by all three methods, with moPepGen being modestly more sensitive (**Extended Data Figure 3A**). Next we considered the biological complexity of alternative splicing, RNA editing, RNA circularization and transcript fusion. Only moPepGen was able to evaluate peptides generated by all four of these processes, and therefore 80.2 ± 2.1% of peptides were only identified by moPepGen (median ± MAD; **Extended Data Figure 3B**). By contrast only 3.2% of peptides were not identified by moPepGen, and these corresponded to specific assumptions around the biology of transcription and translation made by other methods (**Extended Data Figure 3C**). An example of the complex proteoforms identified only by moPepGen is a combination of a germline in-frame deletion followed by a substitution in *SYNPO2* (**Figure 1D**). Taken together, this benchmarking results demonstrate the robust and comprehensive nature of moPepGen.

To illustrate the utility of moPepGen for proteogenomic studies, we first evaluated it across multiple proteases. Using independent conservative FDR control across canonical and custom databases (**Online Methods**; **Supplementary Figure 3**)[7,36], we focused on detection of novel ORFs (*i.e.*, polypeptides from transcripts canonically annotated as noncoding) across seven proteases[36] in a deeply fractionated human tonsil sample[37] (**Supplementary Table 1**). moPepGen identified peptides representing 1,787 distinct ORFs previously thought to be noncoding. These tended to be shorter than canonical peptides (**Extended Data Figure 4A**) and easier to detect with the Arg-C protease (**Extended Data Figure 4B**), suggesting alternative proteases may enhance noncoding ORF detection (**Extended Data Figure 4C**). Fully 184 noncoding ORFs were detected by four or more preparation methods in this single sample, demonstrating moPepGen can identify novel proteins (**Extended Data Figure 4D-E**).

We next sought to demonstrate that moPepGen can benefit analyses in different species by studying germline variation in the C57BL/6N mouse[38,39]. DNA sequencing of the related C57BL/6J strain was used to identify 5,481 germline variants affecting protein coding genes and 15,475 affecting non-coding transcripts (**Extended Data Figure 5A**). Across the proteomes of three bulk tissues (cerebellum, liver and uterus), moPepGen identified 18 unique non-canonical peptides in protein-coding genes and 343 from noncoding ORFs (**Extended Data Figure 5B-D**; **Supplementary Table 2**). Thus moPepGen can support proteogenomics in non-human studies, identifying variants of protein-coding genes and novel proteins.

To evaluate the utility of moPepGen for somatic variation, we analyzed 375 human cancer cell line proteomes with matched somatic mutations and transcript fusions[40,41] (**Supplementary Data**). moPepGen processed each cell line in 2:58 minutes (median ± 1:20 minutes, MAD), generating 2,683 ± 2,513 (median ± MAD) potential non-canonical variant peptides per cell-line. The number of potential variant peptides varied strongly with tissue of origin, ranging from 838 – 16,255 (**Figure 2A**) and was driven largely by somatic mutations in protein-coding genes and by fusion events in non-coding genes (**Extended Data Figure 6A-C**). Searching the cell line proteomes identified 39 ± 27 proteogenomic peptides per cell line (**Online Methods**; **Supplementary Figure 4**). The majority of these were derived from noncoding transcript ORFs (**Extended Data Figure 6D**; **Supplementary Table 3**). Variant peptides arising from coding somatic mutations were less likely to be detected than those arising from transcriptional variation (**Extended Data Figure 6E-F**). Nevertheless, 26 genes had variant peptides detected

in cell lines from three or more tissues of origin, including the cancer driver genes like *TP53, KRAS* and *HRAS* (**Figure 2B**). Peptide evidence was also found for fusion transcripts involving cancer driver genes like *MET* and *STK11* (**Extended Data Figure 6G-H**). We validated non-canonical peptide PSMs by predicting MS2 spectra using Prosit[42] and verifying variant peptide MS2 spectra correlated better with these predictions than did their canonical peptide counterparts (**Online Methods**; **Extended Data Figure 6I**) As a positive control, coding variant peptide PSMs also showed high cross-correlations with Prosit-predicted MS2 spectra, on par with those of canonical PSM controls (**Extended Data Figure 6J**). Thus moPepGen can effectively and rapidly detect variant proteoforms arising from somatic variation.

Finally, to demonstrate the utility of moPepGen on the most complex gene-expression data we analyzed five primary prostate cancer samples with matched DNA whole-genome sequencing, ultra-deep ribosomal-RNA-depleted RNA-sequencing and mass-spectrometry-based proteomics[33–35]. moPepGen generated 1,382,666 ± 64,281 (median ± MAD) unique variant peptides per sample, spanning 115 variant combination categories (**Figure 2C**). Searching this database identified 206 ± 56 (median ± MAD) non-canonical peptides per sample. Of these 138 ± 28 affected protein-coding genes (**Extended Data Figure 7A**; **Supplementary Table 4**). All samples harboured proteins containing multiple variant peptides (9 ± 1.5 proteins per tumour; range of 2-6 variant peptides per protein; **Figure 2D**). Some detected peptides harboured multiple variants, including two from prostate-specific antigen (PSA from the *KLK3* gene; **Extended Data Figure 7B**). Germline SNPs were the major common cause of variant peptides on coding transcripts and alternative splicing events were the most common cause on noncoding transcripts (**Extended Data Figure 7C-E**). Nine genes showed recurrent detection of peptides caused by circRNA back-splicing (**Extended Data Figure 7F-G**), with 36/78 circRNA PSMs validated by *de novo* sequencing (**Supplementary Table 5**)[43]. We also detected four peptides from noncoding transcripts with the recently reported tryptophan-to-phenylalanine substitutants[44]. Thus moPepGen can identify peptides resulting from highly complex layers of gene-expression regulation.

moPepGen is a computationally efficient algorithm that enumerates transcriptome and proteome diversity across arbitrary variant types. It detects variant and novel ORF peptides across species, proteases and technologies. moPepGen integrates into existing proteomic analyses workflows, and can broadly enhance proteogenomics analyses for many applications.

References

1. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014).
2. Sinitcyn, P. *et al.* Global detection of human variants and isoforms by deep proteome sequencing. *Nat Biotechnol* (2023) doi:10.1038/S41587-023-01714-X.
3. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463 (2010).
4. Peng, X. *et al.* A-to-I RNA Editing Contributes to Proteomic Diversity in Cancer. *Cancer Cell* **33**, 817-828.e7 (2018).
5. Chick, J. M. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* **33**, 743–749 (2015).
6. Rodriguez, H., Zenklusen, J. C., Staudt, L. M., Doroshow, J. H. & Lowy, D. R. The next horizon in precision oncology: Proteogenomics to inform cancer diagnosis and treatment. *Cell* **184**, 1661–1670 (2021).
7. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nature Methods 2014 11:11* **11**, 1114–1125 (2014).
8. Creighton, C. J. Clinical proteomics towards multiomics in cancer. *Mass Spectrom Rev* (2022) doi:10.1002/MAS.21827.
9. Edwards, N. J. *et al.* The CPTAC data portal: A resource for cancer proteomics research. *J Proteome Res* **14**, 2707–2713 (2015).
10. Ma, B. & Johnson, R. De novo sequencing and homology searching. *Mol Cell Proteomics* **11**, (2012).
11. Fu, Y. Data analysis strategies for protein modification identification. *Methods in Molecular Biology* **1362**, 265–275 (2016).
12. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506–D515 (2019).
13. Lane, L. *et al.* neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res* **40**, (2012).
14. Kawabata, T., Ota, M. & Nishikawa, K. The Protein Mutant Database. *Nucleic Acids Res* **27**, 355–357 (1999).

15. Salz, R. *et al.* Personalized Proteome: Comparing Proteogenomics and Open Variant Search Approaches for Single Amino Acid Variant Detection. *J Proteome Res* **20**, 3353–3364 (2021).
16. Wang, X. *et al.* Protein identification using customized protein sequence databases derived from RNA-seq data. *J Proteome Res* **11**, 1009–1017 (2012).
17. Wang, X., Zhang, B. & Wren, J. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **29**, 3235–3237 (2013).
18. Wen, B., Li, K., Zhang, Y. & Zhang, B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nature Communications 2020 11:1* **11**, 1–14 (2020).
19. Sinitcyn, P., Gerwien, M. & Cox, J. MaxQuant Module for the Identification of Genomic Variants Propagated into Peptides. *Methods in Molecular Biology* **2456**, 339–347 (2022).
20. Van De Geer, W. S., Van Riet, J. & Van De Werken, H. J. G. ProteoDisco: a flexible R approach to generate customized protein databases for extended search space of novel and variant proteins in proteogenomic studies. *Bioinformatics* **38**, 1437–1439 (2022).
21. Umer, H. M. *et al.* Generation of ENSEMBL-based proteogenomics databases boosts the identification of non-canonical peptides. *Bioinformatics* **38**, 1470–1472 (2022).
22. Ruggles, K. V. *et al.* An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. *Molecular & Cellular Proteomics* **15**, 1060–1071 (2016).
23. Sheynkman, G. M., Shortreed, M. R., Frey, B. L. & Smith, L. M. Discovery and Mass Spectrometric Analysis of Novel Splice-junction Peptides Using RNA-Seq. *Molecular & Cellular Proteomics* **12**, 2341–2353 (2013).
24. Sheynkman, G. M. *et al.* Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics* **15**, 1–9 (2014).
25. Sheynkman, G. M., Shortreed, M. R., Frey, B. L., Scalf, M. & Smith, L. M. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J Proteome Res* **13**, 228–240 (2014).
26. Wen, B. *et al.* sapFinder: an R/Bioconductor package for detection of variant peptides in shotgun proteomics experiments. *Bioinformatics* **30**, 3136–3138 (2014).
27. Wen, B. *et al.* PGA: An R/Bioconductor package for identification of novel peptides using a customized database derived from RNA-Seq. *BMC Bioinformatics* **17**, 1–6 (2016).

28. Woo, S. *et al.* Proteogenomic database construction driven from large scale RNA-seq data. *J Proteome Res* **13**, 21–28 (2014).
29. Woo, S. *et al.* Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics* **14**, 2719–2730 (2014).
30. Cesnik, A. J. *et al.* Spritz: A Proteogenomic Database Engine. *J Proteome Res* **20**, 1826–1834 (2021).
31. Cifani, P. *et al.* ProteomeGenerator: A Framework for Comprehensive Proteomics Based on de Novo Transcriptome Assembly and High-Accuracy Peptide Mass Spectral Matching. *J Proteome Res* **17**, 3681–3692 (2018).
32. Kwok, N. *et al.* Integrative proteogenomics using ProteomeGenerator2. *bioRxiv* (2023) doi:10.1101/2023.01.04.522774.
33. Fraser, M. *et al.* Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359–364 (2017).
34. Chen, S. *et al.* Widespread and Functional RNA Circularization in Localized Prostate Cancer. *Cell* **176**, 831-843.e22 (2019).
35. Sinha, A. *et al.* The Proteogenomic Landscape of Curable Prostate Cancer. *Cancer Cell* **35**, 414-427.e6 (2019).
36. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry–based proteomics beyond trypsin. *Nature Protocols 2016 11:5* **11**, 993–1006 (2016).
37. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol Syst Biol* **15**, e8503 (2019).
38. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
39. Giansanti, P. *et al.* Mass spectrometry-based draft of the mouse proteome. *Nature Methods 2022 19:7* **19**, 803–811 (2022).
40. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature 2019 569:7757* **569**, 503–508 (2019).
41. Nusinow, D. P. *et al.* Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell* **180**, 387-402.e16 (2020).
42. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods 2019 16:6* **16**, 509–518 (2019).

43. Ma, B. Novor: real-time peptide de novo sequencing software. *J Am Soc Mass Spectrom* **26**, 1885–1894 (2015).
44. Pataskar, A. *et al.* Tryptophan depletion results in tryptophan-to-phenylalanine substitutants. *Nature 2022 603:7902* **603**, 721–727 (2022).
45. Andreev, D. E. *et al.* Non-AUG translation initiation in mammals. *Genome Biology 2022 23:1* **23**, 1–17 (2022).

Online Methods

Transcript Variant Graph

A transcript variant graph (TVG) is instantiated once per transcript with all associated variants incorporated. In a TVG, nodes are fragments of the transcript with reference or alternative nucleotide sequences, while edges are the opening or closing of variant nodes connecting them to the reference sequence, or simply the elongation of reference sequences. The TVG starts with three linear nodes of the entire transcript sequence representing the three reading frames, with *x* number of nucleotides removed from the transcript N-terminus, where *x* equals 0, 1, or 2. A variant is incorporated into the graph by cutting nodes in the graph at the variant's start and end positions and attaching a new node with the alternative sequence to the new upstream and downstream nodes. An in-frame variant is represented as a node that has incoming and outgoing nodes in the same reading frame subgraph, while frameshifting variants have incoming nodes and outgoing nodes in different subgraphs representing different reading frames. The outgoing reading frame index equals to $(S_{ref} - S_{alt})\ mod\ 3$, where S_{ref} is the length of the reference sequence and S_{alt} is the length of the alternative sequence. For transcripts with an annotated known canonical open reading frame (ORF), variants are only incorporated into the subgraph with the appropriate reading frame index (**Extended Data Figure 2A**). If frameshifting variants are present, variants are also incorporated into the subgraphs with the reading frame indices of the outgoing frameshift nodes. For transcripts without an annotated ORF, all variants are incorporated into all three subgraphs (**Extended Data Figure 2A**). Large insertions and substitutions as the result of alternative splicing events (*e.g.*, retained introns, alternative 3'/5' splicing, *etc.*) are represented as subgraphs that can carry additional variants (**Extended Data Figure 2B)**.

Variant Bubbles and Peptide Variant Graph

After the TVG has been populated with all variants, nodes that overlap with each other in transcriptional coordinates are aligned to create variant bubbles within which all nodes point to the same upstream and downstream nodes (**Extended Data Figure 1B**). This is done by first finding connection nodes in the TVG. A connection node is a reference node without any variants in its transcriptional coordinates that connects two variant bubbles after they are aligned. The root node is the first connection node, and the next connection node is found by looking for the first commonly connected downstream node with length of five or more

nucleotides that is outbound to more than one node (**Supplementary Note 1**; **Supplementary Figure 5**). Nodes between the two connection nodes are then aligned and a variant bubble is formed by generating all combinations of merged nodes so that they all point to the same upstream and downstream nodes (**Extended Data Figure 1B**). Overlapping variants in the variant bubble are eliminated because they are disjoint. The sequence lengths of nodes in the variant bubble are also adjusted by taking nucleotides from the commonly connected upstream and downstream nodes to ensure that they are multiples of three. A peptide variant graph (PVG) is then instantiated by translating the nucleotide sequence of each TVG node into amino acid sequences.

Peptide Cleavage Graph

The peptide variant graph is converted into the peptide cleavage graph (PCG), where each edge represents an enzymatic cleavage site (**Extended Data Figure 1C**). For connection nodes, all enzymatic cleavage sites are first identified, and the node is cleaved at each cleavage site. Because enzymatic cleavage sites can span over multiple nodes (for example, the trypsin exception of not cutting given K/P but cutting given WK/P), connection nodes are also merged with all downstream and/or upstream nodes and cut at additional cleavage sites if found. To optimize run time, different *merge-and-cleave* operations are used depending on the number of incoming and outgoing nodes, and the number of cleavage sites in a node (**Supplementary Figure 6**). Hypermutated regions where variant bubbles contain many variants and/or the lack of cleavage sites in connection nodes can result in an exponential increase in the number of nodes in the aligned variant bubble. To deal with this, we used a *pop-and-collapse* strategy, such that when *merge-and-cleave* is applied to a connection node, *x* number of amino acids are popped from the end of each node in the variant bubble. The popped nodes are collapsed if they share the same sequence. The pop-and-collapse operation is only applied when the number of nodes in a variant bubble exceed a user-defined cutoff.

Calling Variant Peptides

Variant peptides with the permitted number of miscleavages are called by traversing through the peptide cleavage graph. A node in a PCG can be in a different ORF when arriving at the node by traversing from a different incoming node. Thus, we use a *stage-and-call* approach that first visits all incoming nodes to determine the valid ORFs of a peptide node (**Supplementary Note 2**). *Stage-and-call* also allows cleavage-gain mutations and upstream

frameshift mutations to be carried over to the downstream peptide nodes. Peptide nodes are then extended by merging with downstream nodes to call variant peptides with miscleavages (**Supplementary Note 3**). For noncoding transcripts, novel ORF start sites, including those caused by start-gain mutations, are found by looking for any methionine (M) in all three subgraphs.

Fusion and Circular Transcripts

Most fusion transcript callers detect the fusion events between genes, thus causing ambiguity as to which transcripts of the genes are involved in a particular fusion event. We took the most comprehensive approach and endeavored to capture all possible variant peptides by assuming that the fusion event could happen between any transcript of the donor and accepter genes. Fusion transcripts are considered as novel backbones in graph instantiation, with an individual graph instantiated for each donor and acceptor transcript pair. Single nucleotide variants (SNV) and small insertion/deletions (indels) of both donor and acceptor transcripts are incorporated into the TVG. The translated and cleaved PCG is then traversed to call variant peptides, identical to a canonical transcript backbone. If the fusion breakpoint is in an intronic region, the intronic nucleotide sequence leading up to or following the breakpoint is retained as unspliced, and any SNVs or indels that it carries are also incorporated into the graph. The ORF start site of the donor transcript, if exists, is used when calling variant peptides. The fusion transcript is treated as a noncoding transcript if the donor transcript is annotated as noncoding.

Similar to fusion transcripts, circular RNAs (circRNAs) are also treated as novel backbones, and an individual graph is instantiated for each circRNA (**Extended Data Figure 2C**). A circular variant graph (CVG, a counterpart to a TVG) is first instantiated by connecting the linear sequence of the circRNA onto itself at the back-splice junction and incorporating SNVs and indels into the graph. Novel peptides could theoretically be translated from circRNAs, in the existence of a start codon, by ribosome readthrough across the back-splicing junction site. Thus, the circular graph is then extended into a linear format such that for each reading frame subgraph, three additional copies are appended to account for the potential of a maximum of three frameshifts due to linear sequence length. The extended circular variant graph is then translated to a PVG and converted to a PCG. Variant peptides are called by treating every circRNA as a noncoding transcript and all novel start codons are explored in all three open reading frames.

GVF File Format and Parsers

Genomic (single nucleotide polymorphisms [SNP], SNV, indel) and transcriptomic variants (fusion transcripts, RNA editing sites, alternative splicing transcripts, circRNAs) are first converted into gene-centric entries for each transcript that they impact. We defined the GVF (genetic variant format) file format, derived from the VCF (variant calling format) file format, to store all relevant information for each variant. The location of each variant is represented by gene ID along with offset from the start of the gene. Parsers have been implemented as part of moPepGen to parse the outputs of a variety of variant calling tools and convert them into the GVF file format. moPepGen was implemented in Python and supports easy extension and addition of new parsers.

Fuzz Testing and Brute Force Algorithm

A simulation-based fuzz testing framework is used to ensure the validity of moPepGen. In every fuzz test case, a transcript is simulated with a transcript model of various parameters (*e.g.*, coding or noncoding, positive or negative strand, selenoprotein or not, unknown start or stop codon position) and artificial sequence. Variant records associated with the transcript are then simulated, covering all supported variant types, including SNV, indel, fusion, alternative splicing, RNA editing and circRNA. Variant peptides produced by moPepGen are then compared to the output of the brute force algorithm, which iterates through all possible variant combinations to identify non-canonical peptides. The brute force algorithm also performs three-frame translation for noncoding transcripts. Software encoding this algorithm is available in the same repository as moPepGen.

Datasets

Cancer Cell Line Encyclopedia proteome

Proteomics characterization of 375 cell lines from the Cancer Cell Line Encyclopedia (CCLE) was obtained from Nusinow *et al.*, 2020[41]. Fractionated raw mass spectrometry data were downloaded from MassIVE (project ID: MSV000085836). Somatic SNVs and indels, and fusion transcript calls from the CCLE project were downloaded from the DepMap portal (https://depmap.org/portal, 22Q1). Somatic mutations were converted to GRCh38 coordinates from hg19 using CrossMap (v0.5.2)[46]. Gene and transcript IDs were assigned to each SNV/indel using the Variant Effect Predictor (VEP)[47] (v104) with genomic annotation GTF downloaded from GENCODE (v34)[48]. Fusion results were aligned to the GENCODE v34

reference by first lifting over the fusion coordinates to GRCh38 using CrossMap (v0.5.2). After lift-over, the records were removed if the donor or acceptor breakpoint location was no longer associated with the gene, if either breakpoint dinucleotides did not match with the reference, or if either gene ID was not present in GENCODE (v34).

Mouse proteome

Mass spectrometry-based characterization of the proteome of mouse strain C57BL/6N was obtained from Giansanti *et al.*, 2022[39]. Fractionated raw mass spectrometry data of the liver, uterus and cerebellum proteomes was downloaded from the PRIDE repository (project ID: PXD030983). Germline single nucleotide polymorphisms (SNPs) and indels were obtained from the Mouse Genomes Project[38] with GRCm38 VCFs downloaded from the European Variation Archive (accession: PRJEB43298). Germline SNPs and indels were mapped using VEP (v102) to Ensembl GRCm38 GTF (v102)[49].

Alternative protease and fragmentation proteome

A human tonsil tissue processed using ten different combinations of proteases and peptide fragmentation methods (ArgC_HCD, AspN_HCD, Chymotrypsin_CID, Chymotrypsin_HCD, GluC_HCD, LysC_HCD, LysN_HCD, Trypsin_CID, Trypsin_ETD, Trypsin_HCD) was obtained from Wang *et al.*, 2019[37]. Fractionated raw mass spectrometry data were downloaded from the PRIDE repository (project ID: PXD010154).

Prostate cancer proteome

The proteomics characterization of five prostate cancer tissues were obtained from Sinha *et al.*, 2019[35]. Raw mass spectrometry data were downloaded from MassIVE (project ID: MSV000081552). Germline SNPs and indels, as well as somatic SNVs and indels were obtained from the ICGC Data Portal (Project code: PRAD-CA). Variants were indexed using vcftools (v0.1.16)[50] and lifted over to GRCh38 using Picard tools (v2.19.0), followed by chromosome name mapping from the Ensembl to the GENCODE system using BCFtools (v1.9-1)[51]. Gene and transcript IDs were mapped by VEP (v104) to the GENCODE (v34) GTF. Raw mRNA sequencing data were obtained from Gene Expression Omnibus (accession: GSE84043). Transcriptome alignment was performed using STAR (v2.7.2) to reference genome GRCh38.p13 with GENCODE (v34) GTF and junctions were identified by setting the parameter --chimSegmentMin 10[52]. CIRCexplorer2 (v.2.3.8) was used to parse and annotate junctions for circular RNA detection[53]. Fusion transcripts were called using STAR-Fusion

(v1.9.1)[54]. RNA-editing sites were called using REDItools2 using paired RNA and DNA BAMs (v1.0.0)[55]. Alternative splicing transcripts were called using rMATS (v4.1.1)[56].

Canonical Database Search

All mass spectrometry raw files (.raw) were converted to the open format mzML using ProteoWizard (3.0.21258)[57]. The human GRCh38 canonical proteome database was obtained from GENCODE (v34), concatenated with common contaminants[58], and appended with reversed sequences to enable target-decoy false discovery rate (FDR) control. Mouse GRCm38 canonical proteome database was obtained from Ensembl (v102) and similarly processed. Database search was performed using Comet (v2019.01r5)[59]. All searches were performed with static modifications of cysteine carbamidomethylation, and up to three variable modifications of methionine oxidation, protein N-terminus acetylation and peptide N-terminus pyroglutamate formation. All searches were performed with fully specific digestion on both peptide ends for peptide lengths of 7-35 amino acids. Except for the tonsil samples processed with alternative enzymes, searches were performed with trypsin digestion and up to two miscleavages. For CCLE, static modification of tandem mass tag (TMT; 10plex) on the peptide N-terminus and lysine residues and variable modification of TMT on serine residues were additionally included, in accordance with the original study. For CCLE, searches were set to low resolution with 20 ppm precursor mass tolerance, 0.5025 Da fragment mass tolerance, and clear *m/z* range corresponding to TMT10plex, in accordance with the original publication. All other searches were of high-resolution label-free quantification (LFQ), with precursor mass tolerance of 20 ppm, 10 ppm, 30 ppm for the mouse proteome, tonsil proteome and prostate cancer proteome, respectively, and fragment mass tolerances of 0.025 Da for the tonsil proteome and 0.01 Da otherwise, in accordance with original publications. Tonsil proteomes were searched with the appropriate protease used in sample preparation, with maximum two miscleavages for Lys-C and Arg-C, three miscleavages for Glu-C and Asp-N and four miscleavages for chymotrypsin, as in the original publication[37].

Peptide level target-decoy FDR calculation was performed using the FalseDiscoveryRate module from OpenMS (v3.0.0-1f903c0)[60] using the formula (D+1)/(T+D), where D is the number of decoy peptide spectrum matches (PSMs) and T is the number of target PSMs. Peptides were filtered at 1% FDR, and all PSMs were removed from the corresponding mzML for subsequent non-canonical database search. Peptide quantification was performed using

FeatureFinderIdentification as part of OpenMS (v3.0.0-1f903c0)[61], using the “internal IDs only” strategy and with adjusted precursor mass tolerances as above, and otherwise default parameters for LFQ proteomics. The IsobaricAnalyzer module from OpenMS (v3.0.0-1f903c0) was used for the quantification of channel intensities for TMT proteomics, with no isotope correction due to the lack of correction matrix.

Non-canonical Database Generation

Human GRCh38 proteome reference files were obtained from GENCODE (v34) while mouse GRCm38 proteome reference files were obtained from Ensembl (v102). All non-canonical peptide databases were generated with trypsin digestion of up to two miscleavages and peptide lengths 7-25, except for databases used for alternative protease samples. Peptides from alternative translation were generated using *callAltTranslation*, including those with selenocysteine termination[62] or tryptophan-to-phenylalanine substitutants[44]. Peptides from noncoding ORFs were generated using *callNoncoding*, with ORF order as min and with or without alternative translation. Noncoding ORF peptide databases were also generated for each alternative protease used in processing of the tonsil proteome, with appropriate number of maximum miscleavages as outlined above.

Non-canonical peptide databases were generated for 376 cell lines from CCLE, 375 of which have non-reference channel proteomics characterization. This included all 10 cell lines in the bridge line and 366 non-reference cell lines with mutation data. Of the 378 non-reference channels across 42 plexes, three cell lines were duplicated, seven were in the bridge line, two didn’t have mutation or fusion information and additional eight didn’t have fusion information. Variant databases from all cell lines in a TMT plex, including the ten cell lines in the reference channel, were merged along with noncoding ORF peptides to generate plex-level databases. Plex-level databases were split into “Coding” (coding point mutations and transcript fusions), “Noncoding” (noncoding transcript novel ORFs) and “Noncoding Variant” (point mutations and transcript fusions on noncoding transcript ORFs) databases for tiered database search.

Non-canonical peptide databases for the proteome of mouse strain C57BL/6N was generated by calling variant peptides based on germline SNPs and indels, followed by merging with the noncoding ORF peptides database, and splitting into the three databases. The “Germline” database contained coding germline variations, “Noncoding” the novel ORFs, and “Noncoding-Germline” the peptides from germline variations on noncoding transcript ORFs. Genomic and

transcriptomic variants (*i.e.*, germline SNPs, germline indels, somatic SNVs, somatic indels, RNA editing sites, transcript fusions, alternative splicing and circRNA) from five prostate tumour samples were similarly used to call variant peptides, which were merged with the noncoding ORF peptide database and alternative translation peptide database and split into five databases. The "Variant" database included non-canonical peptides from coding transcripts with SNVs, indels, RNA editing, alternative splicing or transcript fusions. The "Noncoding" database included all peptides from noncoding transcript novel ORFs and noncoding peptides with any variants were in the "Noncoding Variant" database. The "Circular RNA" database included all peptides representing circRNA open reading frames (ORFs) with or without other sequence changes. The "Alt Translation" database included any peptides with selenocysteine insertion or tryptophan-to-phenylalanine substitutants[44].

Non-canonical Database Search

Non-canonical database searches were performed in nearly identical fashion as canonical proteome searches for each dataset, as described in detail above. Custom databases of peptide sequences were concatenated with the reverse sequence for FDR control. Non-canonical peptide searches with Comet (v2019.01r5) were set to "no cleavage" and did not permit protein N-terminus modifications or clipping of N-terminus methionine. Peptide-level FDR was set to 1% independently for each tier of non-canonical database, and PSMs of peptides that passed FDR were removed from the mzML for subsequent searches. Each tier of database thus had independent FDR control using database-specific decoy peptides, and a spectra is excluded from subsequent searches after finding its most probable match. This strategy has been shown to both alleviate the detection of false-positive non-canonical peptides due to joint FDR calculation with canonical peptides and to enable a highly conservative approach to non-canonical peptide detection[7,63]. For CCLE specifically, peptide detection and quantification were only considered for a cell line when the non-canonical peptide exists in the sample-specific database. For all datasets, quantified peptides were distinguished by charge and variable modifications, and detected but not quantified peptides were excluded from subsequent analysis.

Statistical Analysis and Data Visualization

All statistical analysis and data visualization were performed in the R statistical environment (v4.0.3), with visualization using BoutrosLab.plotting.general (v6.0.2)[64]. All boxplots show all

data points, the median (center line), upper and lower quartiles (box limits), and whiskers extend to the minimum and maximum values within 1.5 times the interquartile range. Schematics were created in Inkscape (v1.0) and Adobe Illustrator (27.8.1), and figures were assembled using Inkscape (v1.0).

Spectrum Visualization and Validation

Target PSM experimental spectra were extracted from mzML files using custom scripts based on pyOpenMS[65] and visualized in R. Theoretical spectra were generated from the target peptide sequences using the TheoreticalSpectrumGenerator module of OpenMS, and hyperscores between the experimental and theoretical spectra were calculated using the HyperScore module with the same parameters (*e.g.*, fragment mass tolerance) used during database search. Fragment ion matching between the experimental and theoretical spectra was performed using a similar approach to IPSA[66]. Theoretical spectra with predicted fragment peak intensities were generated using Prosit through the Oktoberfest[67] Python package using parameters (*e.g.*, fragmentation method and energy) in accordance with the original publication[42]. Similarities between the experimental spectra and the Prosit predicted spectra were estimated using cross-correlation[68], using the same parameters (*e.g.*, fragment_bin_offset) during database search with Comet. To assess the distribution of cross-correlation values for variant peptide PSMs, we randomly selected 1,000 PSMs from the canonical database search of each of the 42 TMT-plexes as control. Spectra matched to peptides derived from circRNA events were validated using the Novor algorithm through app.novor.cloud, using parameters (*e.g.*, fragmentation method, MS2 analyzer, enzyme, precursor and fragment mass tolerance) consistent with database searches[43].

Online Methods References

46. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
47. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
48. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res* **49**, D916–D923 (2021).
49. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res* **50**, D988–D995 (2022).
50. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
51. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).
52. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
53. Zhang, X. O. *et al.* Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res* **26**, 1277–1287 (2016).
54. Haas, B. J. *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol* **20**, 1–16 (2019).
55. Lo Giudice, C., Tangaro, M. A., Pesole, G. & Picardi, E. Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. *Nature Protocols 2020 15:3* **15**, 1098–1131 (2020).
56. Shen, S. *et al.* rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A* **111**, E5593–E5601 (2014).
57. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology 2012 30:10* **30**, 918–920 (2012).
58. Mellacheruvu, D. *et al.* The CRAPome: a Contaminant Repository for Affinity Purification Mass Spectrometry Data. *Nat Methods* **10**, 730 (2013).
59. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: An open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
60. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature Methods 2016 13:9* **13**, 741–748 (2016).
61. Weisser, H. & Choudhary, J. S. Targeted Feature Detection for Data-Dependent Shotgun Proteomics. *J Proteome Res* **16**, 2964–2974 (2017).

62. Berry, M. J., Harney, J. W., Ohama, T. & Lhatfield, D. Selenocysteine insertion or termination: factors affecting UGA codon fate and complementary anticodon:codon mutations. *Nucleic Acids Res* **22**, 3753–3759 (1994).
63. Wen, B., Li, K., Zhang, Y. & Zhang, B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nature Communications 2020 11:1* **11**, 1–14 (2020).
64. P'ng, C. *et al.* BPG: Seamless, automated and interactive visualization of scientific data. *BMC Bioinformatics* **20**, 42 (2019).
65. Röst, H. L., Schmitt, U., Aebersold, R. & Malmström, L. pyOpenMS: a Python-based interface to the OpenMS mass-spectrometry algorithm library. *Proteomics* **14**, 74–77 (2014).
66. Brademan, D. R., Riley, N. M., Kwiecien, N. W. & Coon, J. J. Interactive Peptide Spectral Annotator: A Versatile Web-based Tool for Proteomic Applications. *Mol Cell Proteomics* **18**, S193–S201 (2019).
67. Picciani, M. *et al.* Oktoberfest: Open-source spectral library generation and rescoring pipeline based on Prosit. *Proteomics* 2300112 (2023) doi:10.1002/PMIC.202300112.
68. Eng, J. K., Fischer, B., Grossmann, J. & MacCoss, M. J. A fast SEQUEST cross correlation algorithm. *J Proteome Res* **7**, 4598–4602 (2008).

Acknowledgments

The authors thank members of the Boutros and Kislinger labs for their continued support, particularly Dr. Amanda Khoo, Meinusha Govindarajan and Dr. Matthew Waas. The authors also thank Dr. James Wohlschlegel from UCLA Proteome Research Center and Dr. Mathias Wilhelm from Technical University of Munich. This work was supported by the NIH *via* awards P30CA016042, U01CA214194, U2CCA271894, U24CA248265, P50CA092131 and R01CA244729, by the Canadian Cancer Society *via* an Impact Grant (705649) and by the Canadian Institute of Health Research *via* a Project Grant (PJT156357). LYL was supported by a CIHR Vanier Fellowship and Ontario Graduate Scholarship. HZ was supported by a CIHR Doctoral Award. TK is supported through the Canadian Research Chair program. University Health Network was supported by the Ontario Ministry of Health and Long-Term care.

Data Availability

Data supporting the conclusions of this paper is included within it and its supplementary files.

Code Availability

moPepGen is publicly available at: https://github.com/uclahs-cds/package-moPepGen. Data processing, analysis and visualization scripts are available upon request.

Conflicts of Interest

PCB sits on the Scientific Advisory Boards of Intersect Diagnostics Inc., BioSymetrics Inc. and previously sat on that of Sage Bionetworks. All other authors declare no conflicts of interest.

Figure Legends

Figure 1: moPepGen is a graph-based algorithm that uncovers non-canonical peptides with variant combinations

a) moPepGen algorithm schematic. moPepGen is a graph-based algorithm that generates databases of non-canonical peptides that harbour genomic and transcriptomic variants (*e.g.,* single nucleotide variants [SNVs], small insertions and deletions [indels], RNA editing, alternative splicing, gene fusion, and circular RNAs [circRNAs]) from coding transcripts, as well as from novel open reading frames of noncoding transcripts. **b**) and **c**) moPepGen achieves linear runtime complexity when fuzz testing with SNVs only (**b**) and with SNVs and indels (**c**). **d**) A variant peptide from *SYNPO2* that harbours a small deletion and an SNV. Fragment ion mass spectrum from peptide spectrum match of the non-canonical peptide harbouring two variants (top, both) is compared against the canonical peptide theoretical spectra (left, theoretical spectra at the bottom) and against the variant peptide theoretical spectra (right, bottom). Fragment ion matches are colored, with b-ions in blue and y-ions in red.

Figure 2: moPepGen generates comprehensive non-canonical databases that support for proteogenomic analysis

a) Sizes of variant peptide databases generated by moPepGen using somatic single nucleotide variants, small insertions and deletions and transcript fusions for 376 cell lines from the Cancer Cell Line Encyclopedia project. Color indicates cell line tissue of origin. **b**) Genes with variant peptides detected in cell lines across three or more tissues of origin (bottom covariate). Barplot shows number of recurrences across tissues and color of heatmap indicates number of cell lines. **c**) Number of non-canonical peptides from different variant combinations (bottom covariate) generated using genomic and transcriptomic data from five primary prostate tumours, shown across four tiers of custom databases and grouped by the number of variant sources in combination. Alternative translation (Alt Translation) sources with $\geqslant$ 10 peptides are visualized.

d) Five variant peptides detected in one prostate tumour (CPCG0183) from the protein plectin (PLEC). Fragment ion matches are colored, with b-ions in blue and y-ions in red.

Extended Data Figure Legends

Extended Data Figure 1: Core graph algorithm of moPepGen

The graph algorithm of moPepGen implements the following key steps: **a**) A transcript variant graph (TVG) is generated from the transcript sequence with all variants associated. All three reading frames are explicitly generated to efficiently handle frameshift variants. **b**) Variant bubbles of the TVG are aligned and expanded to ensure the sequence length of each node is a multiple of three. **c**) Peptide variant graph (PVG) is generated by translating the sequence of each node of the TVG. **d**) Peptide cleavage graph is generated from the PVG in such a way that each node is an enzymatically cleaved peptide.

Extended Data Figure 2: Differential handling of noncoding transcripts, subgraphs, and circular RNAs

a) For coding transcripts, variants are only incorporated into the effective reading frames. For transcripts that are canonically annotated as noncoding, variants are added to all three reading frames to perform comprehensive three-frame translation. **b**) Subgraphs are created for variant types that involve the insertion of large segments of the genome, which can carry additional variants. **c**) The graph of a circular RNA is extended four times in order to capture all possible peptides that span the back-splicing junction site in all three reading frames. In the right panel, the nodes in rose-red harbour the variant 130-A/T and the nodes in yellow harbour 165-A/AC.

Extended Data Figure 3: moPepGen demonstrates comprehensive results and deliberate biological assumptions

a) and **b**) Non-canonical peptide generation results from benchmarking of moPepGen, pyQUILTS and customProDBJ using only point mutations and small insertions and deletions (indels; **a**), and with inputs from point mutations, indels, RNA editing, transcript fusion, alternative splicing, and circular RNAs (**b**). Top boxplot shows number of peptides in each set intersection and right barplot shows the total number of non-canonical peptides generated by each algorithm in five primary prostate tumour samples. **c**) Assumptions made by moPepGen for the handling edge cases that differ from other algorithms. Start-codon altering and splicing site altering variants are omitted. Transcripts with unknown stop codon do not have trailing peptide outputs. Stop codon altering variants do not result in translation beyond the transcript end.

Extended Data Figure 4: Detection of novel open reading frame peptides across proteases

a) Peptide length distributions after *in silico* digestion with seven enzymes, as indicated by color, of the canonical human proteome and three-frame translated noncoding transcript open reading frames (ORFs). Dotted lines indicate the 7-35 amino acids peptide length range used for database search. **b**) Noncoding peptide detection across ten enzyme-fragmentation methods. Top barplot shows number of peptides in each set intersection and right barplot shows the total number of non-canonical peptides from noncoding ORFs detected in each enzyme-fragmentation method, as indicated by covariate color. **c**) Optimal combinations of one to ten enzyme-fragmentation methods for maximizing the number of transcripts detected from the canonical proteome, or the number of ORFs detected from noncoding transcripts. Bottom covariate indicates the optimal combinations of enzyme-fragmentation methods from combinations of one to ten, with color indicating enzyme-fragmentation method. **d**) Noncoding transcript ORFs with peptides detected across four or more enzyme-fragmentation methods, with recurrence count shown in the right barplot. The color of the heatmap indicates the number of peptides detected per ORF per enzyme-fragmentation method. **e**) Example ORFs with coverage by multiple proteases are shown, with peptides tiled according to detection in each enzyme-fragmentation method, as indicated by covariate color. Representative fragment ion mass spectra of peptide spectrum matches shown, with theoretical spectra at the bottom and fragment ion matches colored (blue: b-ions, red: y-ions in).

Extended Data Figure 5: Germline non-canonical peptide detection in mouse strain C57BL/6N

a) Comparison of canonical and custom database sizes for the C57NL/6N mouse. Germline database includes single nucleotide polymorphisms (SNPs) and short insertions and deletions. **b**) Number of non-canonical peptides detected from each database in each tissue, with database indicated by color. **c**) Comparison of a variant peptide spectrum match (PSM) spectra (top, both) with the theoretical spectra of the canonical peptide counterpart (left, bottom) as well as the theoretical spectra of the variant peptide harbouring a SNP (right, bottom). Fragment ion matches are colored, with b-ions in blue and y-ions in red. **d**) Noncoding transcripts with open reading frames yielding two or more non-canonical peptides recurrently detected across tissues, with color indicating the number of peptides detected in each tissue.

Extended Data Figure 6: Proteogenomic investigation of the Cancer Cell Line Encyclopedia

a) Number of non-canonical peptides generated per cell line, with color indicating peptide source. Bottom covariate indicates tissue of origin. **b**) and **c**) Number of variant peptides per cell line with given number of variants in coding (**b**) and noncoding transcripts (**c**). **d**) Number of non-canonical peptides detected in each cell line, with color representing peptide source. Bottom covariate indicates tissue of origin. **e**) Per cell line, number of variant effect predictor (VEP) annotated intragenic coding mutations, mutations predicted to produce detectable non-canonical peptides and mutations detected through proteomics. **f**) Per cell line, number of transcript fusions, fusions theoretically able to produce detectable non-canonical peptides and fusions with detected peptide products. Color indicates tissue of origin. **g**) Fusion transcripts (upstream transcript gene symbol – downstream transcript gene symbol) with detected peptide products, with number of peptides shown across cell lines. Color indicates whether the upstream fusion transcript was coding or noncoding. **h**) Fragment ion mass spectrum from peptide spectrum match (PSM) of the non-canonical peptide at the junction of the *FLNB-SLMAP* fusion transcript. Peptide theoretical spectra is shown at the bottom and fragment ion matches are colored (blue: b-ions, red: y-ions in). **i**) Comparison of mass spectrum (top, both) from PSM of a non-canonical peptide with a single nucleotide variant against Prosit-predicted MS2 mass spectra based on the canonical counterpart peptide sequence (left, bottom) and the detected variant peptide sequence (right, bottom). Fragment ion matches are colored, with b-ions in blue and y-ions in red. **j**) Cross-correlation (Xcorr) distribution of PSMs of coding variant non-canonical peptides against Prosit-predicted MS2 mass spectra (solid lines, color indicate charge), in comparison with Xcorr of control canonical PSMs against Prosit-predicted mass spectra (dotted lines).

Extended Data Figure 7: Detection of non-canonical peptides from genomic variants, alternative splicing and circular RNAs

a) Number of detected non-canonical peptides in five primary prostate tumour samples per database tier (colored by database). **b**) Peptides as the result of a combination of two variants, with variant type indicated in left covariate and gene on the right. Heatmap shows presence of peptide across samples. **c**-**f**) Non-canonical peptide detection results across genes, with color of heatmap representing number of peptides detected per gene per sample. Barplot indicates

recurrence across samples, and when colored indicate variant type associated with the gene entry. The Variant database includes non-canonical peptides from coding transcripts with single nucleotide variant, small insertion and deletion, RNA editing, alternative splicing or transcript fusion (**c**). Noncoding database includes all peptides from noncoding transcript three-frame translation open reading frames (**d**) and noncoding peptides with any variants are included in the Noncoding Variant database (**e**). The Circular RNA database includes all peptides representing circular RNA open reading frames (ORFs) with or without other variants (**f**). Bottom covariate indicates prostate cancer sample. **g**) Mass spectrum from peptide spectrum match of a non-canonical peptide spanning the back-splicing junction between exon 29 and exon 24 of *MYH10*, reflective of circular RNA translation. Peptide theoretical spectra is shown at the bottom and fragment ion matches are colored (blue: b-ions, red: y-ions in).

Extended Data Tables

Extended Data Table 1: Feature comparison of custom database generation algorithms

Algorithm / Feature	moPepGen	customProDBj[16–18]	MaxQuant module[19]	ProteoDisco[20]	ProteomeGenerator2[31,32]	pypgatk[21]	pyQUILTS[22]	samplespecificDBGenerator[23–25]	sapFinder / PGA[26,27]	spliceDB[28,29]	Spritz[30]
DNA Single Nucleotide Variants	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Short DNA Sequence Variants	✓	✓		✓	✓	✓	✓		✓	✓	✓
Short RNA Sequence Variants	✓										✓
Alternative Splice Junctions	✓			✓	✓		✓	✓	✓	✓	
Fusion Transcripts	✓			✓	✓		✓	✓	✓	✓	
Circular RNAs	✓										
Supported Variant Combinations	Any combi-nations	None	All SNVs only	All SNVs only	All variants only	None	Single SNV with splicing or fusion	None	None	All variants only	None
Modular Support for New Input Formats	✓							✓			
Noncoding Transcript Three-Frame Translation	✓				✓	✓			✓		
Noncoding Transcript and Short Sequence Variants	✓				✓						
Alternative Translation W->P	✓										
Export Only Non-Canonical Proteotypic Peptides	✓			✓				✓	✓		
Summary of Database Generation	✓	✓			✓				✓		
Visualization of Database Generation	✓								✓		
Database Splitting for Tiered False Discovery Rate Control	✓										
Filter FASTA by RNA Abundance	✓				✓			✓			

Figure 1

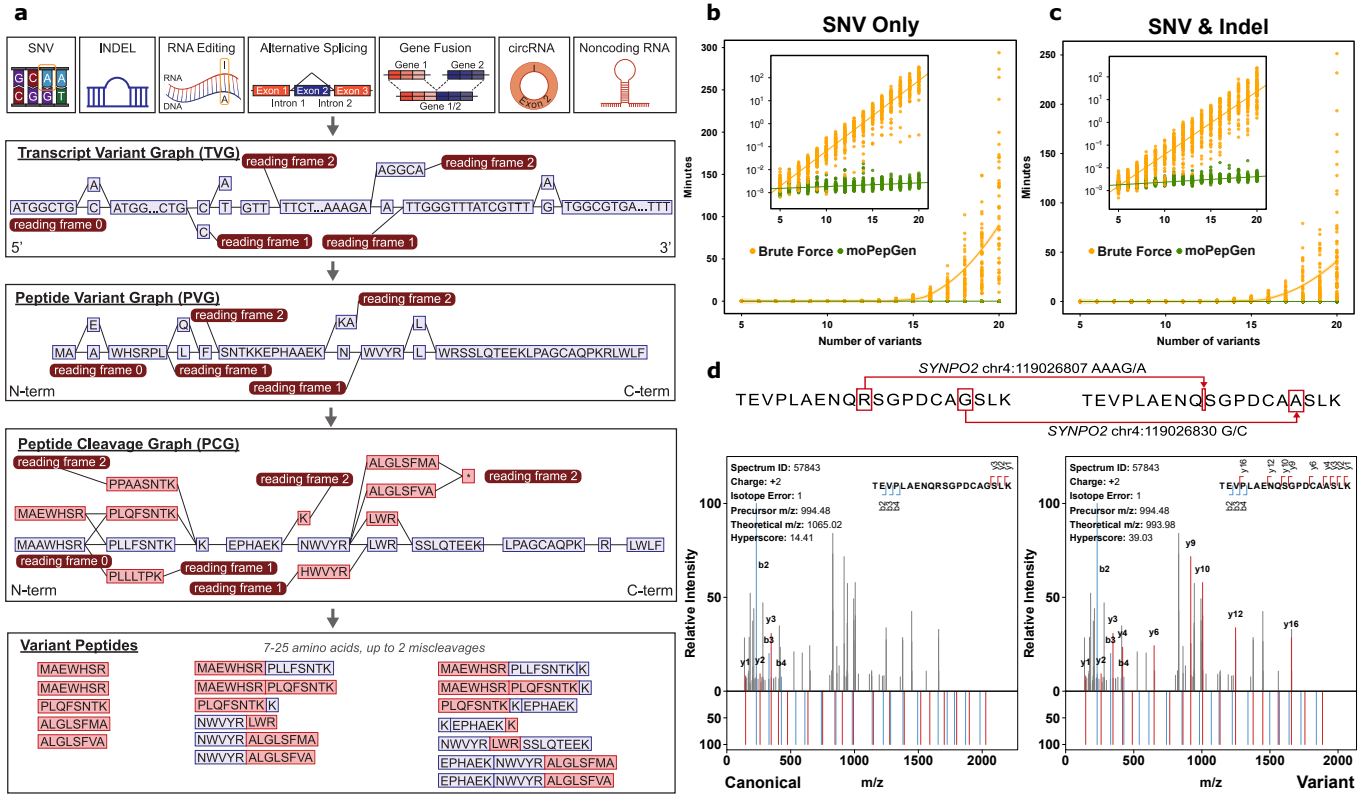
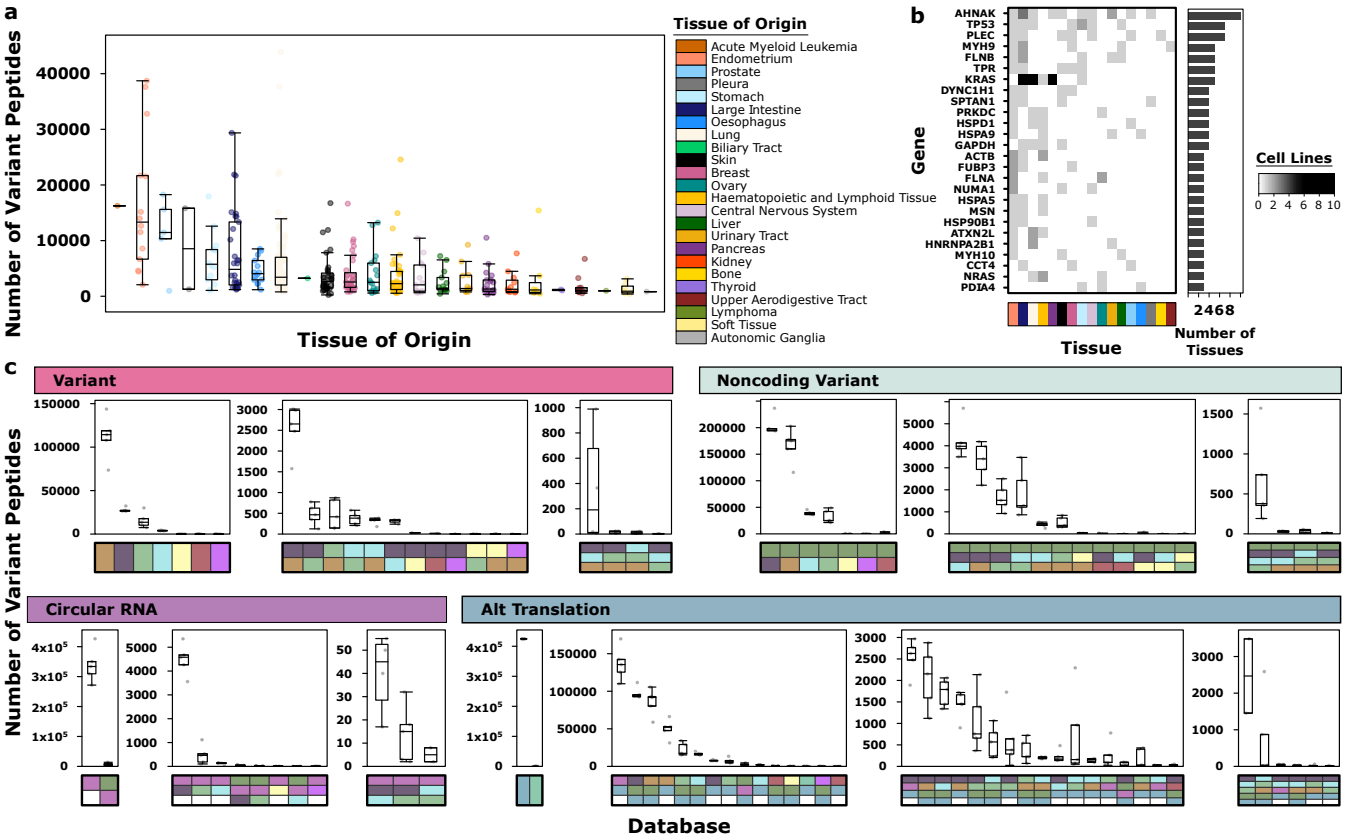
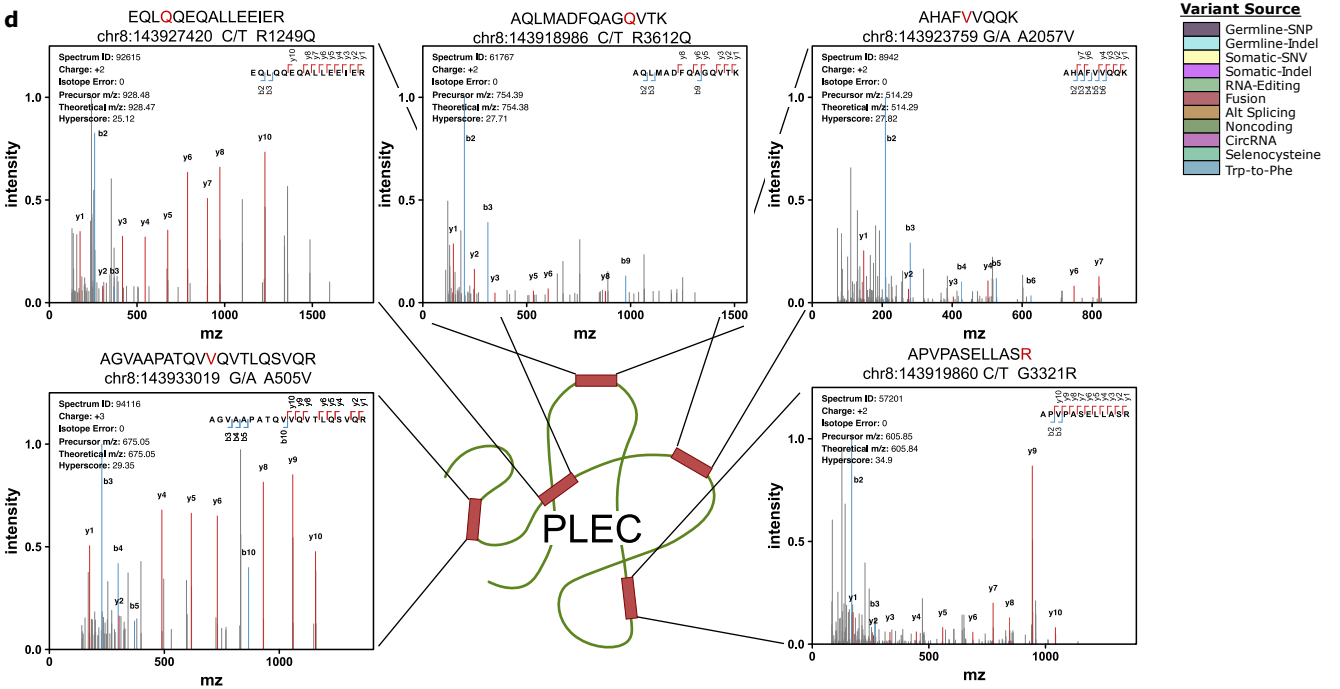

a
SNV
INDEL
RNA Editing
Alternative Splicing
Gene Fusion
circRNA
Noncoding RNA
Transcript Variant Graph (TVG)
reading frame 0
reading frame 1
reading frame 2
5'
3'
Peptide Variant Graph (PVG)
N-term
C-term
Peptide Cleavage Graph (PCG)
Variant Peptides
7-25 amino acids, up to 2 miscleavages
b
SNV Only
c
SNV & Indel
Minutes
Number of variants
Brute Force
moPepGen
d
SYNPO2 chr4:119026807 AAAG/A
TEVPLAENQRSGPDCAGSLK
TEVPLAENQSGPDCAASLK
SYNPO2 chr4:119026830 G/C
Spectrum ID: 57843
Charge: +2
Isotope Error: 1
Precursor m/z: 994.48
Theoretical m/z: 1065.02
Hyperscore: 14.41
Spectrum ID: 57843
Charge: +2
Isotope Error: 1
Precursor m/z: 994.48
Theoretical m/z: 993.98
Hyperscore: 39.03
Relative Intensity
m/z
Canonical
Variant

Figure 2


a
Number of Variant Peptides
Tissue of Origin
Tissue of Origin
Acute Myeloid Leukemia
Endometrium
Prostate
Pleura
Stomach
Large Intestine
Oesophagus
Lung
Biliary Tract
Skin
Breast
Ovary
Haematopoietic and Lymphoid Tissue
Central Nervous System
Liver
Urinary Tract
Pancreas
Kidney
Bone
Thyroid
Upper Aerodigestive Tract
Lymphoma
Soft Tissue
Autonomic Ganglia
b
Gene
AHNAK
TP53
PLEC
MYH9
FLNB
TPR
KRAS
DYNC1H1
SPTAN1
PRKDC
HSPD1
HSPA9
GAPDH
ACTB
FUBP3
FLNA
NUMA1
HSPA5
MSN
HSP90B1
ATXN2L
HNRNPA2B1
MYH10
CCT4
NRAS
PDIA4
Cell Lines
2468
Number of Tissues
Tissue
c
Variant
Noncoding Variant
Circular RNA
Alt Translation
Number of Variant Peptides
Database


d
EQLQQEQALLEEIER
chr8:143927420 C/T R1249Q
AQLMADFQAGQVTK
chr8:143918986 C/T R3612Q
AHAFVVQQK
chr8:143923759 G/A A2057V
AGVAAPATQVVQVTLQSVQR
chr8:143933019 G/A A505V
APVPASELLASR
chr8:143919860 C/T G3321R
PLEC
intensity
mz
Variant Source
Germline-SNP
Germline-Indel
Somatic-SNV
Somatic-Indel
RNA-Editing
Fusion
Alt Splicing
Noncoding
CircRNA
Selenocysteine
Trp-to-Phe