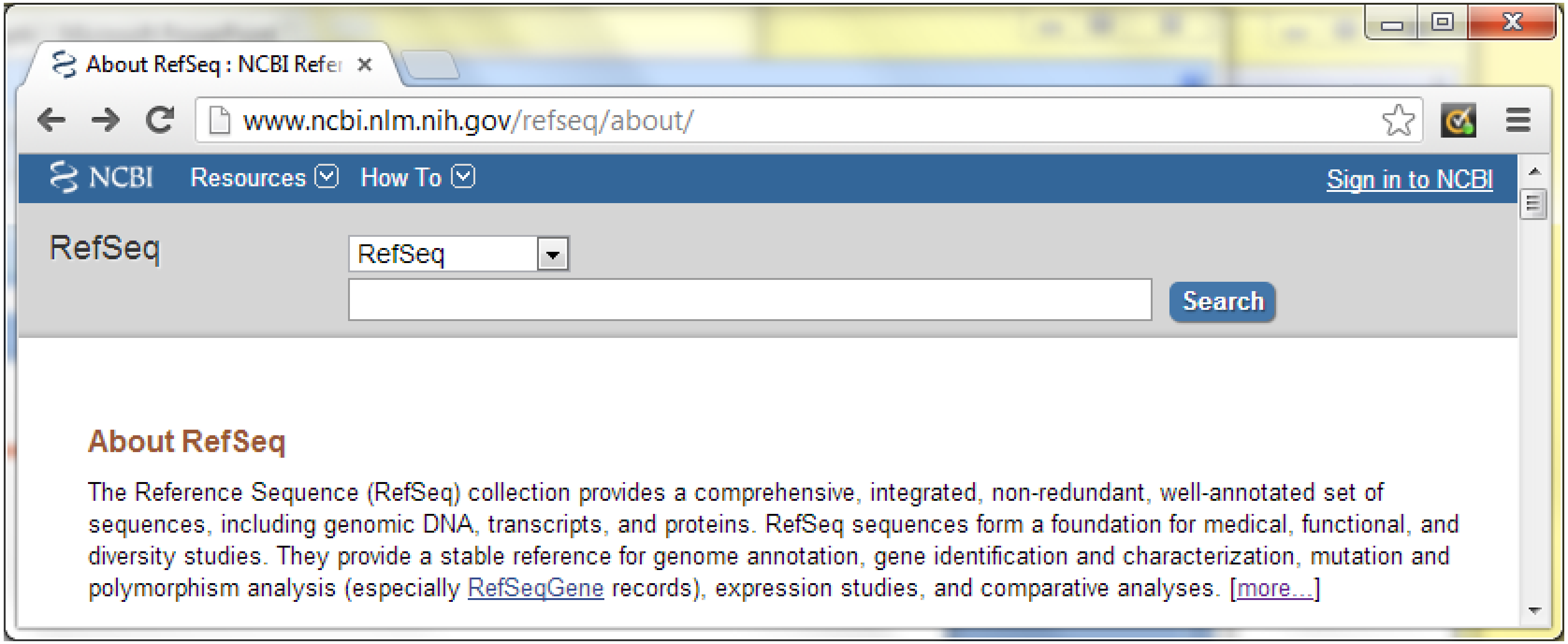
Proteomics Informatics – Databases, data repositories and standardization (Week 8)

Protein Sequence Databases

RefSeq



About RefSeq

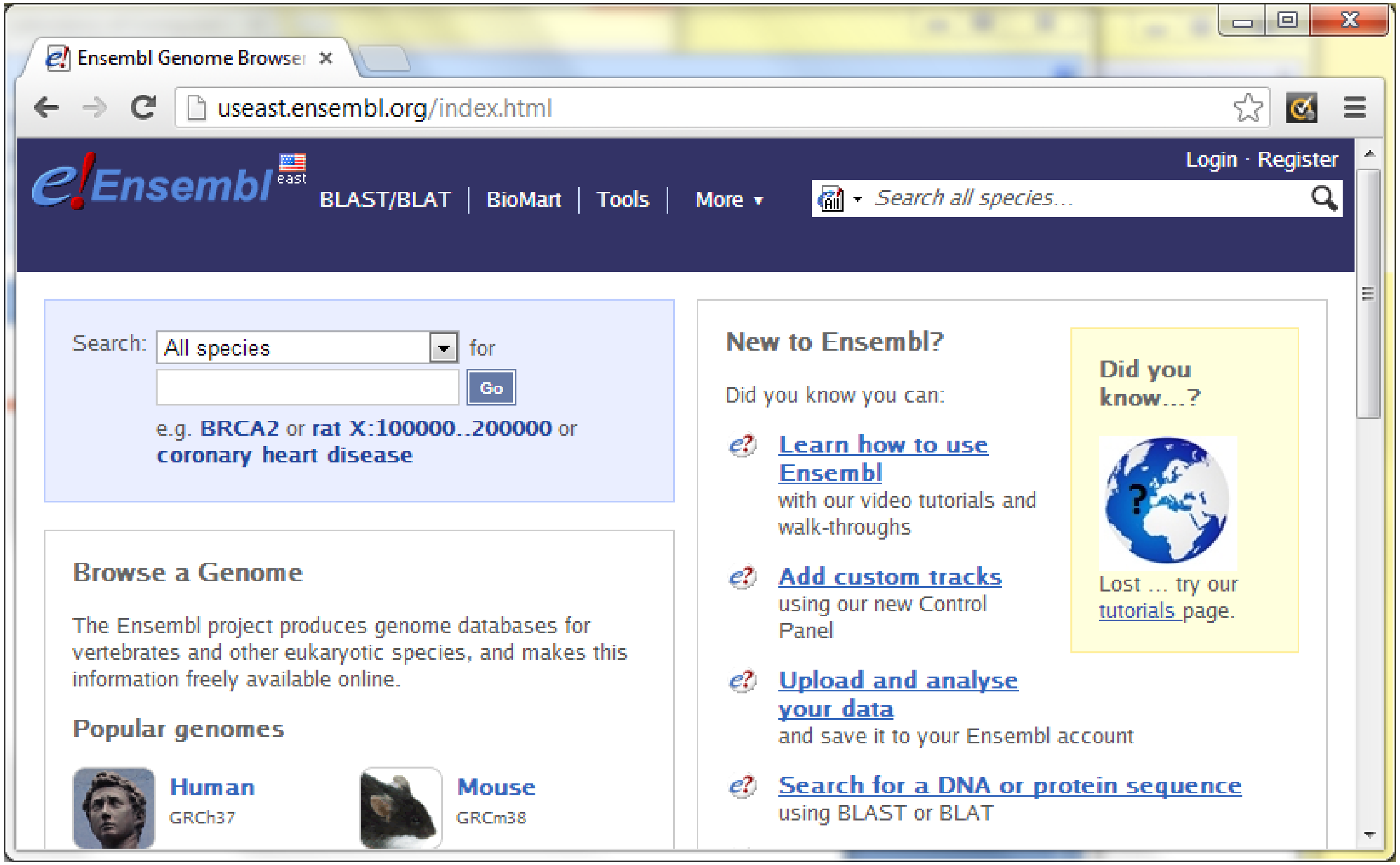
The Reference Sequence (RefSeq) collection provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq sequences form a foundation for medical, functional, and diversity studies. They provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis (especially RefSeqGene records), expression studies, and comparative analyses. [more...]

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



Search: All species for

e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

Human GRCh37

Mouse GRCm38

New to Ensembl?

Did you know you can:

Learn how to use Ensembl with our video tutorials and walk-throughs

Add custom tracks using our new Control Panel

Upload and analyse your data and save it to your Ensembl account

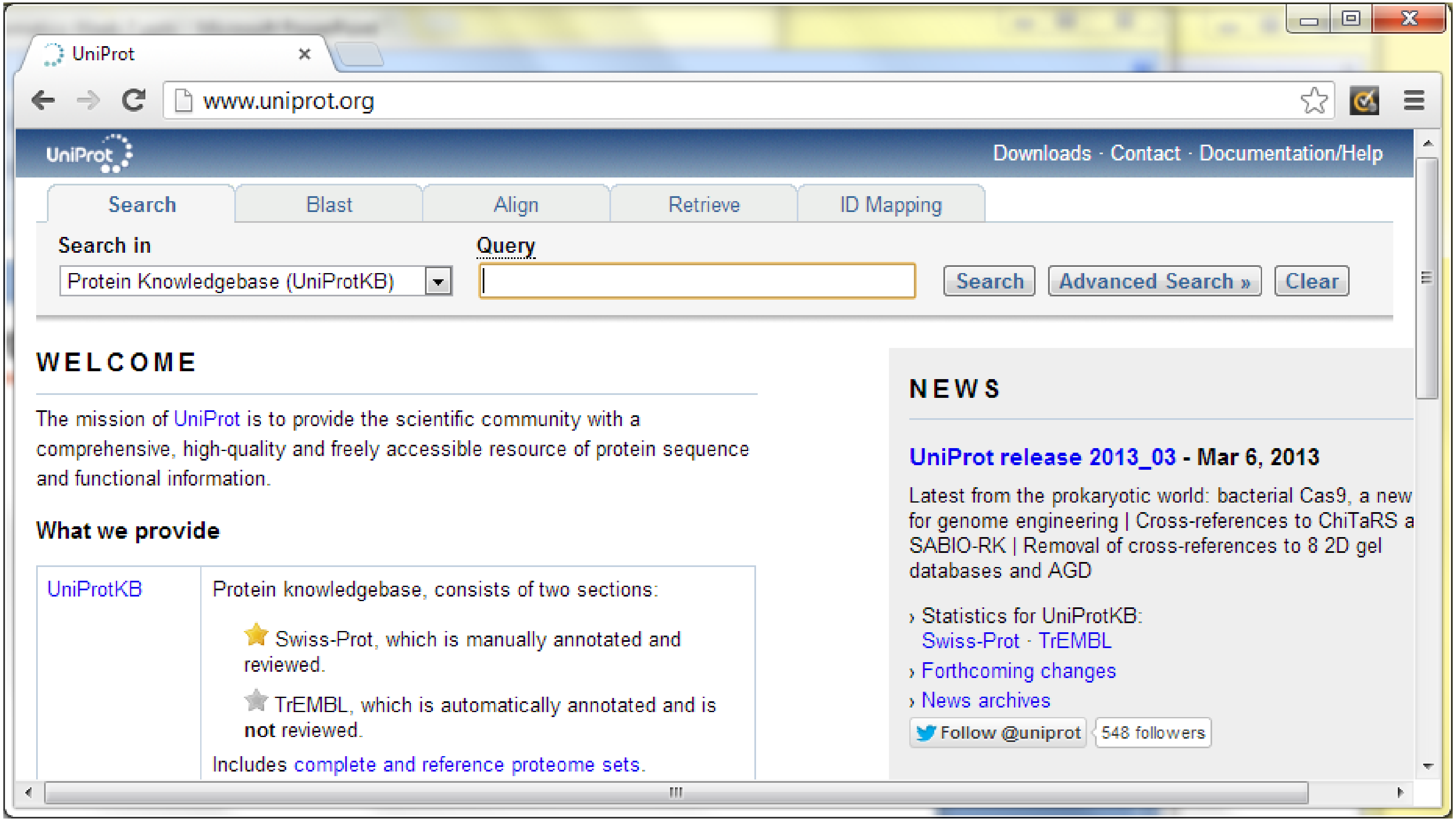
Search for a DNA or protein sequence using BLAST or BLAT

Lost ... try our tutorials page.

- 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



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

UniProtKB	Protein knowledgebase, consists of two sections: ★ Swiss-Prot, which is manually annotated and reviewed. ☆ TrEMBL, which is automatically annotated and is **not** reviewed. Includes complete and reference proteome sets.

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

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
LQNRKSPSPLAFCPCPPVNSISKELPFLLHAFYPGYPLLLPPPHLFTYGALPSDQCPHLLMLPQDPSYPTMAMPSLLMMV
NELGHPSARWETLLPYPGAFQASGQALPSQARNPGAGAAPTDSPGLERGGMASPAKRVPLSSQTGTAALPYPLKKKNGKI
LYECNICGKSFGQLSNLKVHLRVHSGERPFQCALCQKSFTQLAHLQKHHLVHTGERPHKCSVCHKRFSSSSNLKTHLRLH
SGARPFQCSVCRSRFTQHIHLKLHHRLHAPQPCGLVHTQLPLASLACLAQWHQGALDLMAVASEKHMGYDIDEVKVSSTS
QGKARAVSLSSAGTPLVMGQDQNN
```

Ensembl:

>ENSMUSP00000131420 pep:known supercontig:NCBIM37:NT_166407:104574:105272:
gene:ENSMUSG00000092057 transcript:ENSMUST00000167991

```
MFSLMKKRRRKSSSNTLRNIVGCRISHCWKEGNEPVTQWKAIVLGQLPTNPSLYLVKYDGIDSIYGQELYSDDRILNLKVL
PPIVVFPQVRDAHLARALVGRAVQQKFERKDGSEVNWRGVVLAQVPIMKDLFYITYKKDPALYAYQLLDDYKEGNLHMIPD
TPPAEERSGGDSDVLIGNWVQYTRKDGSKKFGKVVYQVLDNPSVFFIKFHGDIHIYVYTMVPKILEVEKS
```

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

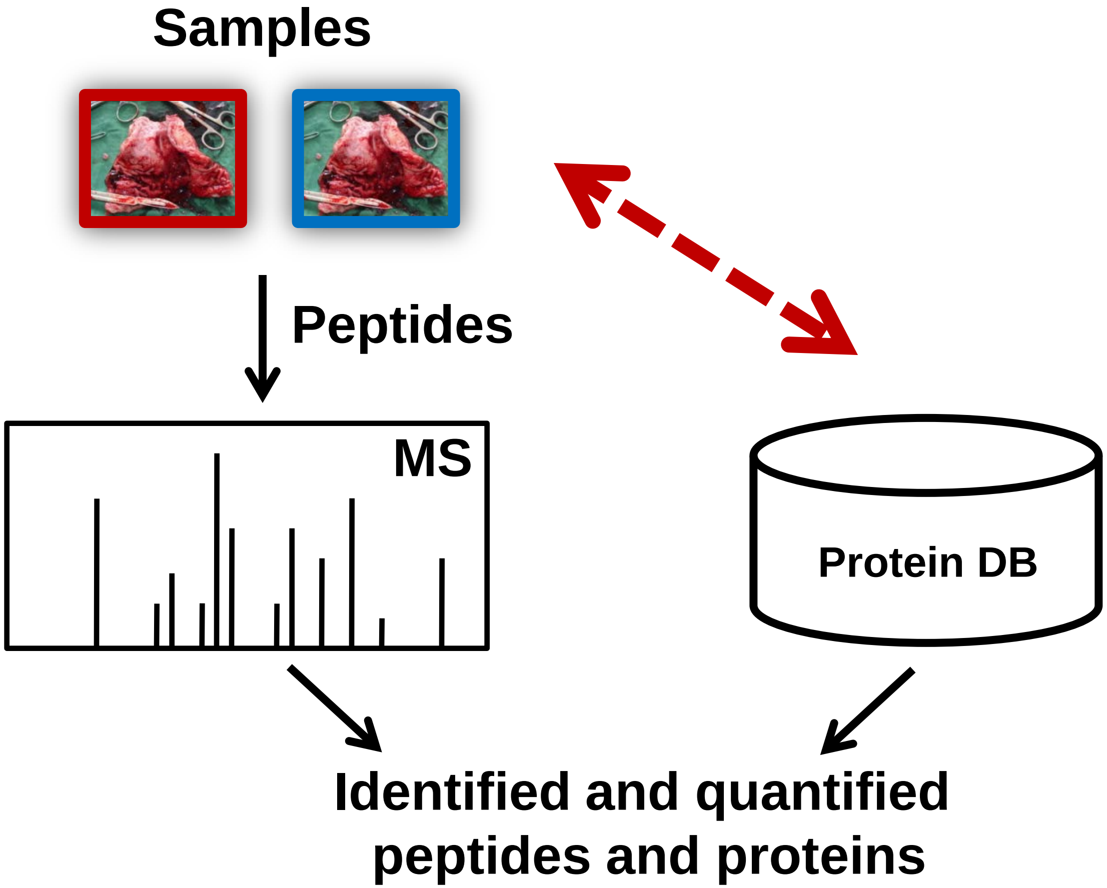
PEFF - PSI Extended Fasta Format

```
>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
```

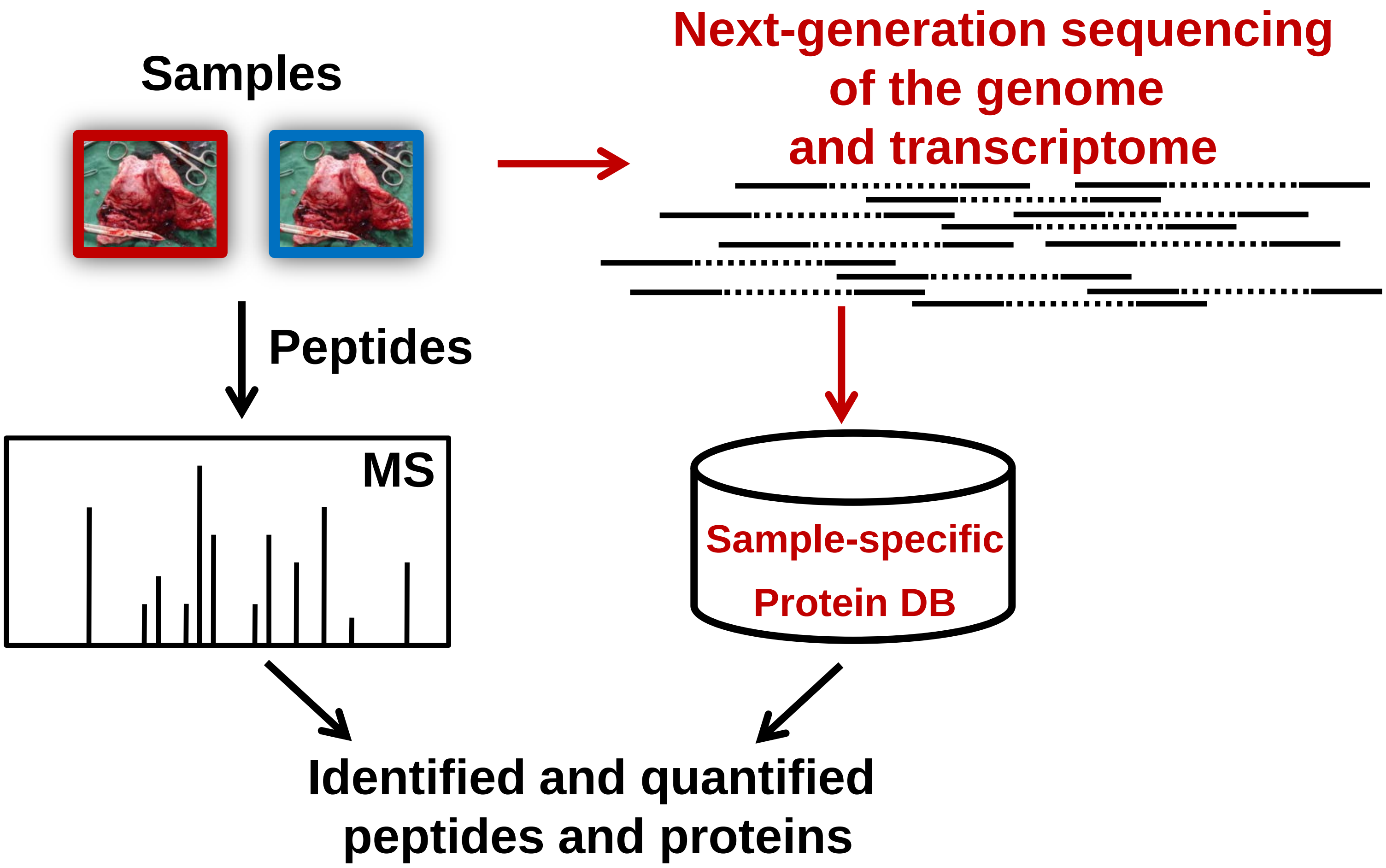
```
>sp:P00761 \ID=TRYP_PIG
\Pname=(Trypsin precursor) (EC 3.4.21.4) \NcbiTaxId=9823
\Variant=(20|20|V)
\Processed=(1|8|PROPEP)(9|231|CHAIN)
\Length=231
```

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

Sample-specific protein sequence databases



Sample-specific protein sequence databases



Samples

Next-generation sequencing of the genome and transcriptome

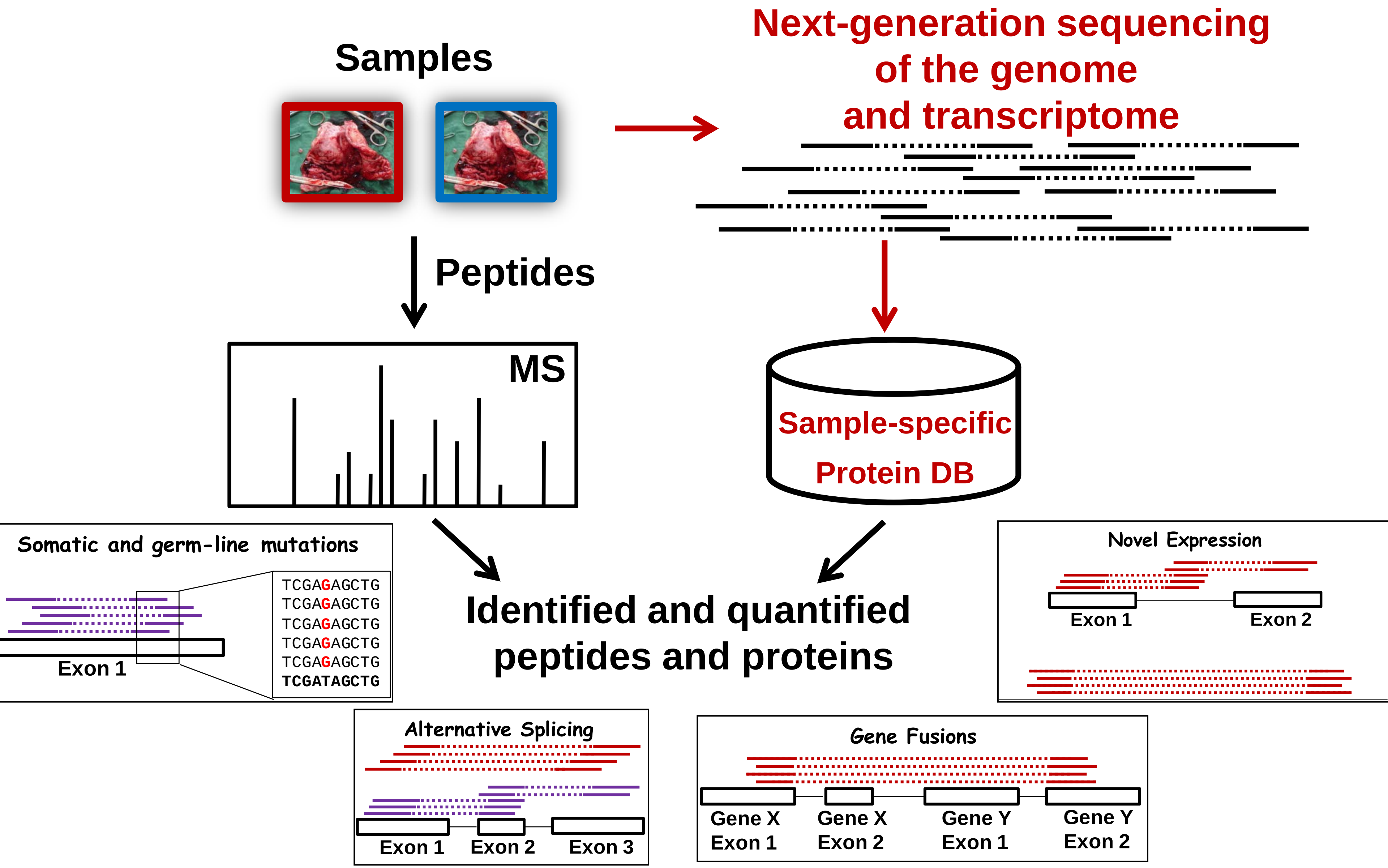
Peptides

MS

Sample-specific Protein DB

Identified and quantified peptides and proteins

Sample-specific protein sequence databases



Samples

Next-generation sequencing of the genome and transcriptome

Peptides

Sample-specific Protein DB

Identified and quantified peptides and proteins

Somatic and germ-line mutations

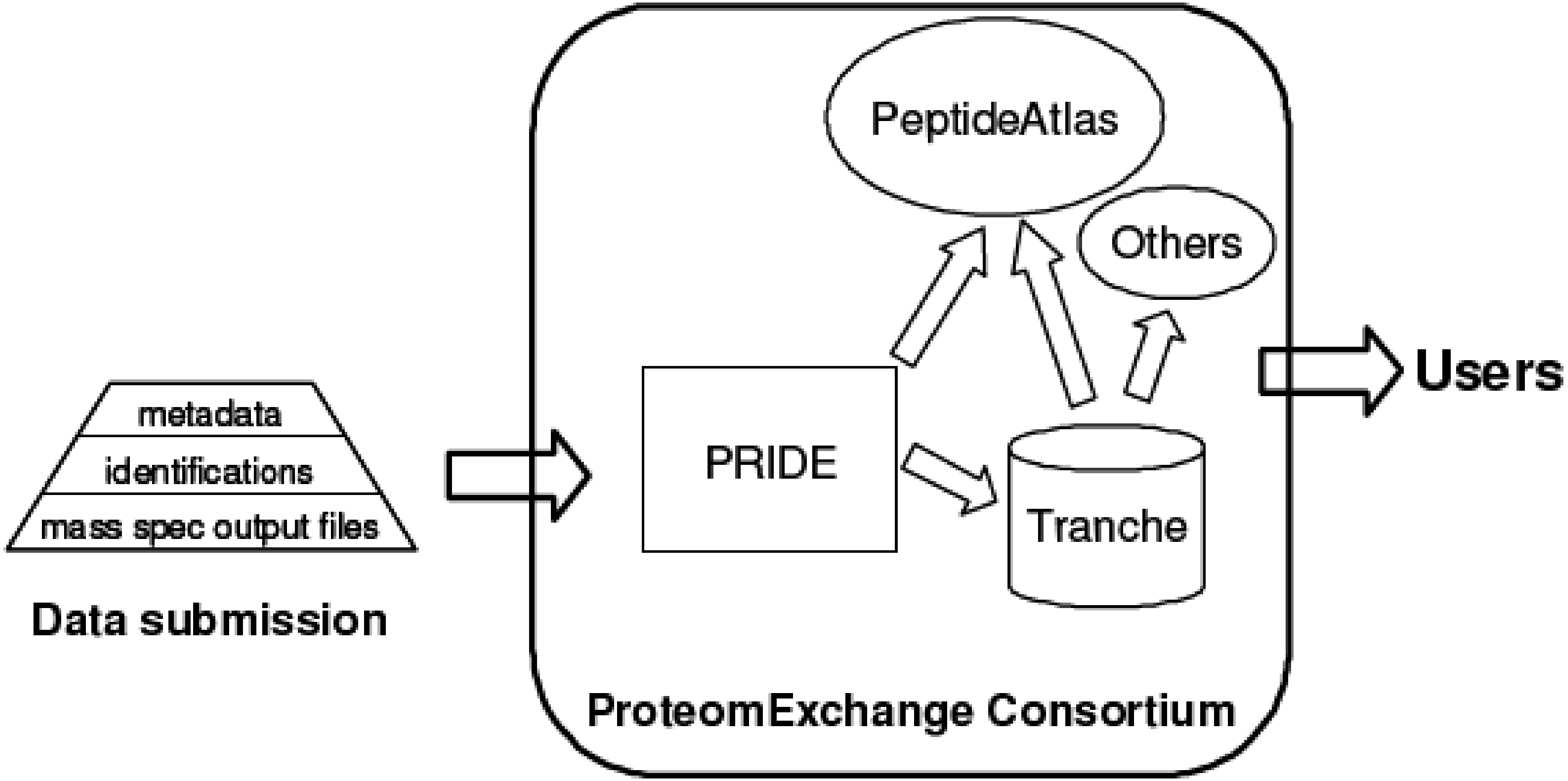
Novel Expression

Alternative Splicing

Gene Fusions

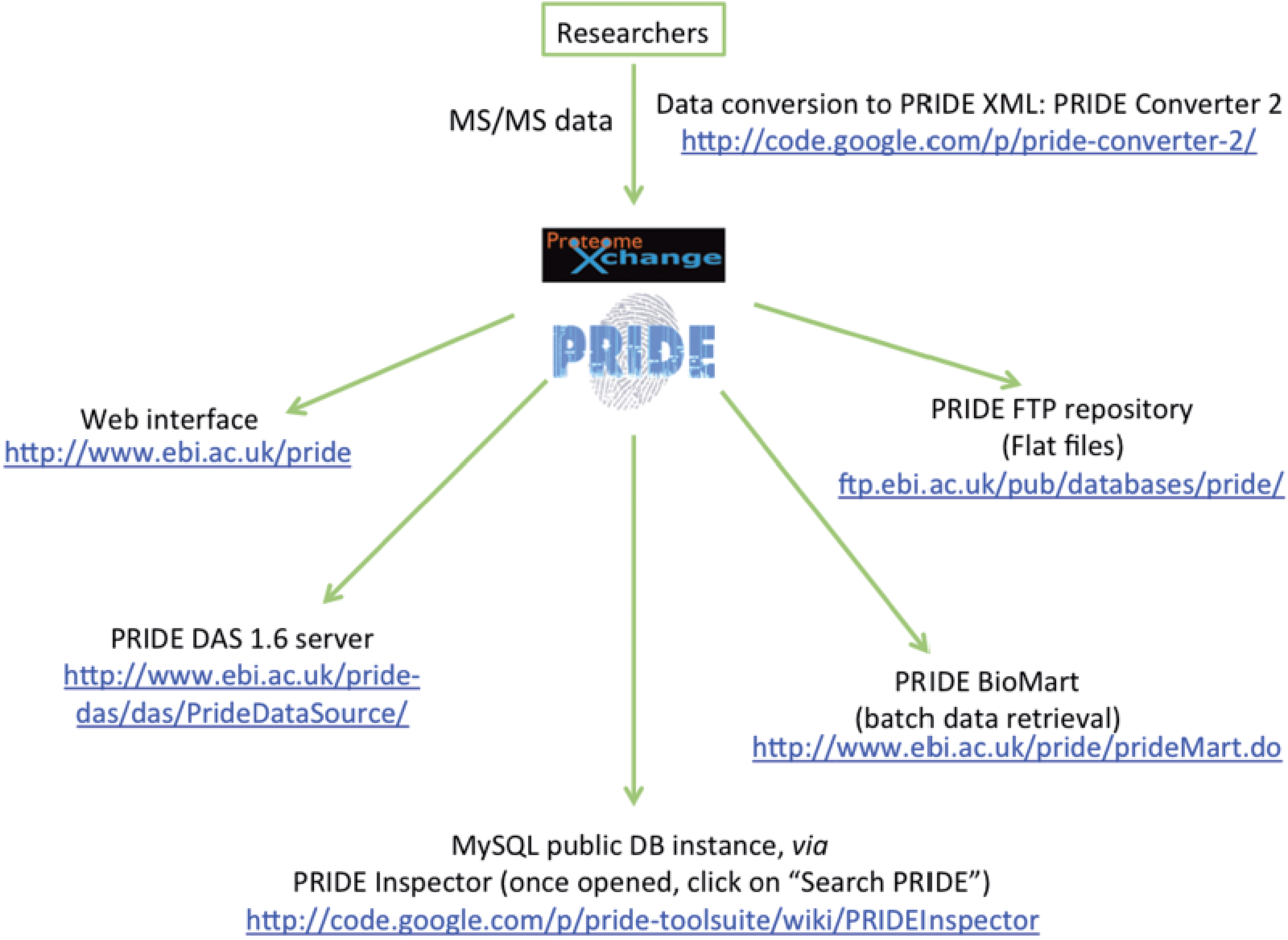
Data Repositories

ProteomeExchange



http://www.proteomeexchange.org/

PRIDE



Researchers

MS/MS data

Data conversion to PRIDE XML: PRIDE Converter 2
http://code.google.com/p/pride-converter-2/

Web interface
http://www.ebi.ac.uk/pride

PRIDE FTP repository
(Flat files)
ftp.ebi.ac.uk/pub/databases/pride/

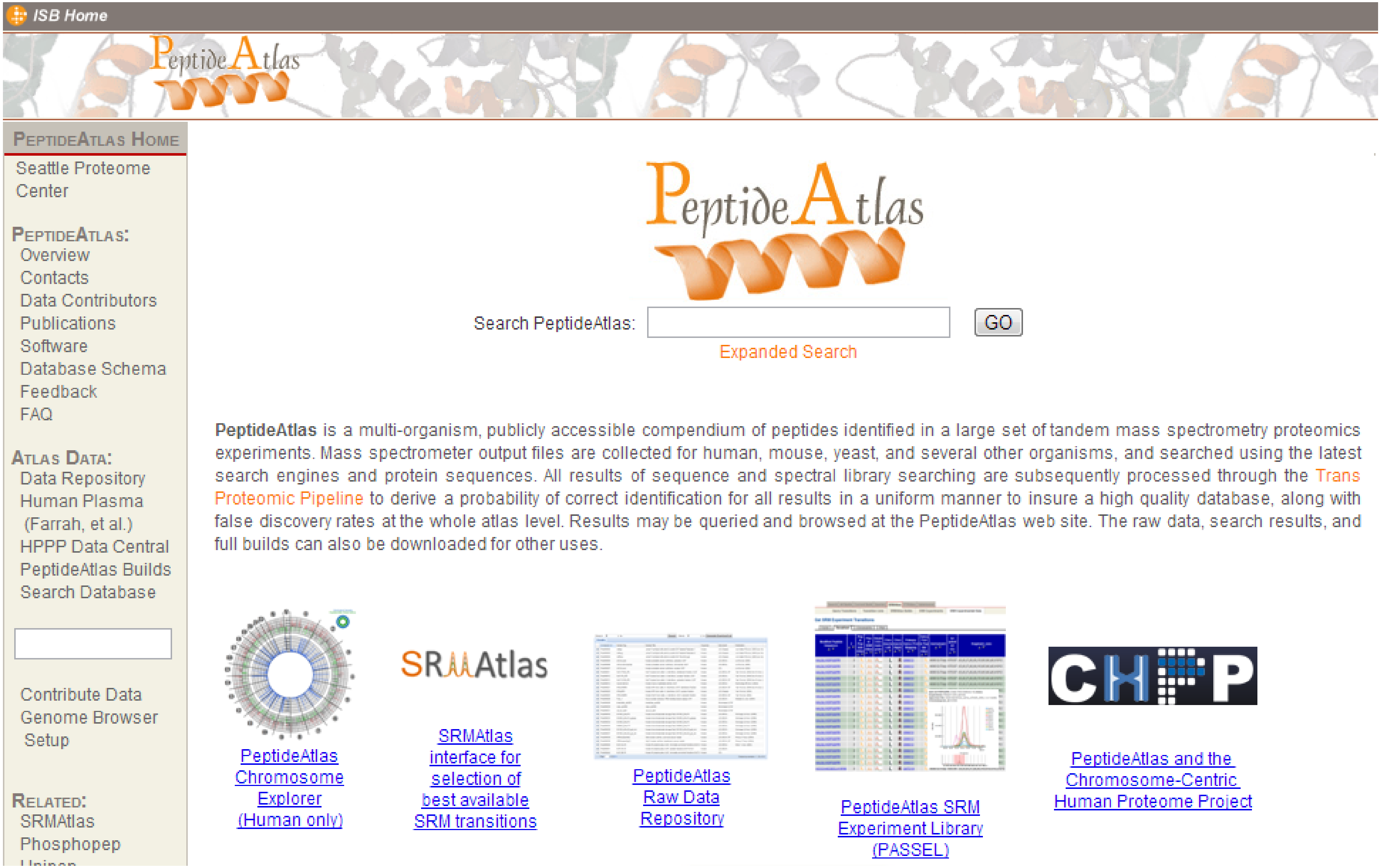
PRIDE DAS 1.6 server
http://www.ebi.ac.uk/pride-das/das/PrideDataSource/

PRIDE BioMart
(batch data retrieval)
http://www.ebi.ac.uk/pride/prideMart.do

MySQL public DB instance, *via*
PRIDE Inspector (once opened, click on "Search PRIDE")
http://code.google.com/p/pride-toolsuite/wiki/PRIDEInspector

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
Unipep

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 interface for selection of best available SRM transitions

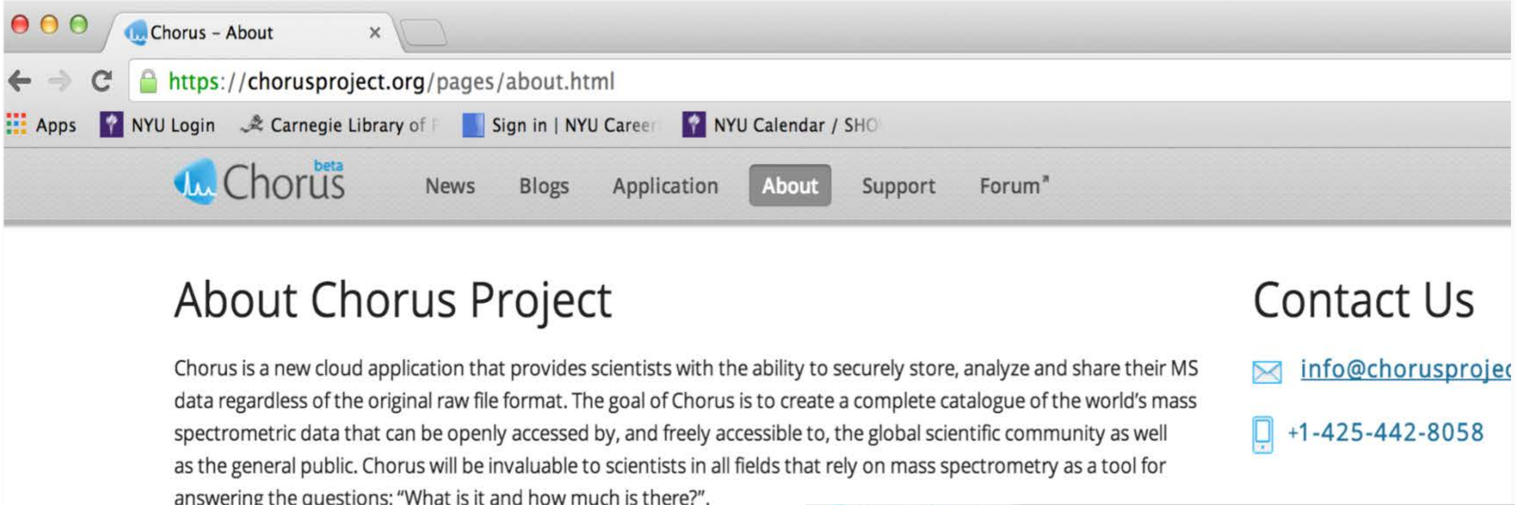
PeptideAtlas Raw Data Repository

PeptideAtlas SRM Experiment Library (PASSEL)

PeptideAtlas and the Chromosome-Centric Human Proteome Project

http://www.peptideatlas.org/

Chorus



About Chorus Project

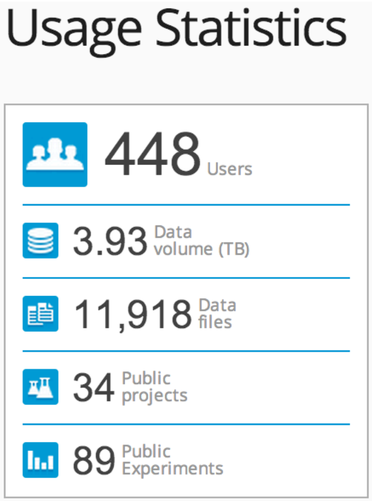
Chorus is a new cloud application that provides scientists with the ability to securely store, analyze and share their MS data regardless of the original raw file format. The goal of Chorus is to create a complete catalogue of the world's mass spectrometric data that can be openly accessed by, and freely accessible to, the global scientific community as well as the general public. Chorus will be invaluable to scientists in all fields that rely on mass spectrometry as a tool for answering the questions: "What is it and how much is there?".

Know more details ↓

Contact Us

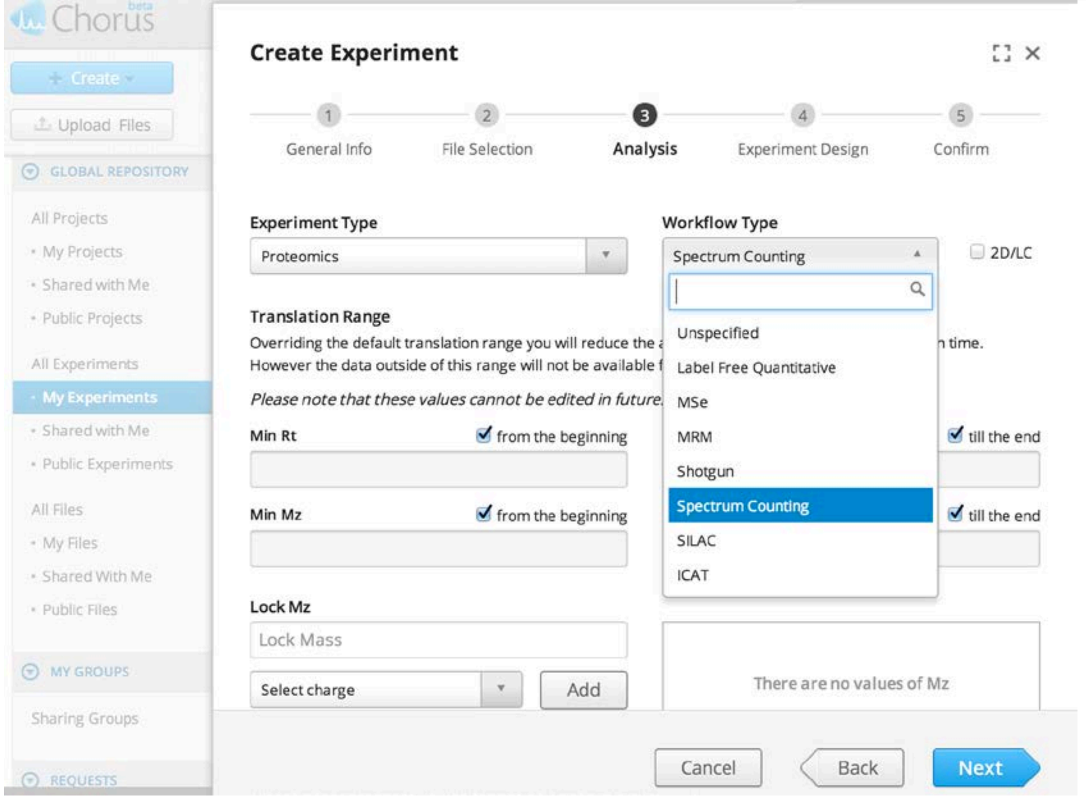
✉ info@chorusprojec

☎ +1-425-442-8058



Usage Statistics

448	Users
3.93	Data volume (TB)
11,918	Data files
34	Public projects
89	Public Experiments



Create Experiment

1 General Info — 2 File Selection — 3 Analysis — 4 Experiment Design — 5 Confirm

Experiment Type: Proteomics

Workflow Type: Spectrum Counting

□ 2D/LC

Translation Range

Overriding the default translation range you will reduce the a … n time. However the data outside of this range will not be available f

Please note that these values cannot be edited in future.

Min Rt ☑ from the beginning ☑ till the end

Min Mz ☑ from the beginning ☑ till the end

Lock Mz

Lock Mass

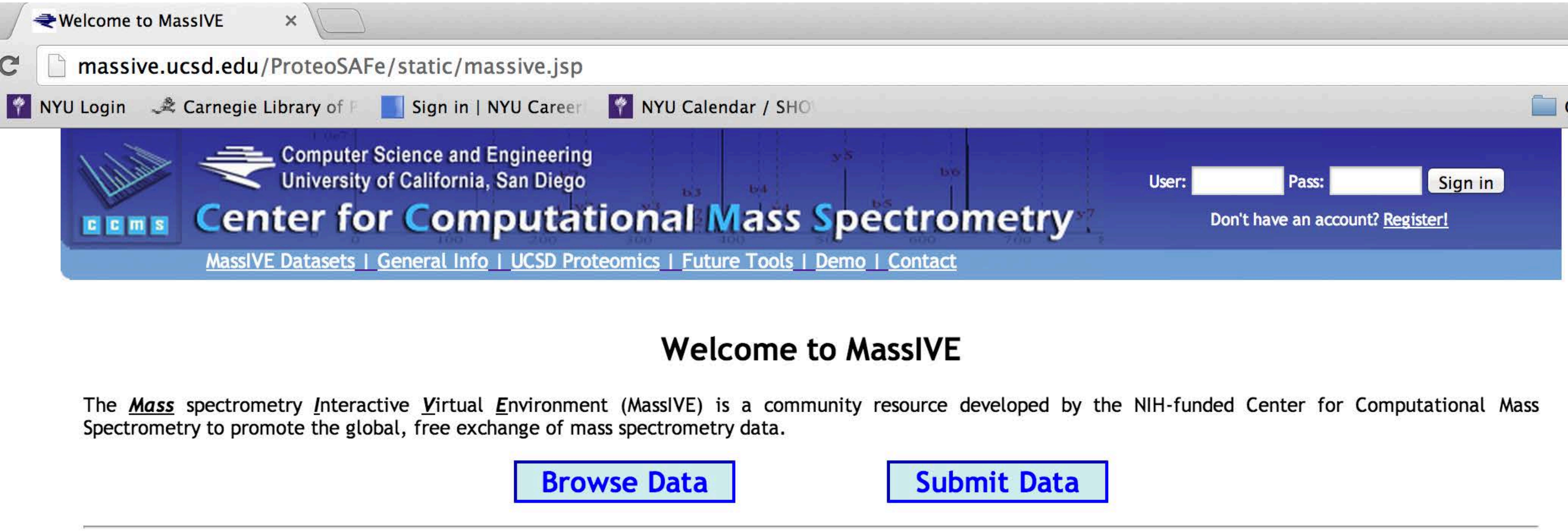
Select charge — Add

There are no values of Mz

Key Aspects:

- Upload and share raw data with collaborators
- Analyze data with available tools and workflows
- Create projects and experiments
- Select from public files and (re-)analyze/visualize
- Download selected files

MassIVE


Welcome to MassIVE
massive.ucsd.edu/ProteoSAFe/static/massive.jsp
NYU Login
Carnegie Library of P
Sign in | NYU Career
NYU Calendar / SHO
c c m s
Computer Science and Engineering
University of California, San Diego
Center for Computational Mass Spectrometry
User:
Pass:
Sign in
Don't have an account? Register!
MassIVE Datasets | General Info | UCSD Proteomics | Future Tools | Demo | Contact

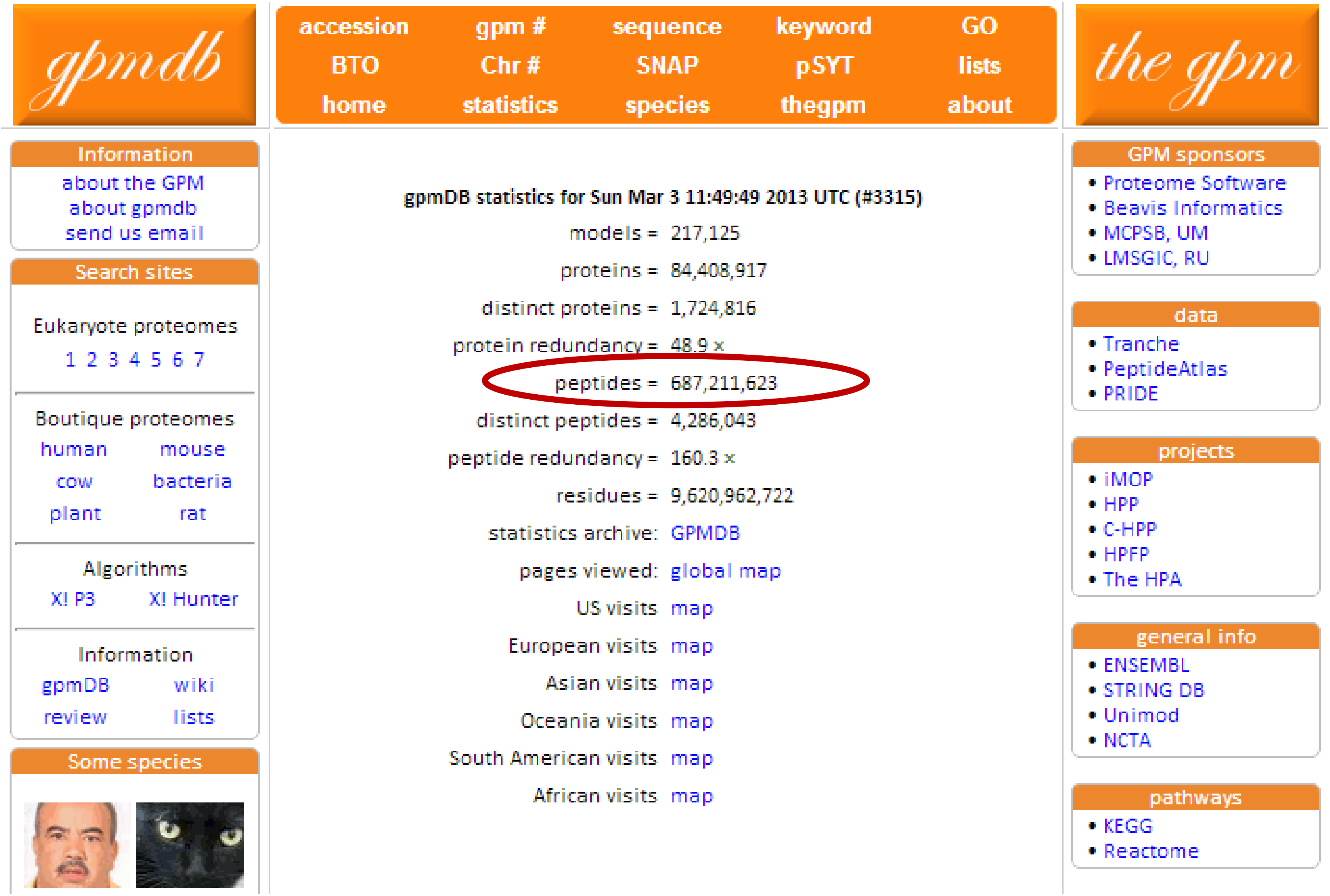
Welcome to MassIVE

The ***Mass*** spectrometry ***I***nteractive ***V***irtual ***E***nvironment (MassIVE) is a community resource developed by the NIH-funded Center for Computational Mass Spectrometry to promote the global, free exchange of mass spectrometry data.

Key Aspects:

- Upload files
    - Spectra and Spectrum libraries, Analysis Results, Sequence Databases, Methods and Protocol)
- Perform analysis using available tools
- Browse public datasets
- Download data

The Global Proteome Machine Databases (GPMDB)

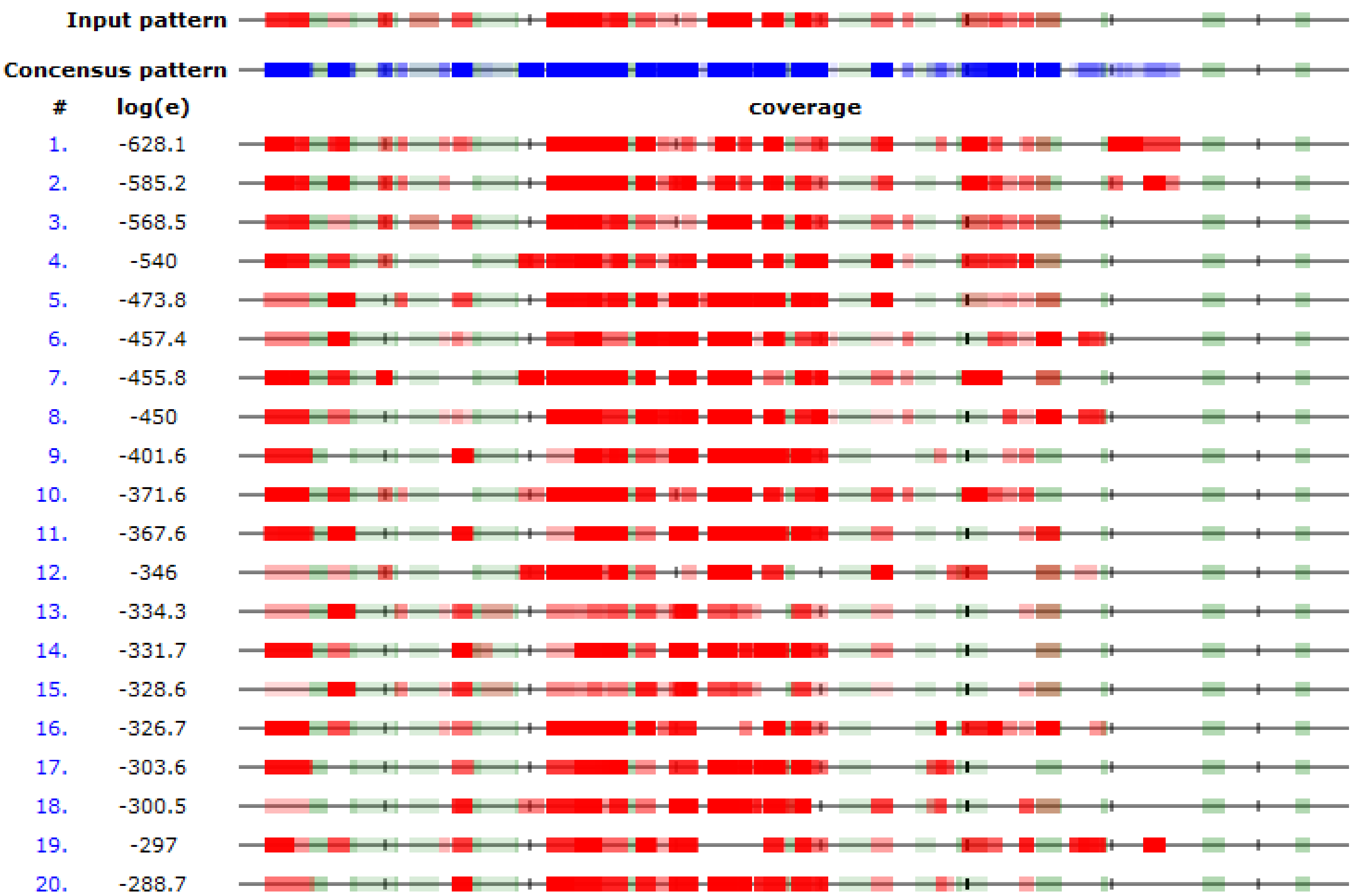


gpmDB statistics for Sun Mar 3 11:49:49 2013 UTC (#3315)

models =	217,125
proteins =	84,408,917
distinct proteins =	1,724,816
protein redundancy =	48.9 ×
peptides =	687,211,623
distinct peptides =	4,286,043
peptide redundancy =	160.3 ×
residues =	9,620,962,722
statistics archive:	GPMDB
pages viewed:	global map
US visits	map
European visits	map
Asian visits	map
Oceania visits	map
South American visits	map
African visits	map

http://gpmdb.thegpm.org

Comparison with GPMDB



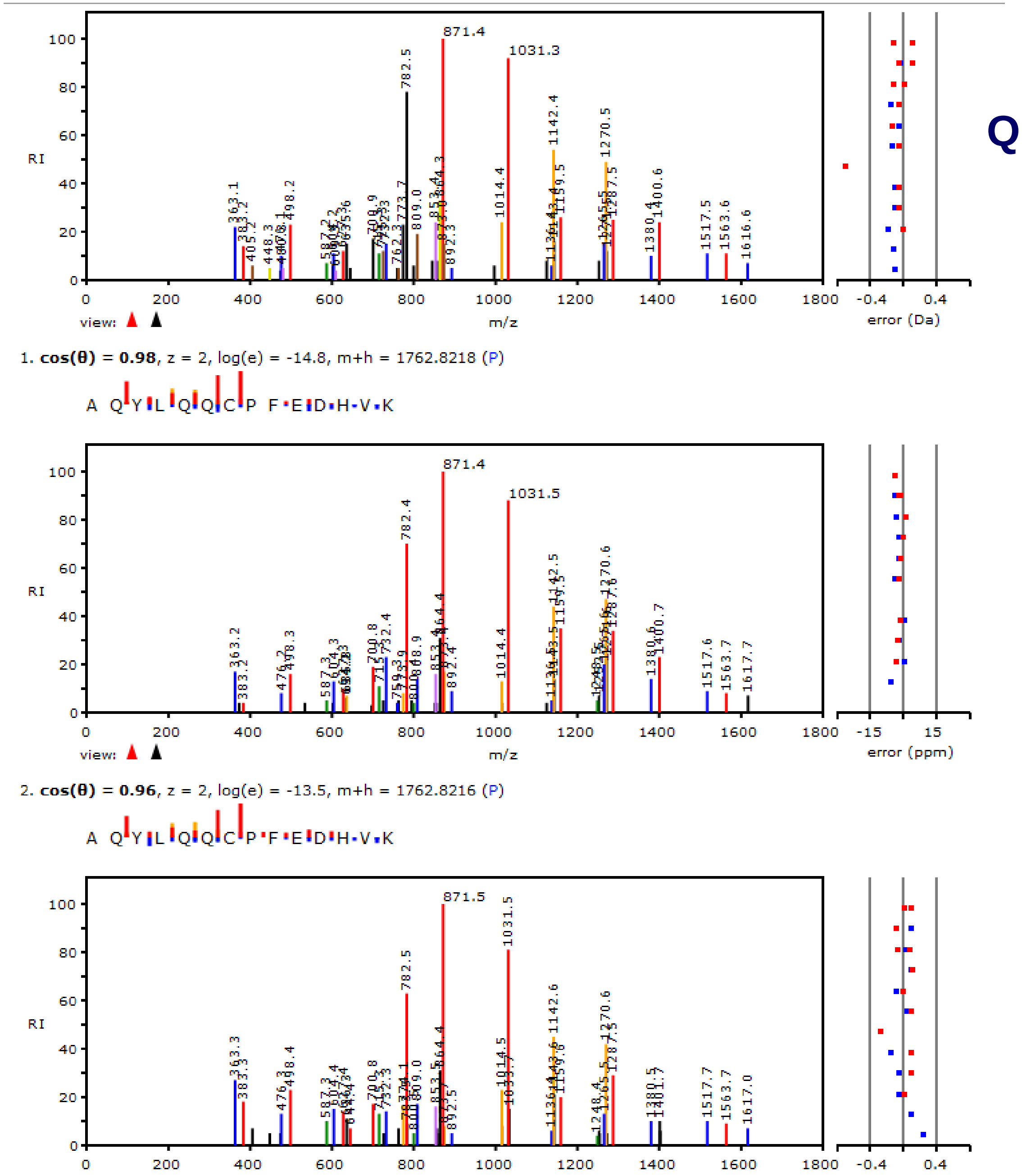
Input pattern

Concensus pattern

#	log(e)	coverage
1.	-628.1	
2.	-585.2	
3.	-568.5	
4.	-540	
5.	-473.8	
6.	-457.4	
7.	-455.8	
8.	-450	
9.	-401.6	
10.	-371.6	
11.	-367.6	
12.	-346	
13.	-334.3	
14.	-331.7	
15.	-328.6	
16.	-326.7	
17.	-303.6	
18.	-300.5	
19.	-297	
20.	-288.7	

Most proteins show very reproducible peptide patterns

Comparison with GPMDB



1. **cos(θ) = 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(θ) = 0.96**, z = 2, log(e) = -13.5, m+h = 1762.8216 (P)

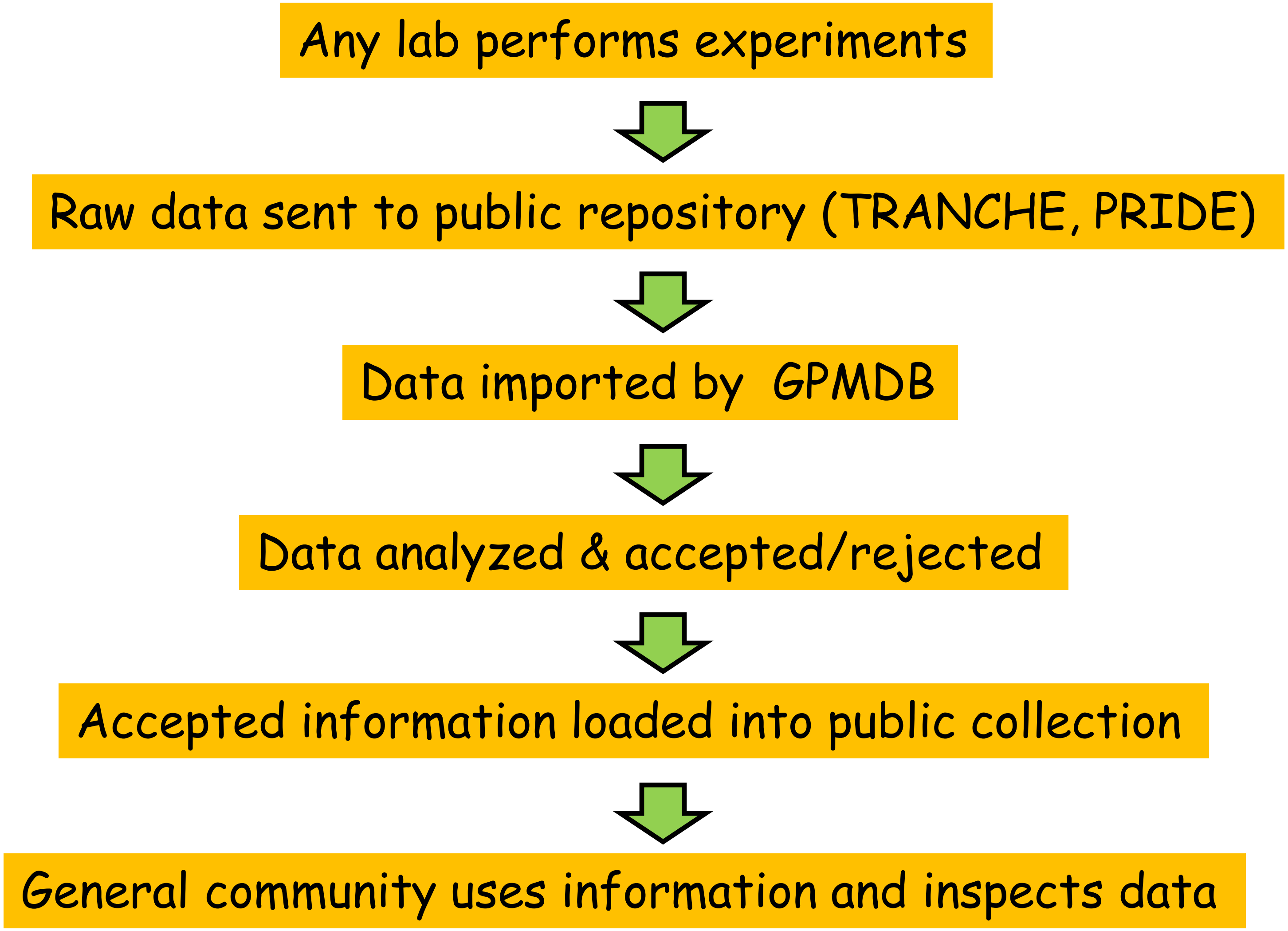
A Q Y L Q Q C P F E D H V K

Query Spectrum

Best match In GPMDB

Second best match In GPMDB

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 GPMDB

1. **MS/MS data (required)**
   1. MS raw data files
   2. ASCII files: mzXML, mzML, MGF, DTA, etc.
   3. Analysis files: DAT, MSF, BIOML

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

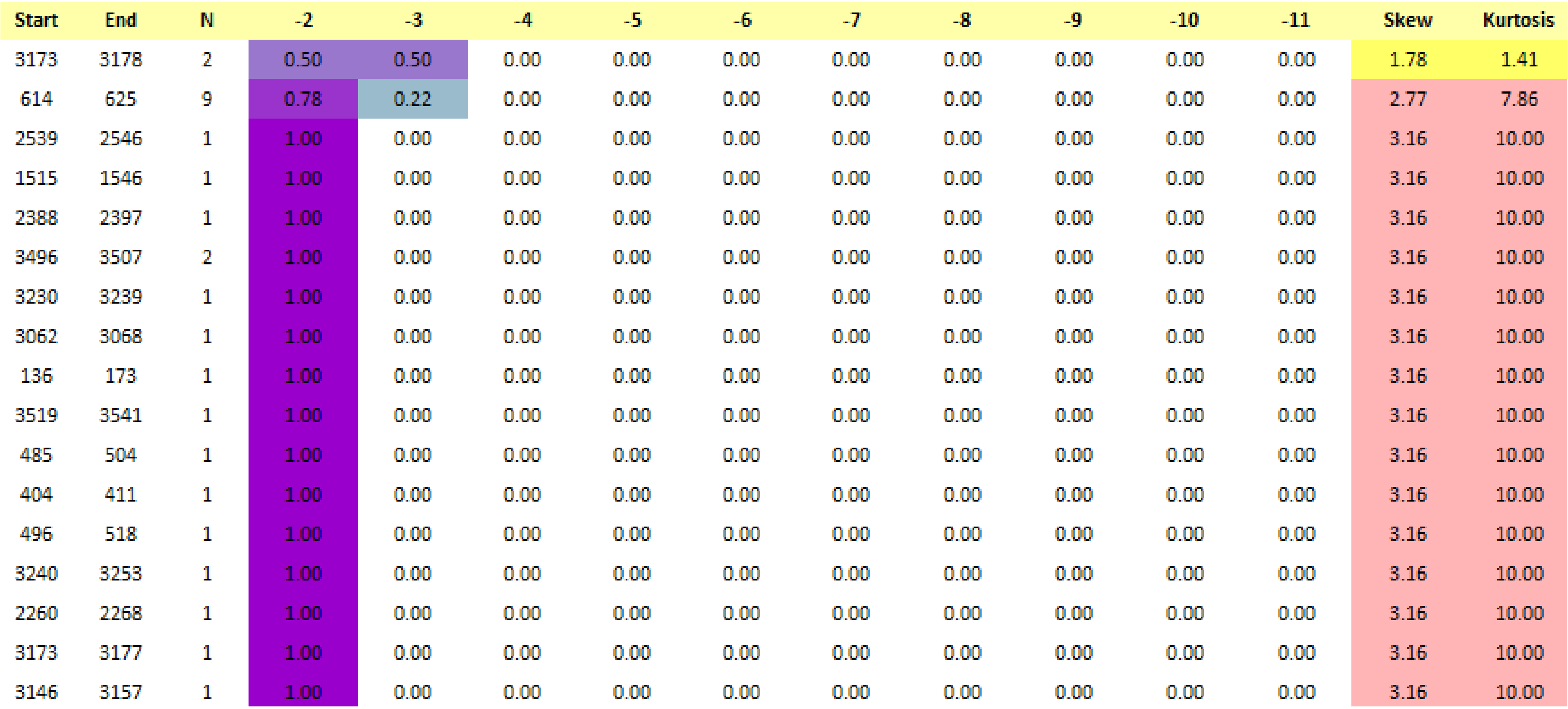
How to characterize the evidence in GPMDB for a protein?

High confidence	🟢
Medium confidence	🟡
Low confidence	🔴
No observation	⚫

Statistical model for 212 observations of TP53

Start	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 DNAH2

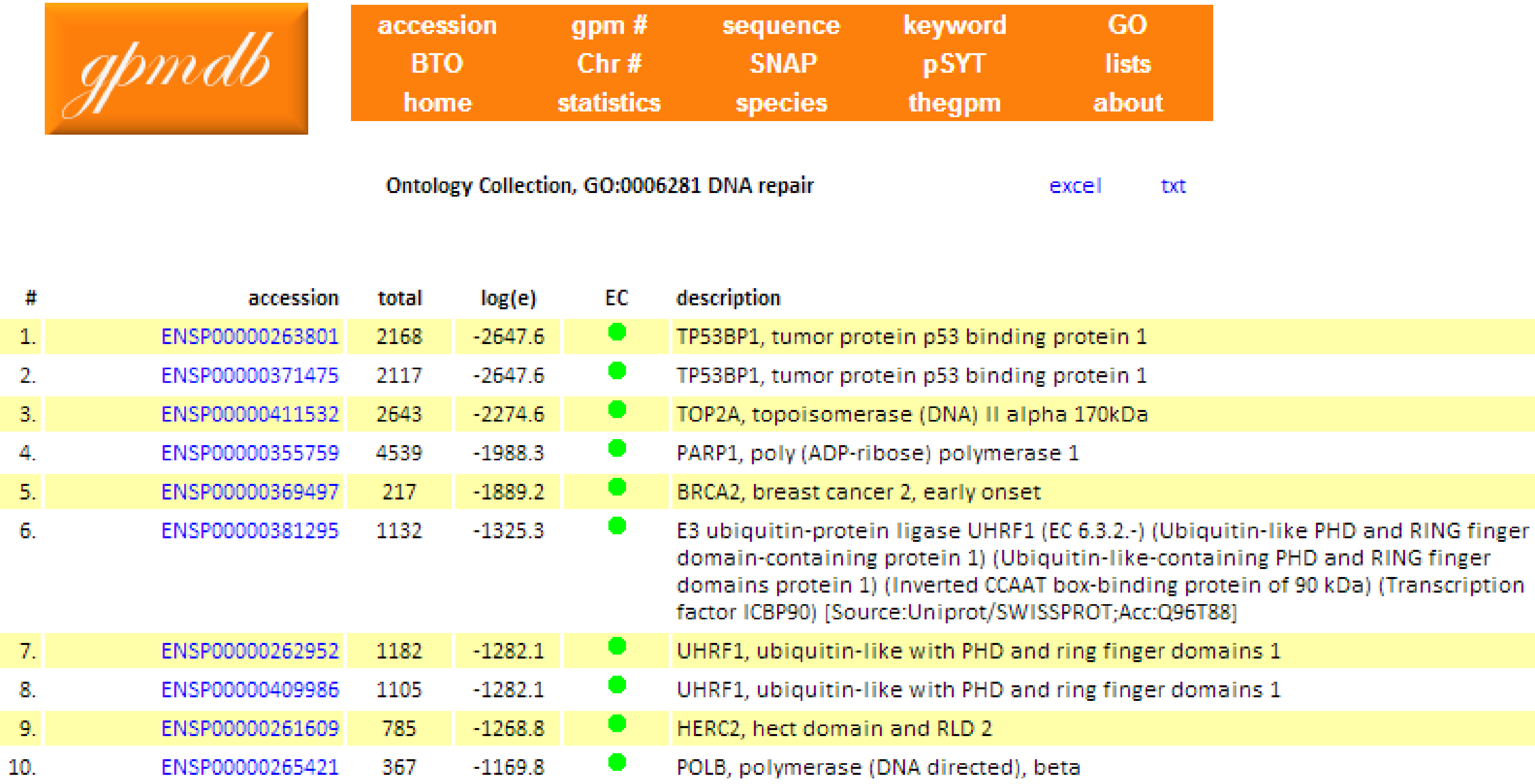


Start	End	N	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	Skew	Kurtosis
3173	3178	2	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.78	1.41
614	625	9	0.78	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.77	7.86
2539	2546	1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.16	10.00
1515	1546	1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.16	10.00
2388	2397	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
3496	3507	2	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.16	10.00
3230	3239	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
3062	3068	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
136	173	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
3519	3541	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
485	504	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
404	411	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
496	518	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
3240	3253	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
2260	2268	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
3173	3177	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
3146	3157	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

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

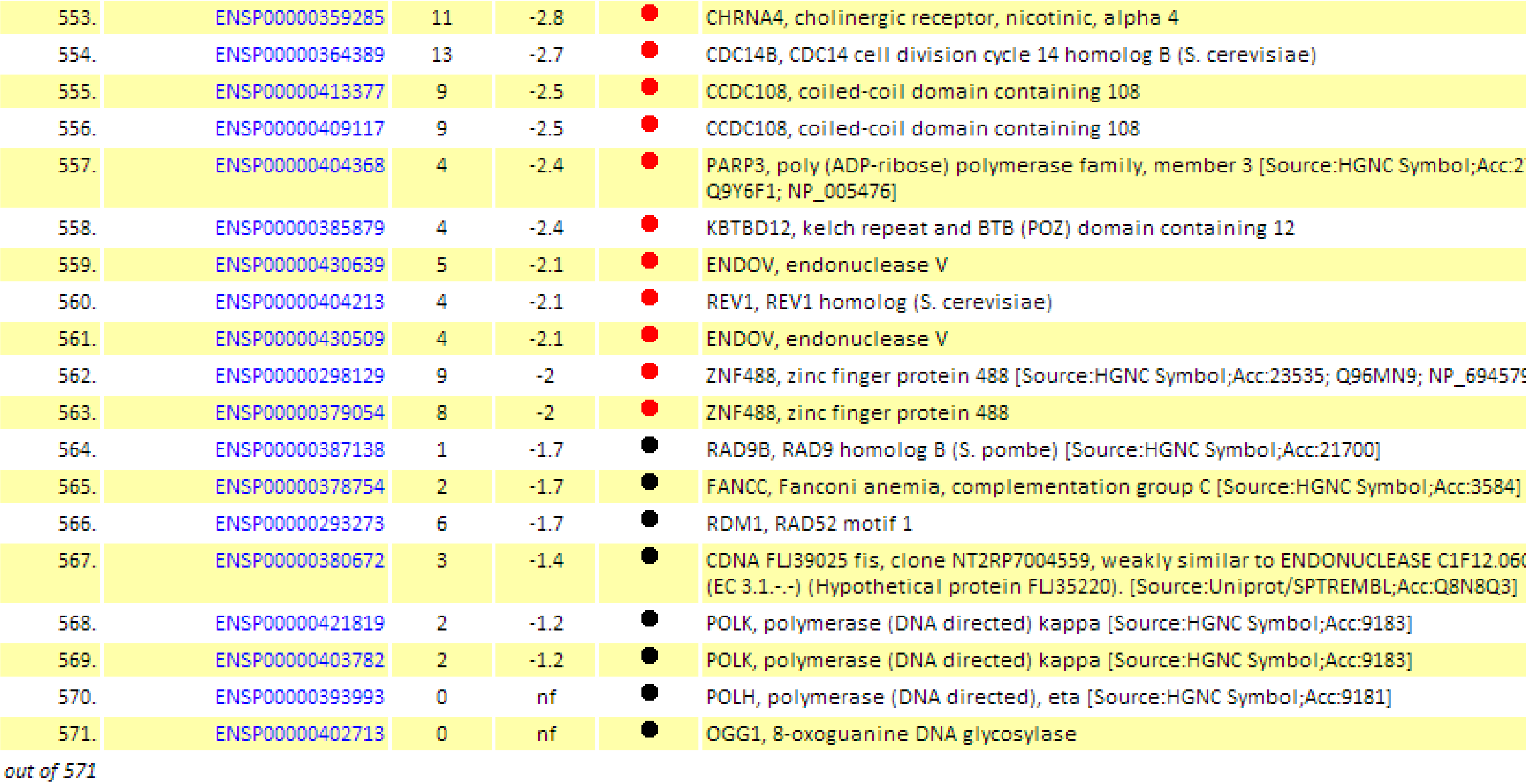
DNA Repair



Ontology Collection, GO:0006281 DNA repair

#	accession	total	log(e)	EC	description
1.	ENSP00000263801	2168	-2647.6	●	TP53BP1, tumor protein p53 binding protein 1
2.	ENSP00000371475	2117	-2647.6	●	TP53BP1, tumor protein p53 binding protein 1
3.	ENSP00000411532	2643	-2274.6	●	TOP2A, topoisomerase (DNA) II alpha 170kDa
4.	ENSP00000355759	4539	-1988.3	●	PARP1, poly (ADP-ribose) polymerase 1
5.	ENSP00000369497	217	-1889.2	●	BRCA2, breast cancer 2, early onset
6.	ENSP00000381295	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.	ENSP00000262952	1182	-1282.1	●	UHRF1, ubiquitin-like with PHD and ring finger domains 1
8.	ENSP00000409986	1105	-1282.1	●	UHRF1, ubiquitin-like with PHD and ring finger domains 1
9.	ENSP00000261609	785	-1268.8	●	HERC2, hect domain and RLD 2
10.	ENSP00000265421	367	-1169.8	●	POLB, polymerase (DNA directed), beta

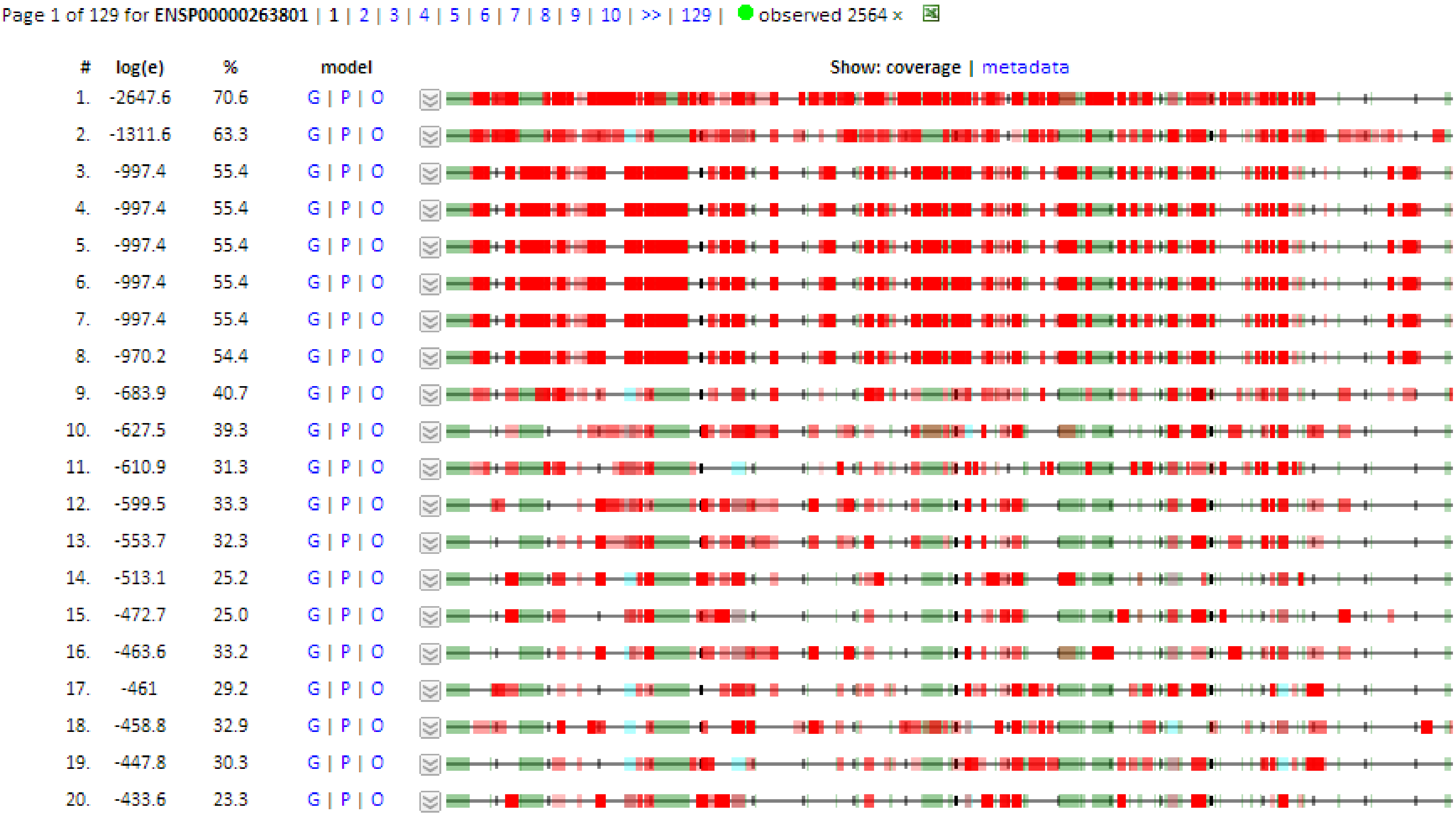
DNA Repair



553.	ENSP00000359285	11	-2.8	🔴	CHRNA4, cholinergic receptor, nicotinic, alpha 4
554.	ENSP00000364389	13	-2.7	🔴	CDC14B, CDC14 cell division cycle 14 homolog B (S. cerevisiae)
555.	ENSP00000413377	9	-2.5	🔴	CCDC108, coiled-coil domain containing 108
556.	ENSP00000409117	9	-2.5	🔴	CCDC108, coiled-coil domain containing 108
557.	ENSP00000404368	4	-2.4	🔴	PARP3, poly (ADP-ribose) polymerase family, member 3 [Source:HGNC Symbol;Acc:2 Q9Y6F1; NP_005476]
558.	ENSP00000385879	4	-2.4	🔴	KBTBD12, kelch repeat and BTB (POZ) domain containing 12
559.	ENSP00000430639	5	-2.1	🔴	ENDOV, endonuclease V
560.	ENSP00000404213	4	-2.1	🔴	REV1, REV1 homolog (S. cerevisiae)
561.	ENSP00000430509	4	-2.1	🔴	ENDOV, endonuclease V
562.	ENSP00000298129	9	-2	🔴	ZNF488, zinc finger protein 488 [Source:HGNC Symbol;Acc:23535; Q96MN9; NP_69457
563.	ENSP00000379054	8	-2	🔴	ZNF488, zinc finger protein 488
564.	ENSP00000387138	1	-1.7	⚫	RAD9B, RAD9 homolog B (S. pombe) [Source:HGNC Symbol;Acc:21700]
565.	ENSP00000378754	2	-1.7	⚫	FANCC, Fanconi anemia, complementation group C [Source:HGNC Symbol;Acc:3584]
566.	ENSP00000293273	6	-1.7	⚫	RDM1, RAD52 motif 1
567.	ENSP00000380672	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.	ENSP00000421819	2	-1.2	⚫	POLK, polymerase (DNA directed) kappa [Source:HGNC Symbol;Acc:9183]
569.	ENSP00000403782	2	-1.2	⚫	POLK, polymerase (DNA directed) kappa [Source:HGNC Symbol;Acc:9183]
570.	ENSP00000393993	0	nf	⚫	POLH, polymerase (DNA directed), eta [Source:HGNC Symbol;Acc:9181]
571.	ENSP00000402713	0	nf	⚫	OGG1, 8-oxoguanine DNA glycosylase

out of 571

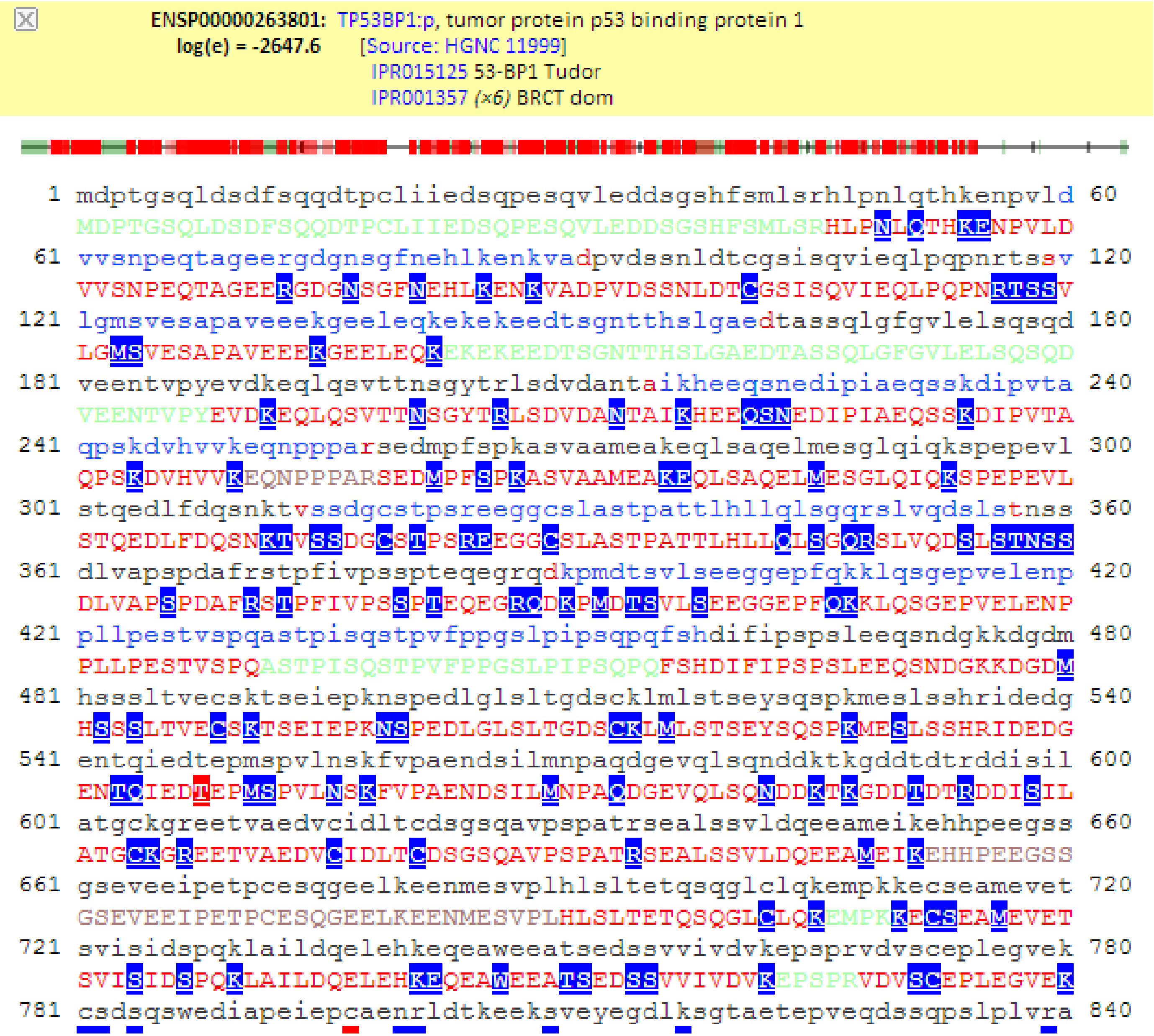
TP53BP1:p, tumor protein p53 binding protein 1



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

#	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	

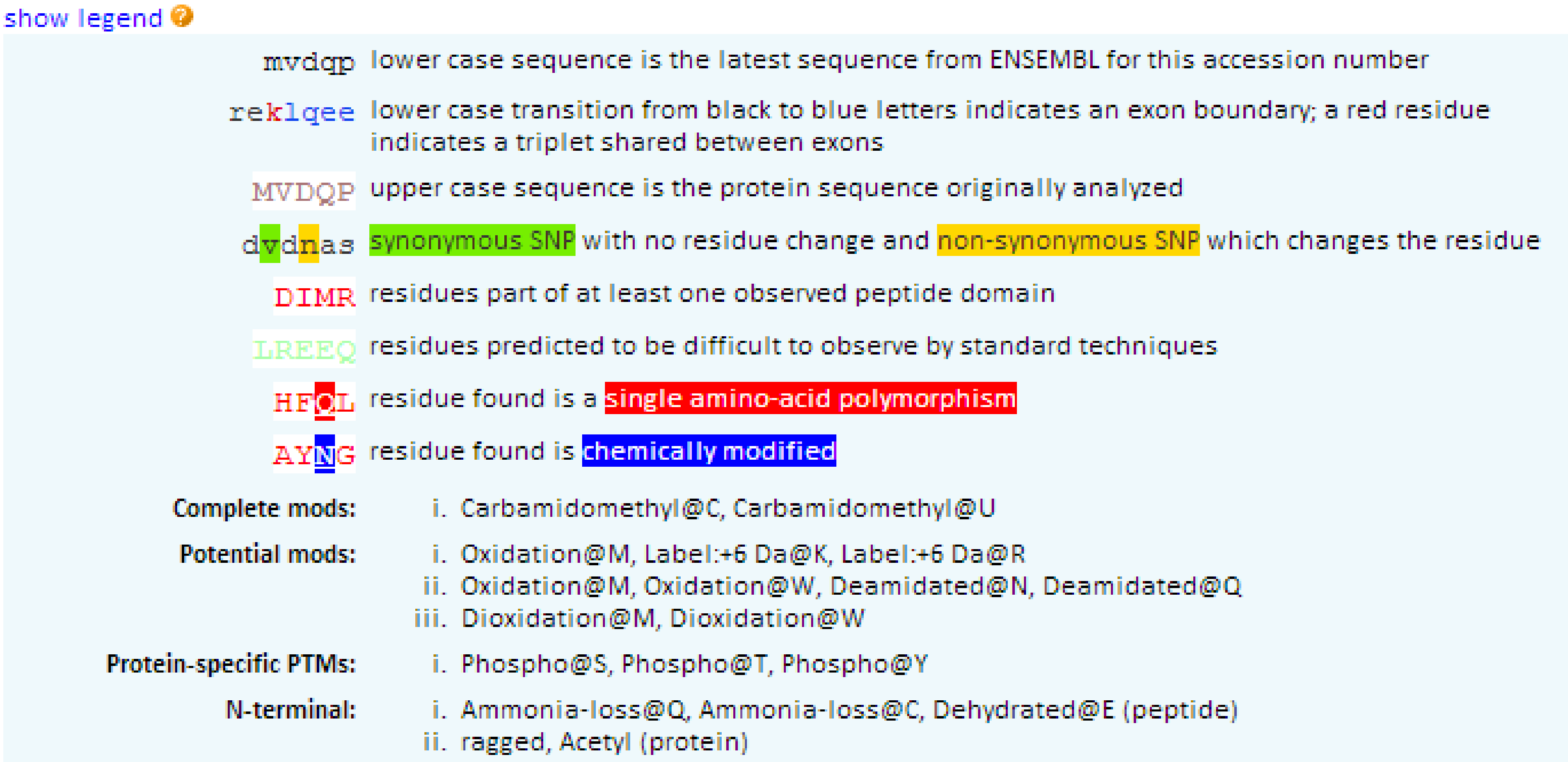
TP53BP1:p, tumor protein p53 binding protein 1



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 mdptgsqldsdfsqqdtpcliiedsqpesqvleddsgshfsmlsrhlpnlqthkenpvld  60
    MDPTGSQLDSDFSQQDTPCLIIEDSQPESQVLEDDSGSHFSMLSRHLPNLQTHKENPVLD
 61 vvsnpeqtageergdgnsgfnehlkenkvadpvdssnldtcgsisqvieqlpqpnrtssv 120
    VVSNPEQTAGEERGDGNSGFNEHLKENKVADPVDSSNLDTCGSISQVIEQLPQPNRTSSV
121 lgmsvesapaveeekgeeleqkekekeedtsgntthslgaedtassqlgfgvlelsqsqd 180
    LGMSVESAPAVEEEKGEELEQKEKEKEEDTSGNTTHSLGAEDTASSQLGFGVLELSQSQD
181 veentvpyevdkeqlqsvttnsgytrlsdvdantaikheeqsnedipiaeqsskdipvta 240
    VEENTVPYEVDKEQLQSVTTNSGYTRLSDVDANTAIKHEEQSNEDIPIAEQSSKDIPVTA
241 qpskdvhvvkeqnppparsedmpfspkasvaameakeqlsaqelmesglqiqkspepevl 300
    QPSKDVHVVKEQNPPPARSEDMPFSPKASVAAMEAKEQLSAQELMESGLQIQKSPEPEVL
301 stqedlfdqsnktvssdgcstpsreeggcslastpattlhllqlsgqrslvqdslstnss 360
    STQEDLFDQSNKTVSSDGCSTPSREEGGCSLASTPATTLHLLQLSGQRSLVQDSLSTNSS
361 dlvapspdafrstpfivpsspteqegrqdkpmdtsvlseeggepfqkklqsgepvelenp 420
    DLVAPSPDAFRSTPFIVPSSPTEQEGRQDKPMDTSVLSEEGGEPFQKKLQSGEPVELENP
421 pllpestvspqastpisqstpvfppgslpipsqpqfshdifipspsleeqsndgkkdgdm 480
    PLLPESTVSPQASTPISQSTPVFPPGSLPIPSQPQFSHDIFIPSPSLEEQSNDGKKDGDM
481 hssltvecsktseiepknspedlglsltgdscklmlstseysqspkmeslsshridedg 540
    HSSSLTVECSKTSEIEPKNSPEDLGLSLTGDSCKLMLSTSEYSQSPKMESLSSHRIDEDG
541 entqiedtepmspvlnskfvpaendsilmnpaqdgevqlsqnddktkgddtdtrddisil 600
    ENTQIEDTEPMSPVLNSKFVPAENDSILMNPAQDGEVQLSQNDDKTKGDDTDTRDDISIL
601 atgckgreetvaedvcidltcdsgsqavpspatrsealssvldqeeameikehhpeegss 660
    ATGCKGREETVAEDVCIDLTCDSGSQAVPSPATRSEALSSVLDQEEAMEIKEHHPEEGSS
661 gseveeipetpcesqgeelkeenmesvplhlsltetqsqglclqkempkkecseamevet 720
    GSEVEEIPETPCESQGEELKEENMESVPLHLSLTETQSQGLCLQKEMPKKECSEAMEVET
721 svisidspqklaildqelehkeqeaweeatsedssvvivdvkepsprvdvsceplegvek 780
    SVISIDSPQKLAILDQELEHKEQEAWEEATSEDSSVVIVDVKEPSPRVDVSCEPLEGVEK
781 csdsqswediapeiepcaenrldtkeeksveyegdlksgtaetepveqdssqpslplvra 840
```

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

HFQL 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

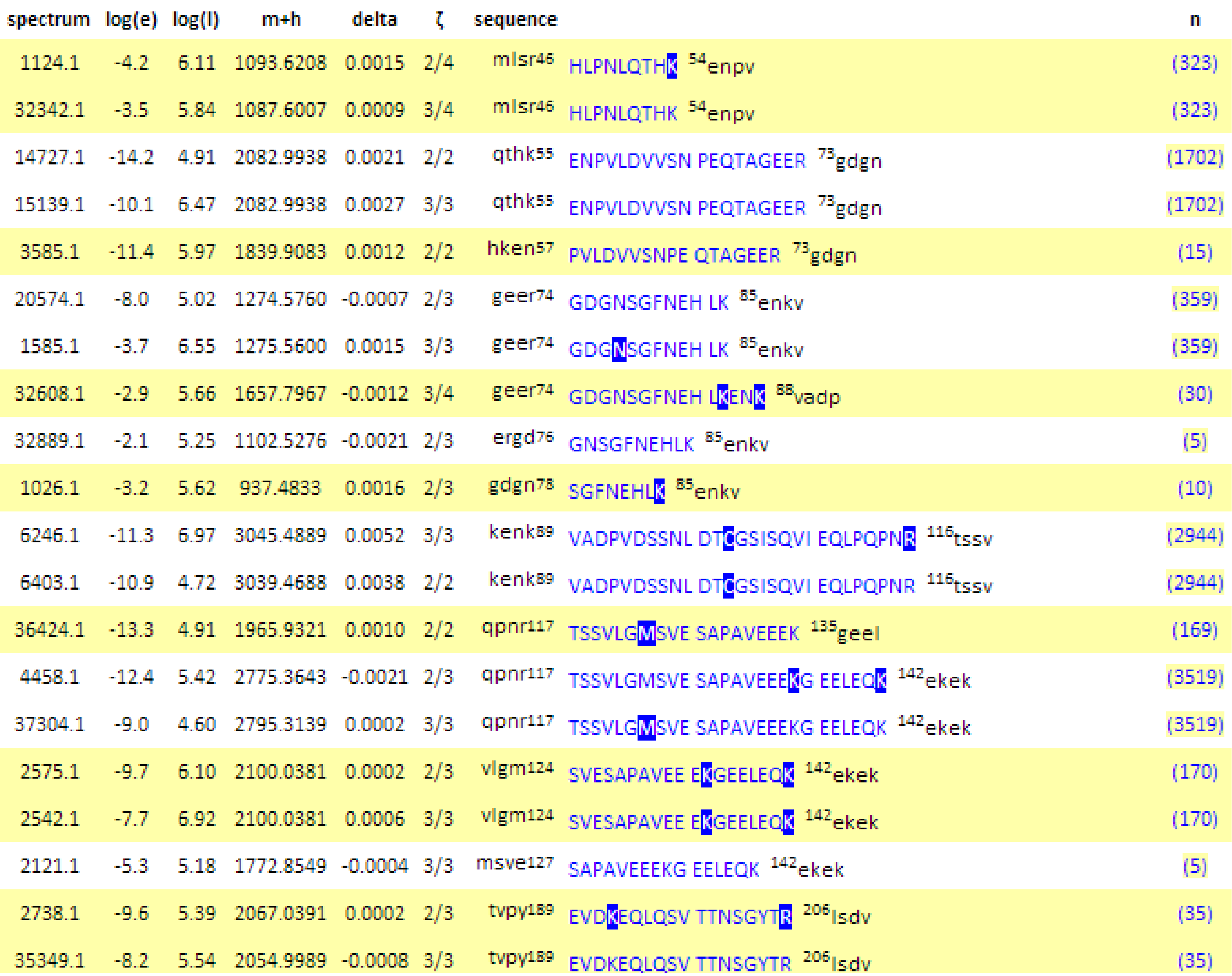
Protein-specific PTMs:

i. Phospho@S, Phospho@T, Phospho@Y

N-terminal:

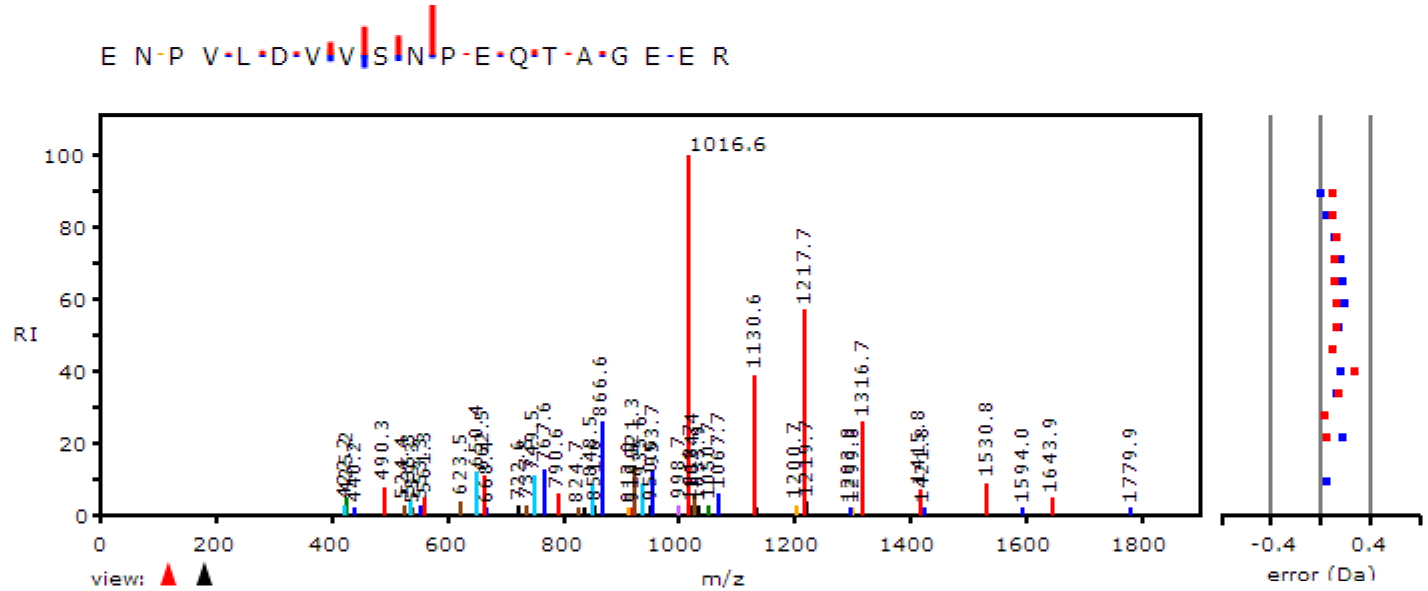
i. Ammonia-loss@Q, Ammonia-loss@C, Dehydrated@E (peptide)
ii. ragged, Acetyl (protein)

TP53BP1:p, tumor protein p53 binding protein 1



spectrum	log(e)	log(I)	m+h	delta	ζ	sequence		n
1124.1	-4.2	6.11	1093.6208	0.0015	2/4	mlsr^{46}	HLPNLQTHK ^{54}enpv	(323)
32342.1	-3.5	5.84	1087.6007	0.0009	3/4	mlsr^{46}	HLPNLQTHK ^{54}enpv	(323)
14727.1	-14.2	4.91	2082.9938	0.0021	2/2	qthk^{55}	ENPVLDVVSN PEQTAGEER ^{73}gdgn	(1702)
15139.1	-10.1	6.47	2082.9938	0.0027	3/3	qthk^{55}	ENPVLDVVSN PEQTAGEER ^{73}gdgn	(1702)
3585.1	-11.4	5.97	1839.9083	0.0012	2/2	hken^{57}	PVLDVVSNPE QTAGEER ^{73}gdgn	(15)
20574.1	-8.0	5.02	1274.5760	-0.0007	2/3	geer^{74}	GDGNSGFNEH LK ^{85}enkv	(359)
1585.1	-3.7	6.55	1275.5600	0.0015	3/3	geer^{74}	GDGNSGFNEH LK ^{85}enkv	(359)
32608.1	-2.9	5.66	1657.7967	-0.0012	3/4	geer^{74}	GDGNSGFNEH LKENK ^{88}vadp	(30)
32889.1	-2.1	5.25	1102.5276	-0.0021	2/3	ergd^{76}	GNSGFNEHLK ^{85}enkv	(5)
1026.1	-3.2	5.62	937.4833	0.0016	2/3	gdgn^{78}	SGFNEHLK ^{85}enkv	(10)
6246.1	-11.3	6.97	3045.4889	0.0052	3/3	kenk^{89}	VADPVDSSNL DTCGSISQVI EQLPQPNR ^{116}tssv	(2944)
6403.1	-10.9	4.72	3039.4688	0.0038	2/2	kenk^{89}	VADPVDSSNL DTCGSISQVI EQLPQPNR ^{116}tssv	(2944)
36424.1	-13.3	4.91	1965.9321	0.0010	2/2	qpnr^{117}	TSSVLGMSVE SAPAVEEEK ^{135}geel	(169)
4458.1	-12.4	5.42	2775.3643	-0.0021	2/3	qpnr^{117}	TSSVLGMSVE SAPAVEEEKG EELEQK ^{142}ekek	(3519)
37304.1	-9.0	4.60	2795.3139	0.0002	3/3	qpnr^{117}	TSSVLGMSVE SAPAVEEEKG EELEQK ^{142}ekek	(3519)
2575.1	-9.7	6.10	2100.0381	0.0002	2/3	vlgm^{124}	SVESAPAVEE EKGEELEQK ^{142}ekek	(170)
2542.1	-7.7	6.92	2100.0381	0.0006	3/3	vlgm^{124}	SVESAPAVEE EKGEELEQK ^{142}ekek	(170)
2121.1	-5.3	5.18	1772.8549	-0.0004	3/3	msve^{127}	SAPAVEEEKG EELEQK ^{142}ekek	(5)
2738.1	-9.6	5.39	2067.0391	0.0002	2/3	tvpy^{189}	EVDKEQLQSV TTNSGYTR ^{206}lsdv	(35)
35349.1	-8.2	5.54	2054.9989	-0.0008	3/3	tvpy^{189}	EVDKEQLQSV TTNSGYTR ^{206}lsdv	(35)

TP53BP1:p, tumor protein p53 binding protein 1



matched/total: # ions: 84% intensity: 95% σ: 0.13 Da

bond	^{+1}y	$^{+1}y^{-17}$	$^{+1}y^{-18}$	^{+1}b	$^{+1}b^{-17}$	$^{+1}b^{-18}$
E_1	1953.951	1936.925	1935.941	130.050	113.023	112.039
N_2	1839.908	1822.882	1821.898	244.093	227.066	226.082
P_3	1742.856	1725.829	1724.845	341.146	324.119	323.135
V_4	1643.787	1626.761	1625.777	440.214	423.187	422.203
L_5	1530.703	1513.676	1512.692	553.298	536.271	535.287
D_6	1415.676	1398.650	1397.666	668.325	651.298	650.314
V_7	1316.608	1299.581	1298.597	767.393	750.367	749.383
V_8	1217.539	1200.513	1199.529	866.462	849.435	848.451
S_9	1130.507	1113.481	1112.497	953.494	936.467	935.483
N_{10}	1016.464	999.438	998.454	1067.537	1050.510	1049.526
P_{11}	919.412	902.385	901.401	1164.590	1147.563	1146.579
E_{12}	790.369	773.342	772.358	1293.632	1276.606	1275.622
Q_{13}	662.310	645.284	644.300	1421.691	1404.664	1403.680
T_{14}	561.263	544.236	543.252	1522.738	1505.712	1504.728
A_{15}	490.226	473.199	472.215	1593.775	1576.749	1575.765
G_{16}	433.204	416.178	415.194	1650.797	1633.770	1632.786
E_{17}	304.162	287.135	286.151	1779.840	1762.813	1761.829
E_{18}	175.119	158.092	157.108	1908.882	1891.856	1890.872

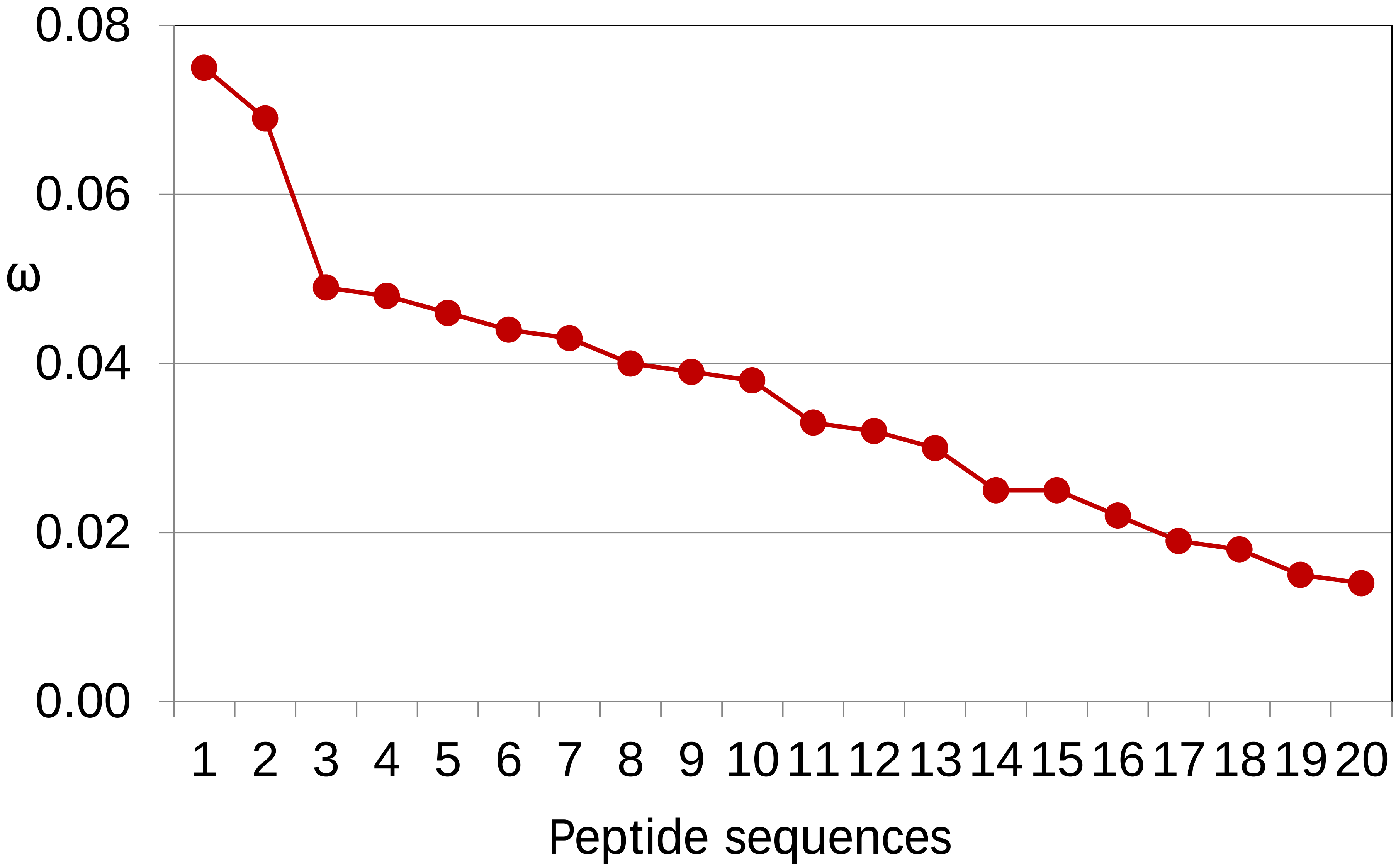
Peptide observations, catalase

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

Peptide frequency (ω), catalase

Peptide Sequence	ω
FSTVAGESGSADTVR	0.08
FNTANDDNVTQVR	0.07
AFYVNVLNEEQR	0.05
LVNANGEAVYCK	0.05
GPLLVQDVVFTDEMAHFDR	0.05
LSQEDPDYGIR	0.04
LFAYPDTHR	0.04
NLSVEDAAR	0.04
FYTEDGNWDLVGNNTPIFFIR	0.04
ADVLTTGAGNPVGDK	0.04

Global frequency of observation (ω), catalase



Omega (Ω) value for a protein identification

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

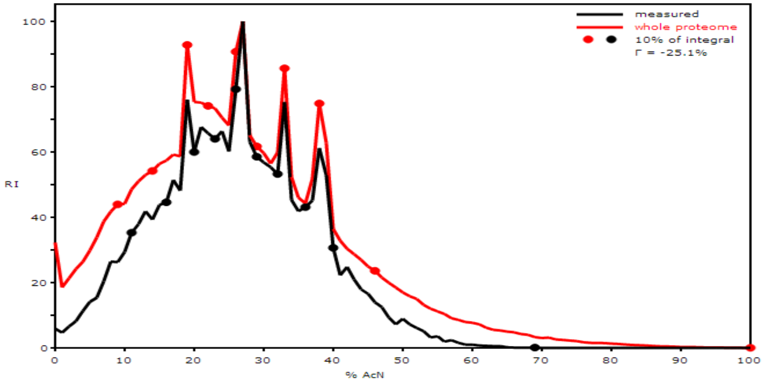
$$\Omega(protein) = \sum_{j} \omega_j$$

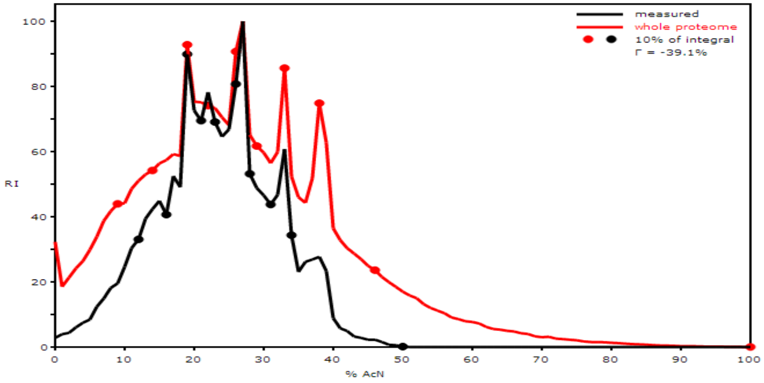
$$\Omega(protein) \leq 1$$

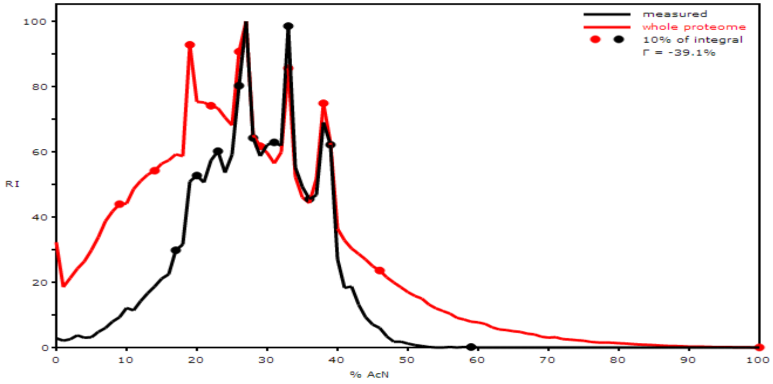
Protein Ω's for a set of identifications

Protein ID	Ω (z=2)	Ω (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

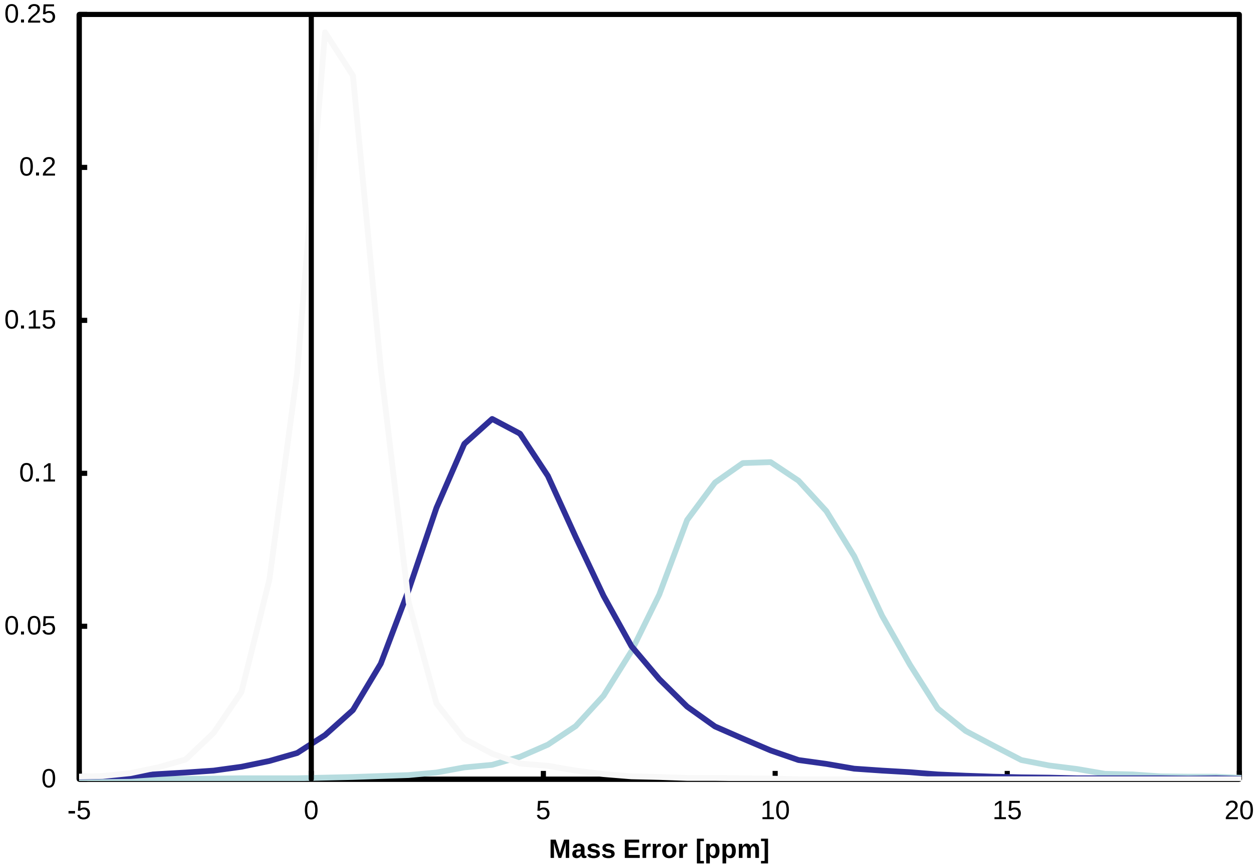
Retention Time Distribution



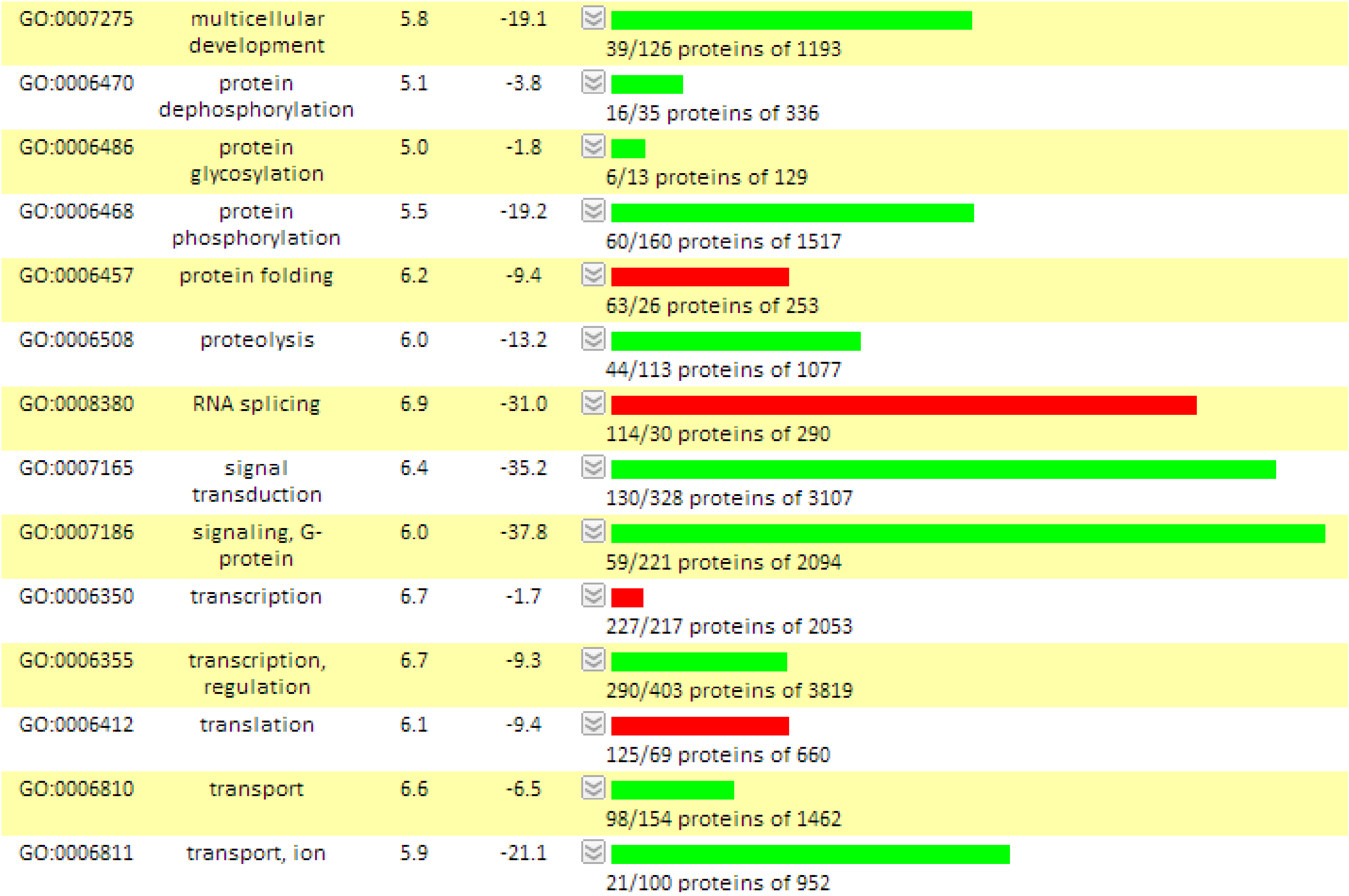




Mass Accuracy

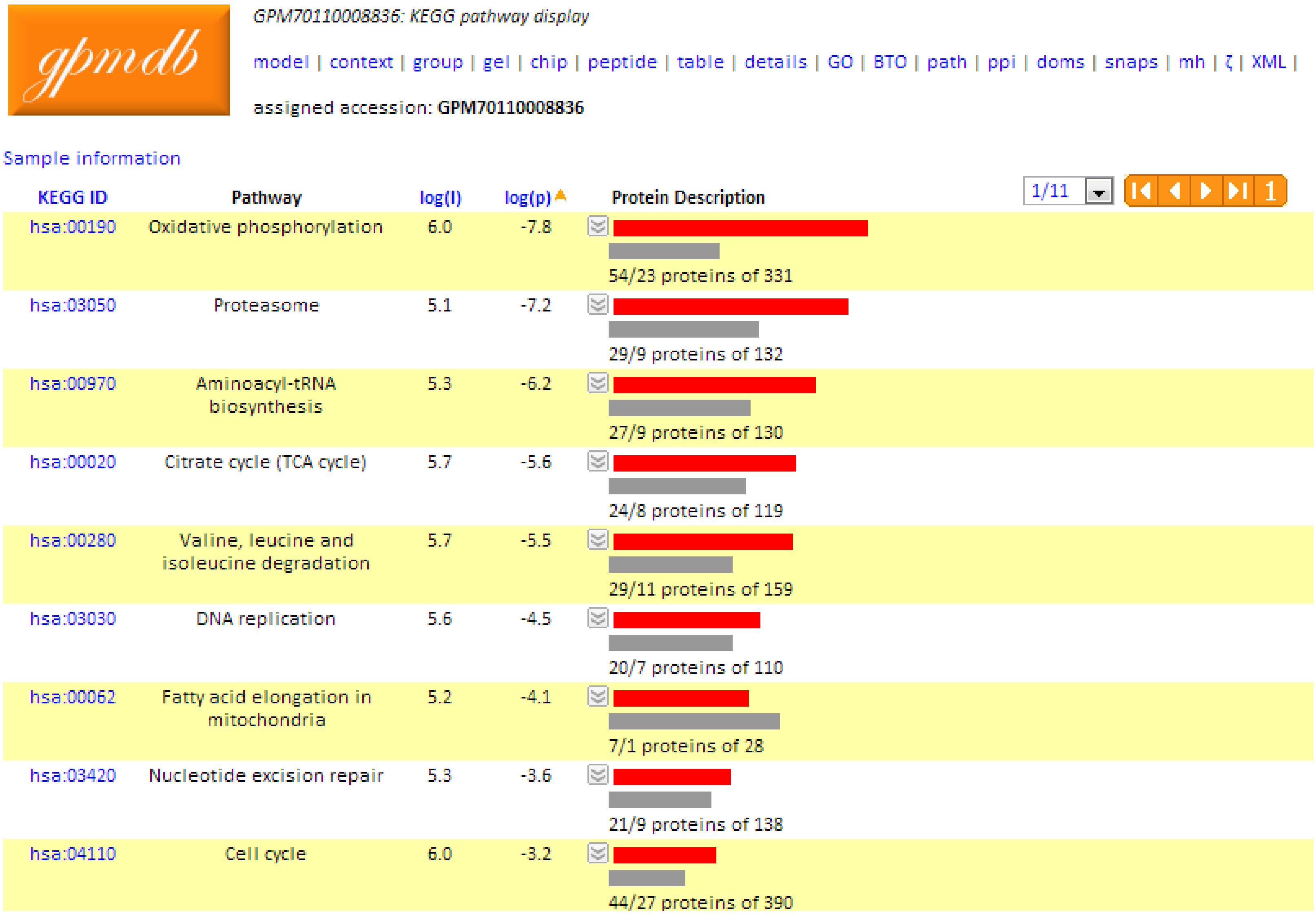


GO Cellular Processes



GO:0007275	multicellular development	5.8	-19.1	🟩	39/126 proteins of 1193
GO:0006470	protein dephosphorylation	5.1	-3.8	🟩	16/35 proteins of 336
GO:0006486	protein glycosylation	5.0	-1.8	🟩	6/13 proteins of 129
GO:0006468	protein phosphorylation	5.5	-19.2	🟩	60/160 proteins of 1517
GO:0006457	protein folding	6.2	-9.4	🟥	63/26 proteins of 253
GO:0006508	proteolysis	6.0	-13.2	🟩	44/113 proteins of 1077
GO:0008380	RNA splicing	6.9	-31.0	🟥	114/30 proteins of 290
GO:0007165	signal transduction	6.4	-35.2	🟩	130/328 proteins of 3107
GO:0007186	signaling, G-protein	6.0	-37.8	🟩	59/221 proteins of 2094
GO:0006350	transcription	6.7	-1.7	🟥	227/217 proteins of 2053
GO:0006355	transcription, regulation	6.7	-9.3	🟩	290/403 proteins of 3819
GO:0006412	translation	6.1	-9.4	🟥	125/69 proteins of 660
GO:0006810	transport	6.6	-6.5	🟩	98/154 proteins of 1462
GO:0006811	transport, ion	5.9	-21.1	🟩	21/100 proteins of 952

KEGG Pathways


gpmdb
model | context | group | gel | chip | peptide | table | details | GO | BTO | path | ppi | doms | snaps | mh | ζ | XML |
1/11
1

GPM70110008836: KEGG pathway display

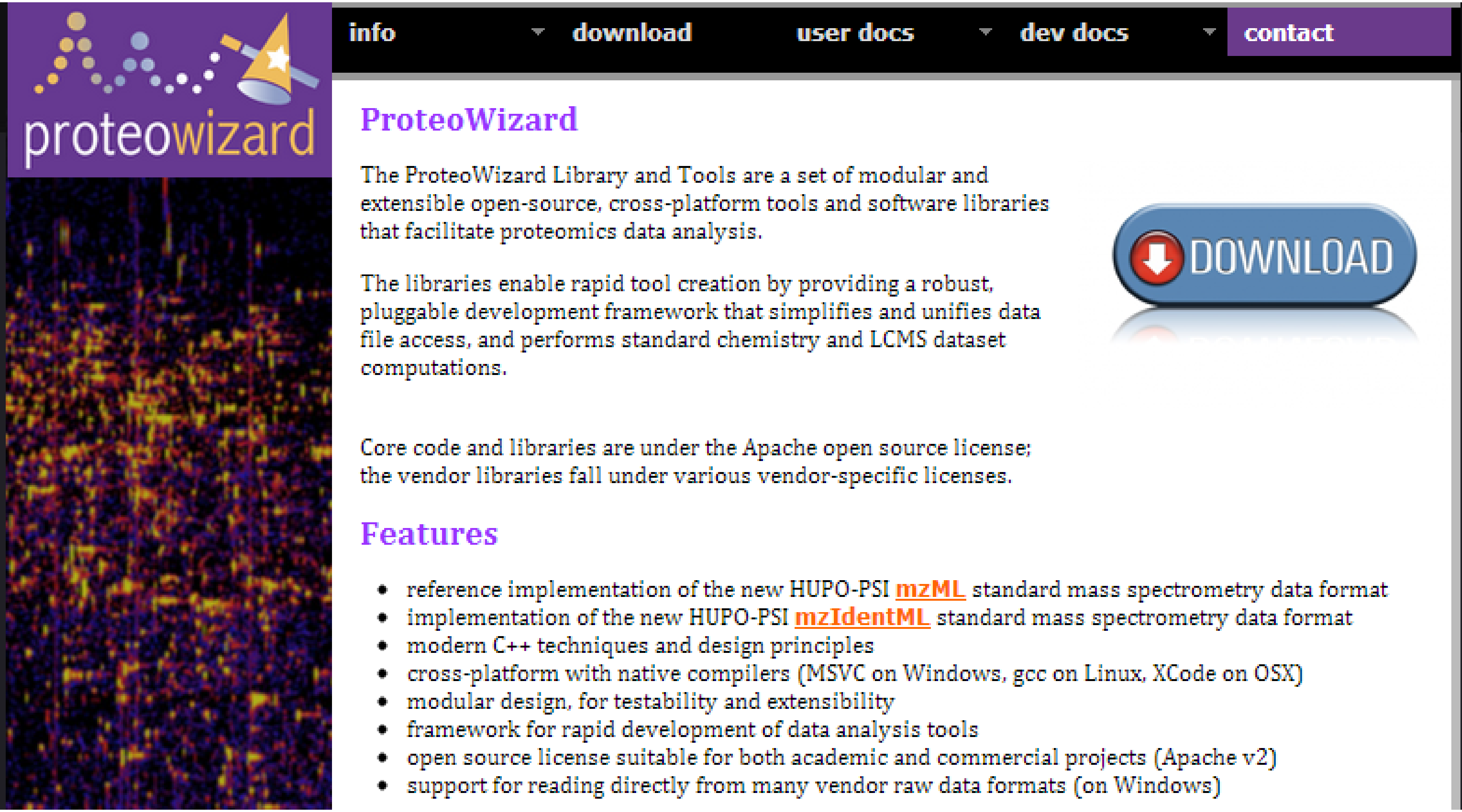
assigned accession: **GPM70110008836**

Sample information

KEGG ID	Pathway	log(I)	log(p)▲	Protein Description
hsa:00190	Oxidative phosphorylation	6.0	-7.8	54/23 proteins of 331
hsa:03050	Proteasome	5.1	-7.2	29/9 proteins of 132
hsa:00970	Aminoacyl-tRNA biosynthesis	5.3	-6.2	27/9 proteins of 130
hsa:00020	Citrate cycle (TCA cycle)	5.7	-5.6	24/8 proteins of 119
hsa:00280	Valine, leucine and isoleucine degradation	5.7	-5.5	29/11 proteins of 159
hsa:03030	DNA replication	5.6	-4.5	20/7 proteins of 110
hsa:00062	Fatty acid elongation in mitochondria	5.2	-4.1	7/1 proteins of 28
hsa:03420	Nucleotide excision repair	5.3	-3.6	21/9 proteins of 138
hsa:04110	Cell cycle	6.0	-3.2	44/27 proteins of 390

Open-Source Resources

ProteoWizard



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.

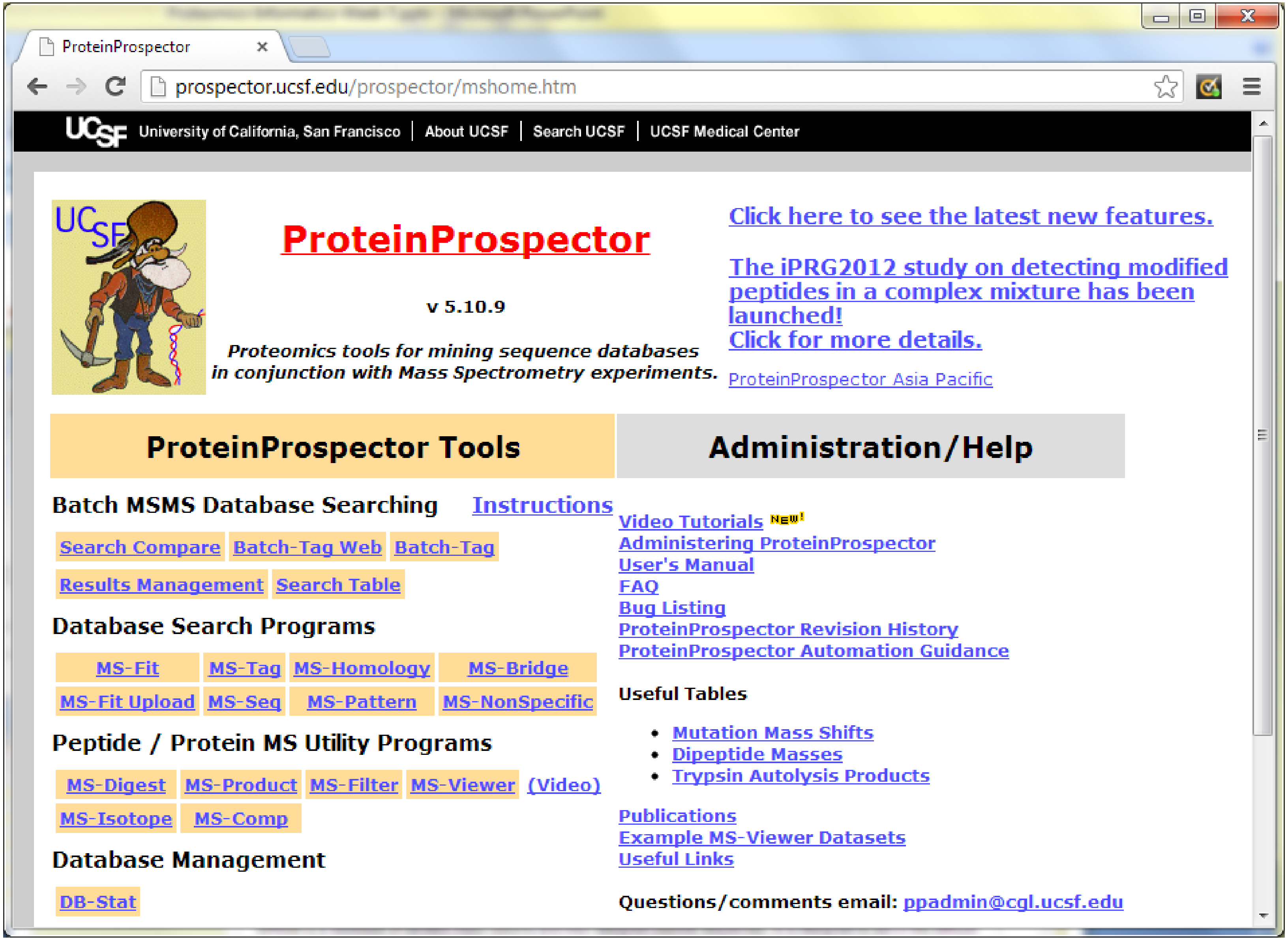
Core code and libraries are under the Apache open source license; the vendor libraries fall under various vendor-specific licenses.

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



ProteinProspector
prospector.ucsf.edu/prospector/mshome.htm
UCSF University of California, San Francisco | About UCSF | Search UCSF | UCSF Medical Center
UCSF

ProteinProspector

v 5.10.9

Proteomics tools for mining sequence databases in conjunction with Mass Spectrometry experiments.

Click here to see the latest new features.

The iPRG2012 study on detecting modified peptides in a complex mixture has been launched!
Click for more details.

ProteinProspector Asia Pacific

ProteinProspector Tools

Batch MSMS Database Searching Instructions

Search Compare | Batch-Tag Web | Batch-Tag

Results Management | Search Table

Database Search Programs

MS-Fit | MS-Tag | MS-Homology | MS-Bridge

MS-Fit Upload | MS-Seq | MS-Pattern | MS-NonSpecific

Peptide / Protein MS Utility Programs

MS-Digest | MS-Product | MS-Filter | MS-Viewer | (Video)

MS-Isotope | MS-Comp

Database Management

DB-Stat

Administration/Help

Video Tutorials NEW!
Administering ProteinProspector
User's Manual
FAQ
Bug Listing
ProteinProspector Revision History
ProteinProspector Automation Guidance

Useful Tables

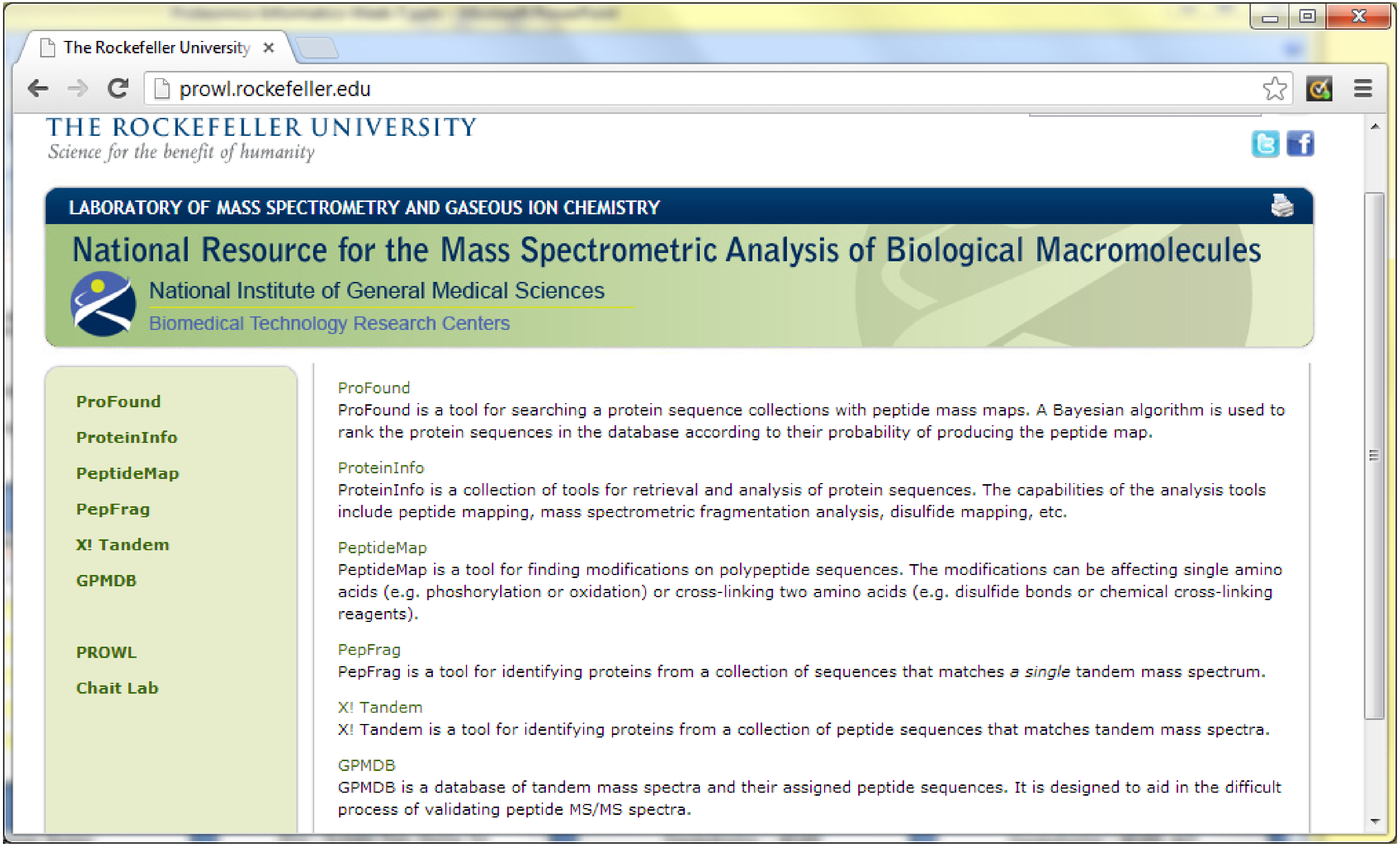
- Mutation Mass Shifts
- Dipeptide Masses
- Trypsin Autolysis Products

Publications
Example MS-Viewer Datasets
Useful Links

Questions/comments email: ppadmin@cgl.ucsf.edu

http://prospector.ucsf.edu/

PROWL



The Rockefeller University
prowl.rockefeller.edu
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
ProteinInfo
PeptideMap
PepFrag
X! Tandem
GPMDB
PROWL
Chait Lab

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. phoshorylation or oxidation) or cross-linking two amino acids (e.g. disulfide bonds or chemical cross-linking reagents).

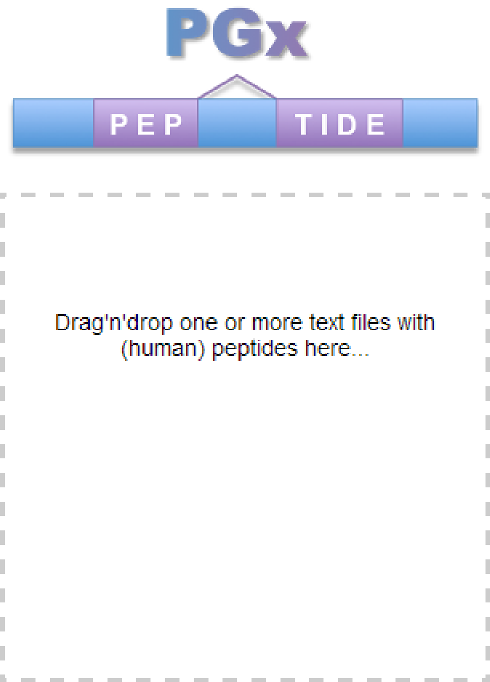
PepFrag
PepFrag is a tool for identifying proteins from a collection of sequences that matches *a single* tandem mass spectrum.

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

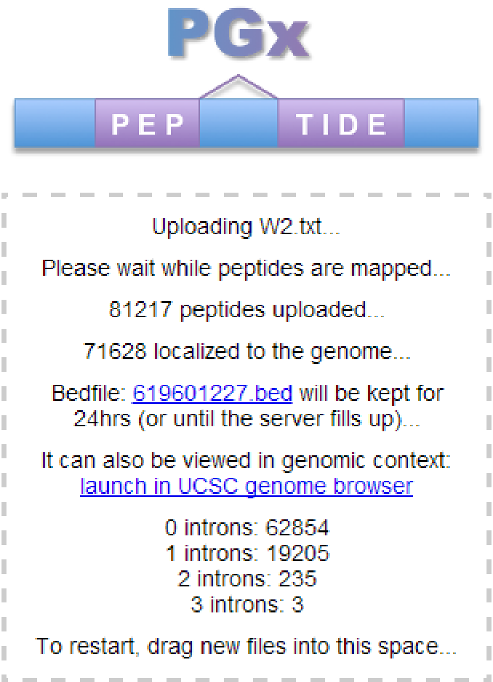
GPMDB
GPMDB is a database of tandem mass spectra and their assigned peptide sequences. It is designed to aid in the difficult process of validating peptide MS/MS spectra.

http://prowl.rockefeller.edu/

Proteogenomics - PGx



Drag'n'drop one or more text files with (human) peptides here...



Uploading W2.txt...

Please wait while peptides are mapped...

81217 peptides uploaded...

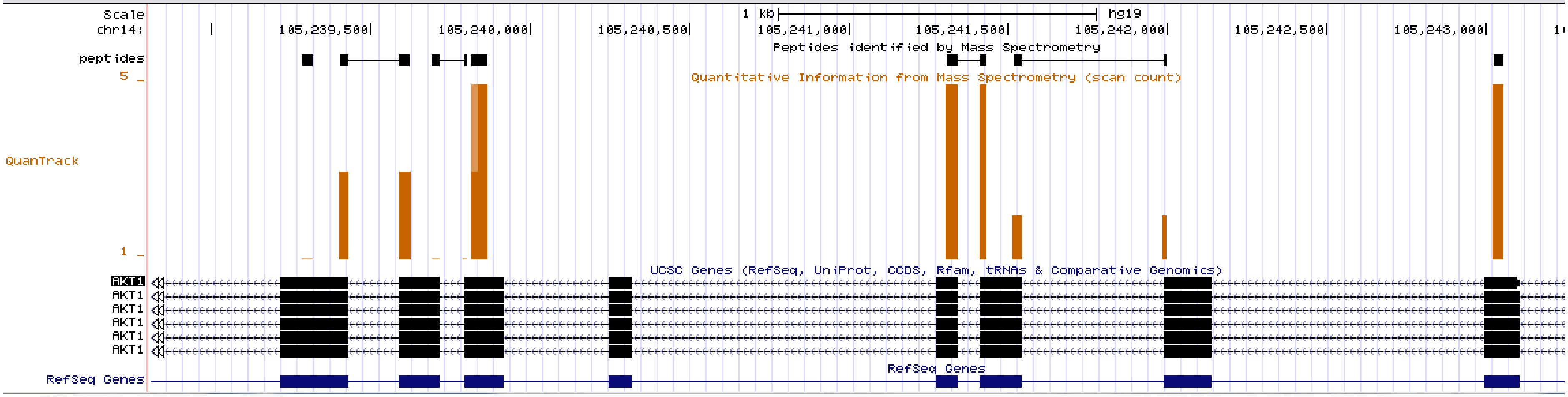
71628 localized to the genome...

Bedfile: 619601227.bed will be kept for 24hrs (or until the server fills up)...

It can also be viewed in genomic context: launch in UCSC genome browser

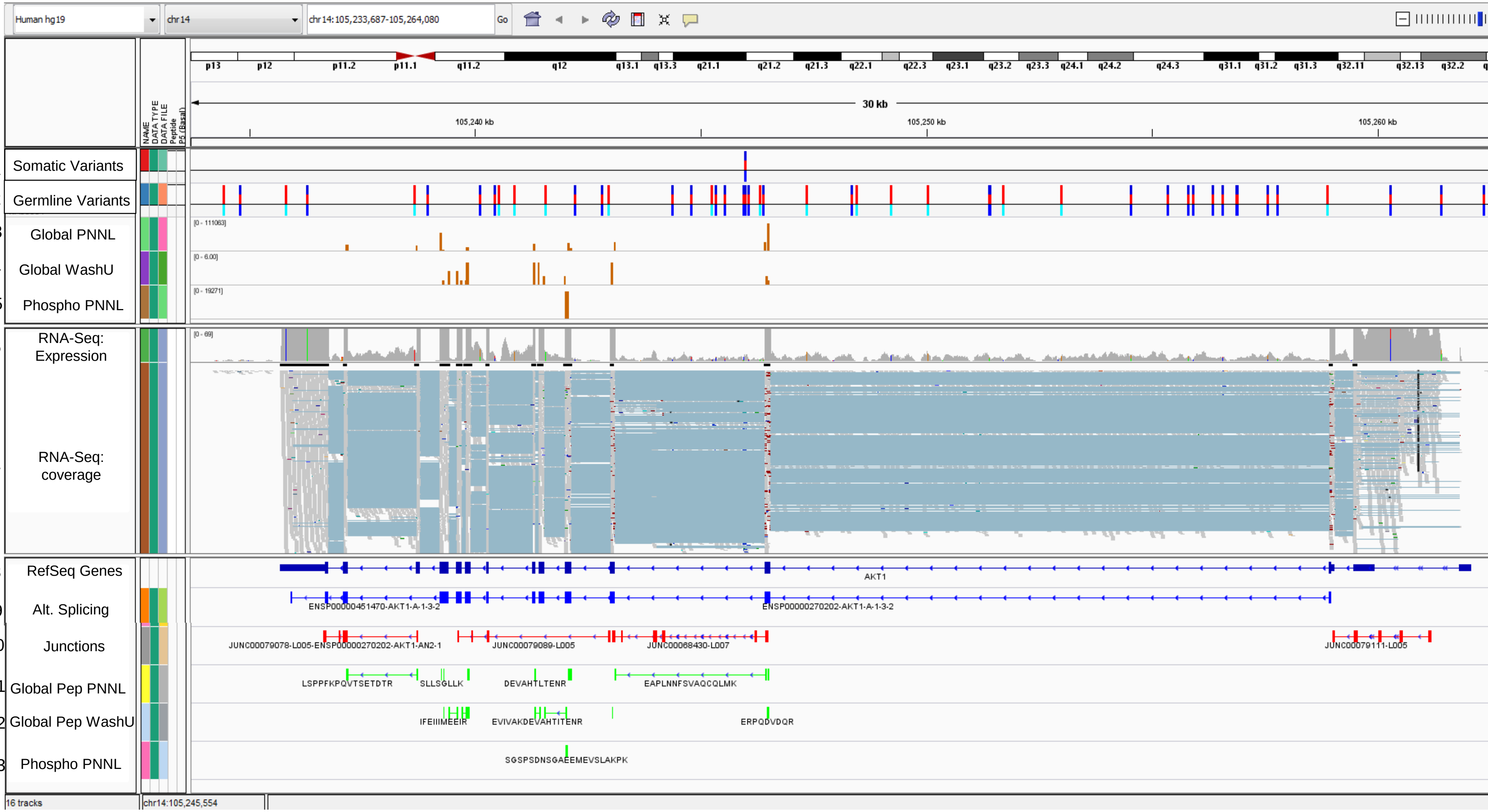
0 introns: 62854
1 introns: 19205
2 introns: 235
3 introns: 3

To restart, drag new files into this space...



http://pgx.fenyolab.org/

UCSC Genome Browser



Human hg19

chr14

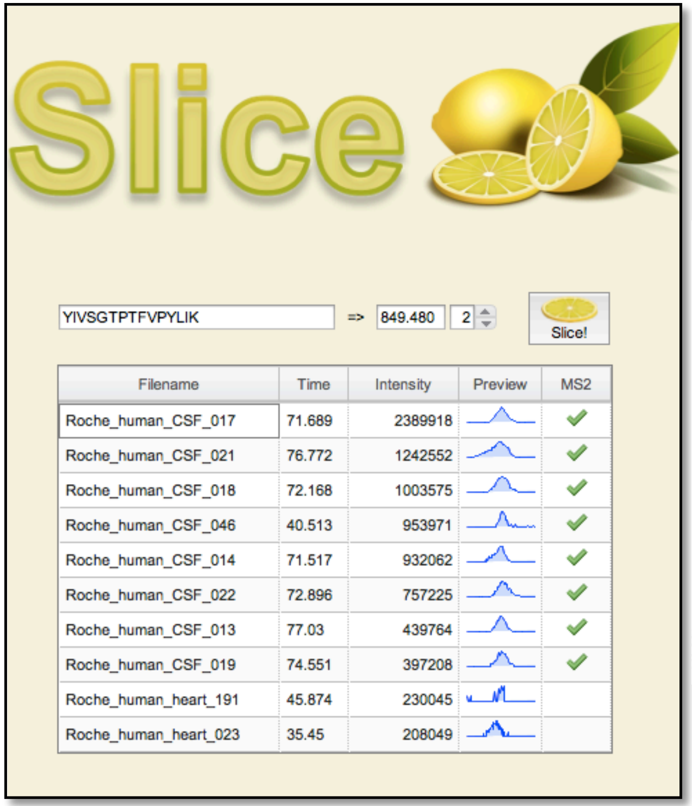
chr14:105,233,687-105,264,080 Go

16 tracks

chr14:105,245,554

http://genome.ucsc.edu/

Slice - Scalable Data Sharing for Remote Mass Informatics

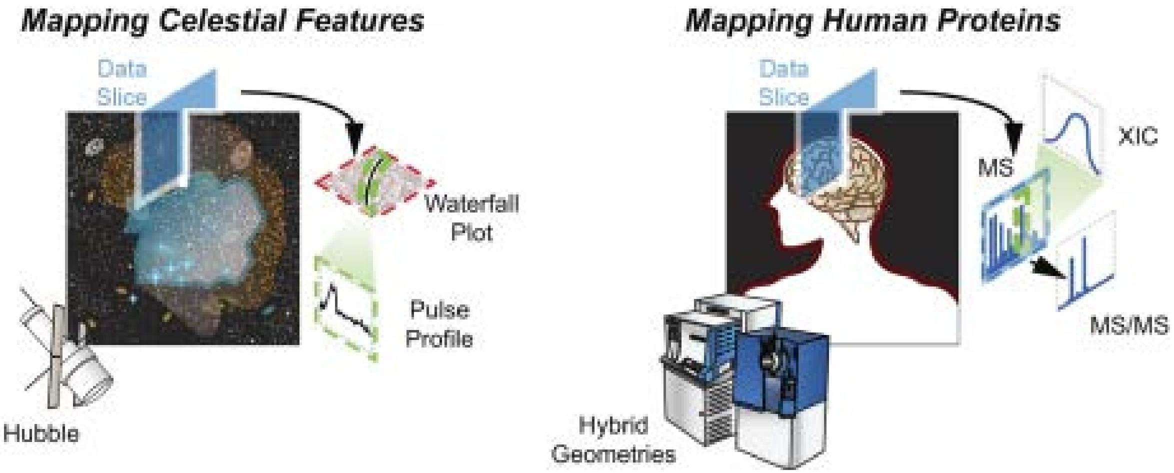


YIVSGTPTFVPYLIK => 849.480 2

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

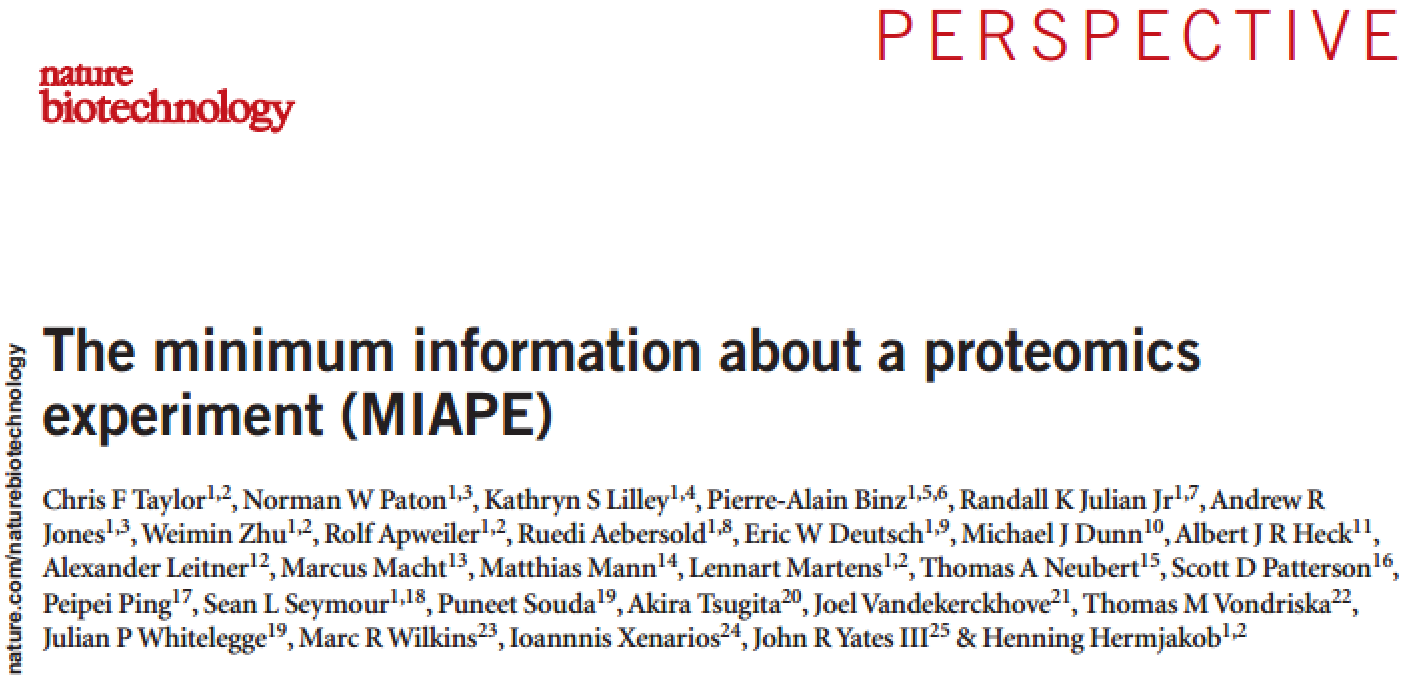
openslice.fenyolab.org



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

Standardization - MIAPE



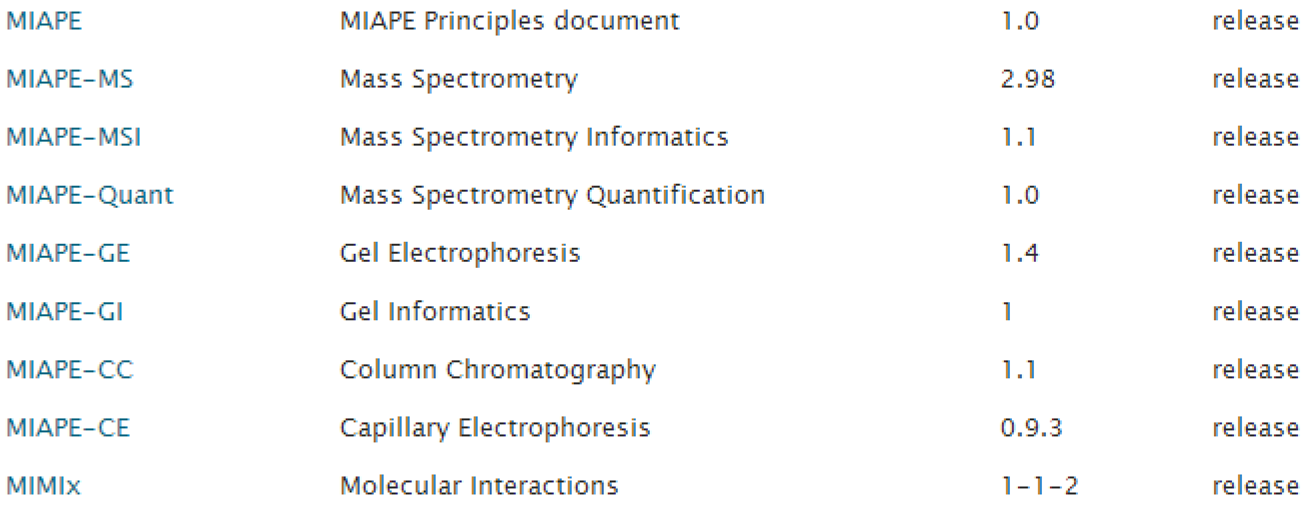
PERSPECTIVE

nature biotechnology

nature.com/naturebiotechnology

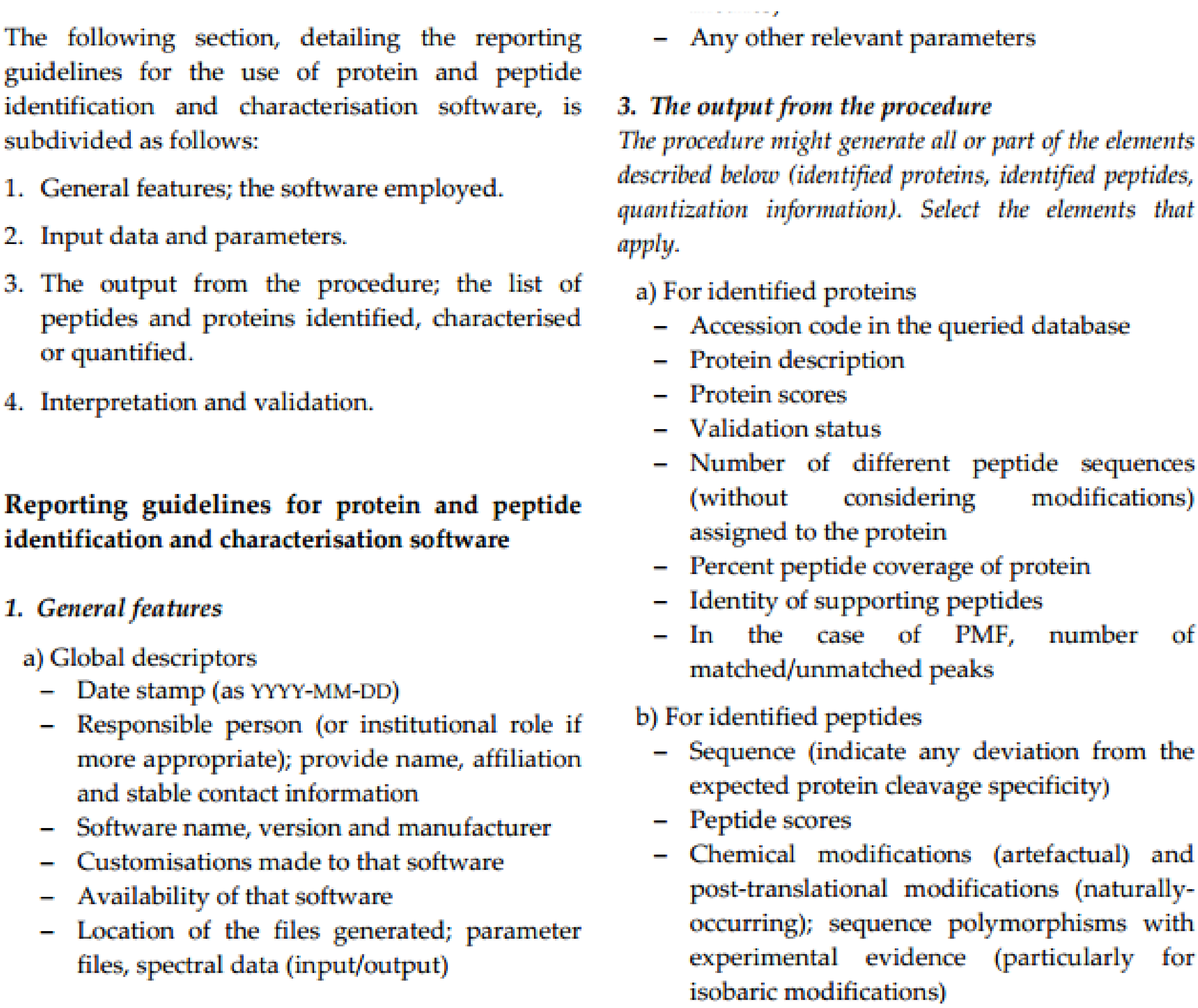
The minimum information about a proteomics experiment (MIAPE)

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



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

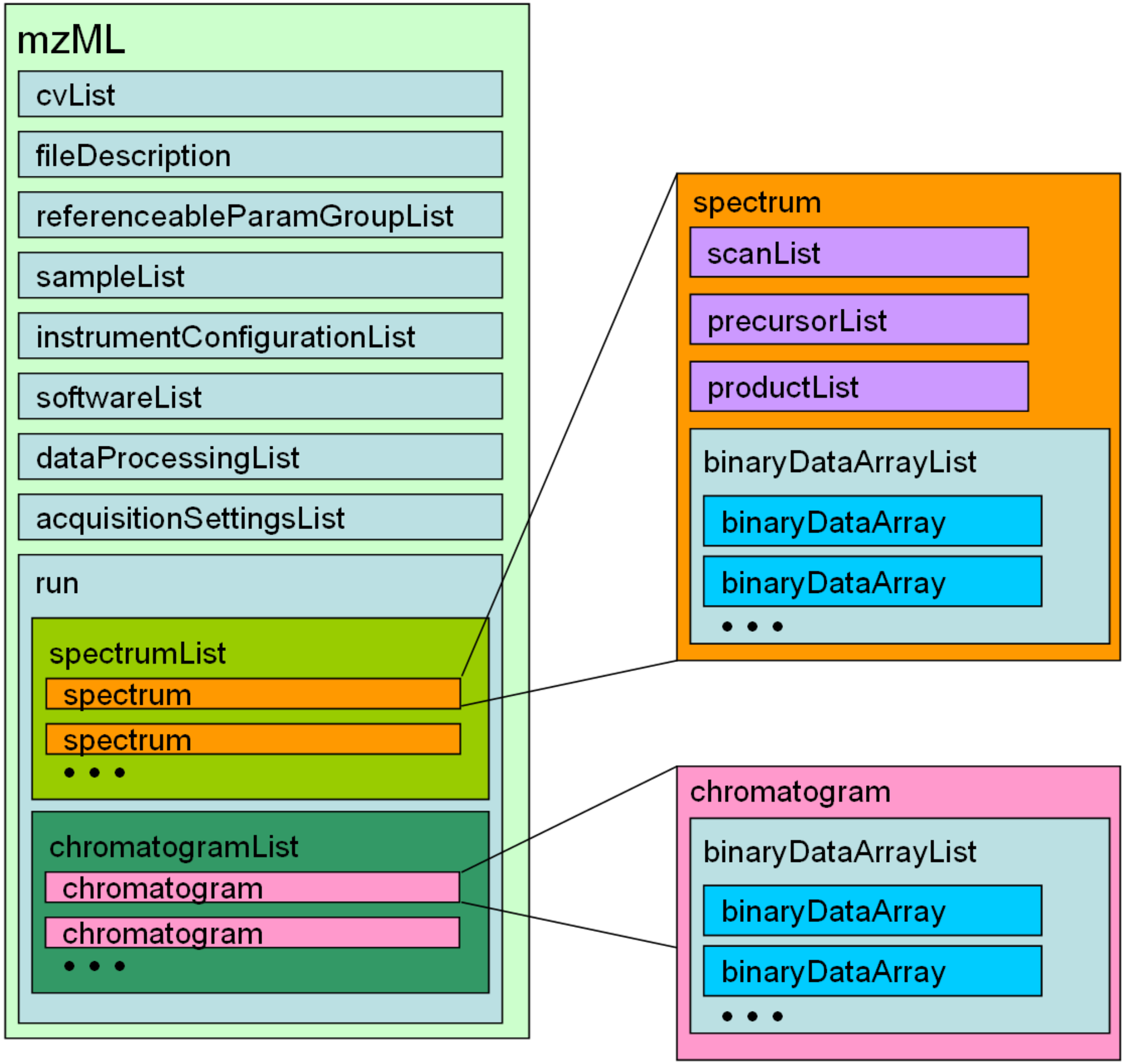
mzQuantML - describe the outputs of quantitation software for proteomics

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

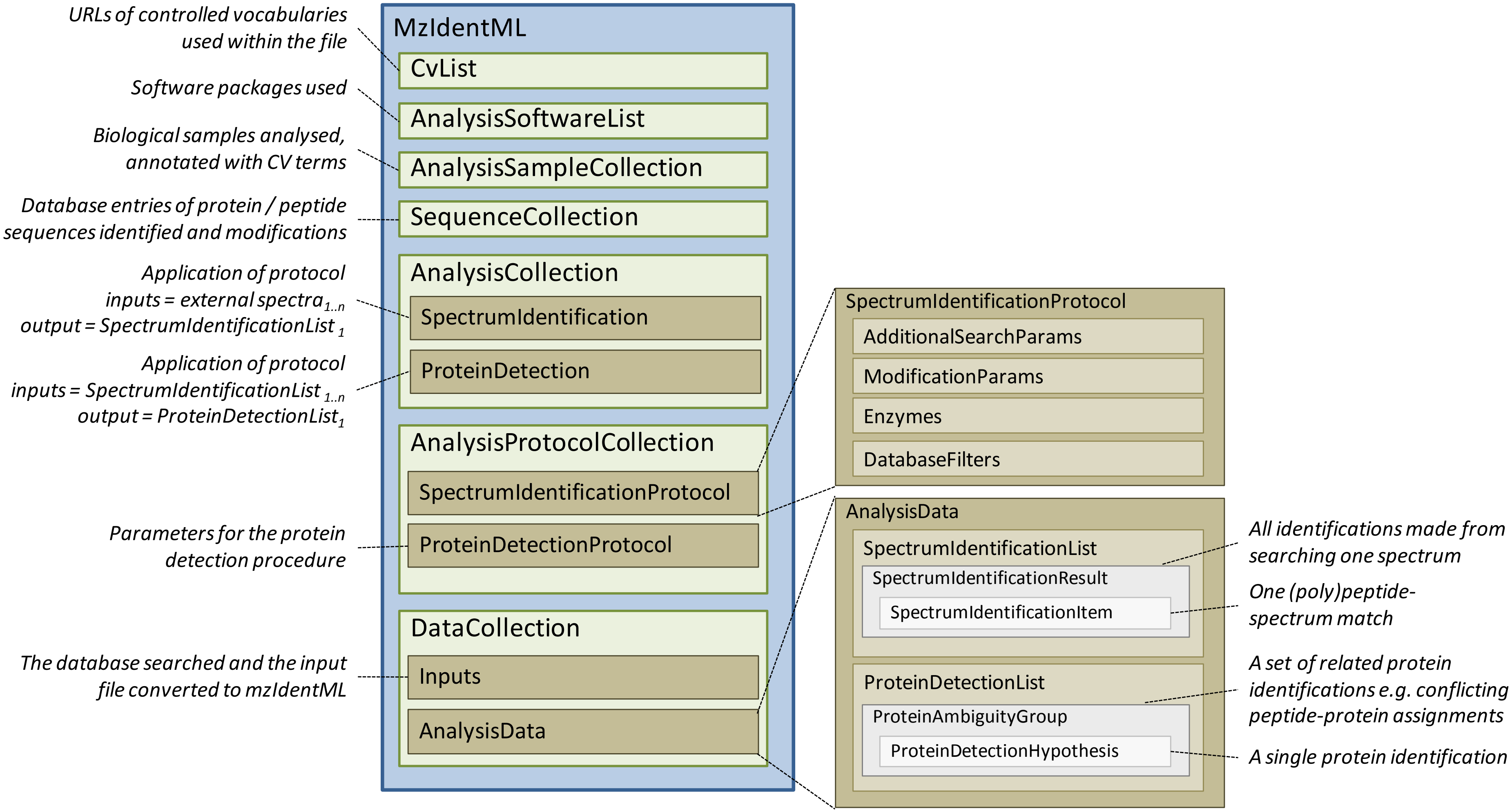
MIF - decribes the molecular interaction data exchange format.

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

Standardization - mzML



Standardization - mzIdentML



MzIdentML

CvList

AnalysisSoftwareList

AnalysisSampleCollection

SequenceCollection

AnalysisCollection

SpectrumIdentification

ProteinDetection

AnalysisProtocolCollection

SpectrumIdentificationProtocol

ProteinDetectionProtocol

DataCollection

Inputs

AnalysisData

URLs of controlled vocabularies used within the file

Software packages used

Biological samples analysed, annotated with CV terms

Database entries of protein / peptide sequences identified and modifications

Application of protocol inputs = external spectra$_{1..n}$ output = SpectrumIdentificationList $_{1}$

Application of protocol inputs = SpectrumIdentificationList $_{1..n}$ output = ProteinDetectionList$_{1}$

Parameters for the protein detection procedure

The database searched and the input file converted to mzIdentML

SpectrumIdentificationProtocol

AdditionalSearchParams

ModificationParams

Enzymes

DatabaseFilters

AnalysisData

SpectrumIdentificationList

SpectrumIdentificationResult

SpectrumIdentificationItem

ProteinDetectionList

ProteinAmbiguityGroup

ProteinDetectionHypothesis

All identifications made from searching one spectrum

One (poly)peptide-spectrum match

A set of related protein identifications e.g. conflicting peptide-protein assignments

A single protein identification

Proteomics Informatics – Databases, data repositories and standardization (Week 8)