Supporting information

“The Personalized Proteome: Comparing Proteogenomics and Open Variant Search Approaches for Single Amino Acid Variant Detection”

Renee Salz[1], Robbin Bouwmeester[2,3], Ralf Gabriels[2,3], Sven Degroeve[2,3], Lennart Martens[2,3], Pieter-Jan Volders[2,3], Peter A.C. ’t Hoen[1,*]

[1]Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, 6525 GA, The Netherlands

[2]VIB-UGent Center for Medical Biotechnology VIB, Technologiepark-Zwijnaarde 75, 9052 Ghent, Belgium

[3]Department of Biomolecular Medicine, Ghent University, Technologiepark-Zwijnaarde 75, 9052 Ghent, Belgium

Table of contents

- Reference counterpart PSMs from the Variant-Free search database
- Reference counterpart PSMs from the Variant-Containing search database

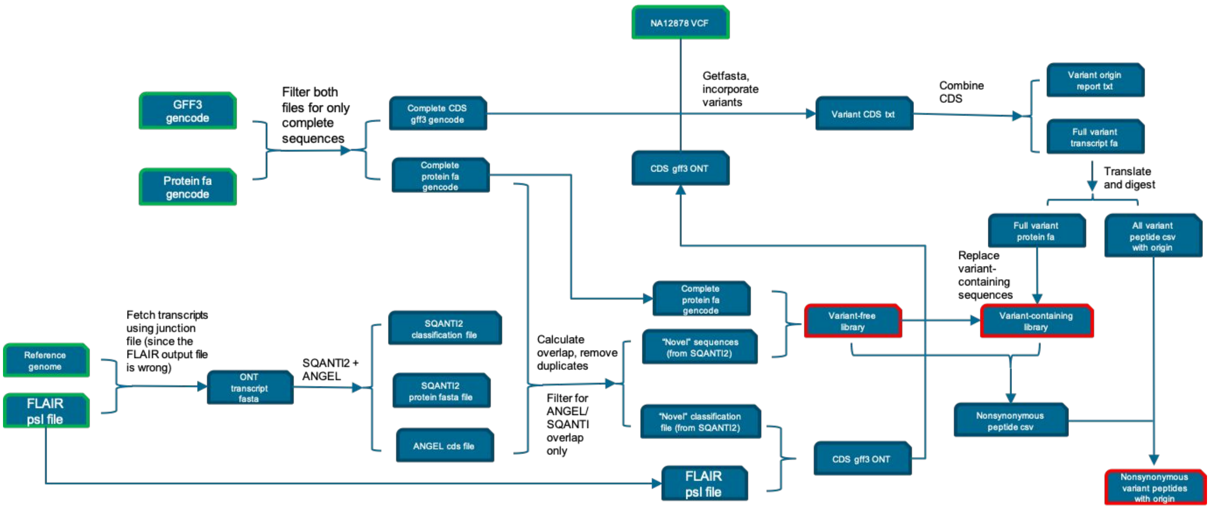

NA12878 VCF
GFF3 gencode
Protein fa gencode
Filter both files for only complete sequences
Complete CDS gff3 gencode
Complete protein fa gencode
Getfasta, incorporate variants
Variant CDS txt
Combine CDS
Variant origin report txt
Full variant transcript fa
Translate and digest
Full variant protein fa
All variant peptide csv with origin
CDS gff3 ONT
Replace variant-containing sequences
Reference genome
FLAIR psl file
Fetch transcripts using junction file (since the FLAIR output file is wrong)
ONT transcript fasta
SQANTI2 + ANGEL
SQANTI2 classification file
SQANTI2 protein fasta file
ANGEL cds file
Calculate overlap, remove duplicates
Filter for ANGEL/ SQANTI overlap only
Complete protein fa gencode
"Novel" sequences (from SQANTI2)
"Novel" classification file (from SQANTI2)
FLAIR psl file
Variant-free library
Variant-containing library
Nonsynonymous peptide csv
CDS gff3 ONT
Nonsynonymous variant peptides with origin

Figure S1. Detailed workflow schematic.

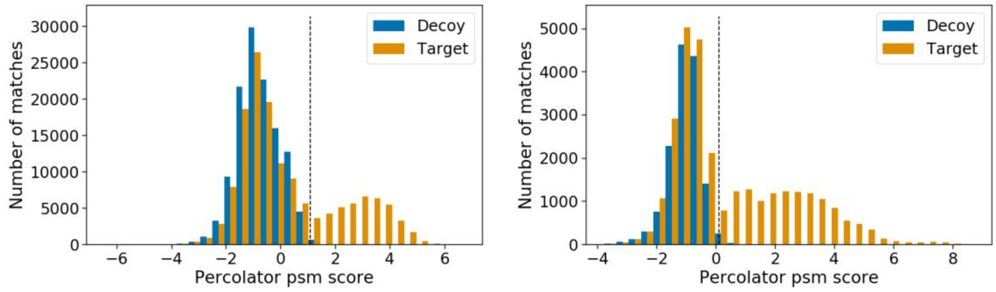
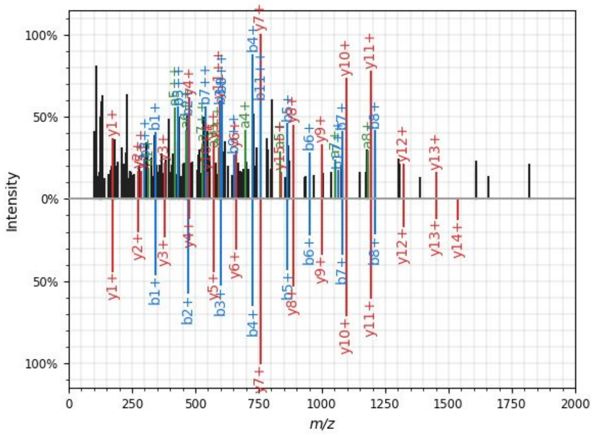
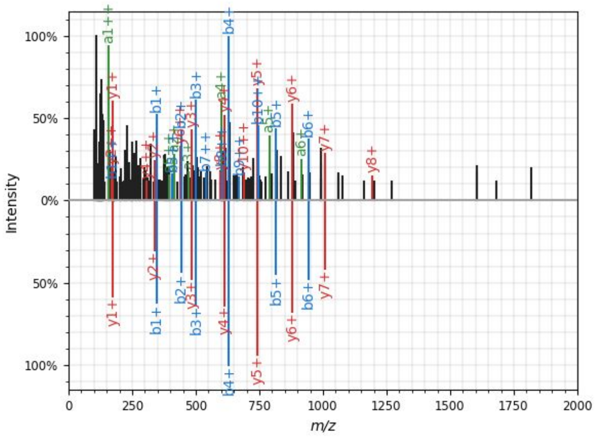


Figure S2. Distribution of target and decoy variant peptides. Variant-containing distribution is on the left, and variant-free is on the right. Separation was made at the dotted line (q<0.01).





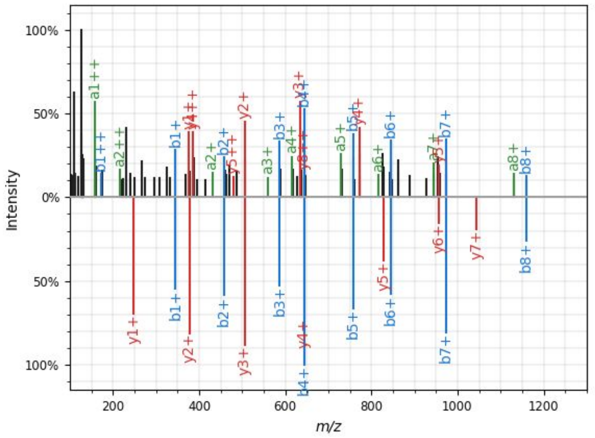


Figure S3. Annotated variant peptide spectra in mirror plots, with theoretical spectra (as predicted by MS2PIP) in the bottom half for reference. Plots made with spectrum_utils python package. Top: variant peptide LQQQHSEQPPLQPSPVTTR, substitution M → T, on chromosome 1 pos 179882939, scan id Linfeng_012511_HapMap39_6.8739.8739.3. Middle: variant peptide DVGEWQHEEFYR, substitution R → G, on chromosome 16 pos 3674464, scan id Linfeng_030911_HapMap46_2.12742.12742.3. This is one of the peptides where no reference counterparts were detected (while 90 variant peptides were identified). Bottom: variant peptide DLEGLSQWHEEK

, substitution W → R, on chromosome 22 pos 36292132, scan id Linfeng_080711_HapMap59_5.15580.15580.3. This is one of the rare variant peptide identifications (AF = 0.001).

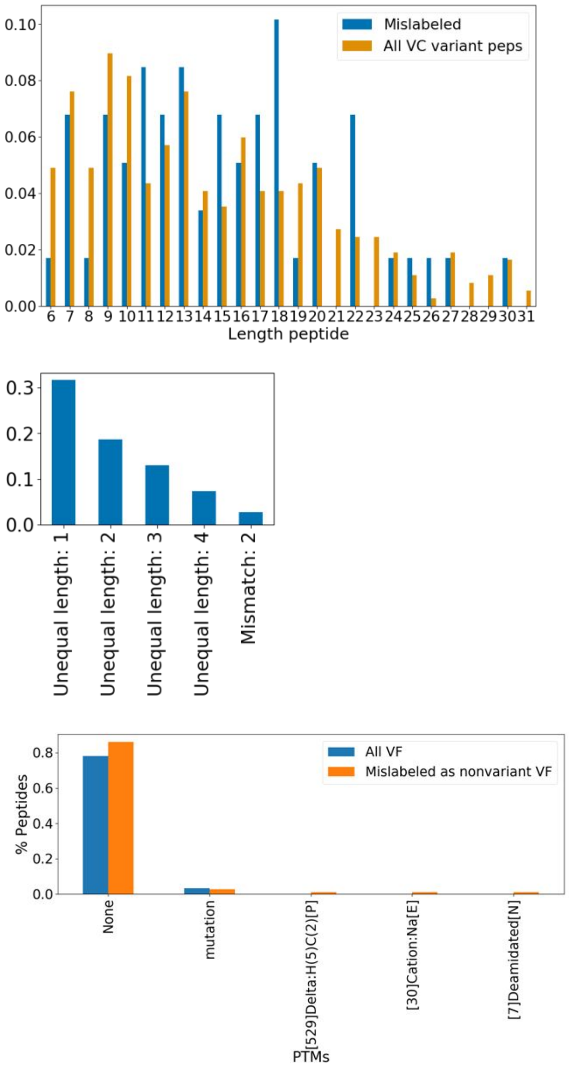
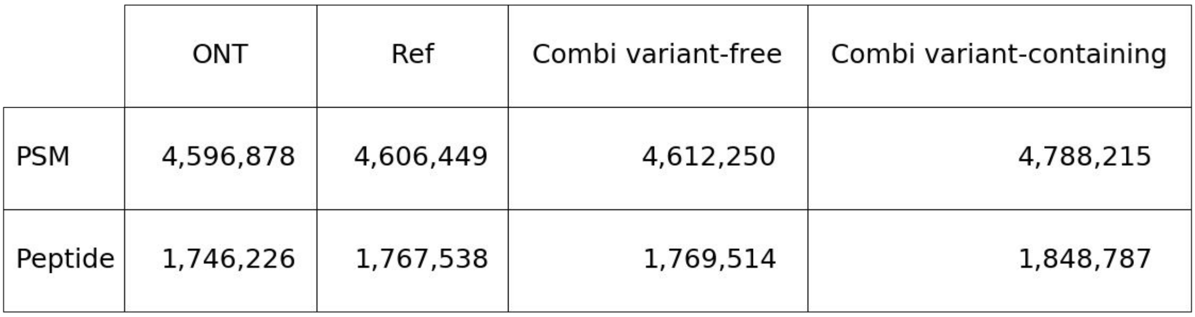


Figure S4. Investigation of false negative ('mislabeled') identifications by *ionbot*^{TM}. Top figure shows the density of mislabeled peptides per length, as compared to lengths of all variant peptides identified by the VC method. Middle figure shows the 5 most common causes of misidentification of variant peptides by *ionbot*^{TM}. Bottom figure shows unexpected modifications of the false negatives versus the unexpected modifications by all VF identifications. Unlabeled y axises refer to density.

Search database contents	**Sequences in GENCODE**	**Sequences in the ONT transcriptome**	**NA12878-specific variants**
ONT	No	Yes	No
Ref	Yes	No	No
VF	Yes	Yes	No
VC	Yes	Yes	Yes

Table S1: Side-by-side comparison of the contents of the search database



	ONT	Ref	Combi variant-free	Combi variant-containing
PSM	4,596,878	4,606,449	4,612,250	4,788,215
Peptide	1,746,226	1,767,538	1,769,514	1,848,787

Table S2. Absolute numbers of PSMs and peptides detected per method.