

1 **Title**

2 PoGo: Jumping from Peptides to Genomic Loci

3 **Authors**

4 Christoph N. Schlaffner*^{1,2}, Georg Pirklbauer¹, Andreas Bender², Jyoti S. Choudhary‡¹

5 **Affiliations**

6 ¹Proteomic Mass Spectrometry, Wellcome Trust Sanger Institute, Cambridge, UK; ²Centre for
7 Molecular Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK

8 **Contact Information**

9 *E-mail: christoph.schlaffner@sanger.ac.uk

10 ‡E-mail: jyoti.choudhary@sanger.ac.uk

11 **Abstract**

12 Current tools for visualization and integration of proteomics with other omics datasets are
13 inadequate for large-scale studies and capture only basic sequence identity information. We
14 developed PoGo for mapping peptides identified through mass spectrometry to a reference
15 genome to overcome these limitations. PoGo exhibited superior performance over other
16 tools on benchmarking with large-scale human tissue and cancer phosphoproteome
17 datasets. Additionally, extended functionality enables representation of single nucleotide
18 variants, post-translational modifications and quantitative features.

19 **Main Text**

20 Mass spectrometry (MS) and Next-generation sequencing (NGS) technologies have vastly
21 improved our understanding of the cross-talk between genome, transcriptome and
22 proteome, and contribute to a better understanding of the variations between healthy and
23 diseases states. Examples are the identification of new therapeutic target kinases in breast
24 cancer ¹ and detection of differentially regulated pathways and functional modules
25 potentially enabling patient stratification in ovarian cancer to inform therapeutic
26 management. ²

27 Substantial advances in MS technologies enable more complete identification and
28 quantification of proteomes, making these data more comparable to transcriptomics. Tools
29 like PGx ³ and iPiG ⁴ to readily visualize proteomics with corresponding RNA-sequencing data
30 on a reference genome are now increasingly indispensable. While iPiG heavily relies on the
31 annotation format used for UCSC genes, PGx uses sample specific protein sequence
32 databases derived from RNA-sequencing experiments and corresponding genomic
33 coordinates. Both tools require reformatting a reference genome annotation in order to
34 enable their mapping.

35 We developed PoGo to allow direct mapping to reference annotations and improve speed
36 and quality of mapping. PoGo leverages the annotated protein coding sequences (CDS)
37 together with a reference protein sequence database (protein-DB) to map peptides to their
38 genomic loci. Firstly, PoGo maps the genomic coordinates of CDSs onto the protein (Figure

39 1), thereby connecting the protein sequences to the genomic coordinate space. Database
40 search tools enable peptides to be identified from MS using a protein-DB.⁵ By using the
41 PoGo-indexed database genomic coordinates of a peptide are retrieved based on the
42 peptide's position within the protein (Figure 1A and Online Methods). PoGo further takes
43 advantage of distinct attribute columns of the output file formats, such as color, to indicate
44 uniqueness of a peptide across the genome, to show positions of post-translational
45 modifications, to allow quantitative comparison between multiple samples and conditions
46 linking this information to transcripts and genetic loci (Figure 2 and Online Methods). The
47 main genome browsers, Ensembl⁶, UCSC⁷, and BioDalliance⁸, however, have file size limits
48 for direct upload insufficient for large-scale proteogenomics. Our track-hub generator
49 application, therefore, enables seamless online visualization directly from PoGo output and
50 is crucial for open access proteomics of large datasets.

51 We first evaluated PoGo's performance on large-scale datasets using the proteogenomic
52 reanalysis of the draft human proteome maps.⁹ We used the filtered high stringency level
53 set comprising 233,055 unique peptides across 59 adult and fetal tissues. The mappings
54 were derived from the gene annotation set and protein coding translation sequences for
55 GENCODE (release 20)⁹ as GTF and FASTA files. All tools were run with standard parameter
56 settings and evaluated based on speed, memory usage and number of unique and correct
57 mappings. PoGo (94 seconds) was 6.9 and 96.4 times faster than PGx (651 seconds) and iPiG
58 (memory error after 9,064 seconds), respectively, and required 20% less memory compared
59 to PGx (9.7 GB and 11.9 GB respectively). These data show a major improvement of speed
60 and memory usage in addition to application with a readily available reference annotation.

61 In total 89% of mappings are common between PoGo and PGx. The 10.5% uniquely reported
62 by PGx can be explained by shifting into the correct frame, indicating incorrect assignment.
63 PoGo resulted in 89 completely unique mappings, 72 of these can be attributed to
64 incompletely annotated transcripts (CDS start/end not found). Additionally, 17 unique
65 mappings correspond to alternative splicing, immunoglobulin genes and multiple
66 overlapping mappings in a repeat region. For example, the peptide 'VPEPGCTKVPEPGCTK'
67 (missed cleavage between repeats of 8 amino acids) was mapped by PGx as two consecutive
68 loci in the *SPRR3* gene (Figure S1). PoGo, on the other hand, mapped the sequence 4 times
69 with the repeats overlapping each other (Figure 1C).

70 The fast and diverse mapping capabilities of PoGo, as shown above, prompted the current
71 integration of the algorithm into the PRIDE¹⁰ tool suite and soon into the OpenMS
72 framework⁹. This dataset also exemplifies the growing need to handle large numbers of
73 peptides. Therefore, we have generated tissue track-hubs at two different significance
74 thresholds from the draft human proteome maps allowing identification of genes and
75 transcripts unique to single tissues. The scaffolding protein CASS4, for example, was only
76 found in platelets (Figure S2). The genomic region of *RBP3*, only identified in retina, shows
77 full peptide support for all splice junctions (Figure S3).

78 The large number of single nucleotide variants in individuals can affect the protein
79 sequences and hinder identification of peptides through database searching against a
80 reference genome.¹⁰ Uniquely compared to other tools, PoGo is able to account for up to 2
81 non-synonymous variants in its mapping (Figure S4). Application with the draft human
82 proteome maps allowing 1 and 2 variants resulted in a 1.5- and 60.8-fold increase in runtime

83 (Figure S5). Unique mappings to single transcripts and single genes were reduced to 94.9%
84 and 84.1% while the number of peptides belonging to multiple genes increased
85 exponentially by 220.9% and 3,175.2% (Figures S5 and S6). The mapping of additional
86 repeats of the sequence 'VPEPGCTK' following application with mismatches were validated
87 through identified peptides in the sample (Figure 1C). This highlights the added value to
88 PoGo for mapping peptides to genomic loci with potential single nucleotide variants.

89 To demonstrate additional PoGo functionalities we chose the phosphoproteome of high-
90 grade serous ovarian cancer with isobaric labelling of 96 tumor samples, identifying 13,646
91 unique peptides with annotated phosphorylation sites (19,156 phosphopeptides).² PoGo
92 mapped 13,617 peptides to 15,944 genomic loci in 66.9 seconds; these could not be mapped
93 by PGx and iPiG. Only a small fraction, 0.2%, of the peptides could not be mapped due to
94 sequence differences of the originating proteins between RefSeq and GENCODE databases.
95 Compared to the other tools PoGo was able to use the annotated post-translational
96 modifications and color code them (Table S1) resulting in mappings for 99.8% of the
97 phosphopeptides with their respective localized phosphorylation sites on the reference
98 genome (Figure S7).

99 PoGo also integrates peptide quantitation with genomic loci through the GCT file format.
100 This allows comparative visualization of multiple samples in the Integrative Genomics
101 Viewer¹¹ and enables downstream quantitative analysis. The log₂-fold changes of
102 phosphopeptides between all 69 ovarian cancer samples and the pooled reference were
103 mapped with PoGo (Figure S7). As an example, *MAPK3* identified with multiple
104 phosphorylated sites in a single peptide and the associated fold changes across samples are
105 shown in Figure 2. To our knowledge PoGo is the only tool directly integrating quantitative
106 information for peptides with genomic coordinates.

107 **Discussion**

108 Our data show that PoGo represents a major advance for peptide-to-genome mapping
109 making it a cornerstone component of proteogenomics workflows. Although the examples
110 used here focus on human tissue and cancer cell lines, PoGo can be applied to any
111 proteomic study for which annotation of coding sequences in GTF format and translated
112 sequences in FASTA format are available. The additional functionalities such as allowing up
113 to two non-synonymous single nucleotide variants, mapping of post-translational
114 modifications and integration of quantitation distinguish it from other tools. Semi-
115 standardized file formats commonly used in genomics for in- and output as well as the
116 scalability for large datasets make PoGo an indispensable component of small and large-
117 scale multi-omics studies. The current integration into the PRIDE tool suite and our track-
118 hub generator application promote open access proteogenomics supporting studies
119 focusing on integration of gene, protein and post-translational modifications expression¹² in
120 the future. PoGo has been developed to cope with the rapid increase of quantitative high-
121 resolution datasets capturing proteomes and global modifications. Integration of orthogonal
122 genomics platforms with these datasets through PoGo will be valuable for large-scale
123 analysis such as personal variation and precision medicine studies.

124 **Acknowledgements**

125 This work is funded by the National Institute of Health grant (U41HG007234) to the
126 GENCODE project and Wellcome Trust grant (WT098051) to the Sanger Institute.

127 **Author Contributions**

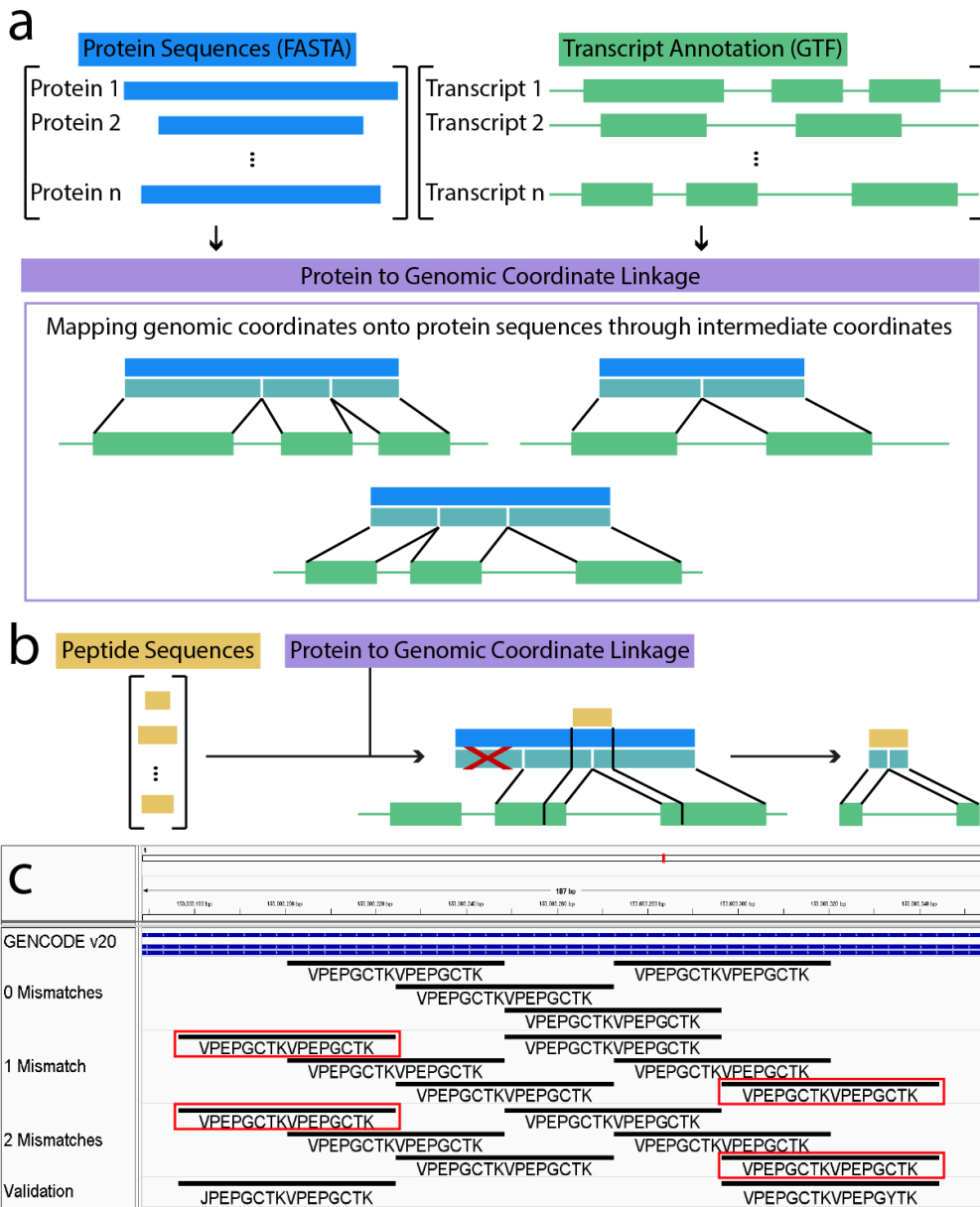
128 C.N.S conceived and designed the algorithms, implemented the genomic mapping algorithm,
129 performed comparisons with other algorithms and wrote the manuscript; G.P. implemented
130 the protein identification algorithm; A.B. and J.S.C. supervised the work and wrote the
131 manuscript.

132 **References**

- 133 1. Mertins, P. *et al. Nature* **534**, 55-62 (2016).
- 134 2. Zhang, H. *et al. Cell* **166**, 755-765 (2016).
- 135 3. Askenazi, M., Ruggles, K.V. & Fenyo, D. *J. Proteome Res.* **15**, 795-799 (2016).
- 136 4. Kuhring, M. & Renard, B.Y. *PLoS One* **7**, e50246 (2012).
- 137 5. Perez-Riverol, Y. *et al. Biochim. Biophys. Acta* **1844**, 63-76 (2014).
- 138 6. Kent, W.J. *et al. Genome Res.* **12**, 996-1006 (2002).
- 139 7. Yates, A. *et al. Nucleic Acids Res.* **44**, D710-716 (2016).
- 140 8. Down, T.A., Piipari, M. & Hubbard, T.J. *Bioinformatics* **27**, 889-890 (2011).
- 141 9. Wright, J.C. *et al. Nat Commun* **7**, 11778 (2016).
- 142 10. Vizcaino, J.A. *et al. Nucleic Acids Res.* **41**, D1063-1069 (2013).
- 143 11. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. *Brief. Bioinform.* **14**, 178-192
144 (2013).
- 145 12. Alvarez, M.J. *et al. Nat. Genet.* **48**, 838-847 (2016).

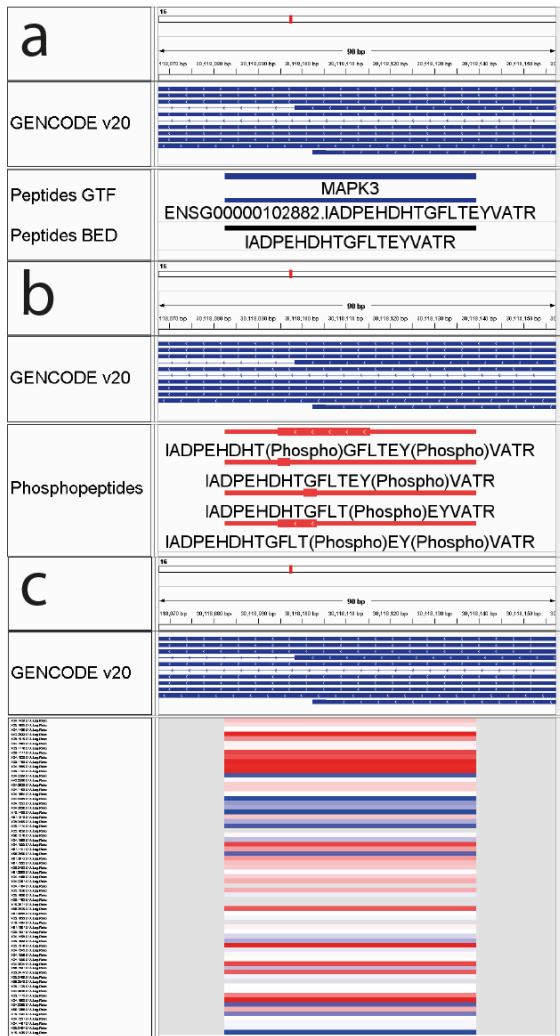
146

147 **Figure Legends**



148

149 **Figure 1** Schema of PoGo algorithm for mapping peptides through proteins to genomic loci.
 150 **(a)** Annotated protein coding transcripts in GTF format and respective translated protein
 151 sequences in FASTA format are integrated by PoGo through intermediate coordinates
 152 (turquoise), representing the exonic structure of the transcript within the protein. **(b)**
 153 Peptides, identified through searching mass spectrometry data against the protein sequence
 154 database, are mapped against the proteins. The position within the proteins then allow
 155 retrieval of overlapping coding exons and enable the calculation of the exact genomic
 156 coordinates. **(c)** Example mappings of PoGo for the overlapping repeat peptide
 157 'VPEPGCTKVPEPGCTK' in a genome browser (0 Mismatches). Application of PoGo allowing
 158 for up to two mismatches results in identification of two additional repeats (1 and 2
 159 Mismatches, red boxes) validated through mass spectrometry identified peptides of the
 160 exact sequence (Validation).



161

162 **Figure 2** Visualization of different PoGo output formats for the peptide
 163 'IADPEHDHTGFLTEYVATR' within the *MAPK3* gene. Genomic coordinates are shown at the
 164 top as x-axis. Gencode (v20) annotations of transcripts are indicated in blue. **(a)** Besides
 165 genomic location of the peptide the GTF format holds additional information, such as the
 166 gene name and gene identifier, while the BED output visualises uniqueness of the mapping
 167 across the genome. Here the black color indicates unique mapping to the gene *MAPK3*. **(b)**
 168 Genomic loci of post-translational modifications within a peptide, here phosphorylation
 169 identified through brackets in the sequence, are depicted through thick blocks spanning
 170 from the first and last modification site. The red color indicated in this output format the
 171 presence of phosphorylation. **(c)** Depiction of log₂-fold changes mapped for the example
 172 peptide to the genomic location across 69 ovarian cancer samples (y-axis). High values are
 173 shown in red while blue indicates low log₂-ratios.

174 **Online Methods**

175 **Software availability.** PoGo is implemented in C++. Executable files for Windows and Linux,
176 instructions for running PoGo, and explanations for each output format and their specific
177 visual attributes are available at <http://www.sanger.ac.uk/science/tools/pogo>. The source
178 code is available through <https://github.com/cschlaffner/PoGo>. The track hub generator
179 application, instructions for running it, explanation of visual components of resulting track
180 hubs, and a list of proteogenomic track hubs generated at the Wellcome Trust Sanger
181 Institute are available at <http://www.sanger.ac.uk/science/projects/proteogenomichubs>.
182 The perl source code is made available through
183 <https://github.com/cschlaffner/TrackHubGenerator>.

184 **PoGo algorithm.** PoGo is a multi-sample peptide-to-genome mapping tool taking as input
185 tab delimited lists of peptides identified through mass spectrometry (MS) with associated
186 number of peptide-to-spectrum matches (PSMs), quantitative value and sample identifier.
187 PoGo also requires a reference genome annotation in the General Transfer Format (GTF)
188 and translated protein coding sequences in FASTA format as input. The genomic coordinates
189 of annotated coding sequences are mapped onto their respective protein sequences.
190 Peptides identified through MS are then mapped against protein sequences accounting for
191 up to two mismatches. The genomic coordinates for each peptide are calculated based on
192 their position within the proteins. Each mapped peptide is additionally assigned the
193 associated sample identifier as well as the number of PSMs and the quantitative value.
194 Furthermore, post-translational modifications annotated in the peptide sequence are
195 mapped to their respective genomic coordinates and color coded for the type of
196 modification.

197 **Connecting protein sequences with genomic coordinates.** PoGo requires protein sequences
198 and gene annotations in FASTA and GTF format, respectively. Protein sequences have to be
199 connected to genes and transcripts through type specific identifiers (IDs). For each protein
200 sequence lines from the GTF file containing the transcript ID and feature-type CDS (coding
201 sequence) are extracted. The order of exons per transcript starts with the first exon in the
202 sequence reflecting the reading direction during translation, regardless of the strand,
203 resulting in a reverse order of genomic coordinates for transcripts on the reverse strand.
204 This way protein sequences and the exons match directionality. The exonic structure is
205 mapped onto the protein sequence through construction of protein exons. Let a transcript T
206 be a set of exons t_1, t_2, \dots, t_n where n is the number of exons and each exon t contains the
207 chromosome identifier, the start and end positions within the chromosome, S_t and E_t
208 respectively, the strand on which the transcript is annotated. The corresponding protein P is
209 defined as a set of protein exons p_1, p_2, \dots, p_n , where each protein exon p contains the start
210 and end positions, s_p and e_p respectively, within the protein sequence so that the protein is
211 mapped onto the transcript as $f: P \rightarrow T, p_i \rightarrow t_i$. For each protein in the FASTA file a map of
212 protein exons to genomic exons is generated in PoGo.

213 To account for frame shifts between genomic exons t_i and t_{i+1} each protein exon p also holds
214 information about the number of base pairs (bp) contributing to the codon of the first (N-
215 term) and last (C-term) amino acid as offsets $O=\{1,2,3\}$. In general, the N-term offset at the
216 beginning of a protein defined as $O(p_1(\text{N-term}))=3$ resulting in $O(p_n(\text{C-term}))=3$ for complete
217 annotations of coding transcripts. In instances where the annotation is missing a start or end
218 codon the offsets may vary and is identified through the annotated frame. C-term offsets

219 $O(p_i(\text{C-term}))$ for each protein exon p are calculated based on the length of the genomic
220 exon $L(t_i)$ and the offset of the N-term $O(p_i(\text{N-term}))$ so that $O(p_i(\text{C-term})) = X = L(t_i) \bmod 3 -$
221 $O(p_i(\text{N-term})) + 3$ with the exception $O(p_i(\text{C-term})) = X \bmod 3$ for $X > 3$. N-term offsets of
222 following protein exons $O(p_{i+1}(\text{N-term}))$ are calculated so that $O(p_i(\text{C-term})) + O(p_{i+1}(\text{N-term}))$
223 $\bmod 3 = 0$.

224 **Identifying proteins of origin for input peptides.** To allow fast lookup of proteins containing
225 any given peptide PoGo creates a dictionary of words with length k (k -mer) overlapping by $k -$
226 1 amino acids from the protein sequences in the FASTA input. Associated with each k -mer is
227 a list of protein entries containing the associated protein with identifiers and the start
228 position of the k -mer in the sequence. The dictionary is designed to consider leucine and
229 isoleucine as equal as they are not distinguishable in MS. Peptides identified through MS are
230 retrieved from the input file and searched against the dictionary. Thereby PoGo allows
231 imperfect matching with up to 2 amino acid substitutions (mismatches m) to also identify
232 proteins with potentially underlying non-synonymous single nucleotide variants. For
233 peptides shorter than $(m + 1) * k$ residues only the first word of length k is used and all
234 combinations with m amino acid substitutions are generated. Each new word is looked up in
235 the dictionary. Peptides longer than $(m + 1) * k$ are split into consecutive k -mers and
236 searched in the dictionary. At most m consecutive k -mers can contain amino acid
237 substitutions leaving one word without any substitutions allowing for perfect matching in
238 the look-up table. The presence of the peptide in each found protein then is validated taking
239 into account the number of mismatches. The gene and transcript identifiers and the
240 respective start position within each protein are retrieved.

241 **Retrieving genomic coordinates for peptides.** Peptides with associated gene and transcript
242 identifiers and the start positions within each protein are used to calculate the genomic
243 coordinates. The length of the peptide sequence A with start position s_A in protein P is used
244 to calculate the end position e_A . To calculate the genomic coordinates for the peptide first
245 the overlapping protein exons p are obtained so that $P(A) = \{x \in P \mid s_x \leq s_A \leq e_x \vee s_x \leq e_A \leq e_x\}$.
246 Through the mapping of protein exons to genomic exons PoGo can now retrieve the
247 genomic exons for the peptide sequence A through $P(A) \rightarrow T(A)$. The genomic coordinates
248 then are calculated as start $S_A = S_E + dS_A$ and end $E_A = S_E + dE_A$ if the gene is on the forward
249 strand or start $S_A = S_E - dS_A$ and end $E_A = S_E - dE_A$ if on the reverse strand with $dS_A = (s_A - s_p -$
250 $1) * 3 + O(P(\text{N-term}))$ and $dE_A = (e_A - s_p) * 3 + O(P(\text{N-term})) - 1$ denoting the distance of the
251 genomic start and end of the peptide, respectively, from the genomic start position S_E of the
252 genomic exon E .

253 **Mapping post-translational modifications.** Besides mapping peptides, PoGo is also capable
254 of mapping post-translational modifications (PTMs) onto the genome. Post-translational
255 modifications are commonly annotated in the peptide sequence through round brackets
256 containing the PSI (Proteomics Standards Initiative) name of the modification following the
257 modified amino acid. With the position of post-translational modifications in the peptide
258 sequence, start s_{PTM} and end e_{PTM} , the mapping of the underlying peptide to the genome the
259 above equations to calculate the genomic positions are adjusted: $dS_{\text{PTM}} = (s_A + s_{\text{PTM}} - s_p - 1) * 3 +$
260 $O(P(\text{N-term}))$ and $dE_{\text{PTM}} = (s_A + e_{\text{PTM}}) * 3 + O(P(\text{N-term})) - 1$. Different types of PTMs are
261 mapped separately and color coded in the output while multiple occurrences of the same
262 PTM type, e.g. phosphorylation, within a single peptide are combined into a single mapping
263 using the first and last PTM sites.

264 **Adding quantitative information for multiple samples.** To allow visualization of quantitative
265 information for peptides on a genome, PoGo records this type of information. Peptide and
266 sample pairings may only occur once in the input file uniquely identifying a quantitation
267 value. PoGo stores the tuples of sample identifier, quantitative value and the number of
268 peptide to spectrum matches (PSMs) for each peptide. This information is used in the
269 different output formats to allow comparative analysis.

270 **Generating different output formats.** PoGo generates output in three formats commonly
271 used in genomics. The first and central output format of PoGo is BED. This format stores
272 each mapped peptide as a single line of twelve tab delimited columns. Besides chromosome
273 coordinates, the peptide sequence, strand as well as start and end coordinates of a thick
274 block the start positions and lengths of peptide blocks mapping to genomic exons are
275 included. Additionally, BED files support individual coloring of each feature. PoGo utilizes
276 this in two different forms. Firstly, in the general peptide centric output of PoGo peptides
277 are colored based on their uniqueness within the genome. Peptides unique to a single
278 transcript are colored in red while peptides shared between multiple transcripts of a single
279 gene are shown in black. Peptides mapping to multiple genes are indicated by their grey
280 color. Secondly, PoGo also generates a separate BED file for peptide forms with post-
281 translational modifications. In this instance the thick block element is used to indicate the
282 position of the post-translational modification. Two or more modifications of the same type
283 within a single peptide sequence are collapsed to indicate the range between the first and
284 last modification site. The coloring of the uniqueness per peptide in the genome is
285 substituted to accommodate color coding of post-translational modifications.

286 The second file format supported by PoGo for mapped peptides is the general transfer
287 format (GTF). PoGo redefines some of the feature types to accommodate mapping of
288 peptides. The feature type 'transcript' is used to indicate a mapped peptide while the
289 feature type 'exon' indicates the concrete mapping of the peptide to underlying genomic
290 exons. PoGo additionally stores information such as the gene identifier, name and biotype
291 for the gene as well as the number of peptide-to-spectrum matches (PSMs) and quantitative
292 values for each sample in which the peptide was identified.

293 For comparative or quantitative analysis PoGo generates the output format GCT which can
294 be visualized in the Integrative Genomics Viewer (IGV).¹¹ This third format is similar to a
295 matrix with rows identifying a peptide with genomic mapping and columns identifying a
296 sample. Each cell holds the quantitative values associated with the peptide and the sample
297 given in the input file.

298 **Human tissue data.** High-resolution MS data from 59 fetal and adult human tissues were
299 used for the validation of PoGo. The raw data of these draft human proteome maps were
300 generated by the Pandey lab¹³, the Kuster lab¹⁴, and Cutler lab.¹⁵ All three datasets were
301 combined and reprocessed by Wright et al. Wright, Mudge, Weisser, Barzine, Gonzalez,
302 Brazma, Choudhary and Harrow⁹ The data were retrieved in a tab delimited format
303 combining all results from mzid files available from PRIDE Archive.¹⁰ Identifications were
304 filtered to the highest stringency level described in Wright et al.⁹ for identification of novel
305 coding regions (q-value \leq 0.01 (1% FDR), a PEP of \leq 0.01, peptide length between 7 and 29
306 residues, full tryptic peptides, a maximum of two missed cleavages).

307 **Phosphoproteomic ovarian cancer data.** We applied PoGo to isobaric labelled
308 phosphoproteome data from an ovarian tumor study comprising 69 samples.²
309 Phosphopeptides with associated iTRAQ quantitation were downloaded as tab separated file
310 from <https://cptac-data-portal.gorgetown.edu>. Lower case characters (s, t and y) in the
311 peptide sequence showing phosphorylation were substituted by upper case characters
312 followed by the PSI name of phosphorylation in brackets.

313 **Protein sequences and gene annotation and PoGo settings.** The annotation of human genes
314 in GTF format and the corresponding protein coding sequence translation as FASTA files
315 were downloaded for GENCODE v20⁹ from <http://www.gencodegenes.org>. Gene and
316 transcript identifiers were set as “ENSG” and “ENST” for genes and transcripts, respectively,
317 followed by 11 digits and the word length for k-mers was set to 5 amino acids. For post-
318 translational modifications 10 biologically relevant types were chosen for easy
319 discriminability of the color code (Table S1).

320 **Comparison of algorithms for performance evaluation.** For the human tissue and the
321 ovarian cancer phosphoproteome data PoGo’s performance was compared against PGx³
322 (downloaded from <https://github.com/FenyoLab/PGx>) and iPiG⁴ (downloaded from
323 <https://sourceforge.net/projects/ipig/>), two standalone tools available to map peptides to
324 their corresponding genomic coordinates. Each dataset was formatted using in-house scripts
325 in R and perl to fit the required input format for each tool. Each program was run using
326 default parameters and the minimum number of required input files. To compare the
327 mappings between the tools, we marked as equal when chromosome name, start and end
328 positions, the exon starts and lengths as well as the peptide sequence were the same.
329 Frameshifts then were identified amongst unique mappings per tool through shifting either
330 start or end position by up to two base pairs and comparing those to the consensus
331 mappings. Remaining unique mappings of the tools then were examined manually by
332 comparing the peptide sequence to the translated sequence of the respective genomic
333 coordinates in the IGV browser.¹¹

334 **Generating track hubs.** Track hubs were generated to visualize different aspects of the
335 human proteome maps. The data was filtered to two stringency levels resulting in two sets.
336 The first result set was filtered to a standard significance (q-value of ≤ 0.01 (1% FDR), a PEP
337 of ≤ 0.05 and a minimum peptide length of 7 residues) while the highest stringency level
338 mentioned in Wright, Mudge, Weisser, Barzine, Gonzalez, Brazma, Choudhary and Harrow⁹
339 (q-value ≤ 0.01 (1% FDR), a PEP of ≤ 0.01 , peptide length between 7 and 29 residues, full
340 tryptic peptides, a maximum of two missed cleavages) was applied to the second set.
341 Additionally, each set was split into subsets for individual tissues, resulting in 60 files per set.
342 PoGo was run with default parameters using the property of passing a comma separated list
343 of input files to be mapped separately. The Track-Hub Generator application then was run
344 using the 60 output files in BED format to generate two track hubs; one for each significance
345 level filter. Folders and files required for track hubs are generated automatically. The script
346 ‘fetchChromSizes.sh’ and tool ‘bedToBigBed’ from UCSC (both downloaded from
347 <http://hgdownload.cse.ucsc.edu>)¹⁶ are used in the Track-Hub Generator to create binary
348 files from the original BED files used for track hubs. The generated track hubs are accessible
349 through ftp and http via <http://www.sanger.ac.uk/science/projects/teogenomichubs>.

350 **References**

- 351 13. Kim, M.S. *et al. Nature* **509**, 575-581 (2014).
352 14. Wilhelm, M. *et al. Nature* **509**, 582-587 (2014).
353 15. Desiere, F. *et al. Nucleic Acids Res.* **34**, D655-658 (2006).
354 16. Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S. & Karolchik, D. *Bioinformatics* **26**,
355 2204-2207 (2010).