Integrating MS-based proteomics and RNA-Seq data

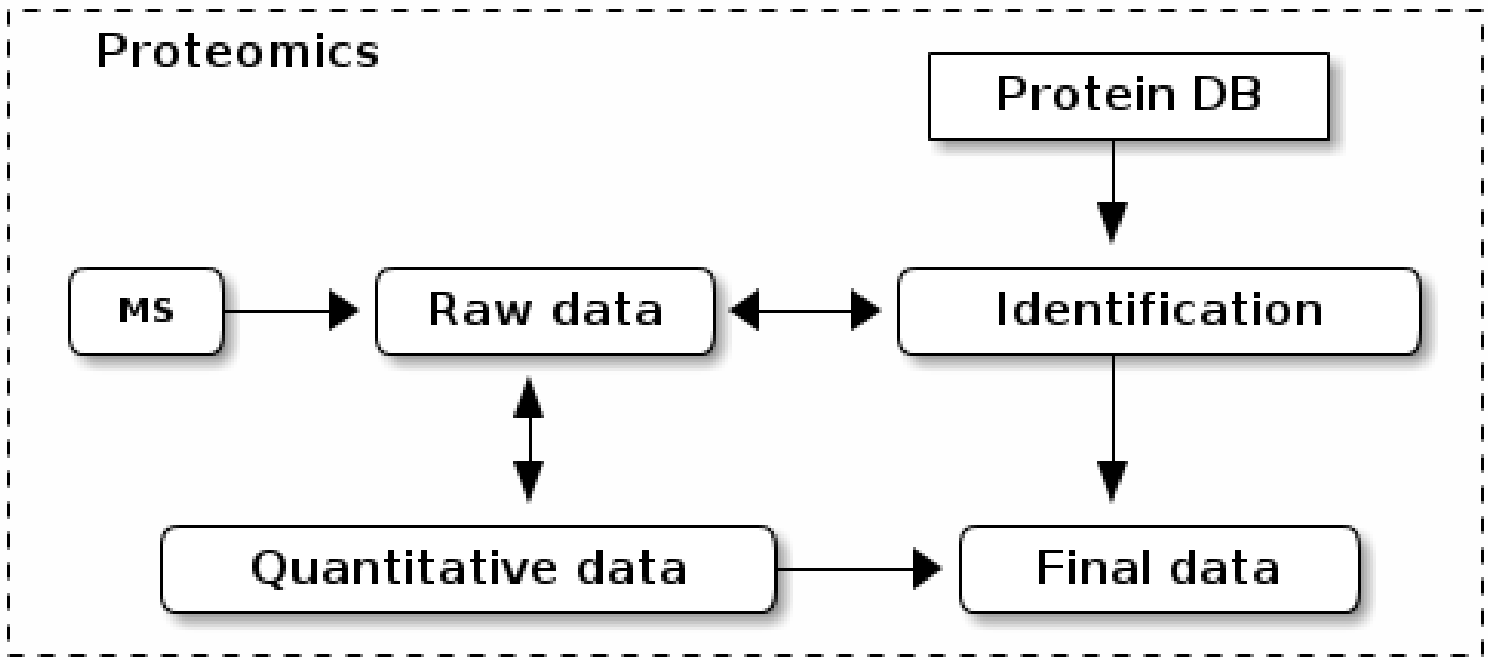
March 7, 2016

Use case 1: mapping peptides

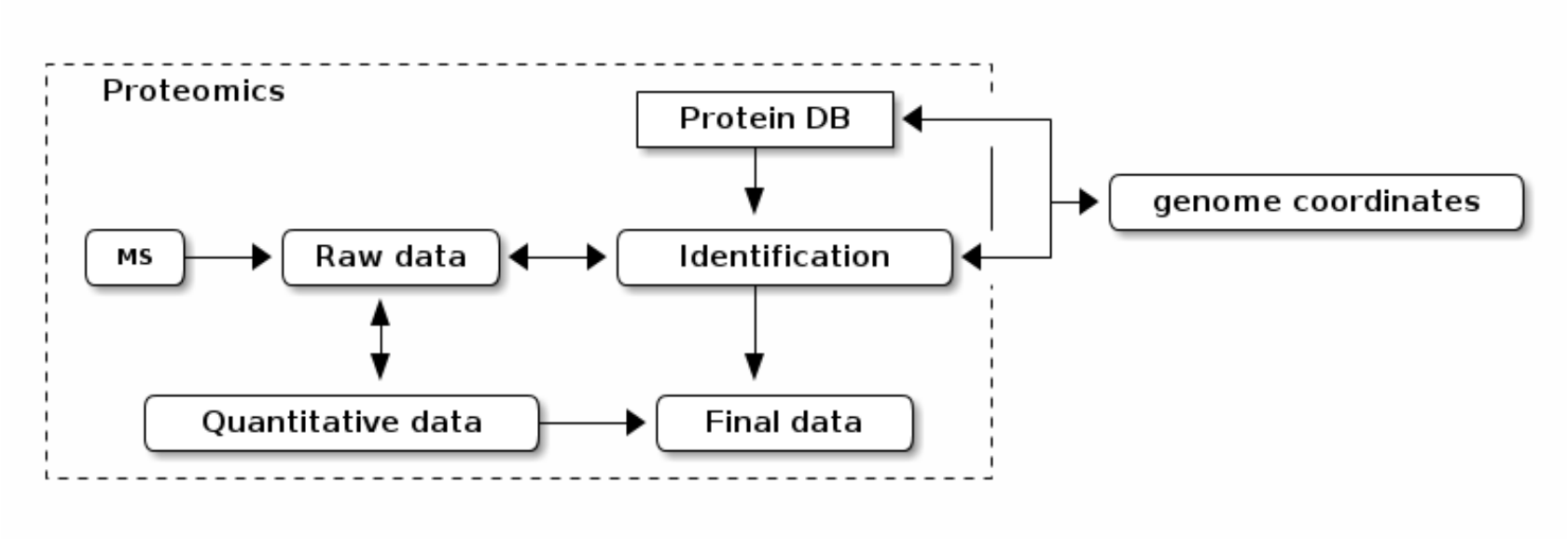
Mapping of *peptides along protein sequences (although not explicitly considered a mapping exercise)* and *short reads along genome coordinates*.

Mapping protein and gene identifers

The protein database and the genome are *independent*, i.e. the proteins do not make explicitly reference to the genome they originate from.



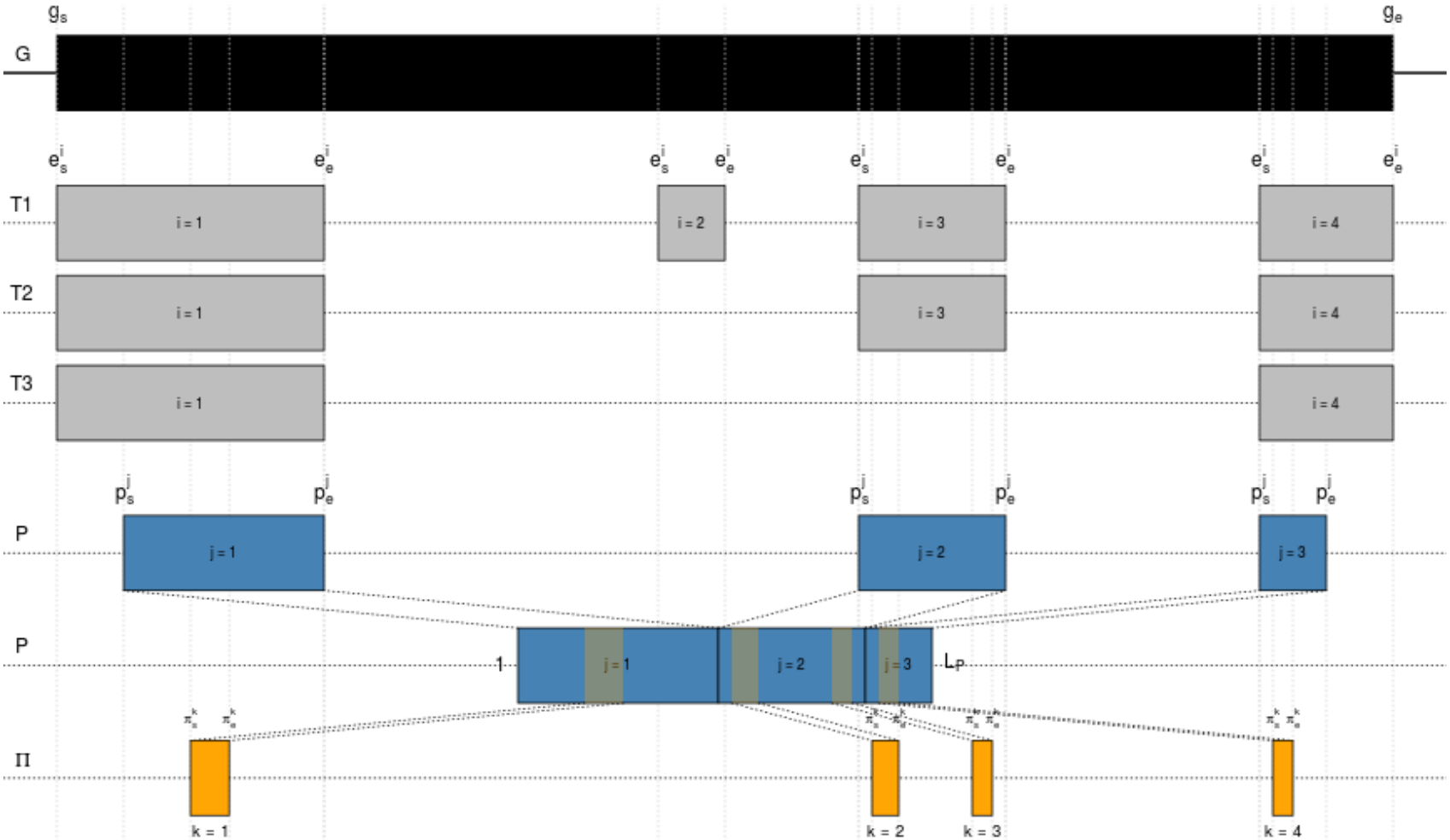
Mapping protein and gene identifers



If we want to map UniProt accession numbers to genomic identifiers (Ensembl transcipt identifiers)

Mapping peptides to genomic coordinates

The **goal** is to map peptides from protein coordinates (1 to *L_p*) to genomic coordinates.



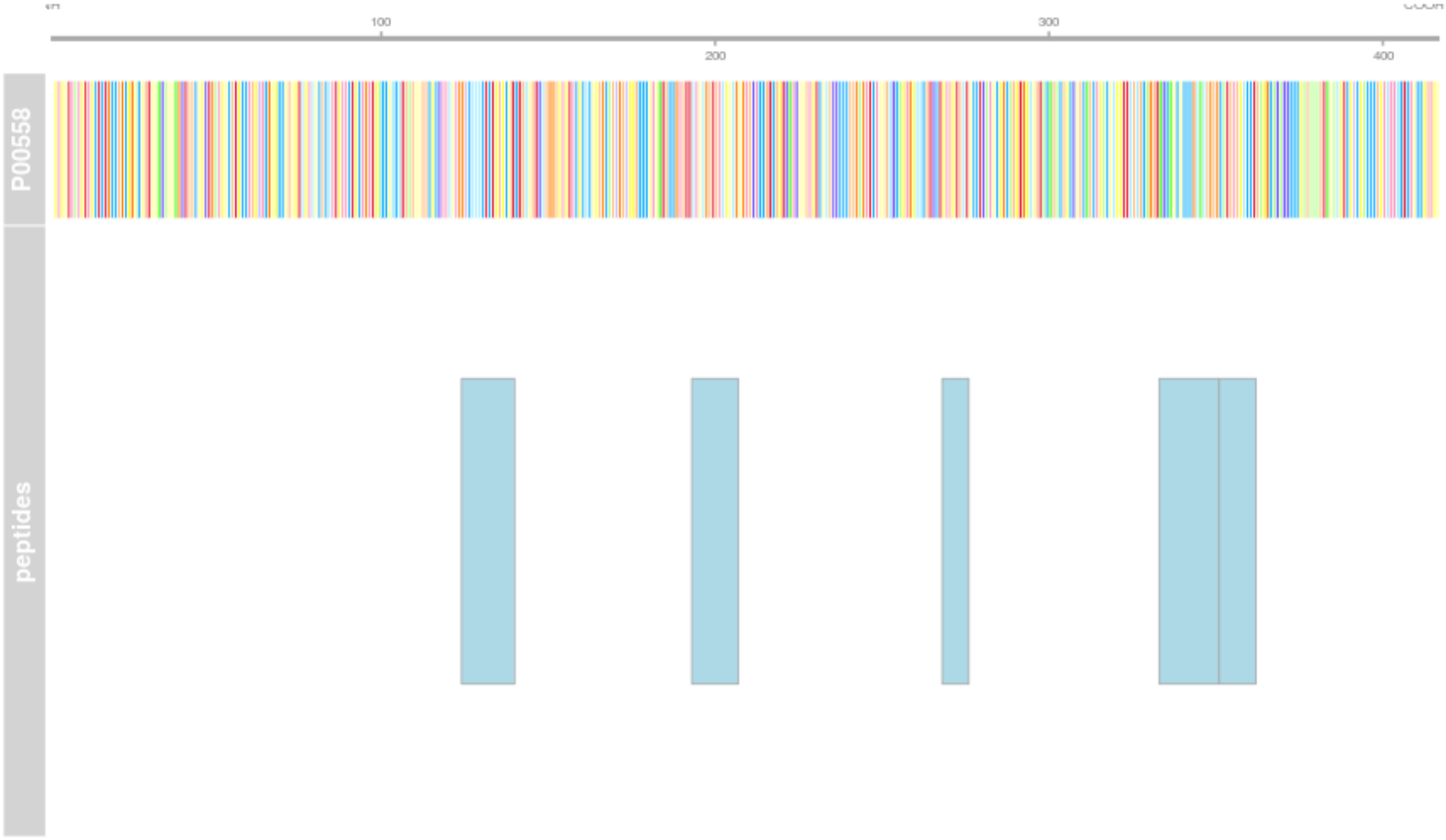
Data

Illustration with the `Pbase` Bioconductor package.

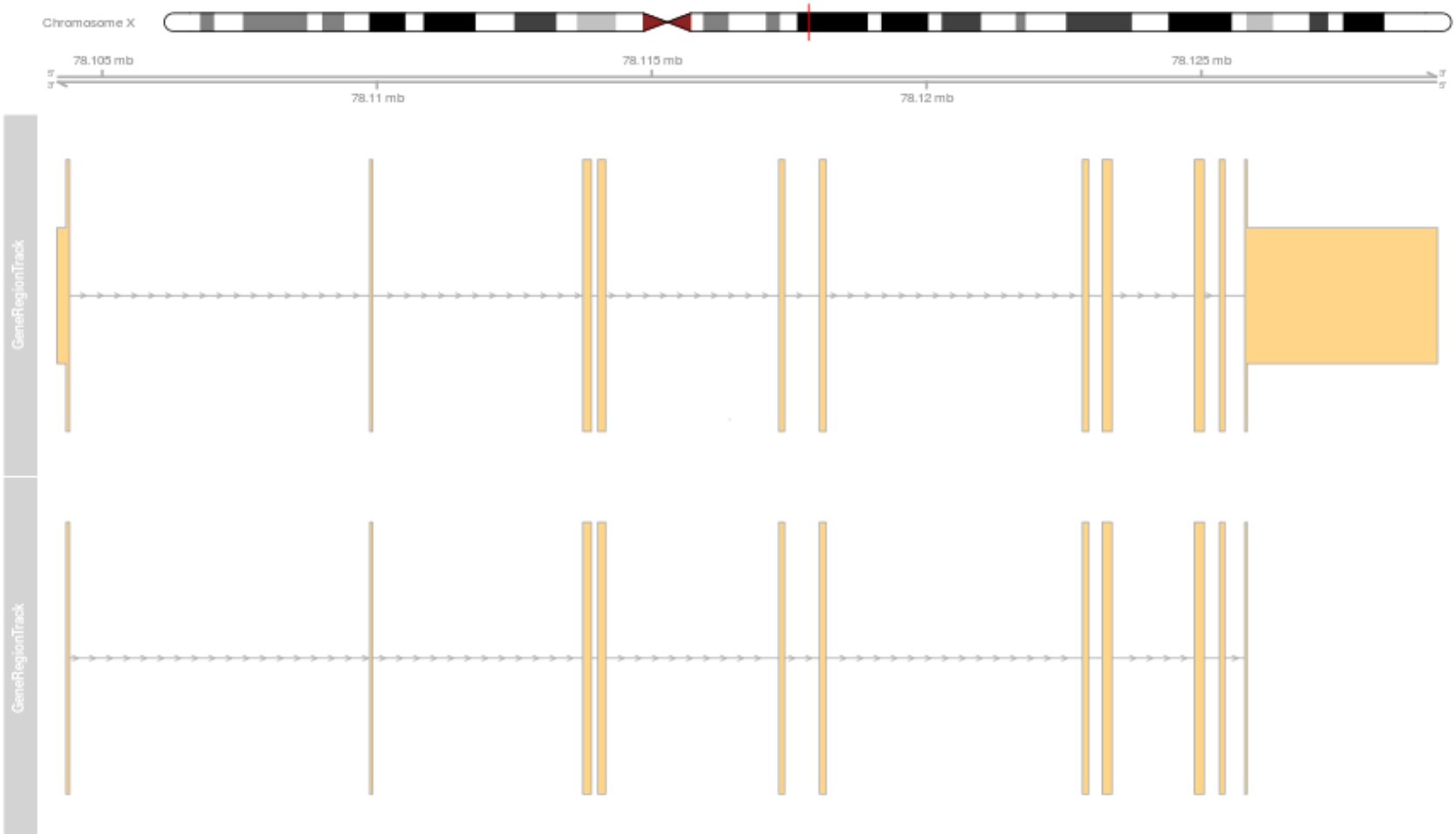
We have an example data composed of 9 proteins, with UniProt accession numbers and Ensembl transcipt identifiers and each protein has a set experimentally observed peptides (see table below). This object was generated from the protein database (fasta file) and the MS identification results (`mzIdentML` file) against this very same protein database.

Acc	ENST	npep
A4UGR9	ENST00000409195	36
A6H8Y1	ENST00000358731	23
O43707	ENST00000252699	6
O75369	ENST00000295956	13
P00558	ENST00000373316	5
P02545	ENST00000368300	12
P04075	ENST00000338110	21
P04075-2	ENST00000395248	20
P60709	ENST00000331789	1

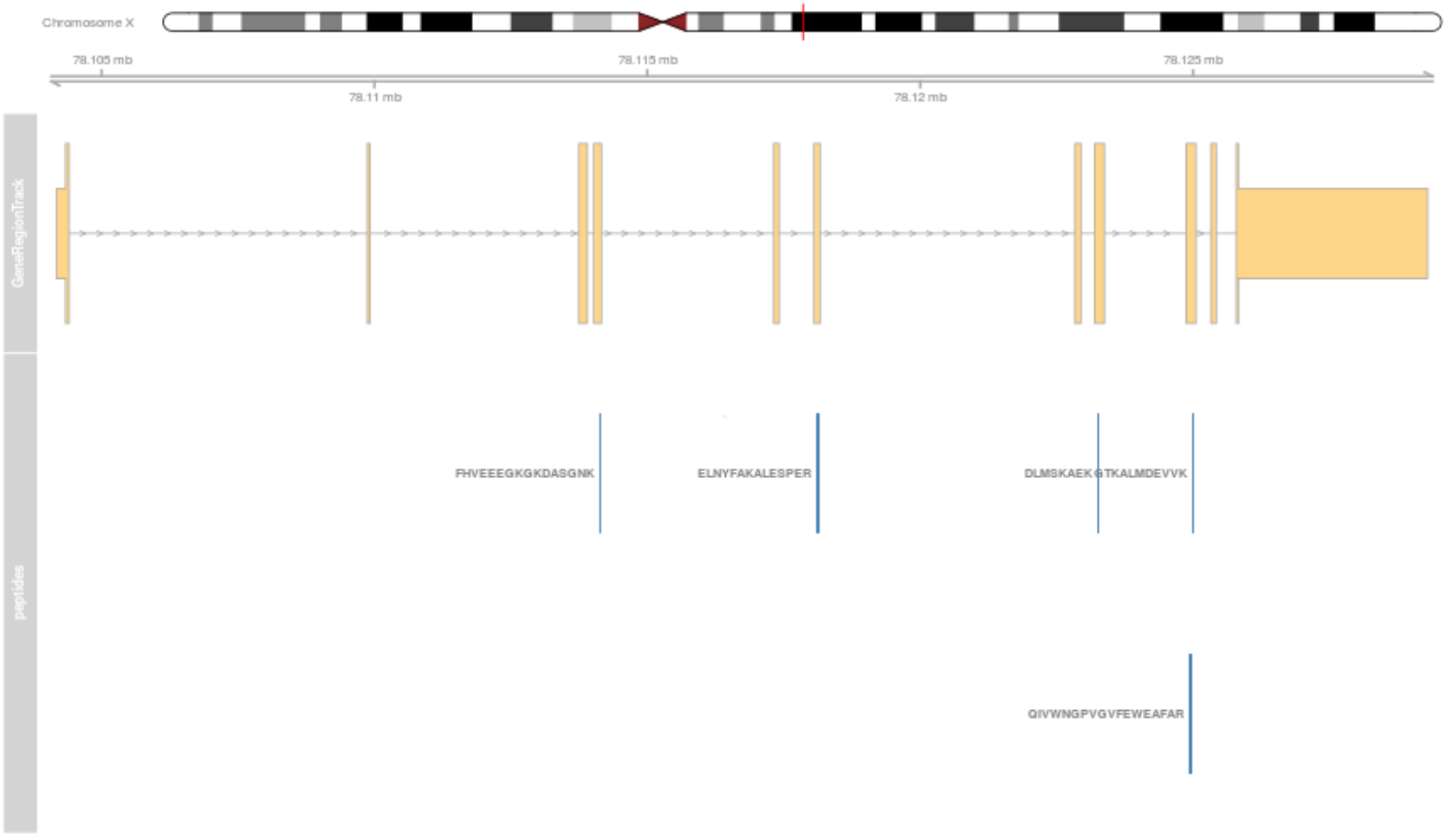
For example, P00558:



Genomic coordinates of the transcripts/exons

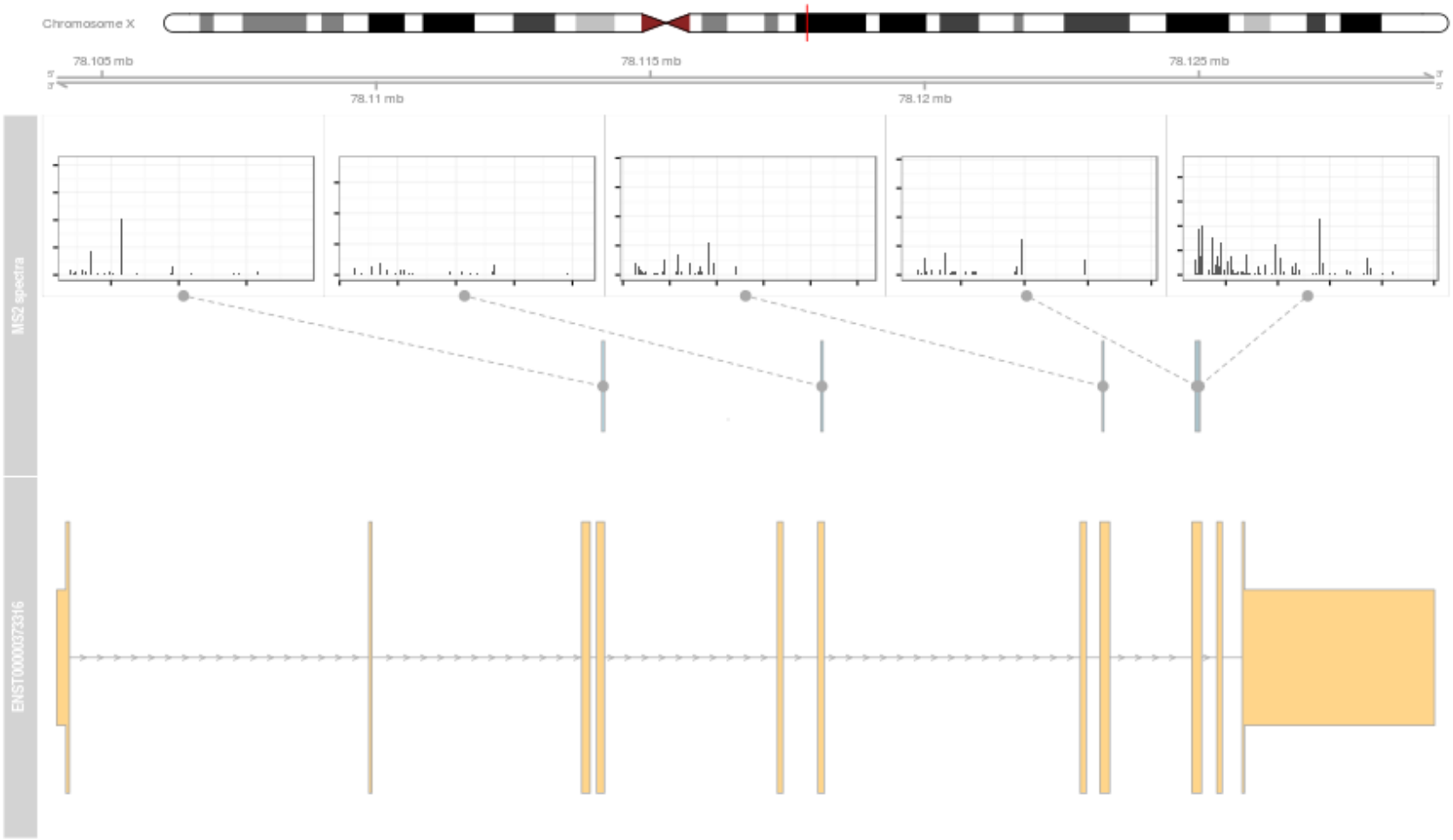


Mapping peptides to the genome



Detailed annotation tracks

Maintaining access to the raw MS data (used as input with the fasta file to generate the identification results).

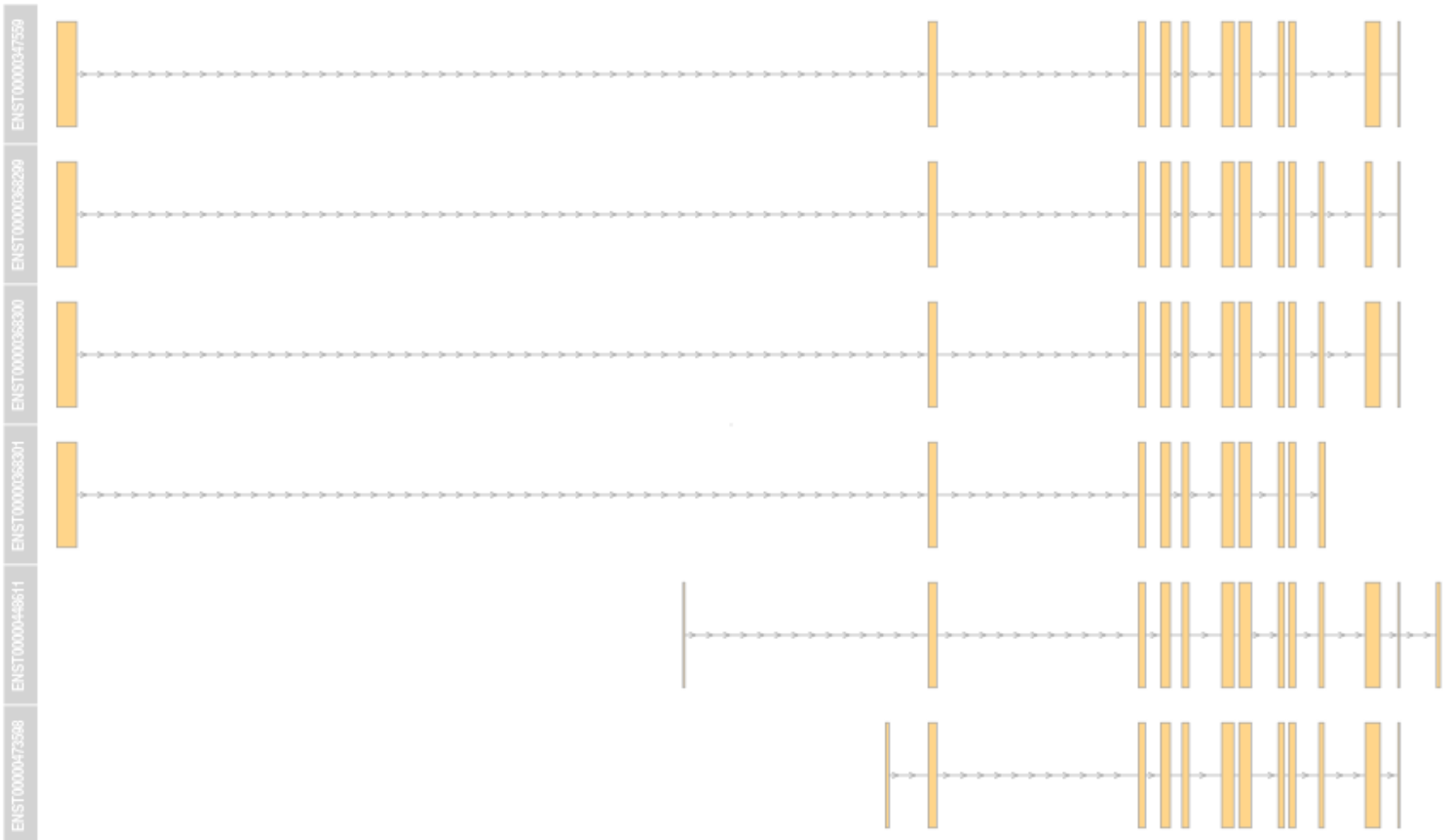


Multiple transcipts per protein

If we hadn't the curated UniProt accession/Ensembl transcript identifer maps, we would, for example, query an online repositary such as the Ensembl Biomart instance. For example

UNIPROTKB	ENSEMBL_TRANSCRIPT
P02545	ENST00000347559
P02545	ENST00000368299
P02545	ENST00000368300
P02545	ENST00000368301
P02545	ENST00000448611
P02545	ENST00000473598

Genomic coordinates

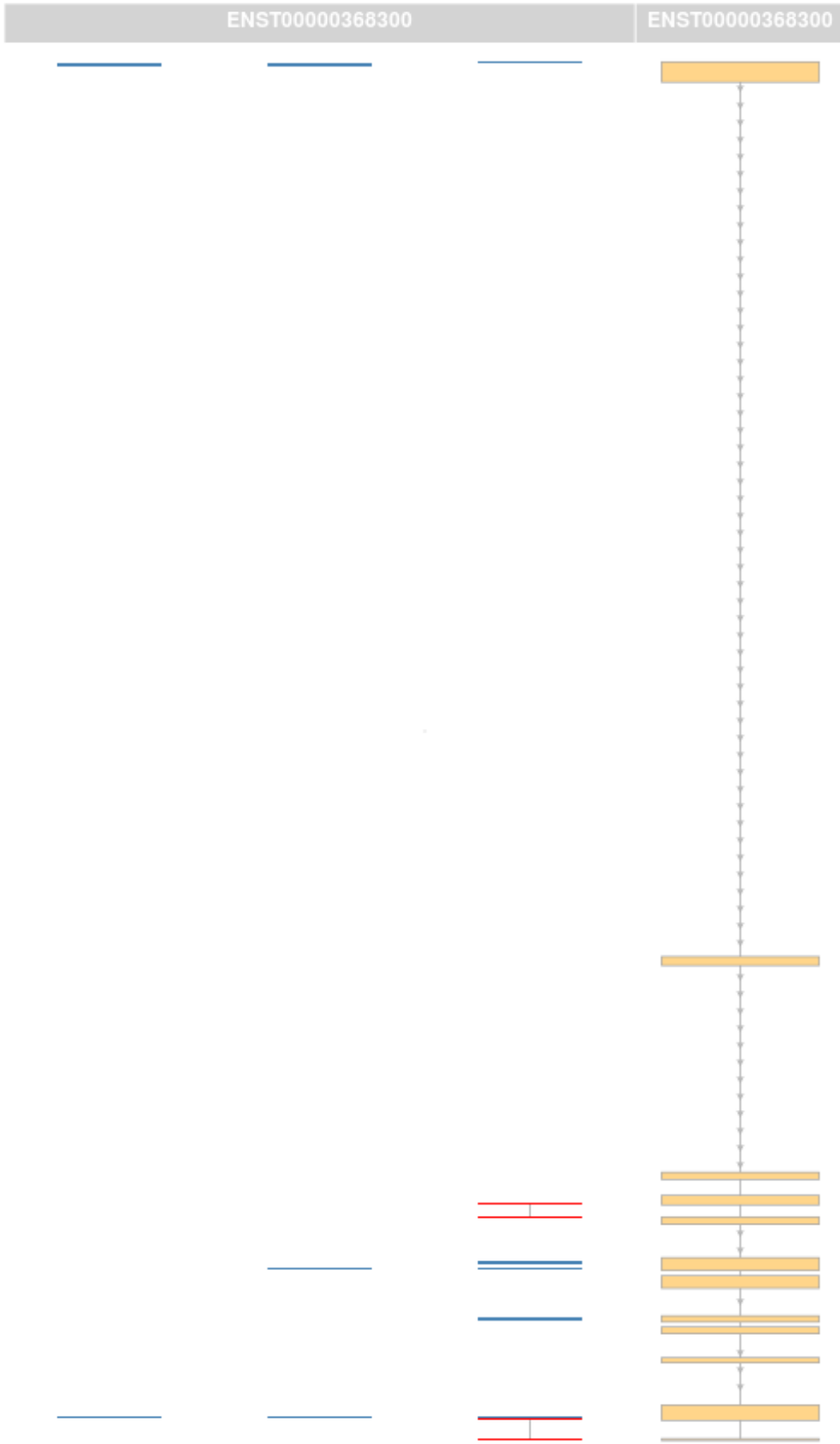


Discriminating transcripts

We extract the transcript sequences, translate them into protein sequences and align each to our original protein sequence.

```
## ENST00000347559 ENST00000368299 ENST00000368300 ENST00000368301
##       0.9548193       0.9246988       1.0000000       0.8614458
## ENST00000448611 ENST00000473598
##       0.8298193       0.8358434
```

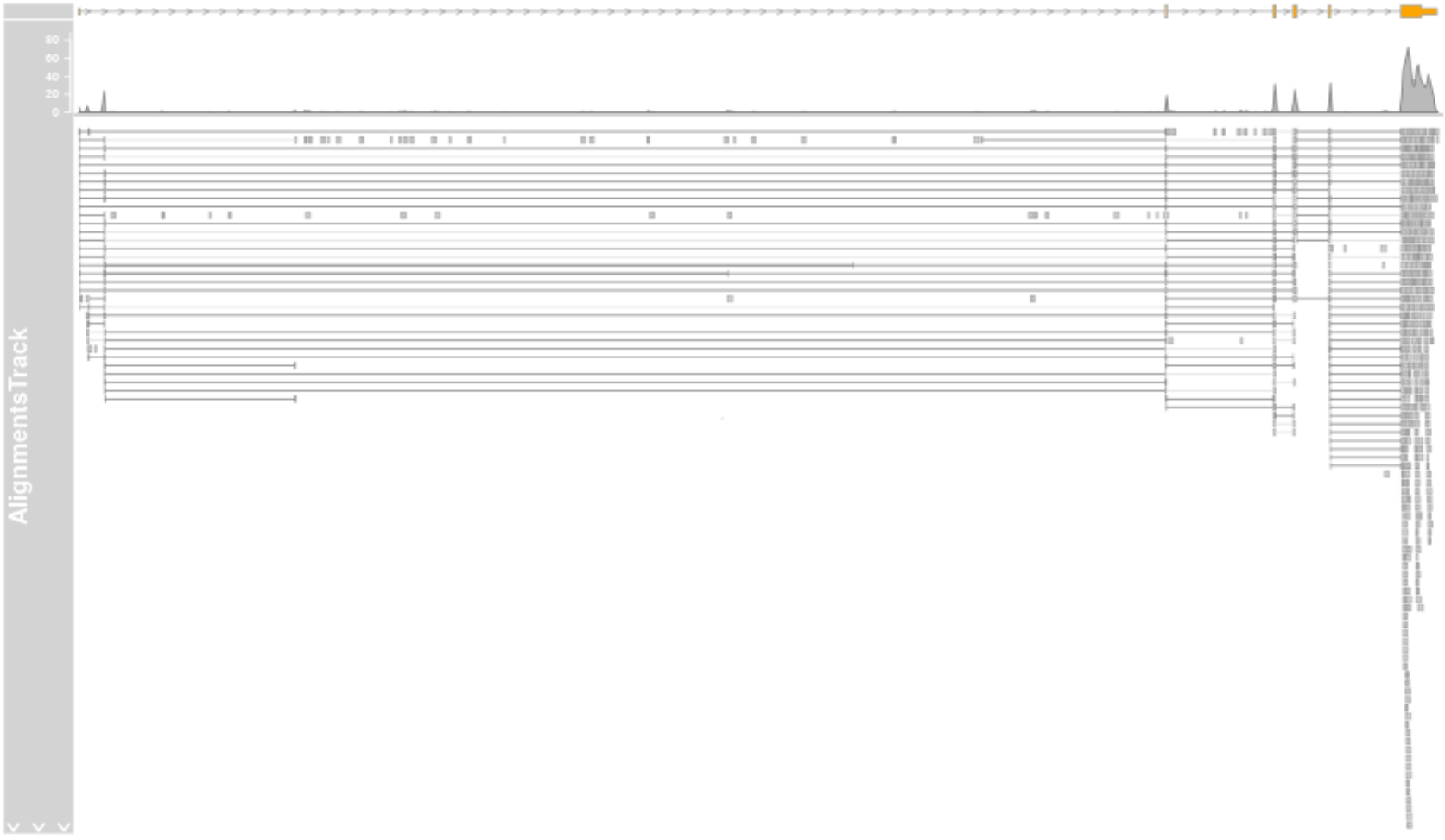
```
## Global PairwiseAlignmentsSingleSubject (1 of 1)
## pattern: [1] METPSQRRATRSGAQASSTPLSPTRITRLQEK...GGGSFGDNLVTRSYLLGNSSPRTQSPQNCSIM
## subject: [1] METPSQRRATRSGAQASSTPLSPTRITRLQEK...GGGSFGDNLVTRSYLLGNSSPRTQSPQNCSIM
## score: 2843.652
```


ENST00000368300
ENST00000368300

Mapping MS peptides and RNA-Seq short reads

The last step of the mapping process is the combine the newly mapped peptides and reads from RNA-Seq experiments. The figures below illustrate this with data from Sheynkman et al. (PMID: 23629695, 25149441) from the Jurkat cell line (TIB-152). The mass spectrometry (PASS00215) and RNA-Seq (SRR791580) where processed with standard pipelines.


Chromosome 17
30.12 mb
30.13 mb
30.14 mb
30.15 mb
30.16 mb
30.17 mb
30.18 mb
ENST00000612959


AlignmentsTrack
80
60
40
20
0

For all details/code see the `Pbase` package mapping vignette
`http://bioconductor.org/packages/Pbase`

Use case 1: mapping peptides

Mapping of *peptides along protein sequences (although not explicitly considered a mapping exercise)* and *short reads along genome coordinates*.

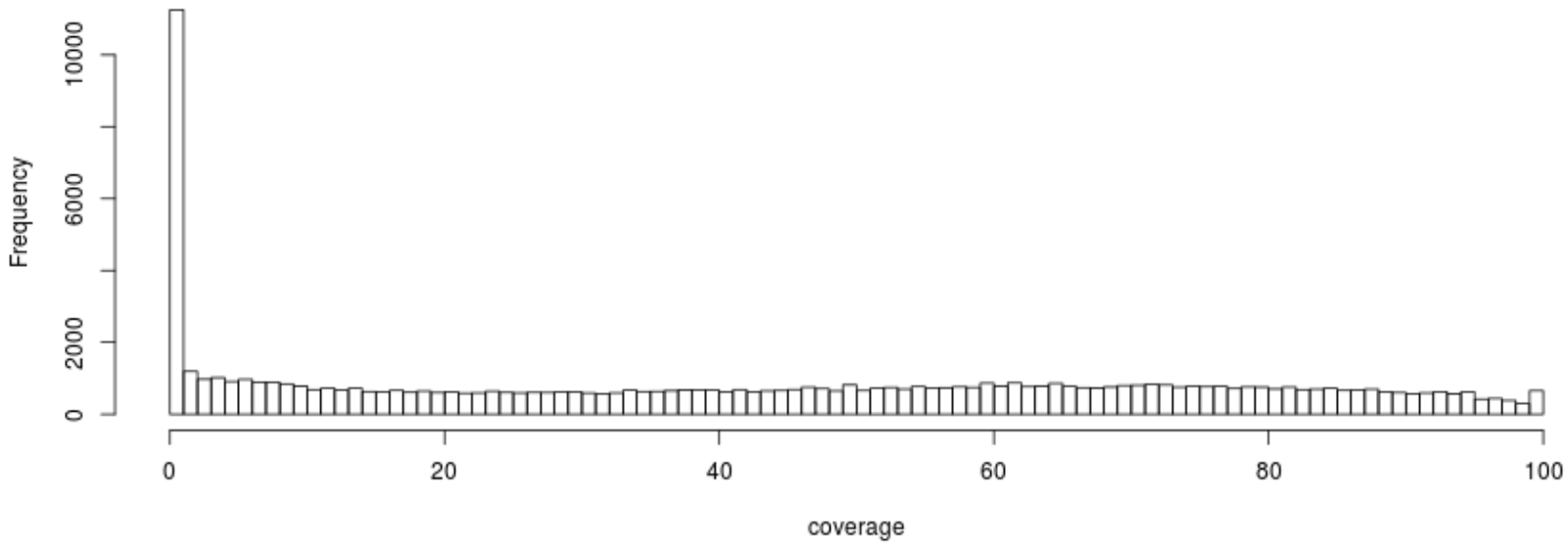
But. . .

- coverage
- protein inference
- identifier mapping
- missing values

Coverage

- Coverage in proteomics in %
- Coverage in RNA-Seq in fold X

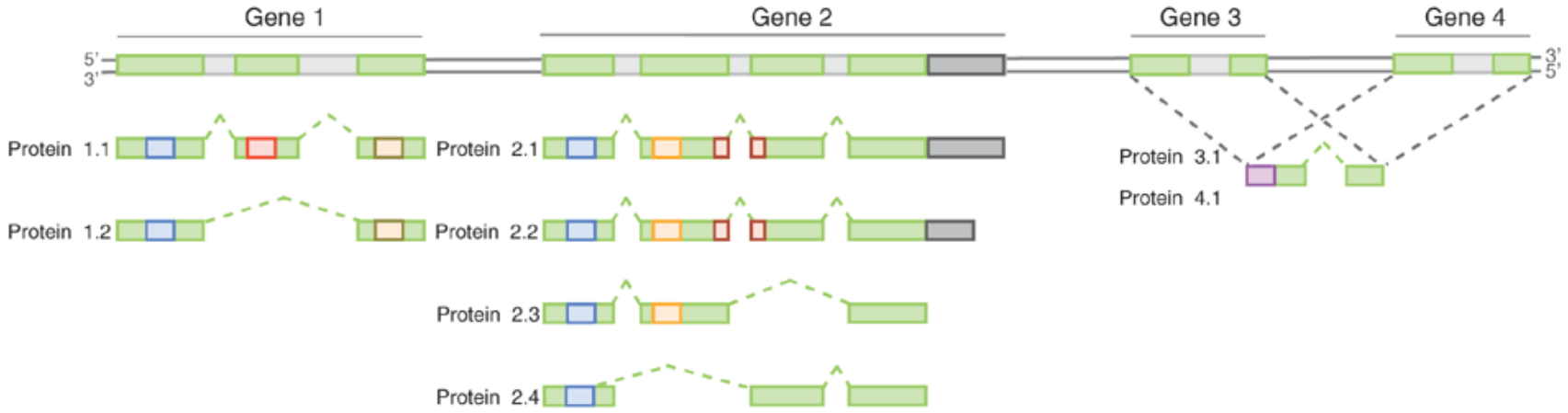
The following values are higher bounds, *without* peptide filtering for about 80000 *gene groups*



And

- the majority of peptides map to a minority of proteins different
- peptides within one protein can be differently detectable in MS acquistions

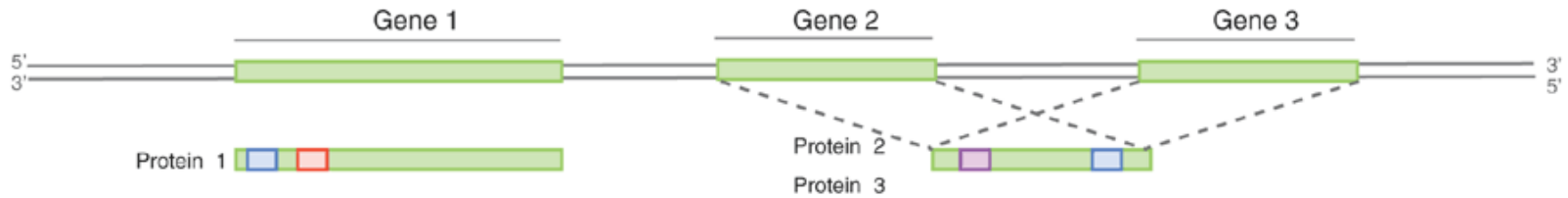
Protein inference



Class	Protein sequence(s)	Protein isoform(s)	Gene(s)
1a	Unambiguous	Unambiguous	Unambiguous
1b	Unambiguous	Ambiguous	Unambiguous
2a	Ambiguous	Ambiguous	Unambiguous
2b	Ambiguous	Ambiguous	Unambiguous
3a	Unambiguous	Ambiguous	Ambiguous
3b	Ambiguous	Ambiguous	Ambiguous

Eukaryotes

Prokaryotes



From Qeli and Ahrens (2010). See also Nesvizhskii and Aebersold (2005).

Protein groups

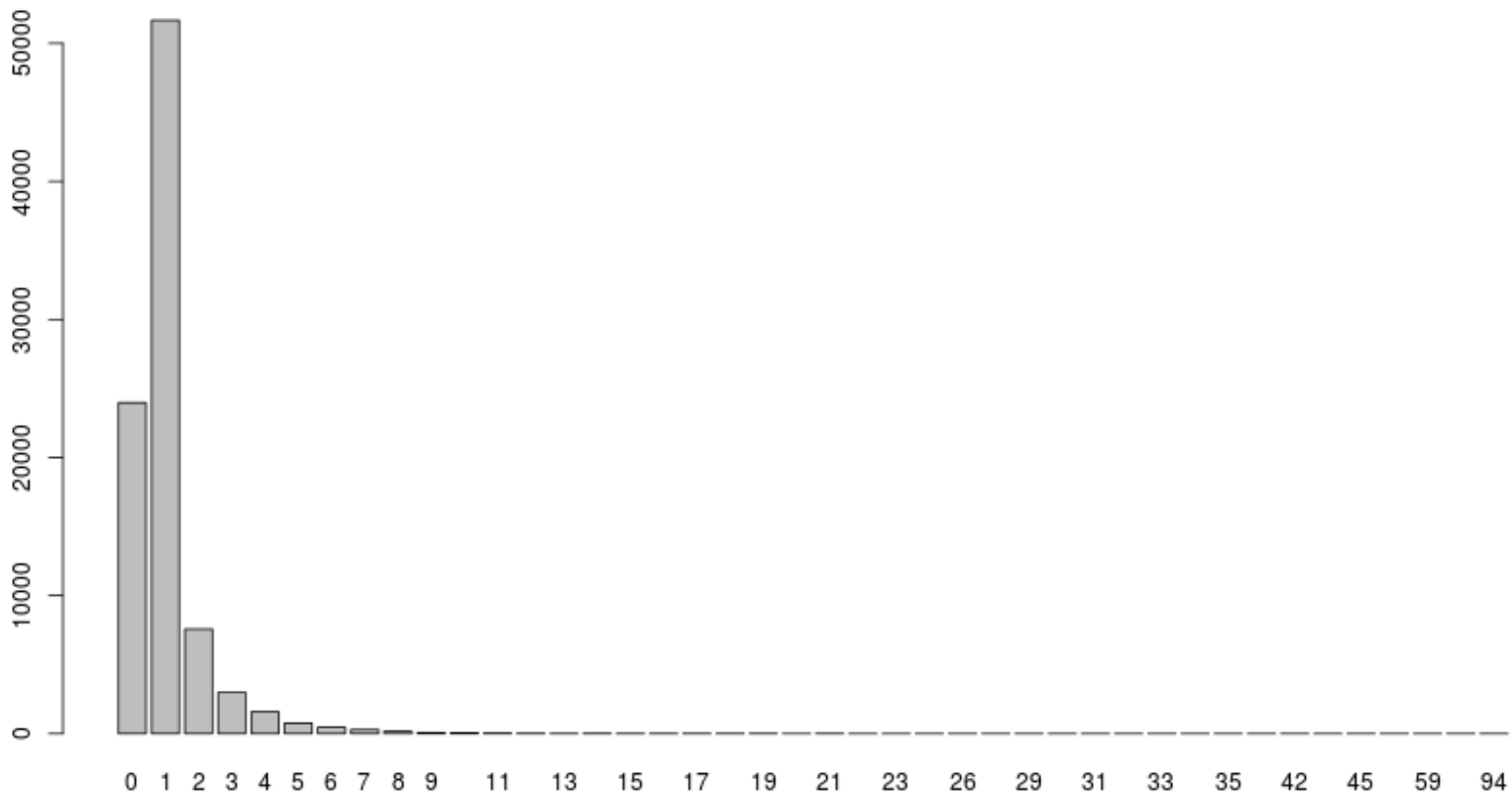
Often, in proteomics experiments, the features represent single proteins and **groups** of indistinguishable or non-differentiable proteins identified by shared (non-unique) peptides.

Caveat: Mapping between protein groups and unique transcripts?

Mapping identifiers

The UniProt human proteome (release 2015_02) has 89796 entries.
Using `UniProt.ws`:

- 23972 have no transcript identifier
- 51673 have a unique transcript identifier
- 14151 have more than one transcript identifier



Using biomaRt:

Mapping 18911 identifiers, of which

Caveat: Mapping between single protein and unique transcripts?

Missing values

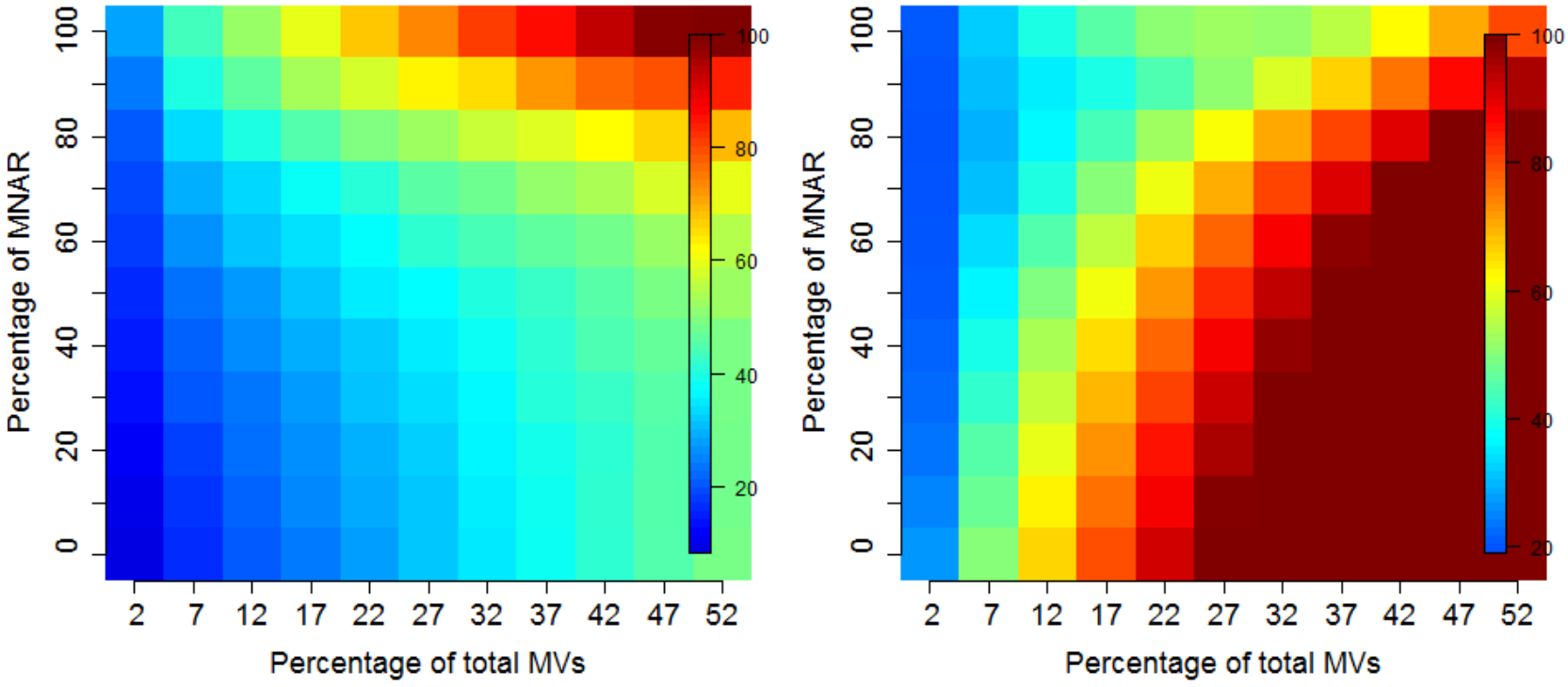
Options are:

- Filtering: Remove missing values, or at least features or samples with excessive number of missing values:
- Data imputation: infering plausible values for missing data.

Data imputation

There are two types of mechanisms resulting in missing values in LC/MSMS experiments.

- Missing values resulting from absence of detection of a feature, despite ions being present at detectable concentrations. For example in the case of ion suppression or as a result from the stochastic, data-dependent nature of the MS acquisition method. These missing value are expected to be randomly distributed in the data and are defined as **missing at random** (MAR) or **missing completely at random** (MCAR).
- Biologically relevant missing values, resulting from the *absence* of the low abundance of ions (below the limit of detection of the instrument). These missing values are not expected to be randomly distributed in the data and are defined as **missing not at random** (MNAR).



MNAR features should ideally be imputed with a **left-censor** (minimum value (right), but not zero, . . .) method. Conversely, it is recommended to use **hot deck** methods (nearest neighbour (left), maximum likelihood, . . .) when data are missing at random.

Summary

- mapping peptides along genomic coordinates
- protein/transcipt mapping
- protein groups
- missing values and coverage

References

Laurent Gatto and Sebastian Gibb (2016). `Pbase`: Manipulating and exploring protein and proteomics data. R package version 0.11.3. `https://github.com/ComputationalProteomicsUnit/Pbase`

Lazar C, Gatto L, Ferro M, Bruley C, and Burger T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. Publication Date: February 23, 2016 DOI: 10.1021/acs.jproteome.5b00981

Pang et al. Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. J Proteome Res. 2014 Jan 3;13(1):84-98. doi: 10.1021/pr400820p. Epub 2013 Nov 12. PubMed PMID: 24152167.

Sheynkman GM, Johnson JE, Jagtap PD, Shortreed MR, Onsongo G, Frey BL, Griffin TJ, Smith LM. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. BMC Genomics. 2014 Aug 22;15:703. doi: 10.1186/1471-2164-15-703. PubMed PMID: 25149441.

Qeli E, Ahrens CH. PeptideClassifier for protein inference and targeted quantitative proteomics. Nat Biotechnol. 2010 Jul;28(7):647-50. doi: 10.1038/nbt0710-647. PubMed PMID: 20622826.

Boekel J, Chilton JM, Cooke IR, Horvatovich PL, Jagtap PD, Käll L, Lehtiö J, Lukasse P, Moerland PD, Griffin TJ. Multi-omic data analysis using Galaxy. Nat Biotechnol. 2015 Feb 6;33(2):137-9. doi: 10.1038/nbt.3134. PubMed PMID: 25658277.

Evans VC, Barker G, Heesom KJ, Fan J, Bessant C, Matthews DA. De novo derivation of proteomes from transcriptomes for transcript and protein identification. Nat Methods. 2012 Dec;9(12):1207-11. doi: 10.1038/nmeth.2227. Epub 2012 Nov 11. PubMed PMID:23142869; PubMed Central PMCID:PMC3581816.