‘RECYCLOMICS’: OLD DATASETS, NEW DATABASES.

Image from:
LOOKING BEYOND THE KNOWN PROTEOME

Mass spectrum

Reference Protein Database from genomic annotation

Peptide Spectral Match

- b1: A NELLNVK, Y8
- b2: AN ELLLNVK, Y7
- b3: ANE LLLLNVK, Y6
- b4: ANEL LLLNVK, Y5
- b5: ANELL LNVK, Y4
- b6: ANELLLL NVK, Y3
- b7: ANELLNN VK, Y2
- b8: ANELLLNV K, Y1

- Y1: 147.1
- Y2: 245.2
- Y3: 298.1
- Y4: 473.3
- Y5: 588.4
- Y6: 699.5
- Y7: 829.5

m/z

100
50

b2: 185.1
b3: 298.1
b4: 411.2
b5: 524.3
b6: 810.5

100
200
300
400
500
600
700
800

LONI Protein Database from genomic annotation
LOOKING BEYOND THE KNOWN PROTEOME

Identification of peptides corresponding to novel proteoforms.

Cancer / Disease related Databases such as COSMIC, IARC p53, OMIM...

Deep genome sequencing data from ICGC, TCGA and CPTAC

RNASeq data (Customized OR Combined)

Mass spectrum

Reference Protein Database from genomic annotation

6-frame DNA sequences.
3-frame cDNA sequences.
XMAN: A HOMO SAPIENS Mutated-Peptide Database for the MS Analysis of Cancerous Cell States.

J. Proteome Res., Article ASAP
DOI: 10.1021/pr5004467

- unknown (X) peptide-level Mutation Analysis.
  - 872,125 sequences in the cancer mutation database.
  - 27,148 sequences in the disease mutation database.

- Mutations derived from various databases. COSMIC database (~1 million mutations), UniProt db (22,354 mutations), OMIM db (14,620 mutations), IARC p53 db (3,917 mutations). Databases were filtered, annotated and made non-redundant.

Figure 1. Venn diagram illustrating the contribution of the four databases to the count of unique peptide entries in the cancer database.
XMAN: A HOMO SAPIENS MUTATED-PePTIDE DATABASE FOR THE MS ANALYSIS OF CANCEROUS CELL STATES.

>CANCER_sp_Q8WUZ0,BCL7C_HUMAN B-cell CLL/lymphoma 7 protein family member C|c.454A>G|p.I152V|PGGVTAG|Missense|COSMIC|Large intestine(1)PVSPAGPPEGVPEEAQPRRLGQERDPGGVTAGSTDEPMLTKEEPVPELLEAEAPEAYPVFE PVPPVPEAAQGDTEDSEGAPPLKR

>DISEASE_sp_Q9HAN9,NMNA1_HUMAN Nicotinamide mononucleotide adenyllyltransferase 1|c.|p.I217N|NEWNAND|Missense|Swissprot|Leber congenital amaurosis 9 (LCA9) (1)FIYESDVLWKRHSNHVNEWNANDISSTKIRR

The description line starts with the symbol “ >” and provides the following information, from left to right, and delimited by “|”: (1) a tag that includes the description “CANCER” or “DISEASE” to mark the entry as being cancer or disease-related, the abbreviations “sp” /“tr” or “co” to indicate the database source that supplied the protein ID (SwissProt, TrEMbL, or COSMIC), and the protein ID and description; (2) the site and type of mutated cDNA; (3) the site and type of amino acid mutation; (4) a short sequence that includes the mutated amino acid encompassed by three amino acids at the left and right of the mutation site (the mutated amino acid is missing in case of deletions; for frameshifts, all amino acids at the right of the mutated site were included until a stop codon was found); (5) the type of mutation (missense, nonsense, deletion−frameshift, insertion−frameshift, complex−frameshift, deletion−inframe, insertion−inframe, or complex−inframe); (6) the originating database(s); and (7) the tissue source and number of hits that generated the mutated entry (provided in parentheses).
**Xman: A Homo Sapiens Mutated-Peptide Database for the MS Analysis of Cancerous Cell States.**

- Yearly updates to the database will be provided through follow-up publications.

- Correlation of mutation ratios “from each amino acid/to each amino acid” in the cancer database and the “1000 Genome Project”.

  - “We also note that while these mutations were detected in cancer tissues they do not necessarily represent cancer-specific mutations.”

  - “Amino acid exchanges that incurred very small differences in mass underscored the need, however, for manual validation or high-end mass spectrometers for data acquisition.”


*J. Proteome Res., Article ASAP DOI: 10.1021/pr5004467*
3D-fractionated salivary dataset.
Reference: J Proteome Res. 2009 doi: 10.1021/pr900675w
Instrument: LTQ/Orbitrap XL
RAW Files: 40
Search parameters: Methionine oxidation.

Oral pre-malignant lesion (OPML) samples.
Instrument: LTQ/Orbitrap XL
RAW Files: 42
Search parameters: Methionine oxidation.

’Snyderome’ dataset.
Reference: Cell. 2012 DOI: 10.1016/j.cell.2012.02.009
Instrument: LTQ/Orbitrap Velos
RAW Files: 16
Search parameters: TMT isobaric tags and Iodoacetamide @ cysteine.

Search against Xman db using GalaxyP workflows

Results and compare it with our earlier studies against 3-frame cDNA db.
Dataset Collection of MGF files

Subset of XMan db with human UniProt db

ProteinPilot™ Software

Peptide Summary

PepUdes

BLAST-P Workflow

Peptides corresponding to novel proteoforms.

Accession numbers
89 peptides (with at least 95% Conf)

- 51 were peptides corresponding to novel proteoforms.
- 10 of these were also identified as high-confidence PSMs using the 3-frame cDNA db.

89 peptides (all from COSMIC db)

88 CANCER db + 1 DISEASE db

72 peptides are from PRB2 locus

75 Frameshifts + 14 missense mutations.

Organs in which these mutations were detected.

- Kidney
- Large Intestine
- Endometrium
- Liver
- Lung
- Prostate
- Stomach
- Urinary Tract
- Hematopoietic

**3D-fractionated salivary dataset.**

**Reference:** J Proteome Res. 2009 doi: 10.1021/pr900675w

**Instrument:** LTQ/Orbitrap XL

**RAW Files:** 40

**Search parameters:** Methionine oxidation.
Oral pre-malignant lesion (OPML) samples.

**Reference:** Proteomics. 2013 doi: 10.1002/pmic.201200352

**Instrument:** LTQ/Orbitrap XL

**RAW Files:** 42

**Search parameters:** Methionine oxidation.

- 18 peptides (all from COSMIC db)
- 18 CANCER db
- 9 peptides are from PRB2 locus
- 10 Frameshifts + 8 missense mutations.

18 peptides (with at least 95% Conf)
5 were peptides corresponding to novel proteoforms.

Organs in which these mutations were detected.
PRB1 and PRB2 Locus

Human chromosome 12

DOI: 10.1021/pr500812t
J. Proteome Res., Just Accepted Manuscript
z.umn.edu/pgfirstlook
‘Snyderome’ dataset.

Reference: Cell. 2012 DOI: 10.1016/j.cell.2012.02.009

Instrument: LTQ/Orbitrap Velos

RAW Files: 16

Search parameters: TMT isobaric tags and Iodoacetamide @ cysteine.

File under investigation with ProteinPilot software team.
WHERE ARE THE DATASETS and DATABASES?


CPTAC:
WHERE ARE THE DATASETS and DATABASES?

Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next generation sequencing data.

Proteomics 2014. DOI: 10.1002/pmic.201400206

(c) Two-stage-FDR

Figure 2: Overlap between novel identifications from unified and single sample database.
### WHERE ARE THE DATASETS and DATABASES?

Construction and assessment of individualized proteogenomic databases for large-scale analysis of non-synonymous single nucleotide variants.


---

#### Table 1: Overview about three different methods used in this study to construct customized nsSNVs containing databases derived from RNA-seq data.

<table>
<thead>
<tr>
<th>Database design</th>
<th>Short name</th>
<th>Remarks</th>
<th>Number of database entries</th>
<th>Number of nsSNVs</th>
<th>Number of detected nsSNVs</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>all frames</td>
<td>Translation of RNA sequences into possible reading frames</td>
<td>51,239</td>
<td>31,354</td>
<td>354</td>
<td>Does not rely on prediction software</td>
<td>Large size</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inflation with artificial sequences</td>
</tr>
<tr>
<td>2</td>
<td>correct frame</td>
<td>Translation into predicted reading frame</td>
<td>10,225</td>
<td>9,234</td>
<td>366</td>
<td>Independent from genome annotation</td>
<td>Compact size</td>
</tr>
<tr>
<td>3</td>
<td>SNV effect</td>
<td>Incorporation of predicted nsSNVs into the canonical protein sequence</td>
<td>2,631</td>
<td>3,480</td>
<td>341</td>
<td>Requires accurate genome annotation</td>
<td>Compact size</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hampered by inconsistent protein annotations in different databases</td>
</tr>
</tbody>
</table>
WHERE ARE THE WORKFLOWS?

Dataset Collection of MGF files

- Peptide Summary

   Galaxy-P Workflow to identify peptides from Xman database

   Peptides

   BLAST-P Workflow

   Peptides corresponding to novel proteoforms.

   PSM Evaluation

   Post-processing of information from identified peptides.

   - QC plots: Examine the quality of the results with Quality Control plots.
   - Reshake PRIDE: re-analyze public datasets in PRIDE as if they were your own.

https://code.google.com/p/peptide-shaker/
CONCLUSIONS

• If you have new datasets, there are databases / resources available to search against so as to get the most out of your data.

• For your older datasets, there are databases / resources available to search against so as to get the most out of your data.

• If you have none of the above - there are datasets and databases available so as to get the most out of your data.