

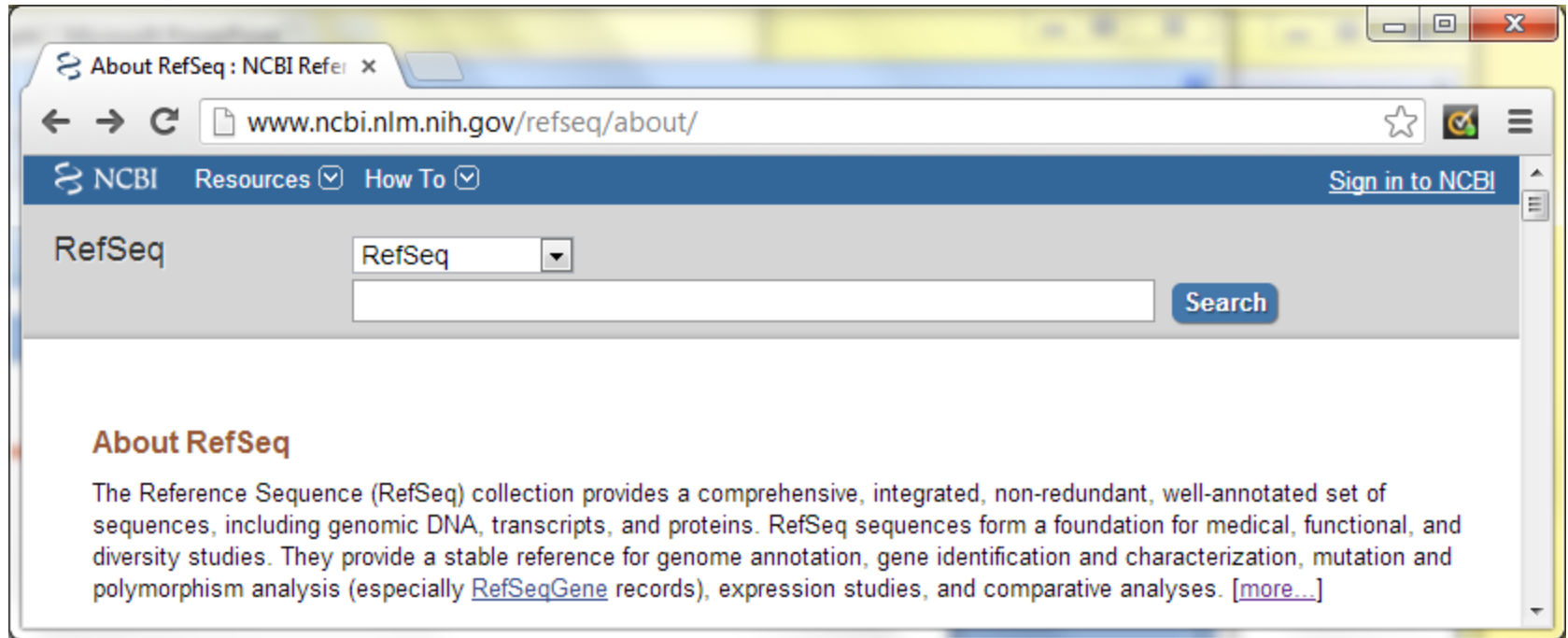
# **Proteomics Informatics - Databases, data repositories and standardization (Week 7)**

---

# Protein Sequence Databases

---

# RefSeq



## Distinguishing Features of the RefSeq collection include:

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current knowledge of sequence data and biology
- data validation and format consistency
- ongoing curation by NCBI staff and collaborators, with reviewed records indicated

<http://www.ncbi.nlm.nih.gov/books/NBK21091/>

# Ensembl

The screenshot shows the Ensembl Genome Browser website. The browser address bar displays 'useast.ensembl.org/index.html'. The page features a dark blue header with the Ensembl logo, navigation links for 'BLAST/BLAT', 'BioMart', 'Tools', and 'More', and a search bar with the placeholder text 'Search all species...'. Below the header, there is a search box with a dropdown menu set to 'All species' and a 'Go' button. A text box below the search box provides examples: 'e.g. BRCA2 or rat X:100000..200000 or coronary heart disease'. To the right, a 'New to Ensembl?' section lists several resources: 'Learn how to use Ensembl' with video tutorials, 'Add custom tracks' using the Control Panel, 'Upload and analyse your data' to save to an account, and 'Search for a DNA or protein sequence' using BLAST or BLAT. A yellow box on the right contains a globe icon and the text 'Did you know...? Lost ... try our tutorials page.' Below the search box, a 'Browse a Genome' section describes the project's goal and lists 'Popular genomes' with icons for 'Human GRCh37' and 'Mouse GRCh38'.

- genome information for sequenced chordate genomes.
- evidenced-based gene sets for all supported species
- large-scale whole genome multiple species alignments across vertebrates
- variation data resources for 17 species and regulation annotations based on ENCODE and other data sets.

<http://www.ensembl.org/>

# UniProt



The screenshot shows the UniProt website homepage in a browser window. The browser's address bar displays "www.uniprot.org". The website's navigation bar includes links for "Downloads", "Contact", and "Documentation/Help". Below this, there are tabs for "Search", "Blast", "Align", "Retrieve", and "ID Mapping". The "Search" tab is active, showing a search interface with a dropdown menu set to "Protein Knowledgebase (UniProtKB)", a "Query" input field, and buttons for "Search", "Advanced Search", and "Clear".

**WELCOME**

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

**What we provide**

<b>UniProtKB</b>	Protein knowledgebase, consists of two sections: <ul style="list-style-type: none"><li>★ Swiss-Prot, which is manually annotated and reviewed.</li><li>★ TrEMBL, which is automatically annotated and is <b>not</b> reviewed.</li></ul> Includes <a href="#">complete and reference proteome sets</a> .
------------------	---

**NEWS**

**UniProt release 2013\_03 - Mar 6, 2013**

Latest from the prokaryotic world: bacterial Cas9, a new for genome engineering | Cross-references to ChiTaRS a SABIO-RK | Removal of cross-references to 8 2D gel databases and AGD

- › Statistics for UniProtKB:
  - [Swiss-Prot](#) · [TrEMBL](#)
- › [Forthcoming changes](#)
- › [News archives](#)

[Follow @uniprot](#) 548 followers

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

<http://www.uniprot.org/>

# Species-Centric Consortia

---

For some organisms, there are consortia that provide high-quality databases:

**Yeast** (<http://yeastgenome.org/>)

**Fly** (<http://flybase.org/>)

**Arabidopsis** (<http://arabidopsis.org/>)

# FASTA

---

## RefSeq:

>gi|168693669|ref|NP\_001108231.1| zinc finger protein 683 [Homo sapiens]

```
MKEESAAQLGCCHRPMALGGTGGSLSPSLDFQLFRGDQVFSACRPLPDMVDAHGPSCASWLCPLPLAPGRSALLACLQDL
DLNLCTPQPAPLGTDLQGLQEDALSMKHEPPGLQASSTDDKKFTVKYPQNKDKLGKQPERAGEGAPCPAFSSHNSSSPPP
LQNRKSPSPLAFPCPPVNSISKELPFLHAFYPGYPLLLPPPHLFTYGALPSDQCPHLLMLPQDPSYPTMAMPSSLMMV
NELGHPSARWETLLPYPGAFAQASGQALPSQARNPGAGAAPTDSPGLERGGMASPAKRVPLSSQTGTAALPYPLKKKNGKI
LYECNICGKSFGQLSNLKVHLRVHSGERPFCALCQKSFTQLAHLQKHHLVHTGERPHKCSVCHKRFSSSSNLKTHLRLH
SGARPFQCSVCRSRFTQHIHLKLHHRLHAPQPCGLVHTQLPLASLACLAQWHQALDLMAVASEKHMGYDIDEVKVSSTS
Q GKARAVSLSSAGTPLVMGQDQNN
```

## Ensembl:

>ENSMUSP00000131420 pep:known supercontig:NCBIM37:NT\_166407:104574:105272

gene:ENSMUSG00000092057 transcript:ENSMUST00000167991

```
MFSLMKRRRKSSSNTLRNIVGCRISHCWKEGNEPVTQWKAIVLGQLPTNPSTLYLVKYDGDIDSIYGQELYSDDRILNLKVL
PPIVVFPQVRDAHLARALVGRAVQQKFERKDGSEVNWRGVVLAQVPIMKDLFYITYKKDPALYAYQLLDDYKEGNLHMIPD
TPPAEERSGGDSVDLIGNWVQYTRKDGSKKFGKVYQVLDNPSVFFIKFHGDIHIYVYTMVPKILEVEKS
```

## UniProt:

>sp|Q16695|H31T\_HUMAN Histone H3.1t OS=Homo sapiens GN=HIST3H3 PE=1 SV=3

```
MARTKQTARKSTGGKAPRKQLATKVARKSAPATGGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLMREIAQDFK
TDLRFQSSAVMALQEACESYLVGLFEDTNLCVIHAKRVTIMPKDIQLARRIRGERA
```

[http://en.wikipedia.org/wiki/FASTA\\_format](http://en.wikipedia.org/wiki/FASTA_format)

# PEFF - PSI Extended Fasta Format

---

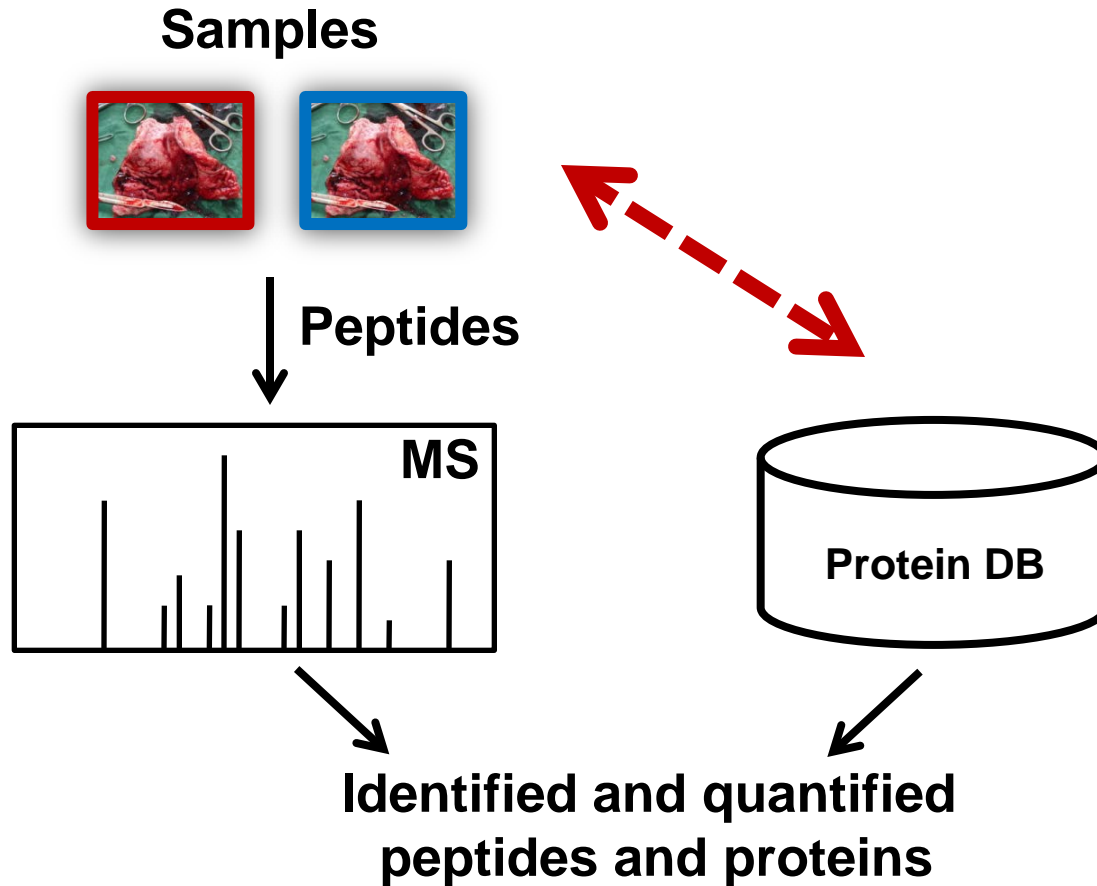
```
>sp:P06748 \ID=NPM_HUMAN
\Pname=(Nucleophosmin) (NPM) (Nucleolar phosphoprotein
B23) (Numatrin) (Nucleolar protein NO38)
\NcbiTaxId=9606
\ModRes=(125|MOD:00046)(199|MOD:00047)
\Length=294
```

```
>sp:P00761 \ID=TRYP_PIG
\Pname=(Trypsin precursor) (EC 3.4.21.4) \NcbiTaxId=9823
\Variant=(20|20|V)
\Processed=(1|8|PROPEP)(9|231|CHAIN)
\Length=231
```

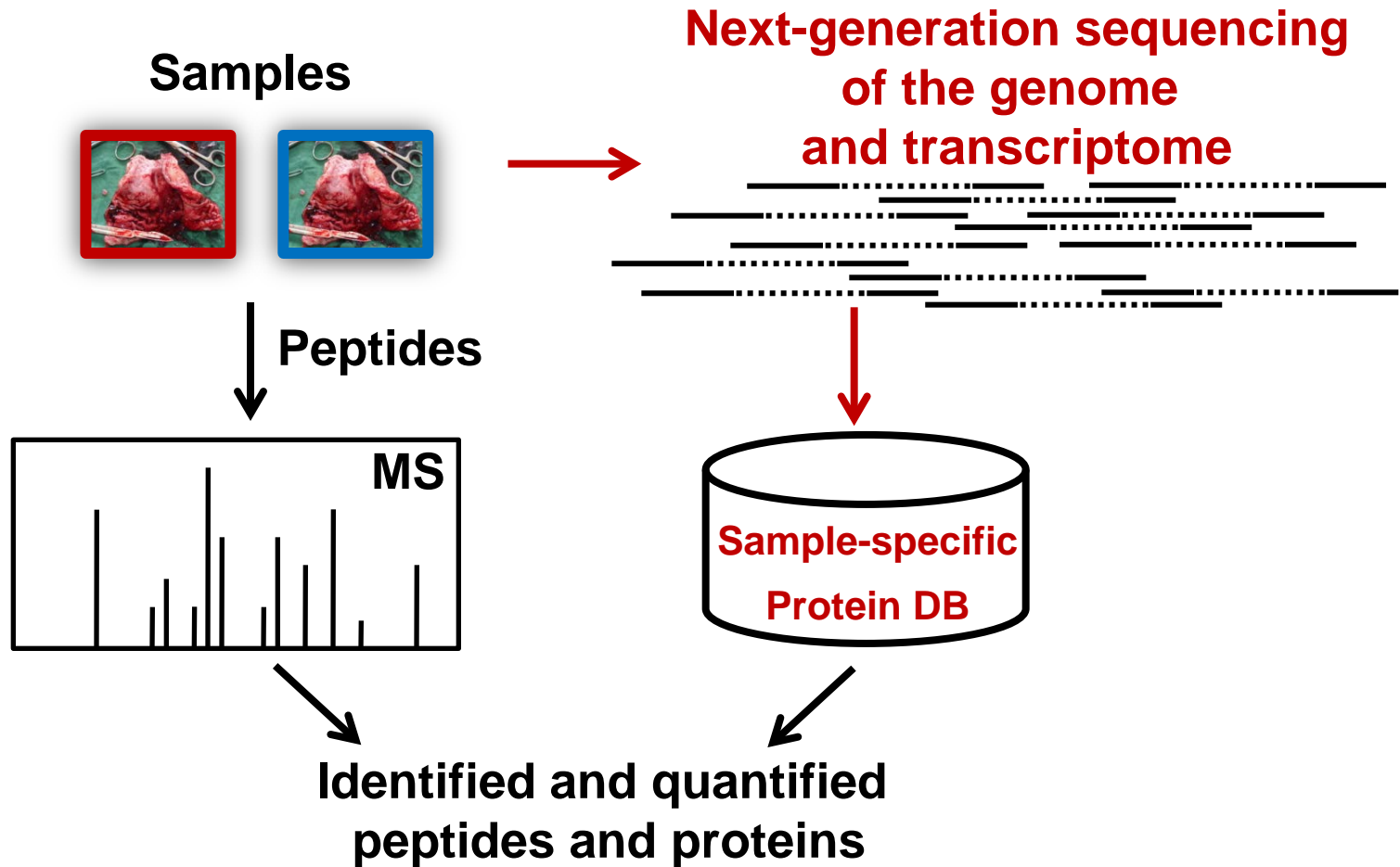
<http://www.psidev.info/node/363>



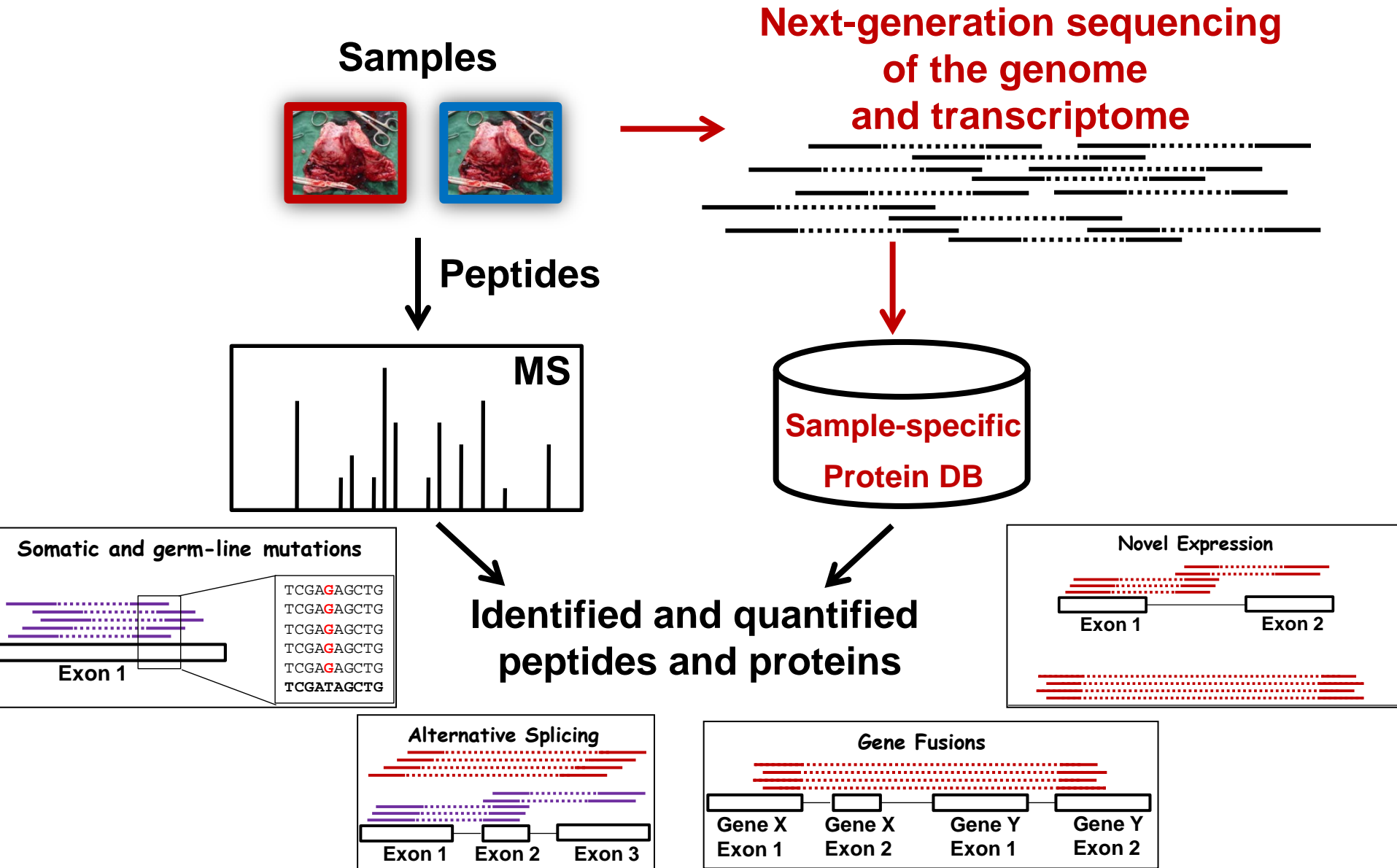
# Sample-specific protein sequence databases



# Sample-specific protein sequence databases



# Sample-specific protein sequence databases

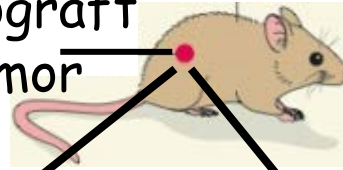


# Proteomics and Transcriptomics of Breast Tumors



Primary Breast tumor

Xenograft tumor



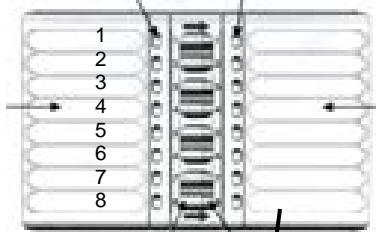
Illumina HiSeq

RNA-Seq

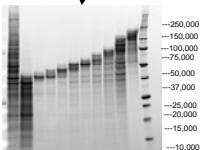


Sample Loading Chamber

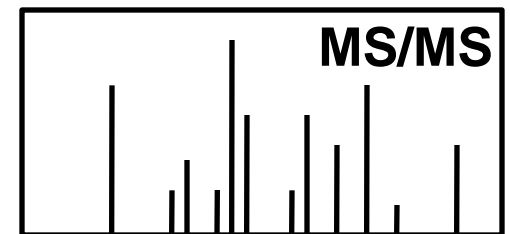
Sample Collection Chamber



Stacking Gel Resolving Gel

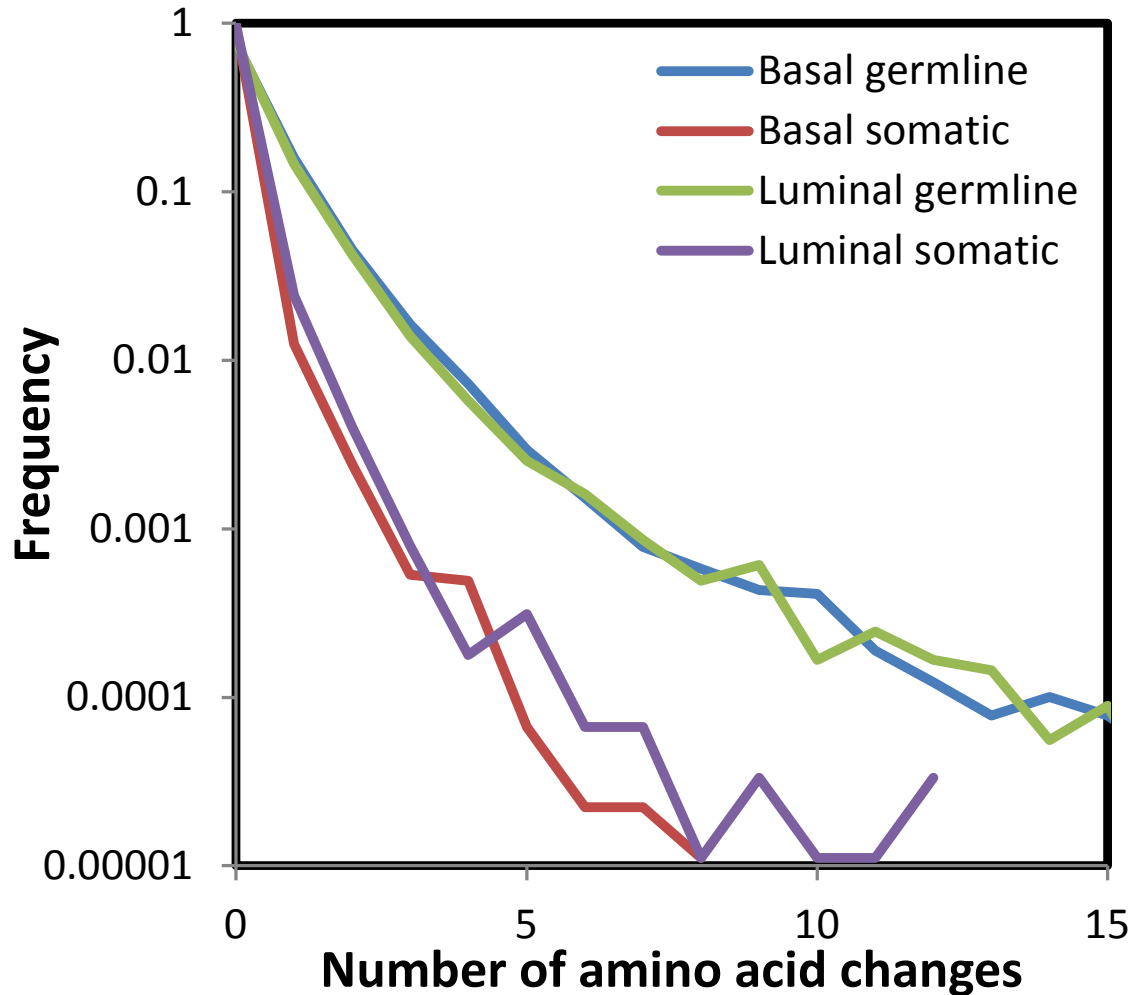


ABI 5600 Triple TOF



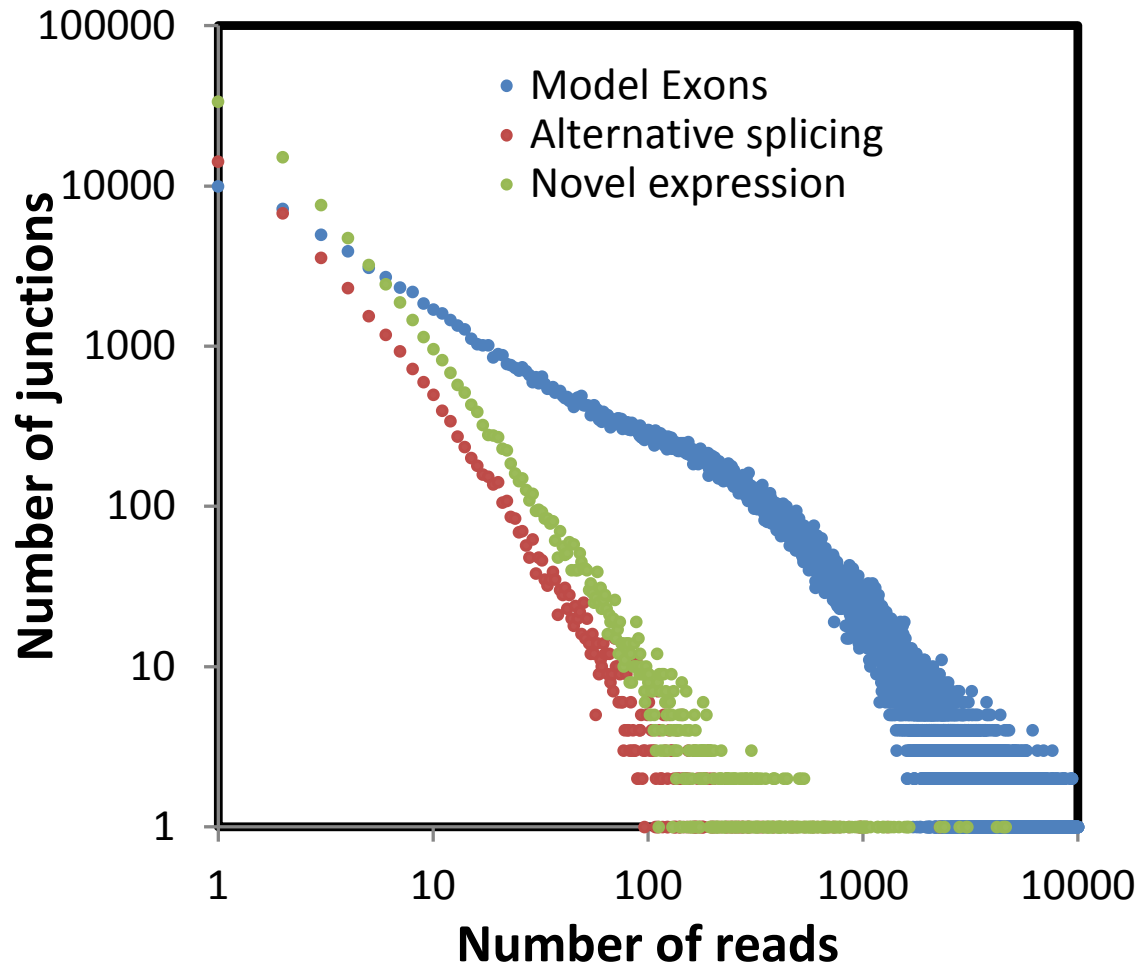
MS/MS

# Germline and Somatic Variants



The frequency of proteins as a function of the number of amino acid changes due to germline and somatic variants for the basal and luminal breast tumor xenografts

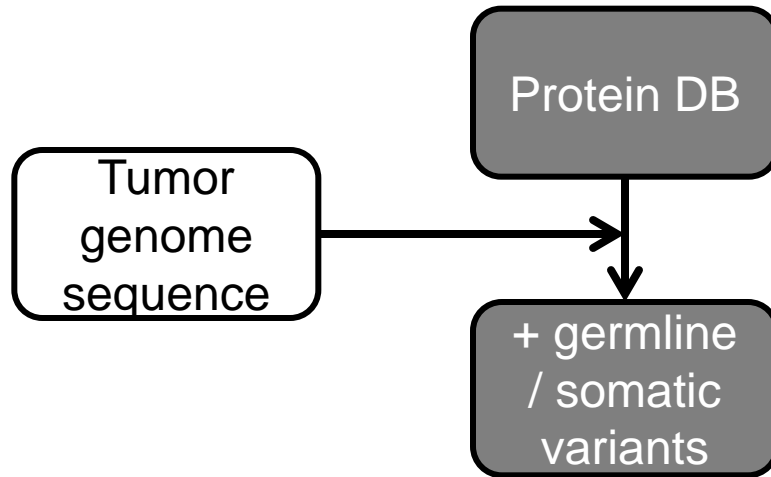
# Alternative Splicing



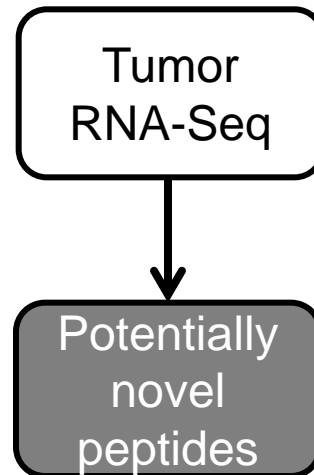
The number of exon/exon junctions as a function of the number of RNA-Seq reads for the basal breast tumor xenograft.

# Protein identification using sample-specific sequence databases

---



Germline variants	362
Somatic variants	9



1114	Potentially novel peptides
70	Spans splice site

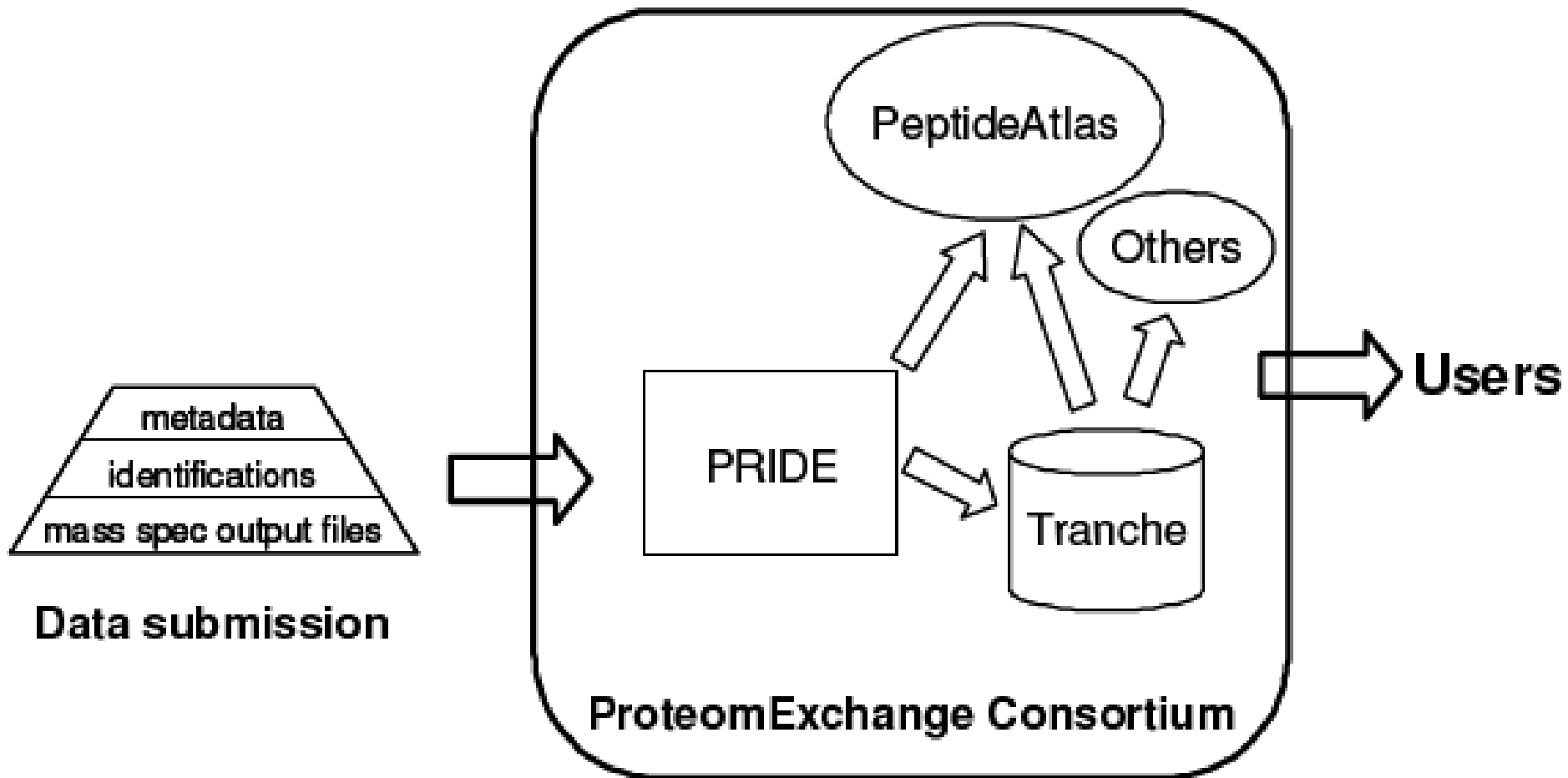
# Data Repositories

---



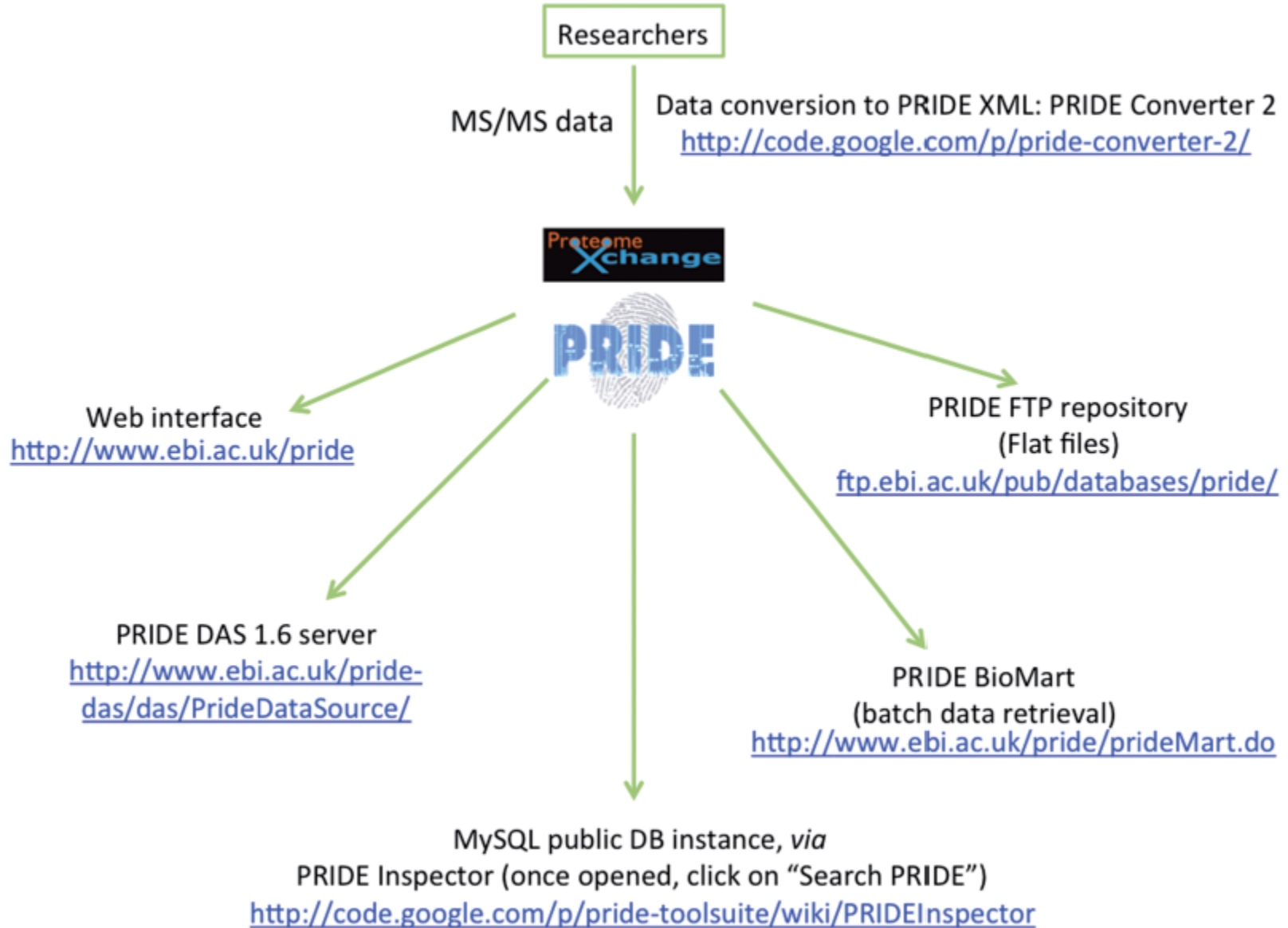
# ProteomeExchange

---



<http://www.proteomeexchange.org/>

# PRIDE



<http://www.ebi.ac.uk/pride/>

# PeptideAtlas

ISB Home

PeptideAtlas

## PEPTIDEATLAS HOME

Seattle Proteome  
Center

### PEPTIDEATLAS:

Overview  
Contacts  
Data Contributors  
Publications  
Software  
Database Schema  
Feedback  
FAQ

### ATLAS DATA:

Data Repository  
Human Plasma  
(Farrah, et al.)  
HPPP Data Central  
PeptideAtlas Builds  
Search Database

Contribute Data  
Genome Browser  
Setup

### RELATED:

SRMATlas  
Phosphopep  
Uniprot

PeptideAtlas

Search PeptideAtlas:

GO

[Expanded Search](#)

**PeptideAtlas** is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments. Mass spectrometer output files are collected for human, mouse, yeast, and several other organisms, and searched using the latest search engines and protein sequences. All results of sequence and spectral library searching are subsequently processed through the [Trans Proteomic Pipeline](#) to derive a probability of correct identification for all results in a uniform manner to insure a high quality database, along with false discovery rates at the whole atlas level. Results may be queried and browsed at the PeptideAtlas web site. The raw data, search results, and full builds can also be downloaded for other uses.



[PeptideAtlas  
Chromosome  
Explorer  
\(Human only\)](#)

SRMATlas

[SRMATlas  
interface for  
selection of  
best available  
SRM transitions](#)

[PeptideAtlas  
Raw Data  
Repository](#)





[PeptideAtlas SRM  
Experiment Library  
\(PASSEL\)](#)

CHP

[PeptideAtlas and the  
Chromosome-Centric  
Human Proteome Project](#)

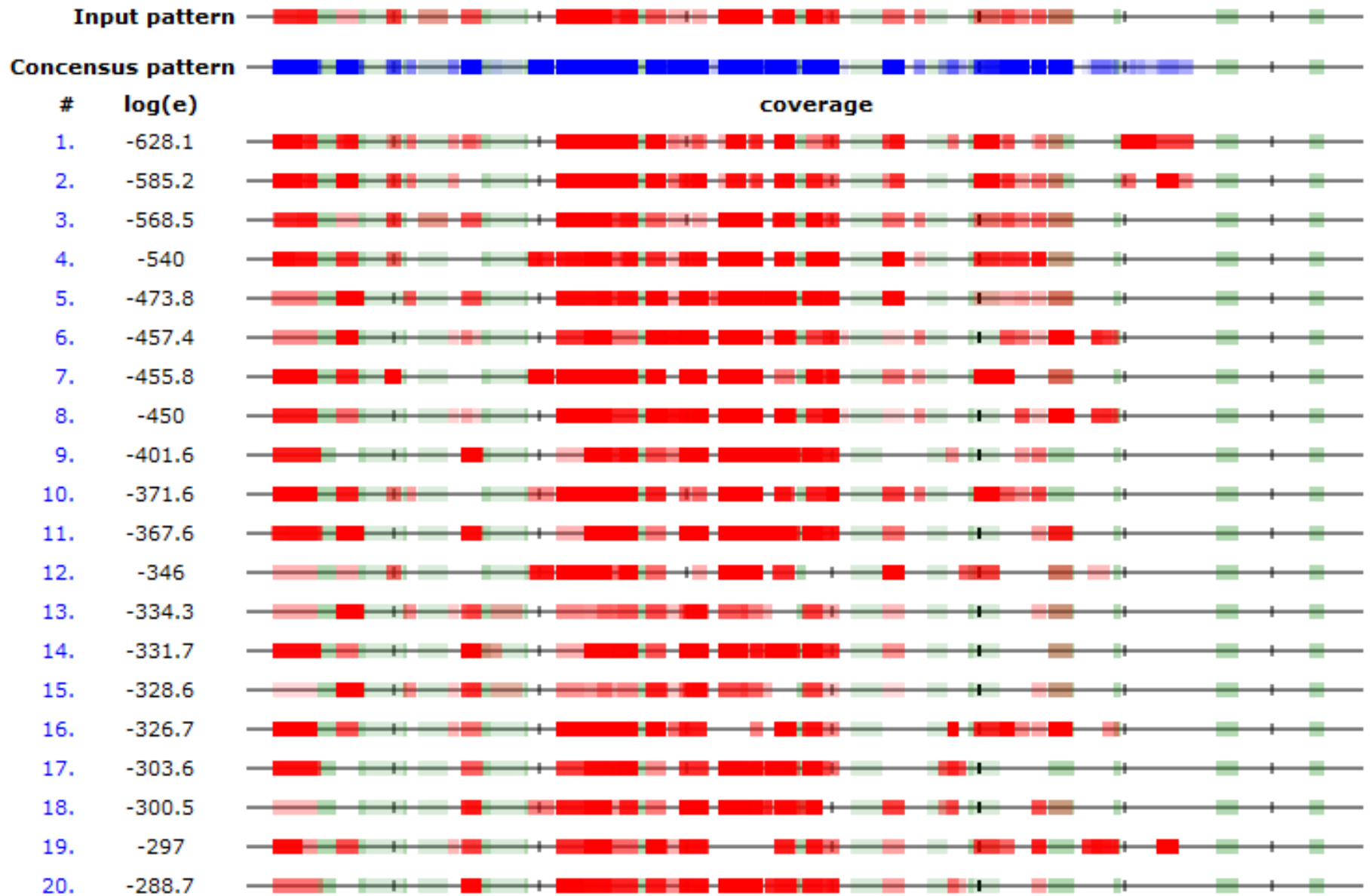
<http://www.peptideatlas.org/>

# The Global Proteome Machine Databases (GPMDB)

<i>gpmdb</i>	accession BTO home	gpm # Chr # statistics	sequence SNAP species	keyword pSYT thegpm	GO lists about	<i>the gpm</i>
<b>Information</b> <a href="#">about the GPM</a> <a href="#">about gpmdb</a> <a href="#">send us email</a>	<b>gpmDB statistics for Sun Mar 3 11:49:49 2013 UTC (#3315)</b>					<b>GPM sponsors</b> <ul style="list-style-type: none"><li>• <a href="#">Proteome Software</a></li><li>• <a href="#">Beavis Informatics</a></li><li>• <a href="#">MCPSB, UM</a></li><li>• <a href="#">LMSGIC, RU</a></li></ul>
<b>Search sites</b>  Eukaryote proteomes <a href="#">1</a> <a href="#">2</a> <a href="#">3</a> <a href="#">4</a> <a href="#">5</a> <a href="#">6</a> <a href="#">7</a>  Boutique proteomes human    mouse cow      bacteria plant    rat	models = 217,125 proteins = 84,408,917 distinct proteins = 1,724,816 protein redundancy = 48.9 × <b>peptides = 687,211,623</b> distinct peptides = 4,286,043 peptide redundancy = 160.3 × residues = 9,620,962,722					<b>data</b> <ul style="list-style-type: none"><li>• <a href="#">Tranche</a></li><li>• <a href="#">PeptideAtlas</a></li><li>• <a href="#">PRIDE</a></li></ul>
<b>Algorithms</b> <a href="#">X! P3</a> <a href="#">X! Hunter</a>	statistics archive: <a href="#">GPMDB</a> pages viewed: <a href="#">global map</a> US visits <a href="#">map</a> European visits <a href="#">map</a> Asian visits <a href="#">map</a> Oceania visits <a href="#">map</a> South American visits <a href="#">map</a> African visits <a href="#">map</a>					<b>projects</b> <ul style="list-style-type: none"><li>• <a href="#">iMOP</a></li><li>• <a href="#">HPP</a></li><li>• <a href="#">C-HPP</a></li><li>• <a href="#">HPFP</a></li><li>• <a href="#">The HPA</a></li></ul>
<b>Information</b> <a href="#">gpmDB</a> <a href="#">wiki</a> <a href="#">review</a> <a href="#">lists</a>						<b>general info</b> <ul style="list-style-type: none"><li>• <a href="#">ENSEMBL</a></li><li>• <a href="#">STRING DB</a></li><li>• <a href="#">Unimod</a></li><li>• <a href="#">NCTA</a></li></ul>
<b>Some species</b>  						<b>pathways</b> <ul style="list-style-type: none"><li>• <a href="#">KEGG</a></li><li>• <a href="#">Reactome</a></li></ul>

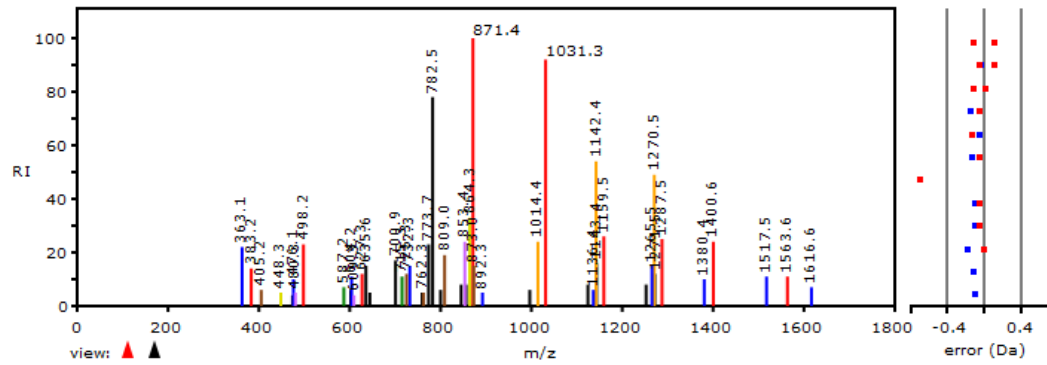
<http://gpmdb.thegpm.org>

# Comparison with GPMDB



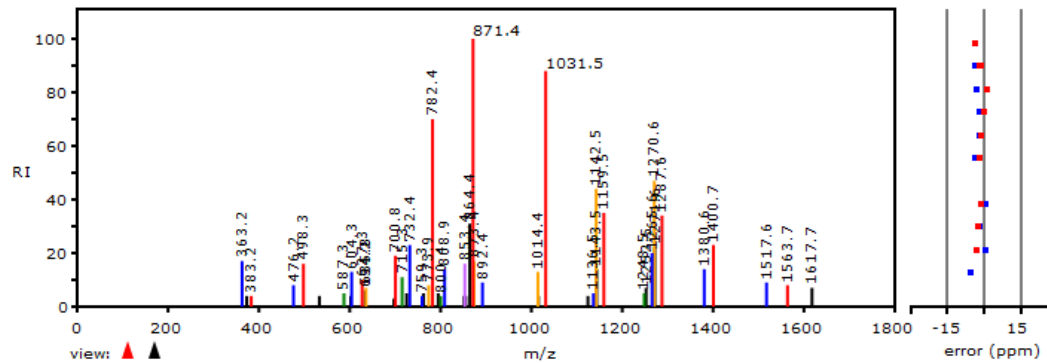
Most proteins show very reproducible peptide patterns

# Comparison with GPMDB



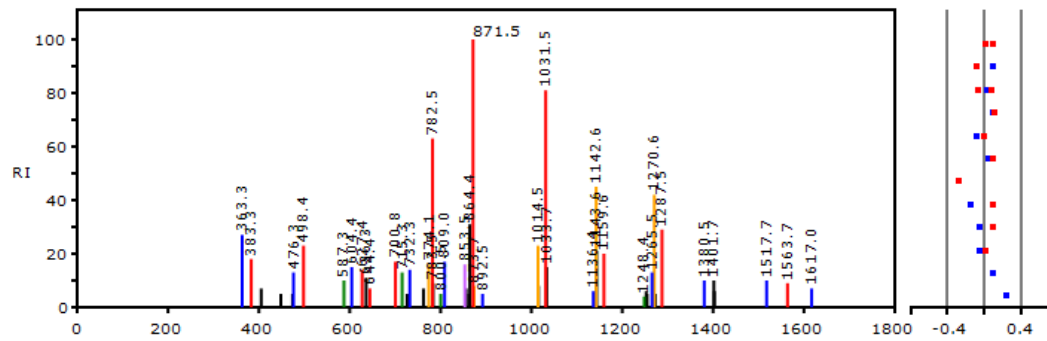
1.  $\cos(\theta) = 0.98$ ,  $z = 2$ ,  $\log(e) = -14.8$ ,  $m+h = 1762.8218$  (P)

A Q Y L Q Q C P F E D H V K

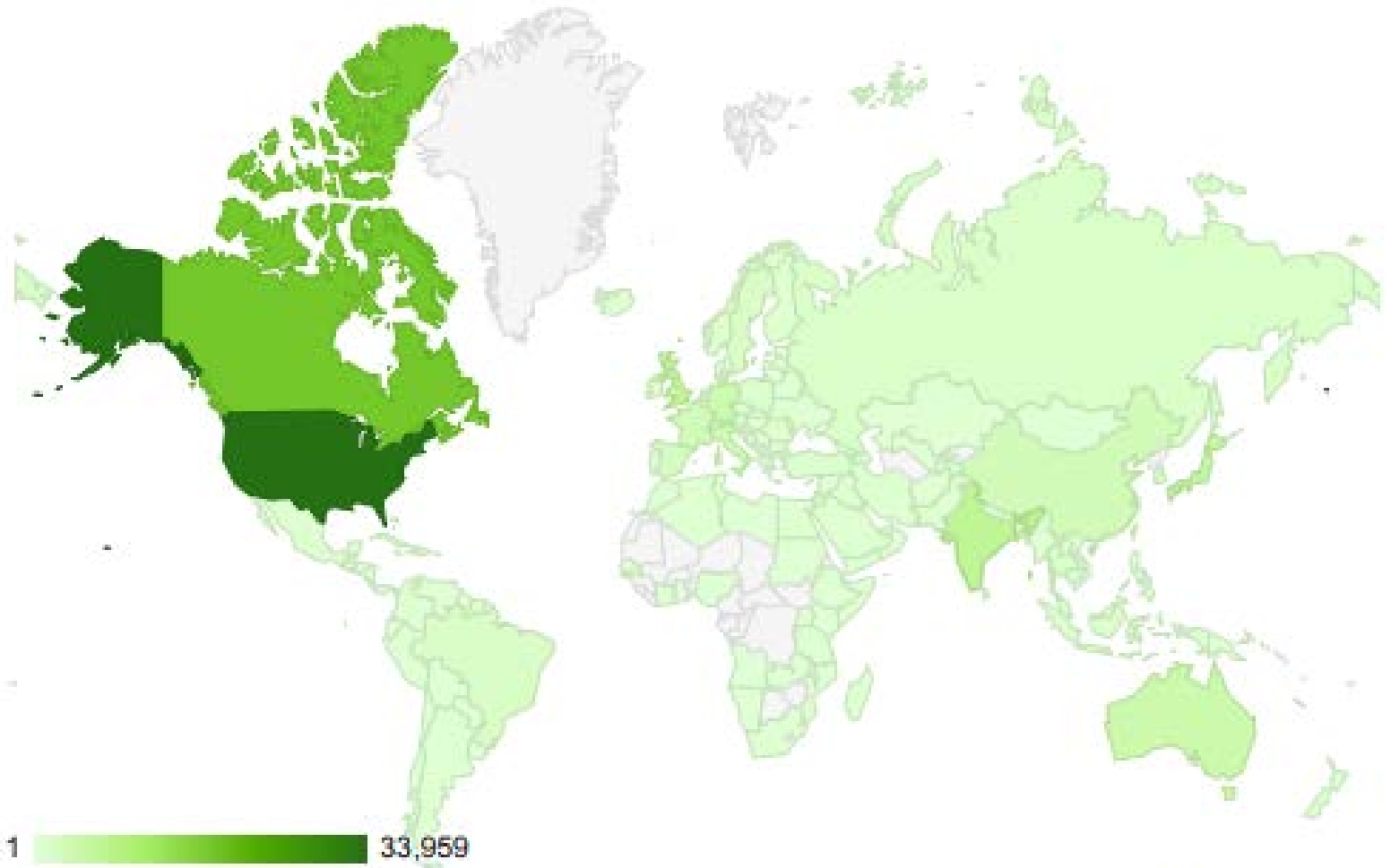


2.  $\cos(\theta) = 0.96$ ,  $z = 2$ ,  $\log(e) = -13.5$ ,  $m+h = 1762.8216$  (P)

A Q Y L Q Q C P F E D H V K



# GPMDB usage last month



Visits

**111,242**

% of Total: 100.00% (111,242)

Pages / Visit

**4.06**

Site Avg: 4.06 (0.00%)

Avg. Visit Duration

**00:04:55**

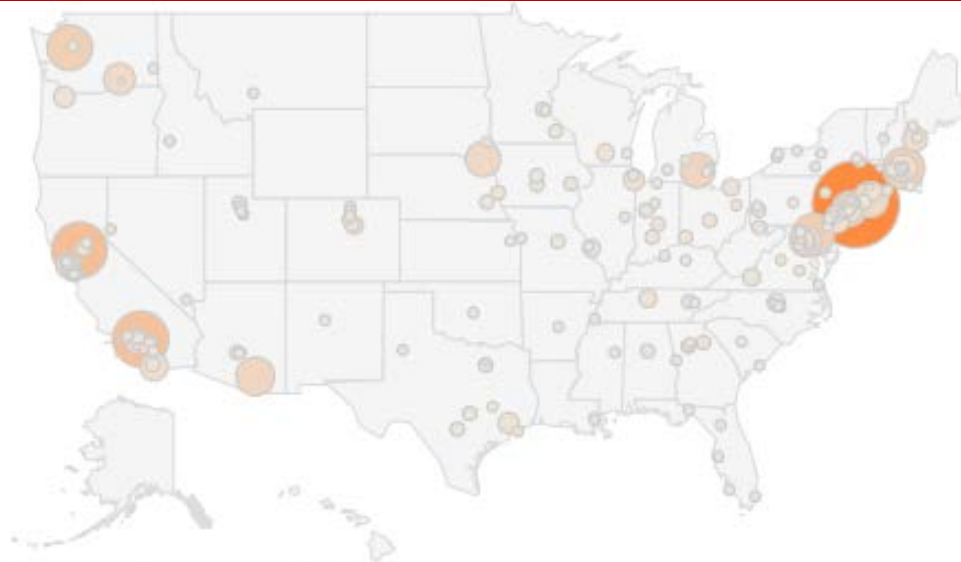
Site Avg: 00:04:55 (0.00%)

% New Visits

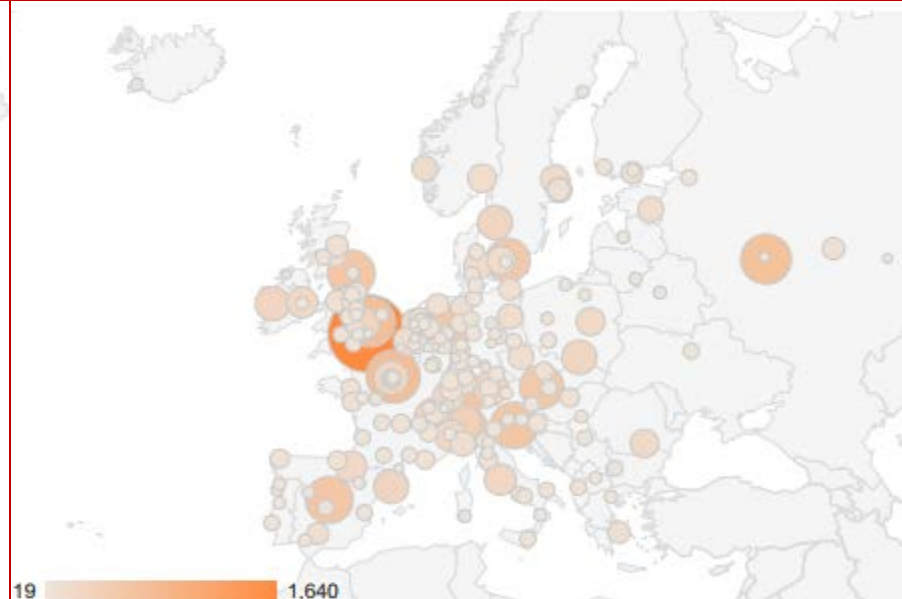
**51.81%**

Site Avg: 51.81% (0.00%)

# GPMDB usage last month



15 4,597



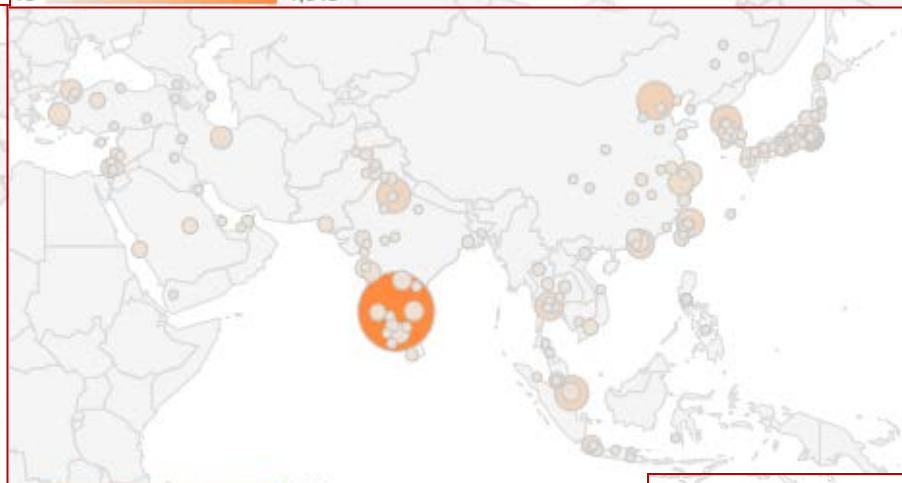
19 1,640



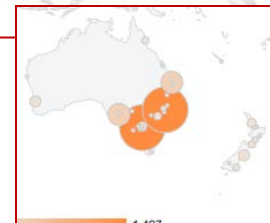
184



179



15 3,453



1,497



# GPMDB Data Crowdsourcing

---

Any lab performs experiments



Raw data sent to public repository (TRANCHE, PRIDE)



Data imported by GPMDB



Data analyzed & accepted/rejected



Accepted information loaded into public collection



General community uses information and inspects data

# Information for including a data set in GPMDDB

---

## 1. MS/MS data (required)

1. MS raw data files
2. ASCII files: mzXML, mzML, MGF, DTA, etc.
3. Analysis files: DAT, MSF, BION

## 2. Sample Information (supply if possible)

1. Species : human, yeast
2. Cell/tissue type & subcellular localization
3. Reagents: urea, formic acid, etc.
4. Quantitation: SILAC, iTRAQ
5. Proteolysis agent: trypsin, Lys-C

## 3. Project information (suggested)

1. Project name
2. Contact information

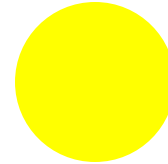
# How to characterize the evidence in GPMDB for a protein?

---

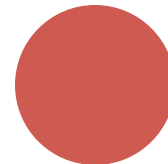
High confidence



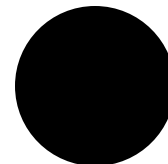
Medium confidence



Low confidence



No observation



# Statistical model for 212 observations of TP53

Star <sub>t</sub>	End	N	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	Skew	Kurt
214	248	539	0.15	0.18	0.22	0.17	0.15	0.07	0.03	0.01	0.01	0.00	-0.01	-2.01
249	267	1010	0.04	0.09	0.13	0.16	0.16	0.14	0.13	0.06	0.04	0.05	-0.08	-1.89
182	196	832	0.09	0.15	0.20	0.19	0.18	0.13	0.05	0.01	0.00	0.00	-0.12	-1.84
250	267	4	0.25	0.00	0.25	0.00	0.25	0.00	0.00	0.00	0.00	0.25	0.48	-2.28
1	24	269	0.10	0.12	0.12	0.17	0.12	0.12	0.14	0.04	0.04	0.03	-0.33	-0.88
24	65	51	0.22	0.22	0.20	0.14	0.06	0.00	0.04	0.08	0.02	0.04	0.47	-1.62
66	101	334	0.09	0.08	0.11	0.11	0.09	0.11	0.09	0.13	0.08	0.12	0.10	-1.21
249	273	60	0.02	0.00	0.20	0.10	0.13	0.25	0.20	0.07	0.03	0.00	0.45	-1.36
214	242	10	0.00	0.10	0.00	0.00	0.00	0.00	0.30	0.20	0.20	0.20	0.54	-1.39
214	239	32	0.03	0.06	0.16	0.16	0.09	0.22	0.09	0.16	0.00	0.03	0.20	-0.99
111	120	117	0.09	0.20	0.15	0.26	0.29	0.01	0.00	0.00	0.00	0.00	0.62	-1.36
251	267	16	0.00	0.00	0.13	0.25	0.19	0.13	0.13	0.13	0.06	0.00	0.24	-0.60
214	241	14	0.00	0.00	0.00	0.07	0.29	0.21	0.07	0.29	0.00	0.07	0.87	-0.97
159	174	100	0.30	0.25	0.31	0.03	0.07	0.03	0.01	0.00	0.00	0.00	0.99	-1.07
68	101	10	0.00	0.00	0.00	0.00	0.00	0.20	0.10	0.10	0.30	0.30	0.86	-0.91
235	248	30	0.00	0.03	0.00	0.00	0.30	0.20	0.23	0.13	0.03	0.07	0.81	-0.82



# Statistical model for observations of GRAP2

Start	End	N	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	Skew	Kurtosis
37	48	167	0.17	0.21	0.23	0.17	0.15	0.04	0.02	0.02	0.00	0.00	0.15	-2.06
84	97	224	0.05	0.15	0.18	0.19	0.18	0.13	0.06	0.04	0.01	0.01	0.08	-1.98
317	330	344	0.09	0.10	0.15	0.13	0.18	0.16	0.12	0.06	0.01	0.01	-0.50	-0.72
222	232	79	0.24	0.25	0.16	0.11	0.16	0.04	0.03	0.00	0.00	0.00	0.44	-1.53
164	184	59	0.27	0.22	0.29	0.12	0.05	0.00	0.02	0.03	0.00	0.00	0.82	-1.20
27	36	67	0.24	0.21	0.12	0.03	0.13	0.10	0.04	0.10	0.01	0.00	0.51	-0.66
278	312	10	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.30	0.20	0.30	0.86	-0.91
260	272	201	0.22	0.23	0.33	0.15	0.05	0.00	0.01	0.00	0.00	0.00	0.86	-0.79
98	106	52	0.33	0.21	0.29	0.10	0.08	0.00	0.00	0.00	0.00	0.00	0.95	-0.75
27	48	11	0.00	0.09	0.18	0.09	0.27	0.09	0.18	0.00	0.09	0.00	0.61	-0.16
7	26	15	0.13	0.33	0.33	0.13	0.07	0.00	0.00	0.00	0.00	0.00	1.17	-0.02
113	127	9	0.33	0.33	0.11	0.00	0.11	0.11	0.00	0.00	0.00	0.00	1.20	0.14
66	75	118	0.13	0.19	0.27	0.37	0.03	0.01	0.01	0.00	0.00	0.00	1.19	0.20
261	272	80	0.36	0.34	0.14	0.04	0.05	0.03	0.03	0.00	0.01	0.01	1.50	0.72
250	259	2	0.00	0.00	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	1.78	1.41
222	233	4	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.78	1.41
186	221	19	0.16	0.05	0.37	0.00	0.05	0.21	0.11	0.05	0.00	0.00	1.50	2.14
317	324	9	0.56	0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.85	1.87
58	65	3	0.33	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.28	4.77
208	221	8	0.13	0.00	0.13	0.13	0.50	0.00	0.13	0.00	0.00	0.00	2.26	5.88
234	259	7	0.00	0.57	0.14	0.14	0.00	0.14	0.00	0.00	0.00	0.00	2.41	6.34
49	57	89	0.69	0.24	0.07	0.01	0.00	0.00	0.00	0.00	0.00	0.00	2.61	6.98
113	121	11	0.64	0.09	0.18	0.09	0.00	0.00	0.00	0.00	0.00	0.00	2.64	7.39
76	83	21	0.95	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.15	9.93
188	221	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	3.16	10.00
66	83	1	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	3.16	10.00
317	328	2	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.16	10.00
110	121	1	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.16	10.00
226	232	1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.16	10.00
1	6	1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.16	10.00
128	133	1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.16	10.00

# DNA Repair



<a href="#">accession</a>	<a href="#">gpm #</a>	<a href="#">sequence</a>	<a href="#">keyword</a>	<a href="#">GO</a>
<a href="#">BTO</a>	<a href="#">Chr #</a>	<a href="#">SNAP</a>	<a href="#">pSYT</a>	<a href="#">lists</a>
<a href="#">home</a>	<a href="#">statistics</a>	<a href="#">species</a>	<a href="#">thegpm</a>	<a href="#">about</a>

Ontology Collection, GO:0006281 DNA repair

[excel](#)

[txt](#)

#	accession	total	log(e)	EC	description
1.	<a href="#">ENSP00000263801</a>	2168	-2647.6	●	TP53BP1, tumor protein p53 binding protein 1
2.	<a href="#">ENSP00000371475</a>	2117	-2647.6	●	TP53BP1, tumor protein p53 binding protein 1
3.	<a href="#">ENSP00000411532</a>	2643	-2274.6	●	TOP2A, topoisomerase (DNA) II alpha 170kDa
4.	<a href="#">ENSP00000355759</a>	4539	-1988.3	●	PARP1, poly (ADP-ribose) polymerase 1
5.	<a href="#">ENSP00000369497</a>	217	-1889.2	●	BRCA2, breast cancer 2, early onset
6.	<a href="#">ENSP00000381295</a>	1132	-1325.3	●	E3 ubiquitin-protein ligase UHRF1 (EC 6.3.2.-) (Ubiquitin-like PHD and RING finger domain-containing protein 1) (Ubiquitin-like-containing PHD and RING finger domains protein 1) (Inverted CCAAT box-binding protein of 90 kDa) (Transcription factor ICBP90) [Source:Uniprot/SWISSPROT;Acc:Q96T88]
7.	<a href="#">ENSP00000262952</a>	1182	-1282.1	●	UHRF1, ubiquitin-like with PHD and ring finger domains 1
8.	<a href="#">ENSP00000409986</a>	1105	-1282.1	●	UHRF1, ubiquitin-like with PHD and ring finger domains 1
9.	<a href="#">ENSP00000261609</a>	785	-1268.8	●	HERC2, hect domain and RLD 2
10.	<a href="#">ENSP00000265421</a>	367	-1169.8	●	POLB, polymerase (DNA directed), beta


# DNA Repair









































553.	<a href="#">ENSP00000359285</a>	11	-2.8	●	CHRNA4, cholinergic receptor, nicotinic, alpha 4
554.	<a href="#">ENSP00000364389</a>	13	-2.7	●	CDC14B, CDC14 cell division cycle 14 homolog B ( <i>S. cerevisiae</i> )
555.	<a href="#">ENSP00000413377</a>	9	-2.5	●	CCDC108, coiled-coil domain containing 108
556.	<a href="#">ENSP00000409117</a>	9	-2.5	●	CCDC108, coiled-coil domain containing 108
557.	<a href="#">ENSP00000404368</a>	4	-2.4	●	PARP3, poly (ADP-ribose) polymerase family, member 3 [Source:HGNC Symbol;Acc:2Q9Y6F1; NP_005476]
558.	<a href="#">ENSP00000385879</a>	4	-2.4	●	KBTBD12, kelch repeat and BTB (POZ) domain containing 12
559.	<a href="#">ENSP00000430639</a>	5	-2.1	●	ENDOV, endonuclease V
560.	<a href="#">ENSP00000404213</a>	4	-2.1	●	REV1, REV1 homolog ( <i>S. cerevisiae</i> )
561.	<a href="#">ENSP00000430509</a>	4	-2.1	●	ENDOV, endonuclease V
562.	<a href="#">ENSP00000298129</a>	9	-2	●	ZNF488, zinc finger protein 488 [Source:HGNC Symbol;Acc:23535; Q96MN9; NP_694579]
563.	<a href="#">ENSP00000379054</a>	8	-2	●	ZNF488, zinc finger protein 488
564.	<a href="#">ENSP00000387138</a>	1	-1.7	●	RAD9B, RAD9 homolog B ( <i>S. pombe</i> ) [Source:HGNC Symbol;Acc:21700]
565.	<a href="#">ENSP00000378754</a>	2	-1.7	●	FANCC, Fanconi anemia, complementation group C [Source:HGNC Symbol;Acc:3584]
566.	<a href="#">ENSP00000293273</a>	6	-1.7	●	RDM1, RAD52 motif 1
567.	<a href="#">ENSP00000380672</a>	3	-1.4	●	CDNA FLJ39025 fis, clone NT2RP7004559, weakly similar to ENDONUCLEASE C1F12.06 (EC 3.1.-.-) (Hypothetical protein FLJ35220). [Source:Uniprot/SPTREMBL;Acc:Q8N8Q3]
568.	<a href="#">ENSP00000421819</a>	2	-1.2	●	POLK, polymerase (DNA directed) kappa [Source:HGNC Symbol;Acc:9183]
569.	<a href="#">ENSP00000403782</a>	2	-1.2	●	POLK, polymerase (DNA directed) kappa [Source:HGNC Symbol;Acc:9183]
570.	<a href="#">ENSP00000393993</a>	0	nf	●	POLH, polymerase (DNA directed), eta [Source:HGNC Symbol;Acc:9181]
571.	<a href="#">ENSP00000402713</a>	0	nf	●	OGG1, 8-oxoguanine DNA glycosylase

out of 571



# TP53BP1:p, tumor protein p53 binding protein 1

Page 1 of 129 for ENSP00000263801 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >> | 129 | ● observed 2564 x 

#	log(e)	%	model		Show: coverage   metadata
1.	-2647.6	70.6	G   P   O		
2.	-1311.6	63.3	G   P   O		
3.	-997.4	55.4	G   P   O		
4.	-997.4	55.4	G   P   O		
5.	-997.4	55.4	G   P   O		
6.	-997.4	55.4	G   P   O		
7.	-997.4	55.4	G   P   O		
8.	-970.2	54.4	G   P   O		
9.	-683.9	40.7	G   P   O		
10.	-627.5	39.3	G   P   O		
11.	-610.9	31.3	G   P   O		
12.	-599.5	33.3	G   P   O		
13.	-553.7	32.3	G   P   O		
14.	-513.1	25.2	G   P   O		
15.	-472.7	25.0	G   P   O		
16.	-463.6	33.2	G   P   O		
17.	-461	29.2	G   P   O		
18.	-458.8	32.9	G   P   O		
19.	-447.8	30.3	G   P   O		
20.	-433.6	23.3	G   P   O		

# TP53BP1:p, tumor protein p53 binding protein 1



ENSP00000263801: TP53BP1:p, tumor protein p53 binding protein 1

log(e) = -2647.6 [Source: HGNC 11999]

IPR015125 53-BP1 Tudor

IPR001357 (x6) BRCT dom



```
1 mdptgsqldsdfsqqdtpcliiedsqpesqvleddsgshfsmlsruhlpnlqthkenpvl 60
  MDPTGSQLDSDFSQQDTPCLIIEDSQPESQVLEDDSGSHFMSLSRHLPLNLOTHKENPVL
61 vvsnpeqtageergdgnsgfnehlkenkvadpvdssnldtcgsisqvieqlpqpnrts 120
  VVSNPEQTAGEERG DGN SGF NEHLKENKVADPVDSSNLDTCGSISQVIEQLPQPNRTSSV
121 lgmsvesapaveeekgeeleqkekekeedtsgnthslgaedtassqlgfgvlelsqsq 180
  LGMSVESAPAVEEEKGEELEQKEKEKEEDTSGNTHSLGAEDTASSQLGFGVLELSQSQD
181 veentvpyevdkeqlqsvttnsgytrlsdvdantaikheeqsnedipiaeqsskdipvt 240
  VEENTVVPYEVDK EQ LQSVTTNSGYTRLSDVDANTAIKHEEQSNEDIPIAEQSSKDIPVTA
241 qpskdvhv vkeqnp ppar sedmp fspkasvaameakeqlsaqelmesglqiqkspepe 300
  QPSKDVHVVK EQNPPPARSEDMPFS PKASVAAMEAKEQLSAQELMESGLQIQKSPEPEVL
301 stqedlfdqsnktvssdgcstpsreeggcslastpattlhllqsgqrslvqdslstnss 360
  STQEDLFDQSNKTVSSDGCSTPSREEGGCSLASTPATTLHLLQLSGQRSLVQDSLSTNSS
361 dlvapspdafirstpfivps spteqegrqdkpmdtsvlseeggepfqkklqsgepvele 420
  DLVAPSPDAFRSTPFIVPSSPTEQEGRODKPMDTSVLSEEGGEPFQKKLQSGEPVELENP
421 pllpestvspqastp isqstp vfp pgs l p i p s q p q f s h d i f i p s p s l e e q s n d g k k d g d m
  PLLPESTVSPQASTPISQSTPVFPFPGSLPIPSQPQFSHDIFIPSPSLEEQSDGKKDGM
481 hsssltvecsktsei epkns pedlgl sltgdsc klm l stseysqspkmeslsshridedg 540
  HSSSLTVECSKTSEIEPKNSPEDLGLSLTGDSCKLMLSTSEYSQSPKMESLSSHRIDEDG
541 entqiedtepmspvl nskfvpaendsil mnpa qdgevqlsqnddktkgddtdtrddisil 600
  ENTQIEDTEPMSPVLNSKFVPAENDSILMNPACDGEVQLSQNDDKTKGDDTDTRDDISIL
601 atgckgreetvaedvc idl tcdsgsqavps patr sealssvldqeeameikehhpeegss 660
  ATGCKGREETVAEDVCIDLTCDSGSQAVPSPATRSEALSSVLDQEEAMEIKEHHPEEGSS
661 gseveeipetpcesqgeelkeenmesvplhlsltetqsgglclqkempkcecseamevet 720
  GSEVEEIPETPCESQGEELKEENMESVPLHLSLTETQSQGLCLQKEMPRKECSEAMEVET
721 svisidspqklaildqelehkeqea weeats edssvvi vdvk epsrvdvsc eplegvek 780
  SVISIDSPQKLAILDQELEHKEQEA WEEATSEDES SVVIVDVKEPSRVDVSC EPLEGVEK
781 cdsqswediapeiepcaenrldtkeeksveyegdlksgtaetepveqdssqpslplvra 840
  CDSQSWEDIAPEIEPCAENR LDTKEEKSVEYEGDLKSGTAETEPVEQDSSQPSLPLVRA
```

# Sequence Annotations

show legend ?

`mvdqp` lower case sequence is the latest sequence from ENSEMBL for this accession number

`reklqee` lower case transition from black to blue letters indicates an exon boundary; a red residue indicates a triplet shared between exons

`MVDQP` upper case sequence is the protein sequence originally analyzed

`dvdnas` **synonymous SNP** with no residue change and **non-synonymous SNP** which changes the residue

`DIMR` residues part of at least one observed peptide domain

`LREEQ` residues predicted to be difficult to observe by standard techniques

`HFOL` residue found is a **single amino-acid polymorphism**

`AYNG` residue found is **chemically modified**

**Complete mods:** i. Carbamidomethyl@C, Carbamidomethyl@U

**Potential mods:** i. Oxidation@M, Label:+6 Da@K, Label:+6 Da@R  
ii. Oxidation@M, Oxidation@W, Deamidated@N, Deamidated@Q  
iii. Dioxidation@M, Dioxidation@W

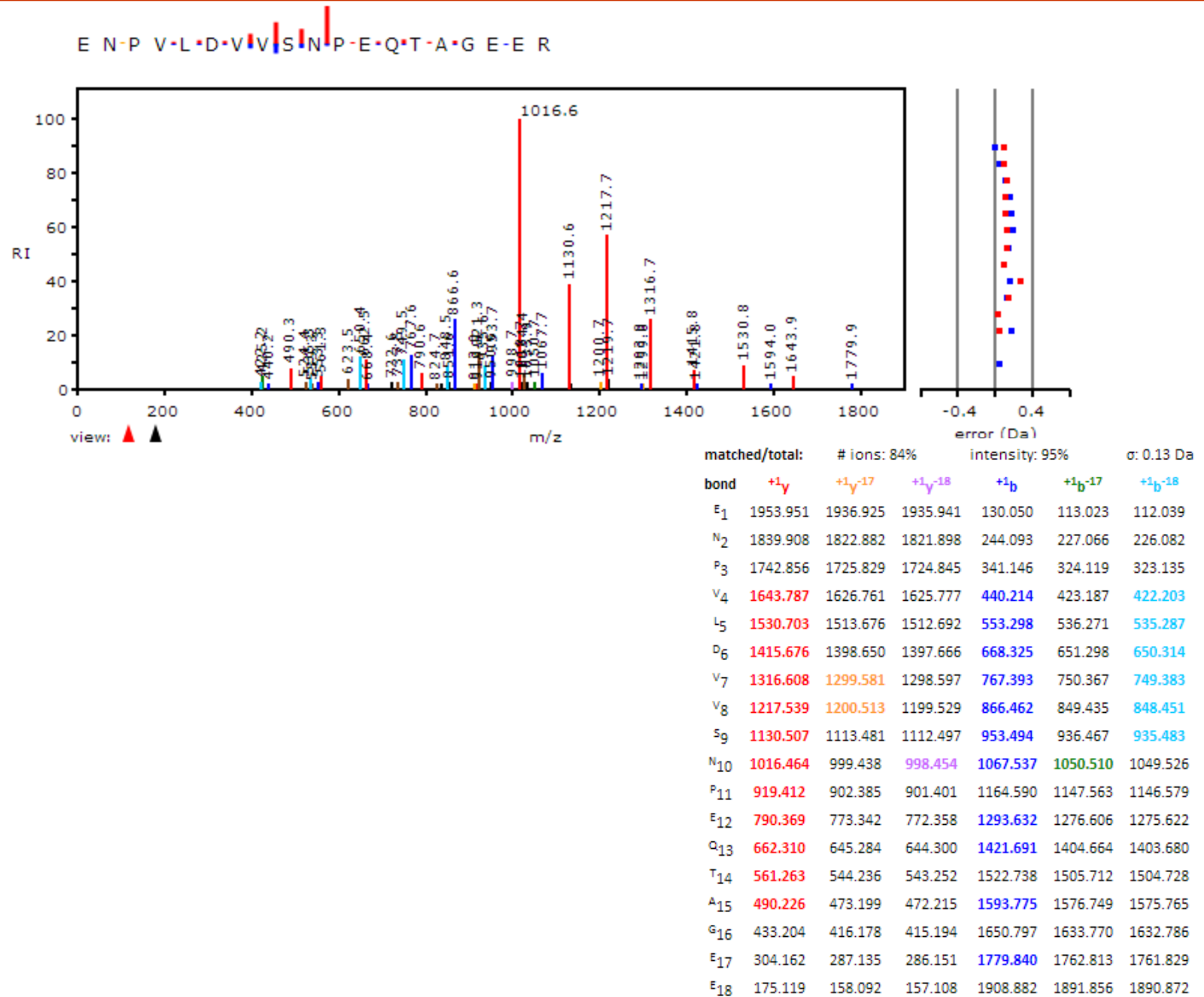
**Protein-specific PTMs:** i. Phospho@S, Phospho@T, Phospho@Y

**N-terminal:** i. Ammonia-loss@Q, Ammonia-loss@C, Dehydrated@E (peptide)  
ii. ragged, Acetyl (protein)

# TP53BP1:p, tumor protein p53 binding protein 1

spectrum	log(e)	log(l)	m+h	delta	ζ	sequence	n
1124.1	-4.2	6.11	1093.6208	0.0015	2/4	mIsr46 HLPNLQTHK <sup>54</sup> enpv	(323)
32342.1	-3.5	5.84	1087.6007	0.0009	3/4	mIsr46 HLPNLQTHK <sup>54</sup> enpv	(323)
14727.1	-14.2	4.91	2082.9938	0.0021	2/2	qthk55 ENPVLVVSN PEQTAGEER <sup>73</sup> gdgn	(1702)
15139.1	-10.1	6.47	2082.9938	0.0027	3/3	qthk55 ENPVLVVSN PEQTAGEER <sup>73</sup> gdgn	(1702)
3585.1	-11.4	5.97	1839.9083	0.0012	2/2	hken57 PVLVVSNPE QTAGEER <sup>73</sup> gdgn	(15)
20574.1	-8.0	5.02	1274.5760	-0.0007	2/3	geer74 GDGNSGFNEH LK <sup>85</sup> enkv	(359)
1585.1	-3.7	6.55	1275.5600	0.0015	3/3	geer74 GDGNSGFNEH LK <sup>85</sup> enkv	(359)
32608.1	-2.9	5.66	1657.7967	-0.0012	3/4	geer74 GDGNSGFNEH LKENK <sup>85</sup> vadp	(30)
32889.1	-2.1	5.25	1102.5276	-0.0021	2/3	ergd76 GNSGFNEHLK <sup>85</sup> enkv	(5)
1026.1	-3.2	5.62	937.4833	0.0016	2/3	gdgn78 SGFNEHLK <sup>85</sup> enkv	(10)
6246.1	-11.3	6.97	3045.4889	0.0052	3/3	kenk89 VADPVDSSNL DTGSGISQVI EQLQPQR <sup>116</sup> tssv	(2944)
6403.1	-10.9	4.72	3039.4688	0.0038	2/2	kenk89 VADPVDSSNL DTGSGISQVI EQLQPQR <sup>116</sup> tssv	(2944)
36424.1	-13.3	4.91	1965.9321	0.0010	2/2	qpnr117 TSSVLGMSVE SAPAVEEEK <sup>135</sup> geel	(169)
4458.1	-12.4	5.42	2775.3643	-0.0021	2/3	qpnr117 TSSVLGMSVE SAPAVEEEKG EELEOK <sup>142</sup> ekek	(3519)
37304.1	-9.0	4.60	2795.3139	0.0002	3/3	qpnr117 TSSVLGMSVE SAPAVEEEKG EELEQK <sup>142</sup> ekek	(3519)
2575.1	-9.7	6.10	2100.0381	0.0002	2/3	vlgm124 SVESAPAVEE EKGEELEOK <sup>142</sup> ekek	(170)
2542.1	-7.7	6.92	2100.0381	0.0006	3/3	vlgm124 SVESAPAVEE EKGEELEOK <sup>142</sup> ekek	(170)
2121.1	-5.3	5.18	1772.8549	-0.0004	3/3	msve127 SAPAVEEEKG EELEQK <sup>142</sup> ekek	(5)
2738.1	-9.6	5.39	2067.0391	0.0002	2/3	tvpy189 EVDKEQLQSV TTNSGYTR <sup>206</sup> Isdv	(35)
35349.1	-8.2	5.54	2054.9989	-0.0008	3/3	tvpy189 EVDKEQLQSV TTNSGYTR <sup>206</sup> Isdv	(35)

# TP53BP1:p, tumor protein p53 binding protein 1



# Peptide observations, catalase

---

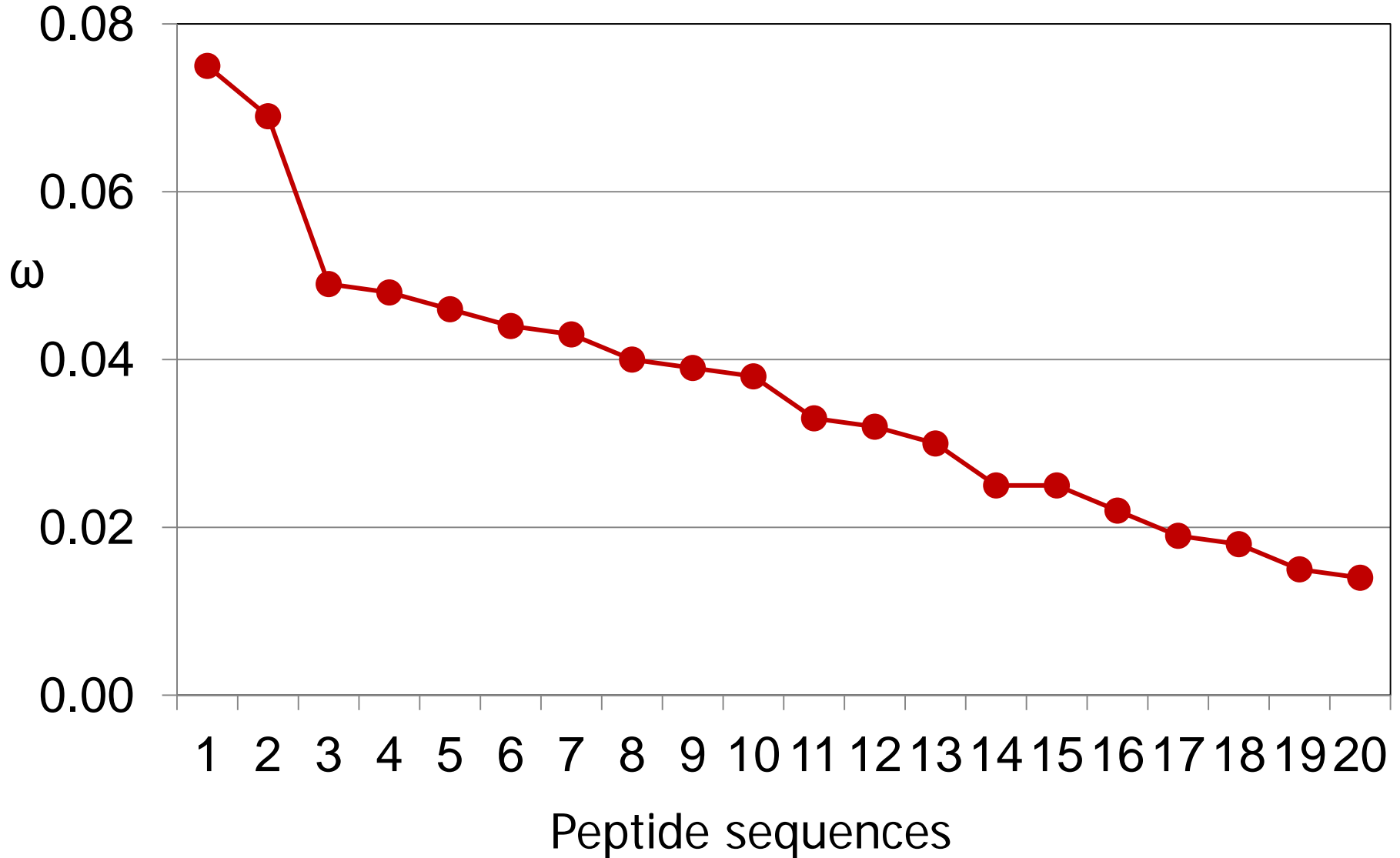
Peptide Sequence	Observations
FSTVAGESGSADTVR	2633
FNTANDDNVTQVR	2432
AFYVNVLNEEQR	1722
LVNANGEAVYCK	1701
GPLLVDVFTDEMAHFDR	1637
LSQEDPDYGIR	1560
LFAYPDTHR	1499
NLSVEDAAR	1400
FYTEDGNWDLVGNNTPIFFIR	1386
ADVLTTGAGNPVGDK	1338

# Peptide frequency ( $\omega$ ), catalase

---

Peptide Sequence	$\omega$
FSTVAGESGSADTVR	0.08
FNTANDDNVTQVR	0.07
AFYVNVLNEEQR	0.05
LVNANGEAVYCK	0.05
GPLLVDVFTDEMAHFDR	0.05
LSQEDPDYGIR	0.04
LFAYPDTHR	0.04
NLSVEDAAR	0.04
FYTEDGNWDLVGNNTPIFFIR	0.04
ADVLTTGAGNPVGDK	0.04

# Global frequency of observation ( $\omega$ ), catalase





# Omega ( $\Omega$ ) value for a protein identification

For any set peptides observed in an experiment assigned to a particular protein (*1 to j*):

$$\Omega(\textit{protein}) = \sum_j \omega_j$$

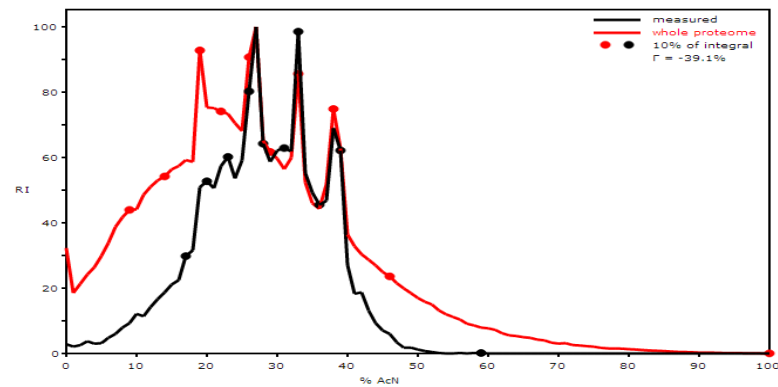
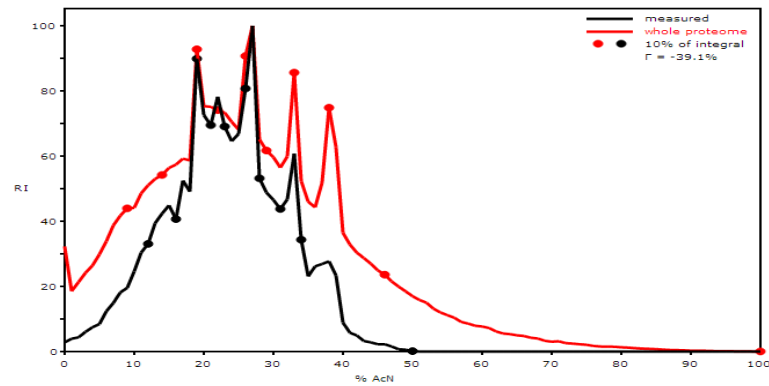
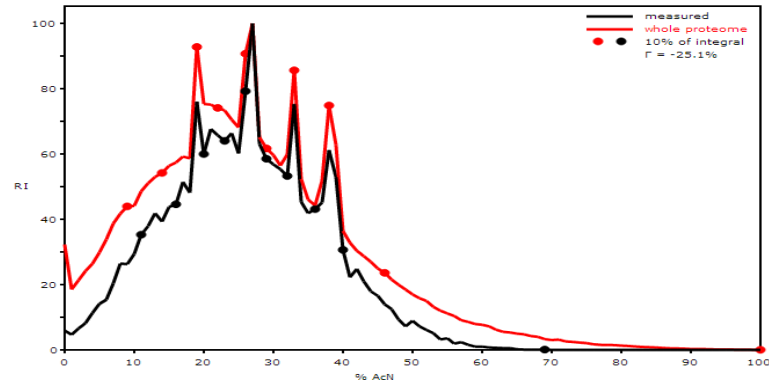
$$\Omega(\textit{protein}) \leq 1$$

# Protein $\Omega$ 's for a set of identifications

---

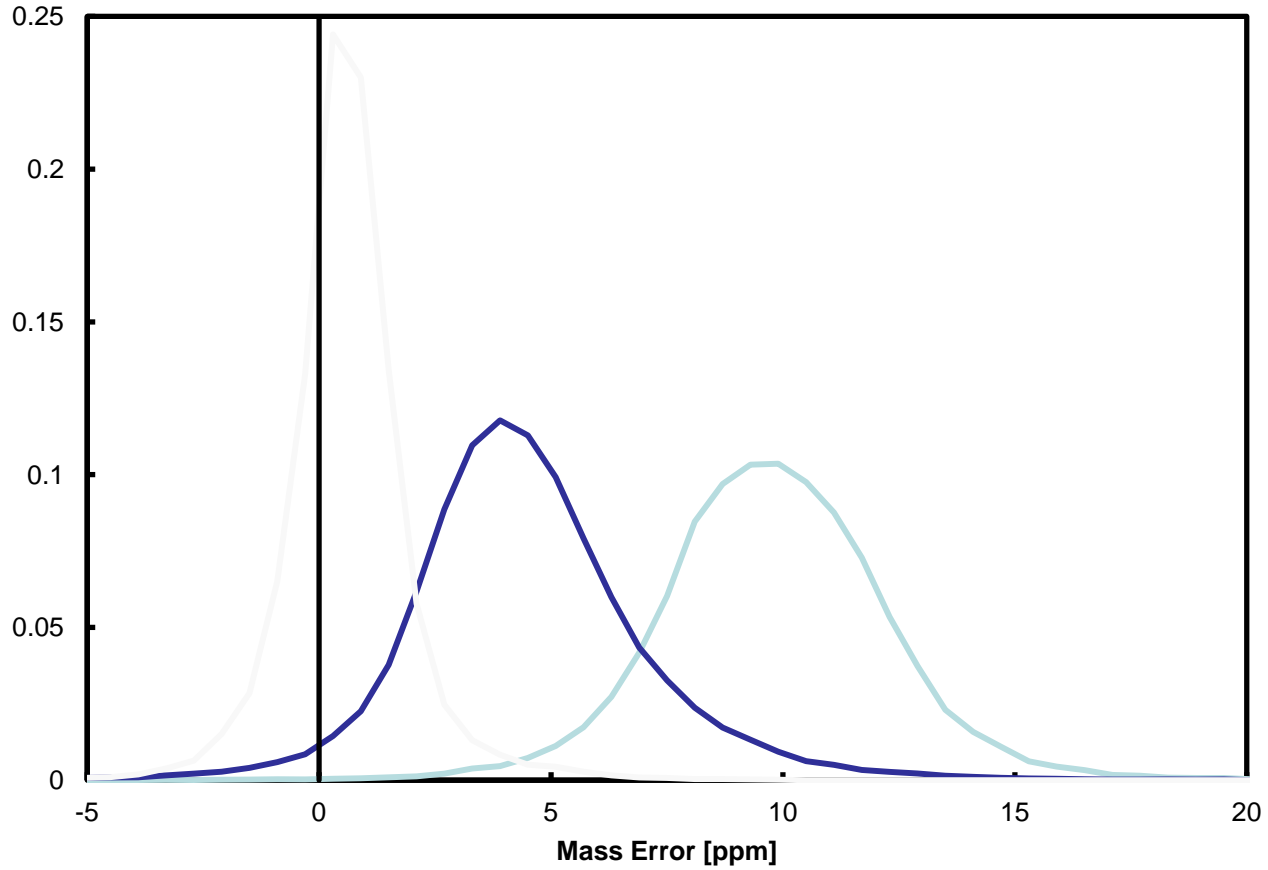
Protein ID	$\Omega$ (z=2)	$\Omega$ (z=3)
SERPINB1	0.88	0.82
SNRPD1	0.88	0.59
CFL1	0.81	0.87
SNRPE	0.8	0.81
PPIA	0.79	0.64
CSTA	0.79	0.36
PFN1	0.76	0.61
CAT	0.71	0.78
GLRX	0.66	0.8
CALM1	0.62	0.76
FABP5	0.57	0.17

# Retention Time Distribution



# Mass Accuracy

---



# GO Cellular Processes

GO:0007275	multicellular development	5.8	-19.1	
GO:0006470	protein dephosphorylation	5.1	-3.8	
GO:0006486	protein glycosylation	5.0	-1.8	
GO:0006468	protein phosphorylation	5.5	-19.2	
GO:0006457	protein folding	6.2	-9.4	
GO:0006508	proteolysis	6.0	-13.2	
GO:0008380	RNA splicing	6.9	-31.0	
GO:0007165	signal transduction	6.4	-35.2	
GO:0007186	signaling, G-protein	6.0	-37.8	
GO:0006350	transcription	6.7	-1.7	
GO:0006355	transcription, regulation	6.7	-9.3	
GO:0006412	translation	6.1	-9.4	
GO:0006810	transport	6.6	-6.5	
GO:0006811	transport, ion	5.9	-21.1	

# KEGG Pathways

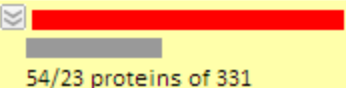
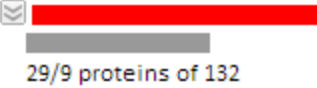
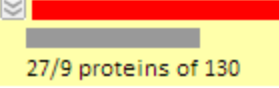
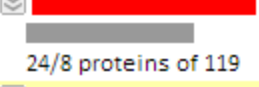
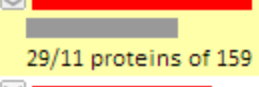
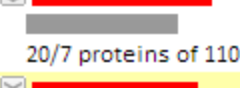
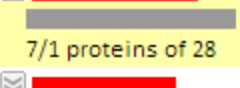
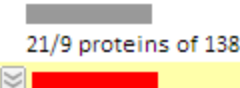
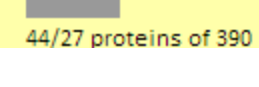


GPM70110008836: KEGG pathway display

[model](#) | [context](#) | [group](#) | [gel](#) | [chip](#) | [peptide](#) | [table](#) | [details](#) | [GO](#) | [BTO](#) | [path](#) | [ppi](#) | [doms](#) | [snaps](#) | [mh](#) | [ζ](#) | [XML](#) |

assigned accession: GPM70110008836

## Sample information

KEGG ID	Pathway	log(I)	log(p)▲	Protein Description
<a href="#">hsa:00190</a>	Oxidative phosphorylation	6.0	-7.8	
<a href="#">hsa:03050</a>	Proteasome	5.1	-7.2	
<a href="#">hsa:00970</a>	Aminoacyl-tRNA biosynthesis	5.3	-6.2	
<a href="#">hsa:00020</a>	Citrate cycle (TCA cycle)	5.7	-5.6	
<a href="#">hsa:00280</a>	Valine, leucine and isoleucine degradation	5.7	-5.5	
<a href="#">hsa:03030</a>	DNA replication	5.6	-4.5	
<a href="#">hsa:00062</a>	Fatty acid elongation in mitochondria	5.2	-4.1	
<a href="#">hsa:03420</a>	Nucleotide excision repair	5.3	-3.6	
<a href="#">hsa:04110</a>	Cell cycle	6.0	-3.2	

# Open-Source Resources

---

# ProteoWizard



info

download

user docs

dev docs

contact

## ProteoWizard

The ProteoWizard Library and Tools are a set of modular and extensible open-source, cross-platform tools and software libraries that facilitate proteomics data analysis.

The libraries enable rapid tool creation by providing a robust, pluggable development framework that simplifies and unifies data file access, and performs standard chemistry and LCMS dataset computations.

Core code and libraries are under the Apache open source license; the vendor libraries fall under various vendor-specific licenses.

## Features

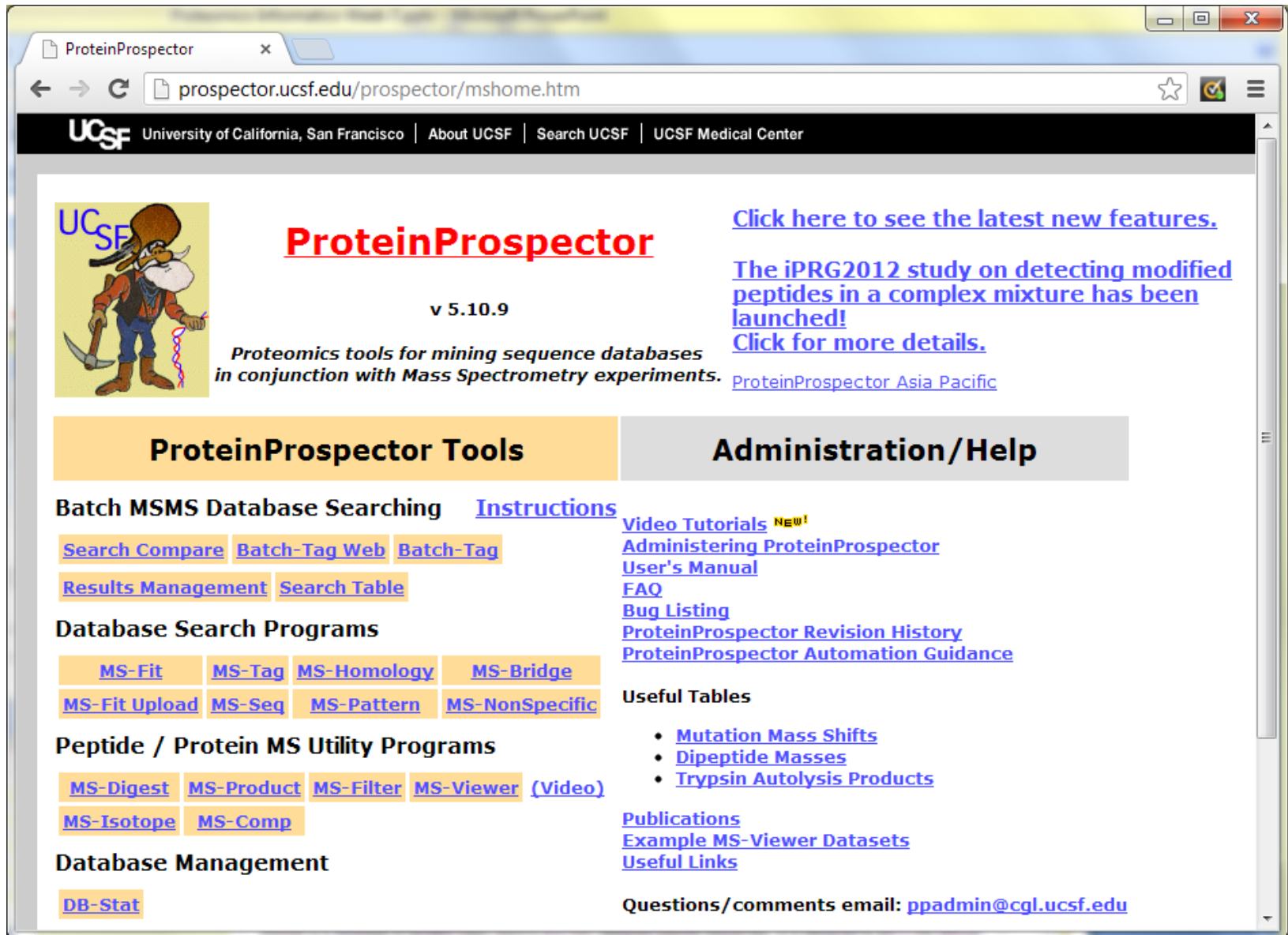
- reference implementation of the new HUPO-PSI [mzML](#) standard mass spectrometry data format
- implementation of the new HUPO-PSI [mzIdentML](#) standard mass spectrometry data format
- modern C++ techniques and design principles
- cross-platform with native compilers (MSVC on Windows, gcc on Linux, XCode on OSX)
- modular design, for testability and extensibility
- framework for rapid development of data analysis tools
- open source license suitable for both academic and commercial projects (Apache v2)
- support for reading directly from many vendor raw data formats (on Windows)



<http://proteowizard.sourceforge.net>



# Protein Prospector



The screenshot shows a web browser window with the URL [prospector.ucsf.edu/prospector/mshome.htm](http://prospector.ucsf.edu/prospector/mshome.htm). The page features the UCSF logo and navigation links. The main content area includes a cartoon prospector character, the title "ProteinProspector v 5.10.9", and a description: "Proteomics tools for mining sequence databases in conjunction with Mass Spectrometry experiments." There are two main navigation tabs: "ProteinProspector Tools" and "Administration/Help". The "ProteinProspector Tools" tab is active and contains several sections: "Batch MSMS Database Searching" with links for "Search Compare", "Batch-Tag Web", "Batch-Tag", "Results Management", and "Search Table"; "Database Search Programs" with links for "MS-Fit", "MS-Tag", "MS-Homology", "MS-Bridge", "MS-Fit Upload", "MS-Seg", "MS-Pattern", and "MS-NonSpecific"; "Peptide / Protein MS Utility Programs" with links for "MS-Digest", "MS-Product", "MS-Filter", "MS-Viewer (Video)", "MS-Isotope", and "MS-Comp"; and "Database Management" with a link for "DB-Stat". The "Administration/Help" tab contains links for "Video Tutorials <sup>NEW!</sup>", "Administering ProteinProspector", "User's Manual", "FAQ", "Bug Listing", "ProteinProspector Revision History", "ProteinProspector Automation Guidance", "Useful Tables" (with sub-links for "Mutation Mass Shifts", "Dipeptide Masses", and "Trypsin Autolysis Products"), "Publications", "Example MS-Viewer Datasets", and "Useful Links". At the bottom, there is a contact email: [ppadmin@cgl.ucsf.edu](mailto:ppadmin@cgl.ucsf.edu). A blue link at the top right says "Click here to see the latest new features." and another blue link below it says "The iPRG2012 study on detecting modified peptides in a complex mixture has been launched! Click for more details." A link for "ProteinProspector Asia Pacific" is also present.

<http://prospector.ucsf.edu/>

# PROWL

The screenshot shows a web browser window with the address bar displaying [prowl.rockefeller.edu](http://prowl.rockefeller.edu). The page header includes the text "THE ROCKEFELLER UNIVERSITY" and "Science for the benefit of humanity". Below this is a dark blue banner for the "LABORATORY OF MASS SPECTROMETRY AND GASEOUS ION CHEMISTRY" and a green banner for the "National Resource for the Mass Spectrometric Analysis of Biological Macromolecules". The page is organized into a sidebar on the left and a main content area on the right. The sidebar lists several tools: ProFound, ProteinInfo, PeptideMap, PepFrag, X! Tandem, GPMDB, PROWL, and Chait Lab. The main content area provides detailed descriptions for ProFound, ProteinInfo, PeptideMap, PepFrag, X! Tandem, and GPMDB.

**THE ROCKEFELLER UNIVERSITY**  
*Science for the benefit of humanity*

LABORATORY OF MASS SPECTROMETRY AND GASEOUS ION CHEMISTRY

**National Resource for the Mass Spectrometric Analysis of Biological Macromolecules**

National Institute of General Medical Sciences  
Biomedical Technology Research Centers

**ProFound**  
ProFound is a tool for searching a protein sequence collections with peptide mass maps. A Bayesian algorithm is used to rank the protein sequences in the database according to their probability of producing the peptide map.

**ProteinInfo**  
ProteinInfo is a collection of tools for retrieval and analysis of protein sequences. The capabilities of the analysis tools include peptide mapping, mass spectrometric fragmentation analysis, disulfide mapping, etc.

**PeptideMap**  
PeptideMap is a tool for finding modifications on polypeptide sequences. The modifications can be affecting single amino acids (e.g. phosphorylation or oxidation) or cross-linking two amino acids (e.g. disulfide bonds or chemical cross-linking reagents).

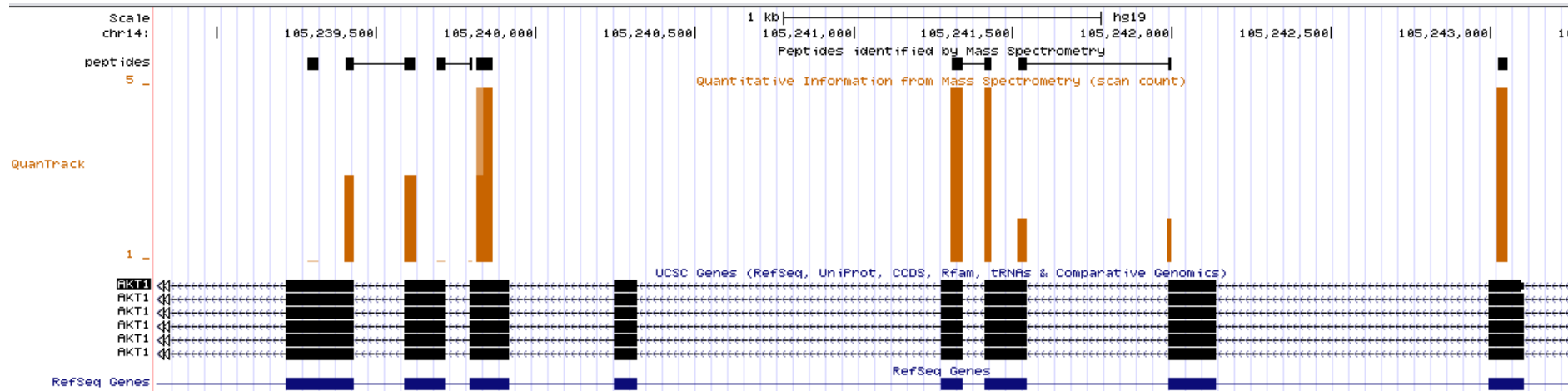
**PepFrag**  
PepFrag is a tool for identifying proteins from a collection of sequences that matches a *single* tandem mass spectrum.

**X! Tandem**  
X! Tandem is a tool for identifying proteins from a collection of peptide sequences that matches tandem mass spectra.

**GPMDB**  
GPMDB is a database of tandem mass spectra and their assigned peptide sequences. It is designed to aid in the difficult process of validating peptide MS/MS spectra.

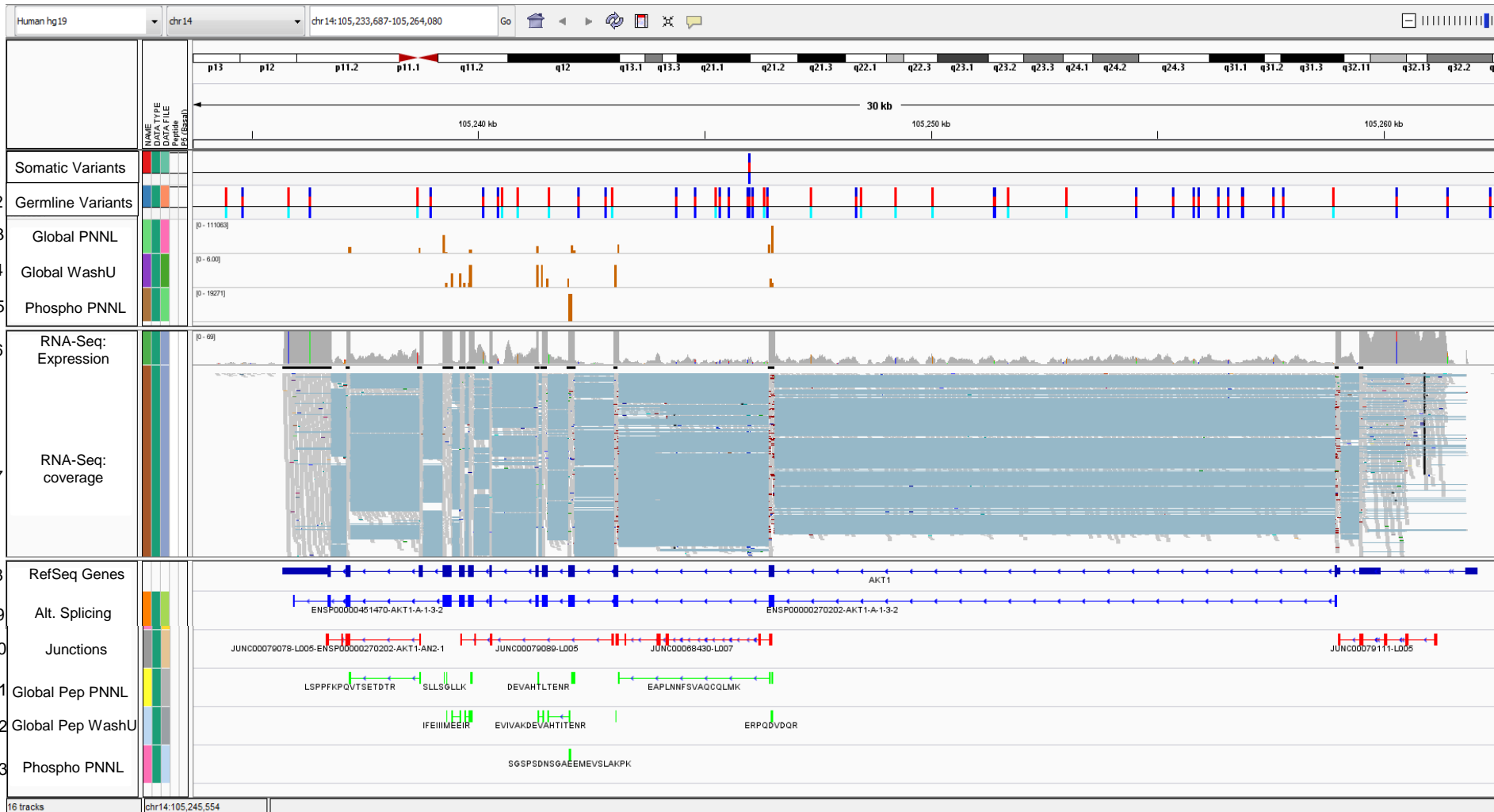
<http://prowl.rockefeller.edu/>

# Proteogenomics - PGx




<http://pgx.fenyolab.org/>

# UCSC Genome Browser



<http://genome.ucsc.edu/>

# Slice - Scalable Data Sharing for Remote Mass Informatics

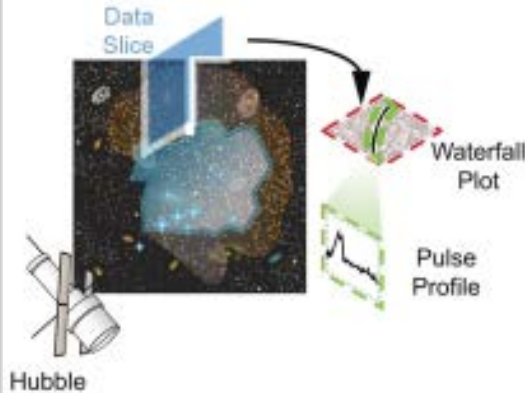


YIVSGTPTFPVYLIK => 849.480 2

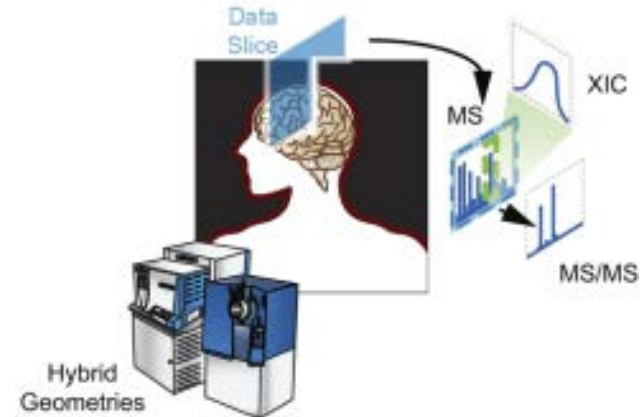
Filename	Time	Intensity	Preview	MS2
Roche_human_CSF_017	71.689	2389918		✓
Roche_human_CSF_021	76.772	1242552		✓
Roche_human_CSF_018	72.168	1003575		✓
Roche_human_CSF_046	40.513	953971		✓
Roche_human_CSF_014	71.517	932062		✓
Roche_human_CSF_022	72.896	757225		✓
Roche_human_CSF_013	77.03	439764		✓
Roche_human_CSF_019	74.551	397208		✓
Roche_human_heart_191	45.874	230045		
Roche_human_heart_023	35.45	208049		

Developed by Manor Askenazi  
[slice.ionomix.com](http://slice.ionomix.com)

Mapping Celestial Features



Mapping Human Proteins



Most mass spectrometry data is acquired in discovery mode, meaning that the data is amenable to open-ended analysis as our understanding of the target biochemistry increases. In this sense, mass spectrometry based discovery work is more akin to an astronomical survey, where the full list of object-types being imaged has not yet been fully elucidated, as opposed to e.g. micro-array work, where the list of probes spotted onto the slide is finite and well understood.

# Standardization

---

## The minimum information about a proteomics experiment (MIAPE)

Chris F Taylor<sup>1,2</sup>, Norman W Paton<sup>1,3</sup>, Kathryn S Lilley<sup>1,4</sup>, Pierre-Alain Binz<sup>1,5,6</sup>, Randall K Julian Jr<sup>1,7</sup>, Andrew R Jones<sup>1,3</sup>, Weimin Zhu<sup>1,2</sup>, Rolf Apweiler<sup>1,2</sup>, Ruedi Aebersold<sup>1,8</sup>, Eric W Deutsch<sup>1,9</sup>, Michael J Dunn<sup>10</sup>, Albert J R Heck<sup>11</sup>, Alexander Leitner<sup>12</sup>, Marcus Macht<sup>13</sup>, Matthias Mann<sup>14</sup>, Lennart Martens<sup>1,2</sup>, Thomas A Neubert<sup>15</sup>, Scott D Patterson<sup>16</sup>, Peipei Ping<sup>17</sup>, Sean L Seymour<sup>1,18</sup>, Puneet Souda<sup>19</sup>, Akira Tsugita<sup>20</sup>, Joel Vandekerckhove<sup>21</sup>, Thomas M Vondriska<sup>22</sup>, Julian P Whitelegge<sup>19</sup>, Marc R Wilkins<sup>23</sup>, Ioannis Xenarios<sup>24</sup>, John R Yates III<sup>25</sup> & Henning Hermjakob<sup>1,2</sup>

MIAPE	MIAPE Principles document	1.0	release
MIAPE-MS	Mass Spectrometry	2.98	release
MIAPE-MSI	Mass Spectrometry Informatics	1.1	release
MIAPE-Quant	Mass Spectrometry Quantification	1.0	release
MIAPE-GE	Gel Electrophoresis	1.4	release
MIAPE-GI	Gel Informatics	1	release
MIAPE-CC	Column Chromatography	1.1	release
MIAPE-CE	Capillary Electrophoresis	0.9.3	release
MIMix	Molecular Interactions	1-1-2	release

# Standardization – MIAPE-MSI

---

The following section, detailing the reporting guidelines for the use of protein and peptide identification and characterisation software, is subdivided as follows:

1. General features; the software employed.
2. Input data and parameters.
3. The output from the procedure; the list of peptides and proteins identified, characterised or quantified.
4. Interpretation and validation.

## **Reporting guidelines for protein and peptide identification and characterisation software**

### *1. General features*

- a) Global descriptors
  - Date stamp (as YYYY-MM-DD)
  - Responsible person (or institutional role if more appropriate); provide name, affiliation and stable contact information
  - Software name, version and manufacturer
  - Customisations made to that software
  - Availability of that software
  - Location of the files generated; parameter files, spectral data (input/output)

- Any other relevant parameters

### *3. The output from the procedure*

*The procedure might generate all or part of the elements described below (identified proteins, identified peptides, quantization information). Select the elements that apply.*

- a) For identified proteins
  - Accession code in the queried database
  - Protein description
  - Protein scores
  - Validation status
  - Number of different peptide sequences (without considering modifications) assigned to the protein
  - Percent peptide coverage of protein
  - Identity of supporting peptides
  - In the case of PMF, number of matched/unmatched peaks
- b) For identified peptides
  - Sequence (indicate any deviation from the expected protein cleavage specificity)
  - Peptide scores
  - Chemical modifications (artefactual) and post-translational modifications (naturally-occurring); sequence polymorphisms with experimental evidence (particularly for isobaric modifications)



# Standardization - XML Formats

---

**mzML** - experimental results obtained by mass spectrometric analysis of biomolecular compounds

**mzIdentML** - describe the outputs of proteomics search engines

**TraML** - exchange and transmission of transition lists for selected reaction monitoring (SRM) experiments

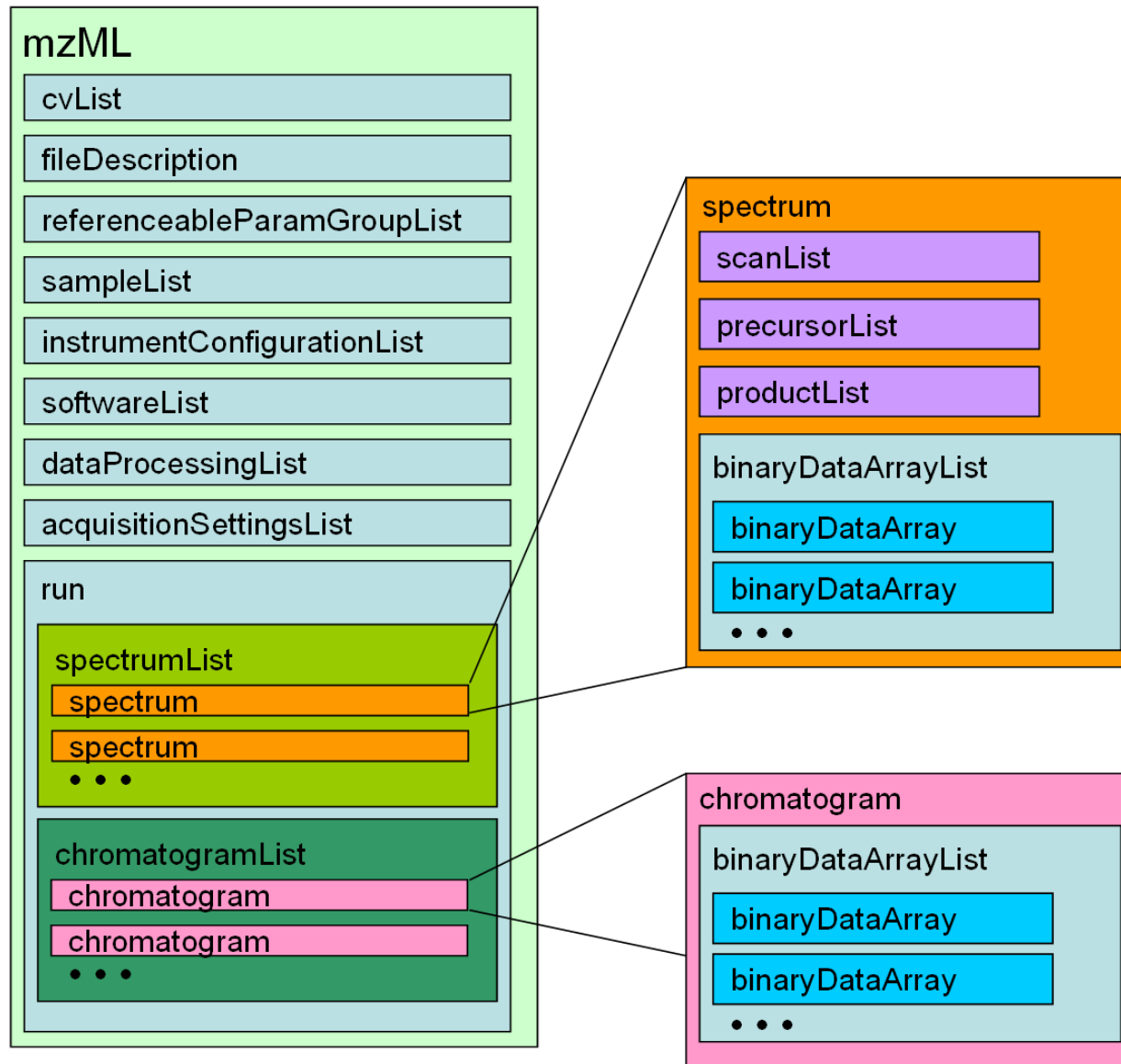
**mzQuantML** - describe the outputs of quantitation software for proteomics

**mzTab** - defines a tab delimited text file format to report proteomics and metabolomics results.

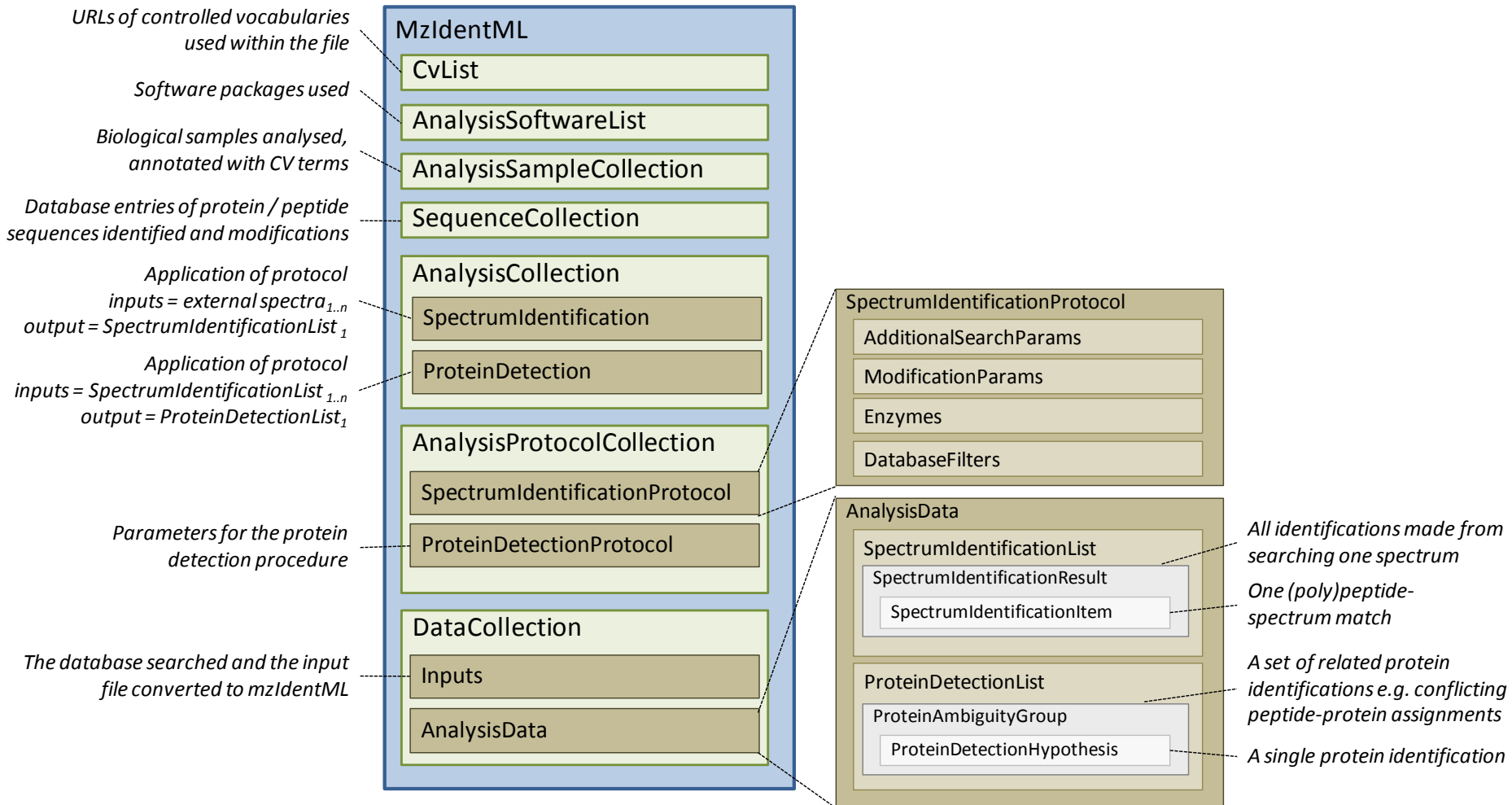
**MIF** - describes the molecular interaction data exchange format.

**GelML** - describes the processing and separations of proteins in samples using gel electrophoresis, within a proteomics experiment.

# Standardization - mzML



# Standardization - mzIdentML



**Proteomics Informatics -  
Databases, data repositories and  
standardization (Week 7)**

---