Databases and repositories in proteomics

Christine Carapito
ccarapito@unistra.fr

Laboratoire de Spectrométrie de Masse BioOrganique
IPHC UMR7178
CNRS / Université de Strasbourg

« Proteomics & Metabolomics data analysis workshop »
28th March 2018

The « chance » of proteomics

Genome —Transcription→ **Transcriptome** —Translation→ **Proteome**

DNA — RNA — Proteins

4 letters code
A, T, G, C

→ **The genetic code** →

20 letters code
20 amino acids

1er nucléotide	2e nucléotide T	2e nucléotide C	2e nucléotide A	2e nucléotide G	3e nucléotide
T	TTT phénylalanine	TCT sérine	TAT tyrosine	TGT cystéine	T
	TTC phénylalanine	TCC sérine	TAC tyrosine	TGC cystéine	C
	TTA leucine	TCA sérine	TAA codon-stop	TGA codon-stop	A
	TTG leucine	TCG sérine	TAG codon-stop	TGG tryptophane	G
C	CTT leucine	CCT proline	CAT histidine	CGT arginine	T
	CTC leucine	CCC proline	CAC histidine	CGC arginine	C
	CTA leucine	CCA proline	CAA glutamine	CGA arginine	A
	CTG leucine	CCG proline	CAG glutamine	CGG arginine	G
A	ATT isoleucine	ACT thréonine	AAT asparagine	AGT sérine	T
	ATC isoleucine	ACC thréonine	AAC asparagine	AGC sérine	C
	ATA isoleucine	ACA thréonine	AAA lysine	AGA arginine	A
	ATG méthionine	ACG thréonine	AAG lysine	AGG arginine	G
G	GTT valine	GCT alanine	GAT acide aspartique	GGT glycine	T
	GTC valine	GCC alanine	GAC acide aspartique	GGC glycine	C
	GTA valine	GCA alanine	GAA acide glutamique	GGA glycine	A
	GTG valine	GCG alanine	GAG acide glutamique	GGG glycine	G

The « Next Generation Sequencing » revolution

First genome sequenced in 1995:
Haemophilus influenza (1,8.10^6 bps)

⇒

Yeast genome in 1996:
Saccharomyces cerevisiae (14.10^6 bps)

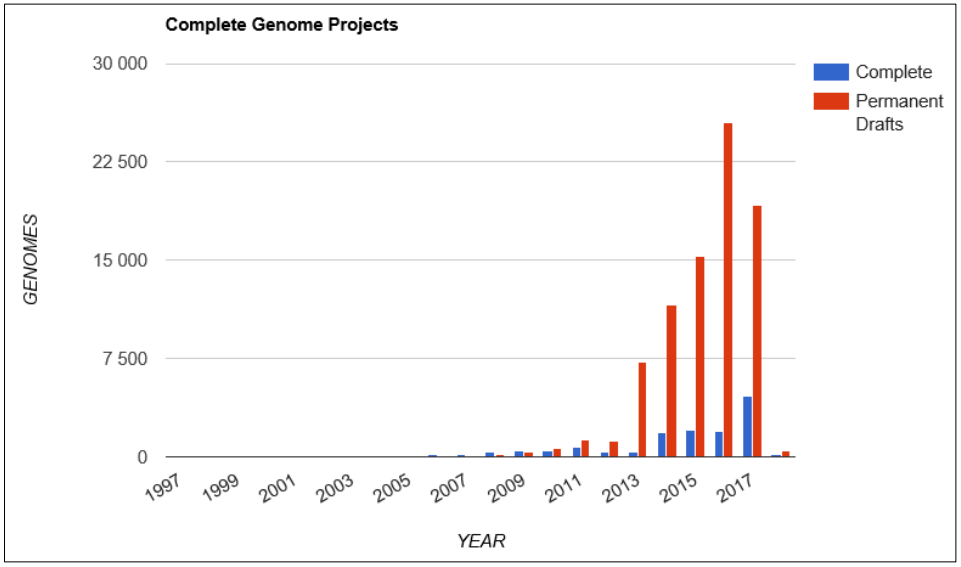
⇒

First draft of the human genome in 2001:
Homo sapiens (3,2.10^9 bps)

Nature, 2001

Science, 2001

Genomes Online website : *http://genomesonline.org*



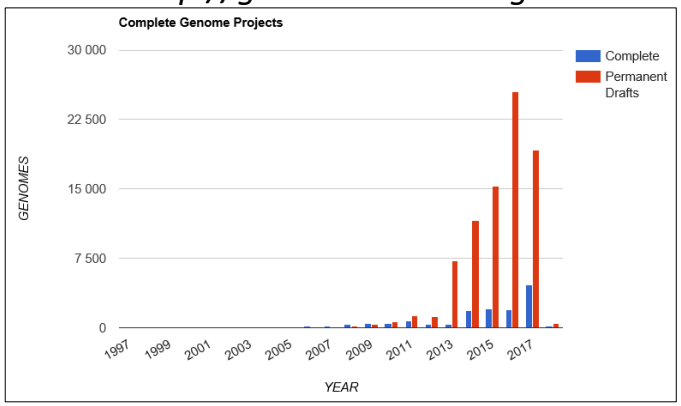
Today:
194 937 ongoing sequencing projects

Large-scale sequencing projects:

- Million Veteran Program (MVP)
https://www.research.va.gov/mvp/
- Human: « 100 000 Genomes Project »
https://www.genomicsengland.co.uk/50000-genomes-landmark/
- iHMP, NIH Integrative Human Microbiome Project (iHMP)
https://www.hmpdacc.org/ihmp/

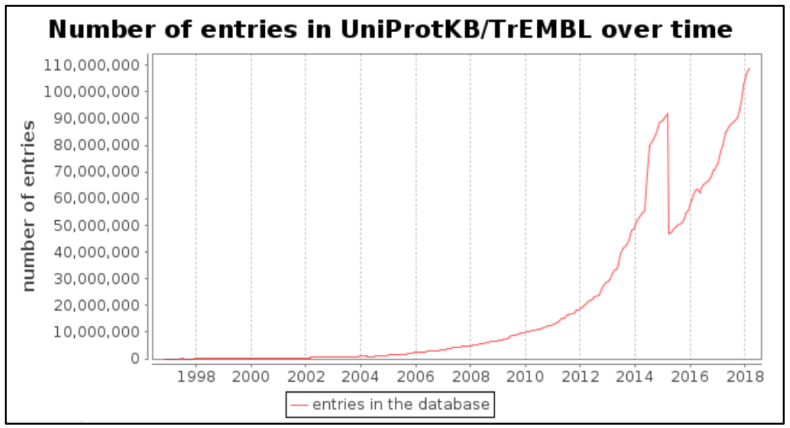
Genome sequences are translated to protein sequences

Genomes Online website :
http://genomesonline.org



***In silico* translation of coding sequences**

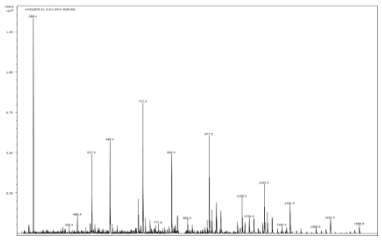
Protein sequence databases



25/03/2018
from http://www.uniprot.org/statistics/TrEMBL

The central role of the database used to interpret proteomics data

MS/MS spectrum



Experimental mass lists

MS	MS/MS
546,45	789,67 876,43 999,12 1018,98 1342,34 1597,09 1678,95 2202,22

Identification algorithms

Mass list comparisons

Protein sequence database

```
>Albumin
ILPMVCCDEKTISHEDAVFRPMLVCKHF
DIYTREHPKSQDCWATTPMKFHLKETI
PRHVVCDETR...
```

Theoretical mass lists

MS	MS/MS
546,43	789,69 876,41 987,50 999,14 1018,97 1342,30 1597,11 1678,99 1987,60 2202,24

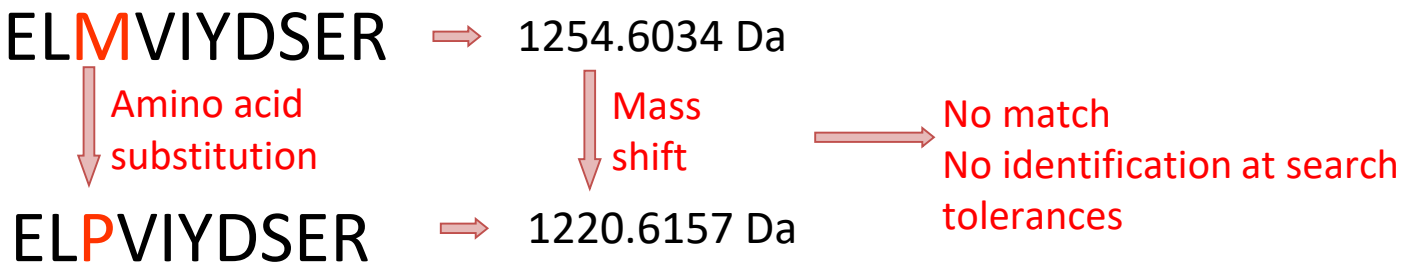
More than 40 different database search algorithms (open-source and proprietary)

Name	Year of publication	website
SEQUEST	1994	fields.scripps.edu/sequest
Mascot	1999	matrixscience.com
OMSSA 2004	2004	ftp.ncbi.nlm.nih.gov/pub/lewisg/omssa
X! Tandem	2004	thegpm.org/TANDEM
Andromeda	2011	maxquant.org
PeaksDB	2011	bioinfor.com/peaks/features/peaksdb.html
MS-GF+	2014	proteomics.ucsd.edu/software-tools/ms-gf

The central role of the database used to interpret proteomics data

The choice of the database used to search MS/MS data against is crucial!

- **Non-error tolerant** searches prevent the identification of peptides absent from the database or containing amino acid substitutions.



- Redundancy is deleterious for the results, especially for quantitative interpretation, as **unique, non-shared, proteotypic** peptide sequences are required.

- The database choice needs to be **adapted according to the taxonomy of interest**, the level of knowledge of its genome, the quality of its annotation and curation.

Protein sequence databases

Different types of protein sequence databases with variable levels of annotation and curation

- Generalist databases, resulting from the automated annotation of sequenced genomes

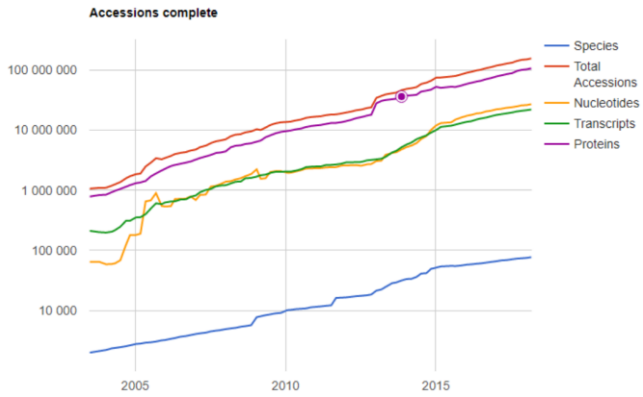
International Nucleotide Sequence Database Collaboration (INSDC) for data exchange between the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena), GenBank nucleotide sequence database (GenBank, https://www.ncbi.nlm.nih.gov/genbank/) and the DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp/)

NCBI reference sequence database, RefSeq

https://www.ncbi.nlm.nih.gov/refseq/

(O'Leary NA, et al., NAR 2016, 44(D1):D733-45)

➡ Derived from the sequence data available in the redundant archival database GenBank



Universal Protein resource, UniProt Knowledgebase, UniProtKB

http://www.uniprot.org/

(The UniProt Consortium, NAR 2017, 45 (D1), D158–D169)

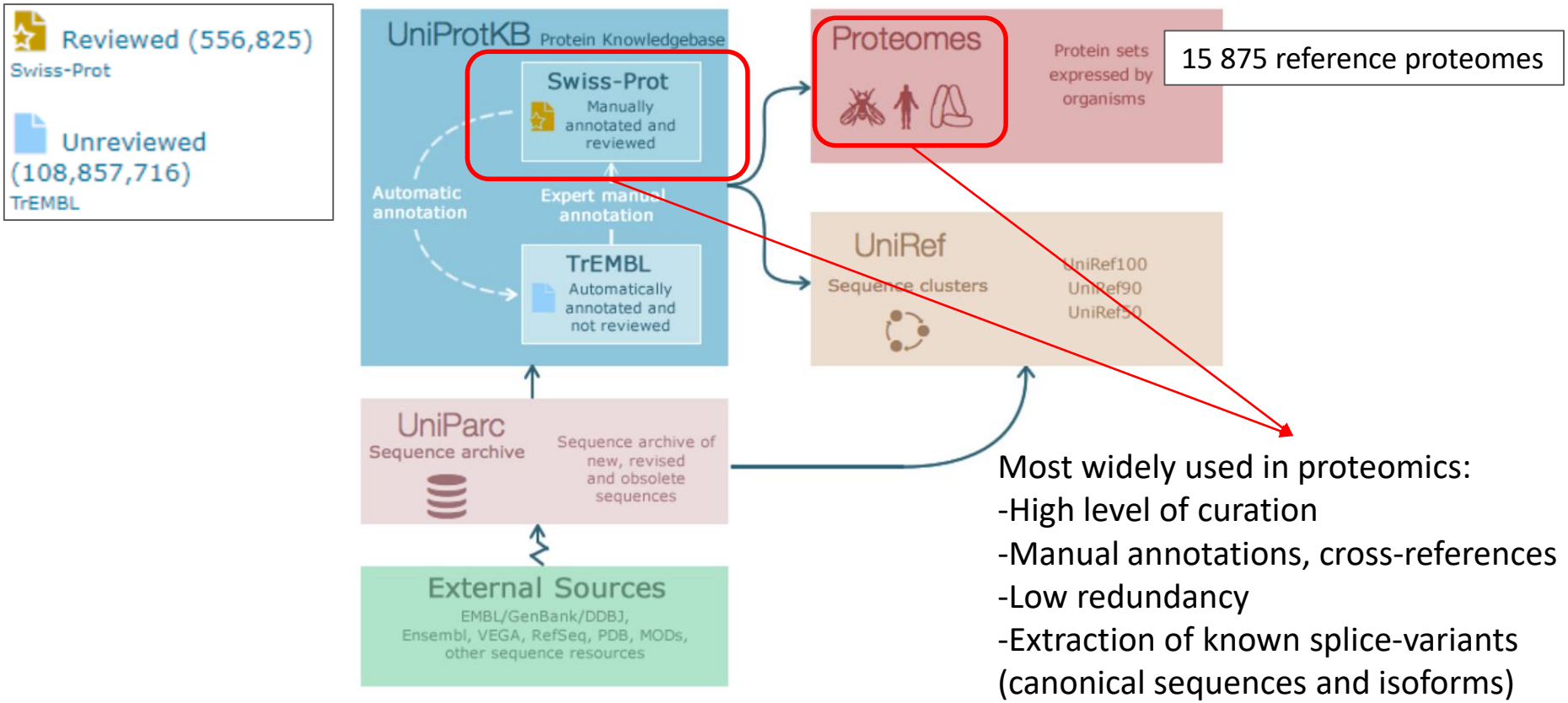
➡ The UniProt Consortium consists of research teams from the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR)

Protein sequence databases

Different types of protein sequence databases with variable levels of annotation and curation

- Generalist databases, resulting from the automated annotation of sequenced genomes

Universal Protein resource, UniProt Knowledgebase, UniProtKB
http://www.uniprot.org/
(The UniProt Consortium, NAR 2017, 45 (D1), D158–D169)



Most widely used in proteomics:
-High level of curation
-Manual annotations, cross-references
-Low redundancy
-Extraction of known splice-variants (canonical sequences and isoforms)

Protein sequence databases

Different types of protein sequence databases with variable levels of annotation and curation

- Taxonomy-specific databases, benefiting from organism's experts annotations
  - FlyBase dedicated to *Drosophila melanogaster* *(The FlyBase Consortium, NAR 2017, 45(D1):D663-D671)*
  - The Mouse Genome Database (MGD) dedicated to *Mus musculus* *(Mouse Genome Database Group, NAR 2017, 45(D1):D723-D729)*
  - The TAIR database dedicated to *Arabidopsis thaliana* *(The Arabidopsis Information Resource, Genesis 2015, 53: 474-485)*

➡ Curated organism-specific post-translational modification databases

- PhosphoGrid database of experimentally verified *in vivo* protein phosphorylation sites of Sa*ccharomyces cerevisiae, https://phosphogrid.org/*
- The human DEPhOsphorylation Database (DEPOD), *http://depod.bioss.uni-freiburg.de/*
- The Arabidopsis Protein Phosphorylation Site Database (PhosPhAt), *http://phosphat.uni-hohenheim.de*

Protein sequence databases

Different types of protein sequence databases with variable levels of annotation and curation

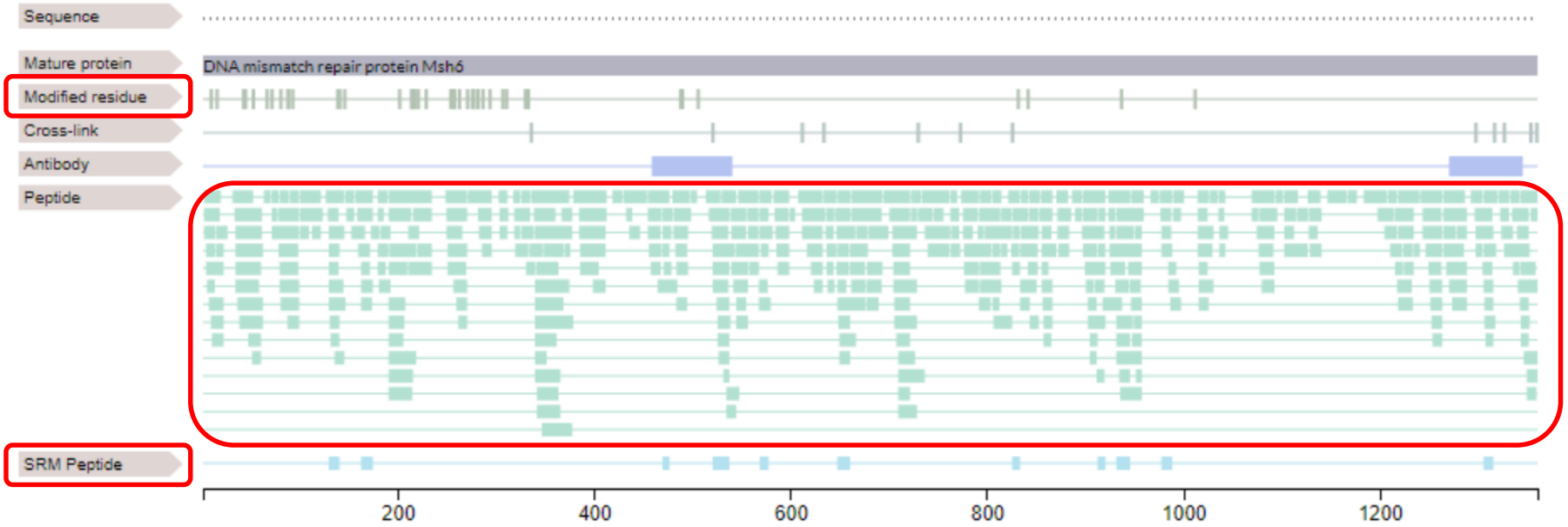
- Taxonomy-specific databases, benefiting from organism's experts annotations

neXtprot dedicated to the human proteome

Gaudet P, et al., NAR 2017, 45(D1), D177–D182

neXtProt integrates high-quality and manually curated UniProt/Swiss-Prot entries with large amount of additional information from other resources such as Human Protein Atlas, ArrayExpress, UniGene, PeptideAtlas, Gene Ontology Annotation, Ensembl, dbSNP, …

- Reference proteome for the *Human Proteome Project HPP*
- Numerous very useful proteomists-oriented information and tools
- Peptide unicity checker, including more than 5millions sequence variants



Protein sequence databases

Different types of protein sequence databases with variable levels of annotation and curation

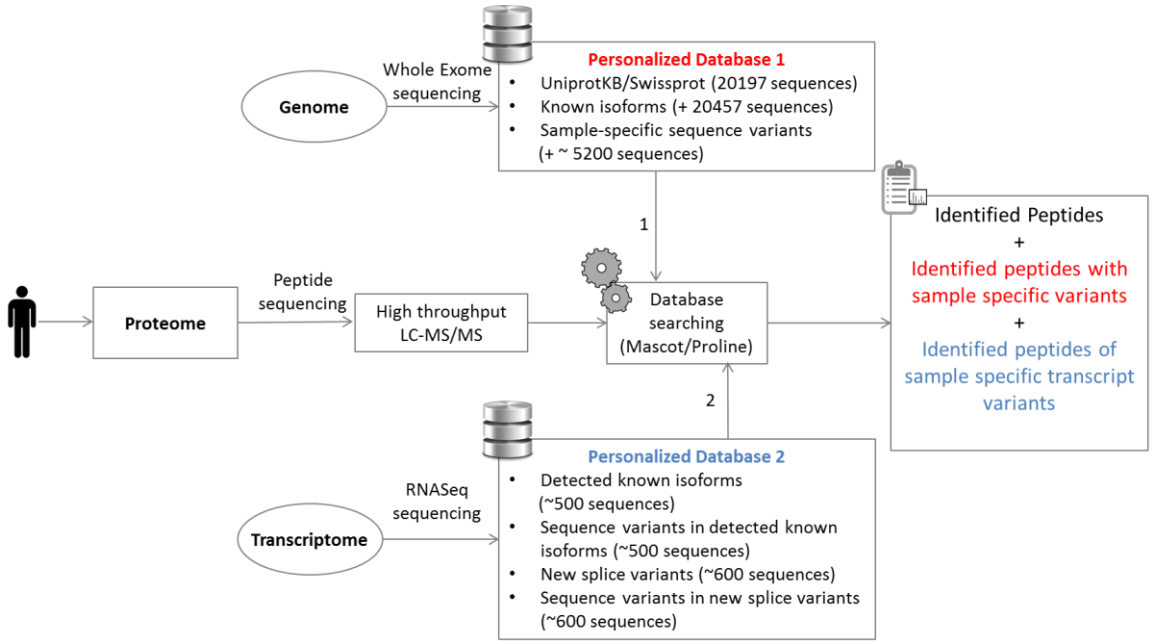
- Towards personalized databases for personnalized medecine (multi-omics approaches)

  - Taking into account disease-specific mutations identified by NGS

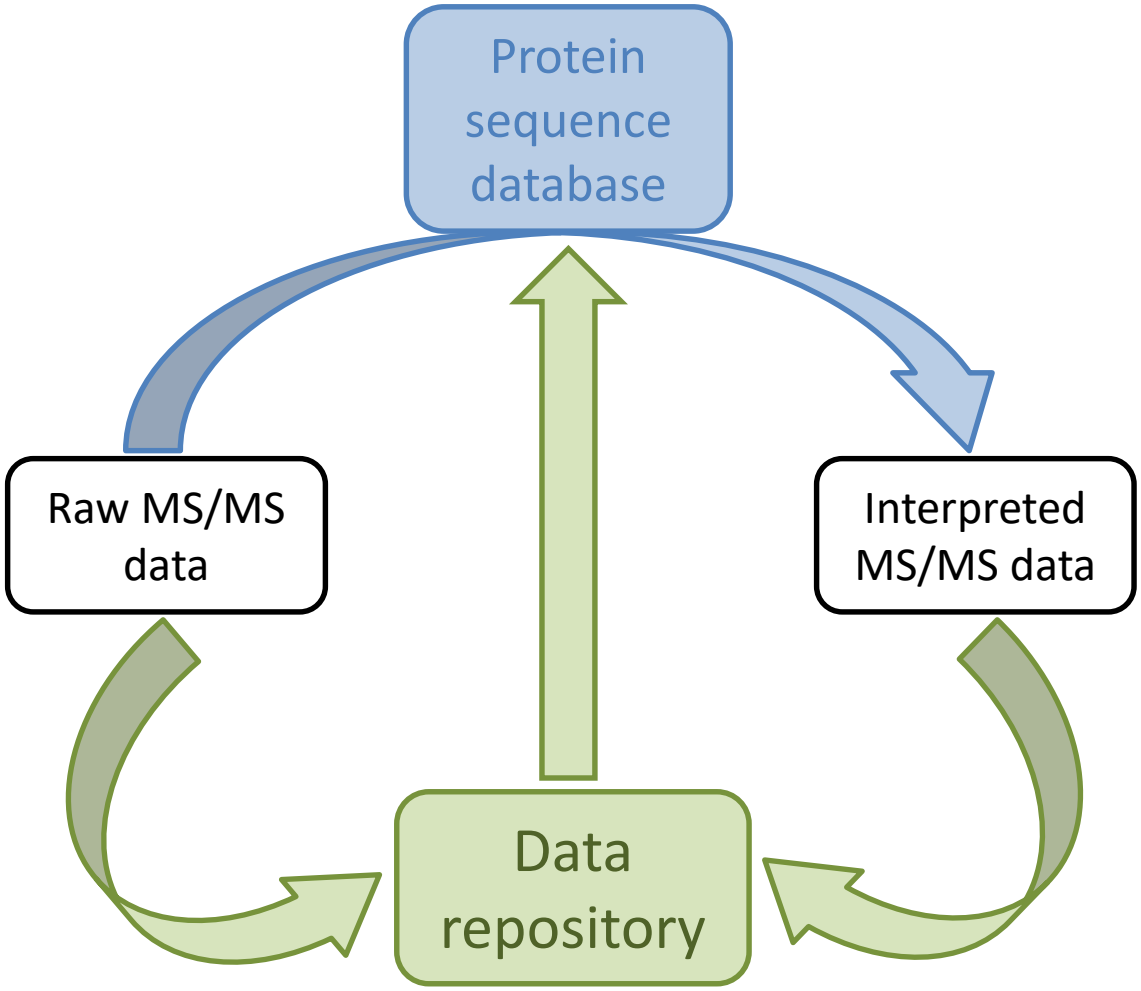
    COSMIC, the Catalogue Of Somatic Mutations In Cancer (*http://cancer.sanger.ac.uk/cosmic*), a comprehensive resource for exploring the impact of somatic mutations in human cancer integrated in neXtprot

  - Ultimate goal: generate one database for each individual sample

Multi-omics data acquisition on the **same samples** allows the development of **proteogenomics** database search strategies



Proteomics data repositories



The protein database is the starting point and should be the ultimate end point.

The reference protein database should include all valuable information and links to experimental evidence of protein expression, abundance levels, identified isoforms and PTMs, interactions with other proteins, …

Proteomics data repositories

The move to a general opinion of data sharing in proteomics was slow!

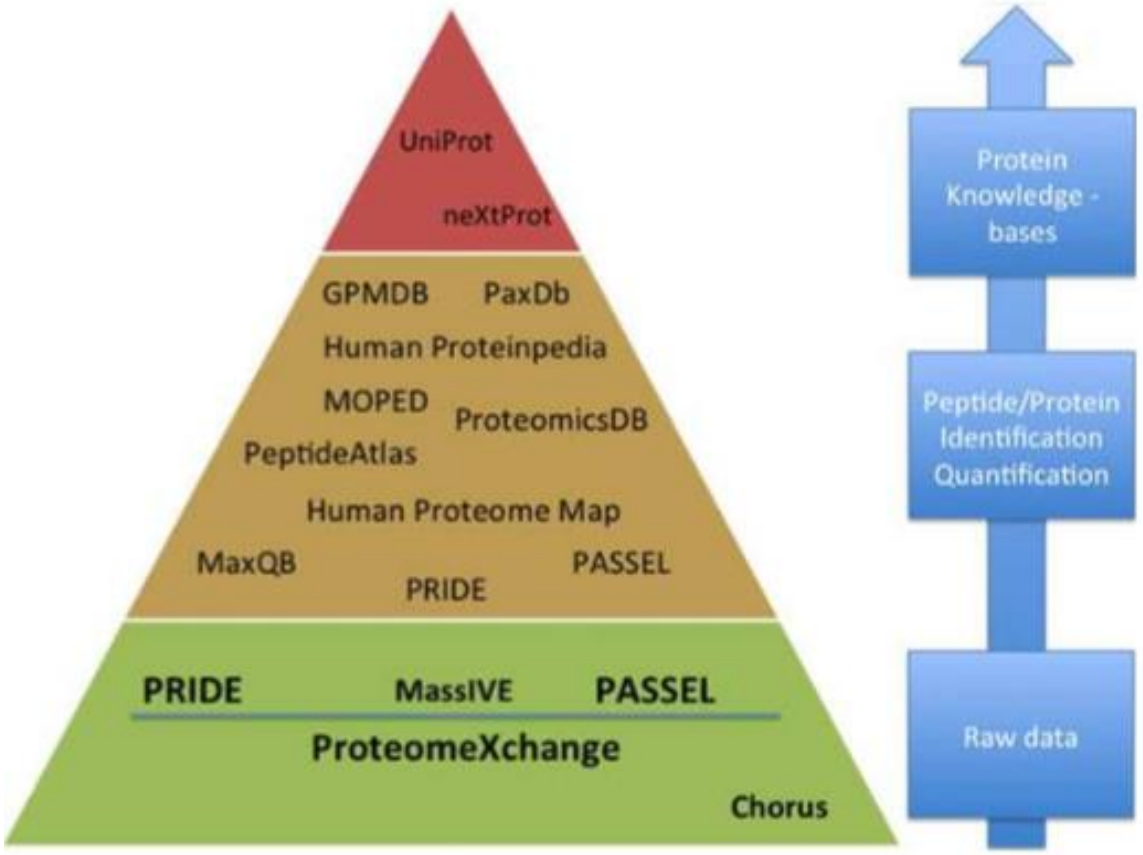
Proteomics data are often more complex than NGS data due to complex sample preparations, a wide variety of analytical approaches, bioinformatics tools and pipelines and related statistical analysis.

Two driving forces have enabled the move:

- **Requirements** promoted by leading scientific journals and funding agencies (MCP guidelines started in 2004-2005).

- Development of **data standard formats** (mzML, mzIdentML, mzQuantML, TraML, …) and open source tools facilitating data deposition, reuse and comparative analyses.
  Designed and maintained thanks to the Proteomics Standards Initiative PSI (*http://www.psidev.info/*) funded at the Washington HUPO meeting in April 2002.

Hierarchy of proteomics data repositories



Interpreted data
Re-interpretation of the data

From Perez-Riverol Y., et al., Proteomics 2015, 15, 930–949

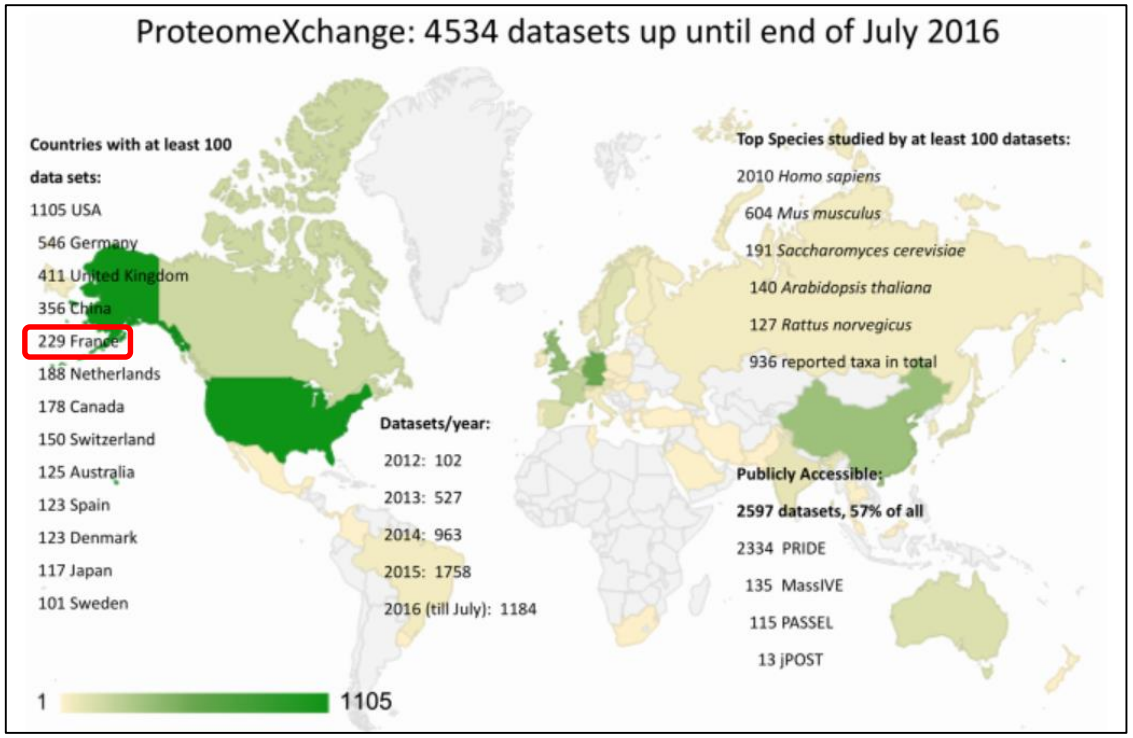
Proteomics data repositories

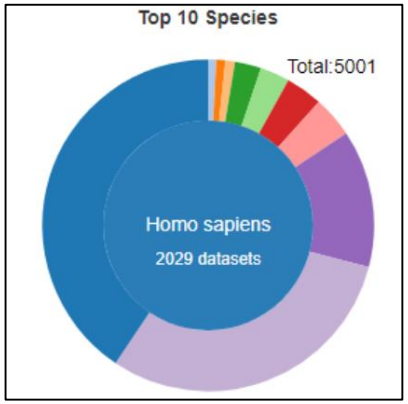
Success of the ProteomeXchange Consortium:

It has been set up to provide a globally coordinated submission of mass spectrometry proteomics data to the main existing proteomics repositories, and to encourage optimal data dissemination.

(Vizcaino, A. J., et al., 2014, Nature Biotechnology 32, 223–226)

Prevent break down due to lack of funding (Tranche and Peptidome)





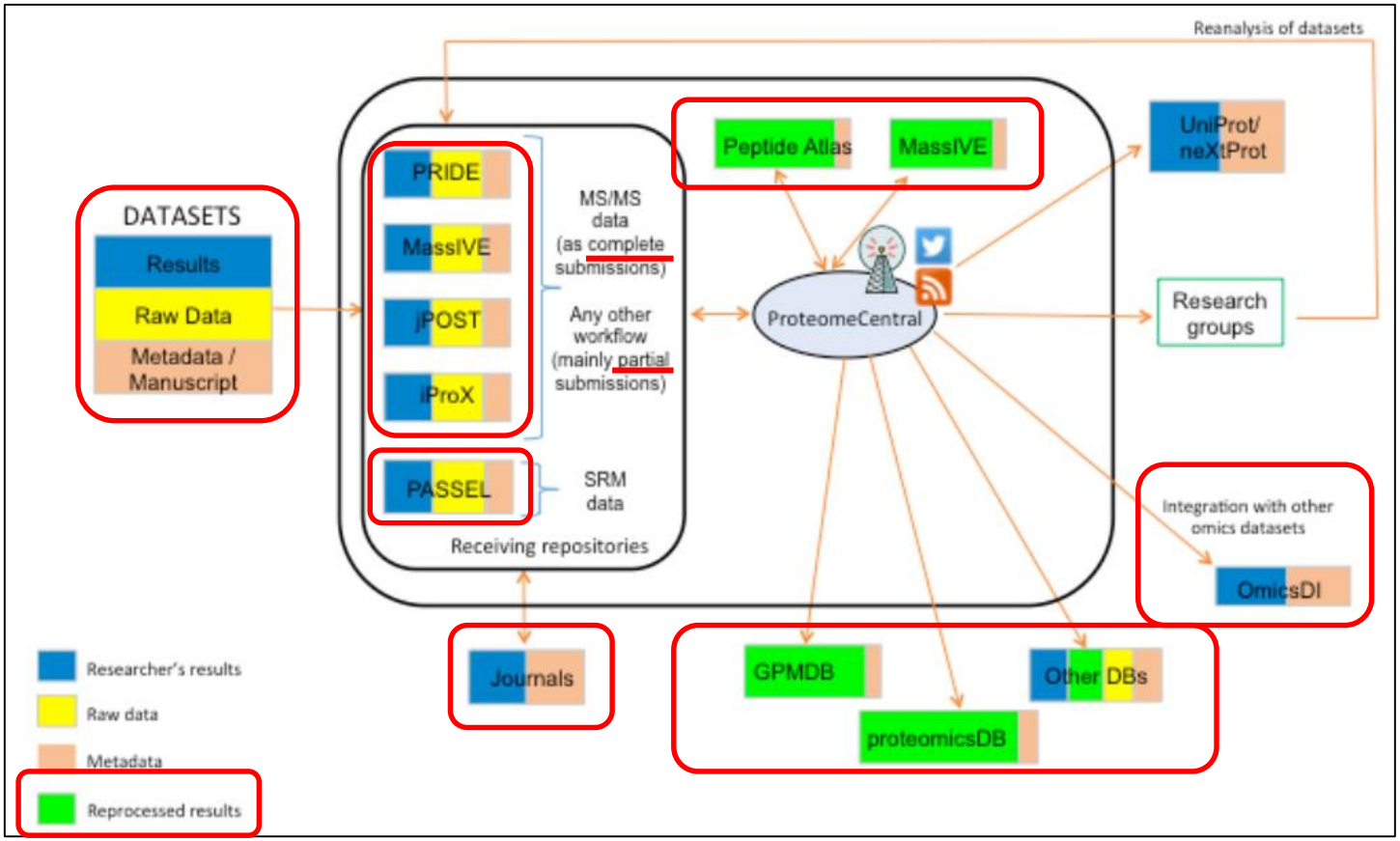
27/03/2018

Deutsch, E., et al., The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition

Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D1100–D1106

Proteomics data repositories

- Three data levels: Raw data, Results (MzIdentML), Metadata
- Different entry points (global and targeted datasets)
- Two submission modes : Complete and Partial
- A unique universal identifier PXD with DOI (link in publications)
- Vizualisation tools (Pride Inspector, PASSEL Browser, PeptideAtlas, …)
- Data reprocessing and reuse by any external tool
- Link to other –omics datasets



From Deutsch, E., et al., NAR, 2017, 45 (D1), D1100–D1106.

Proteomics data repositories

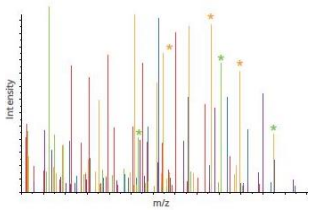
- Three data levels: Raw data, Results (MzIdentML), Metadata
- Different entry points (global and targeted datasets)
- Two submission modes : Complete and Partial
- A unique universal identifier PXD with DOI (link in publications)
- Vizualisation tools (Pride Inspector, PASSEL Browser, PeptideAtlas, …)
- Data reprocessing and reuse by any external tool
- Link to other –omics datasets

Crucial ressource enabling the development and support of new strategies

(*Griss, J., et al., Nat Methods. 2016 13(8): 651–656*: 256 million spectra, 190 million unidentified and 66 million identified spectra)

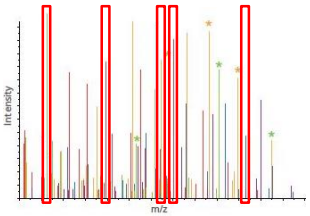
- Spectral library searches
- Data Independent Acquisition (DIA) approach

Data analysis *(peptide centric)*



Spectral library

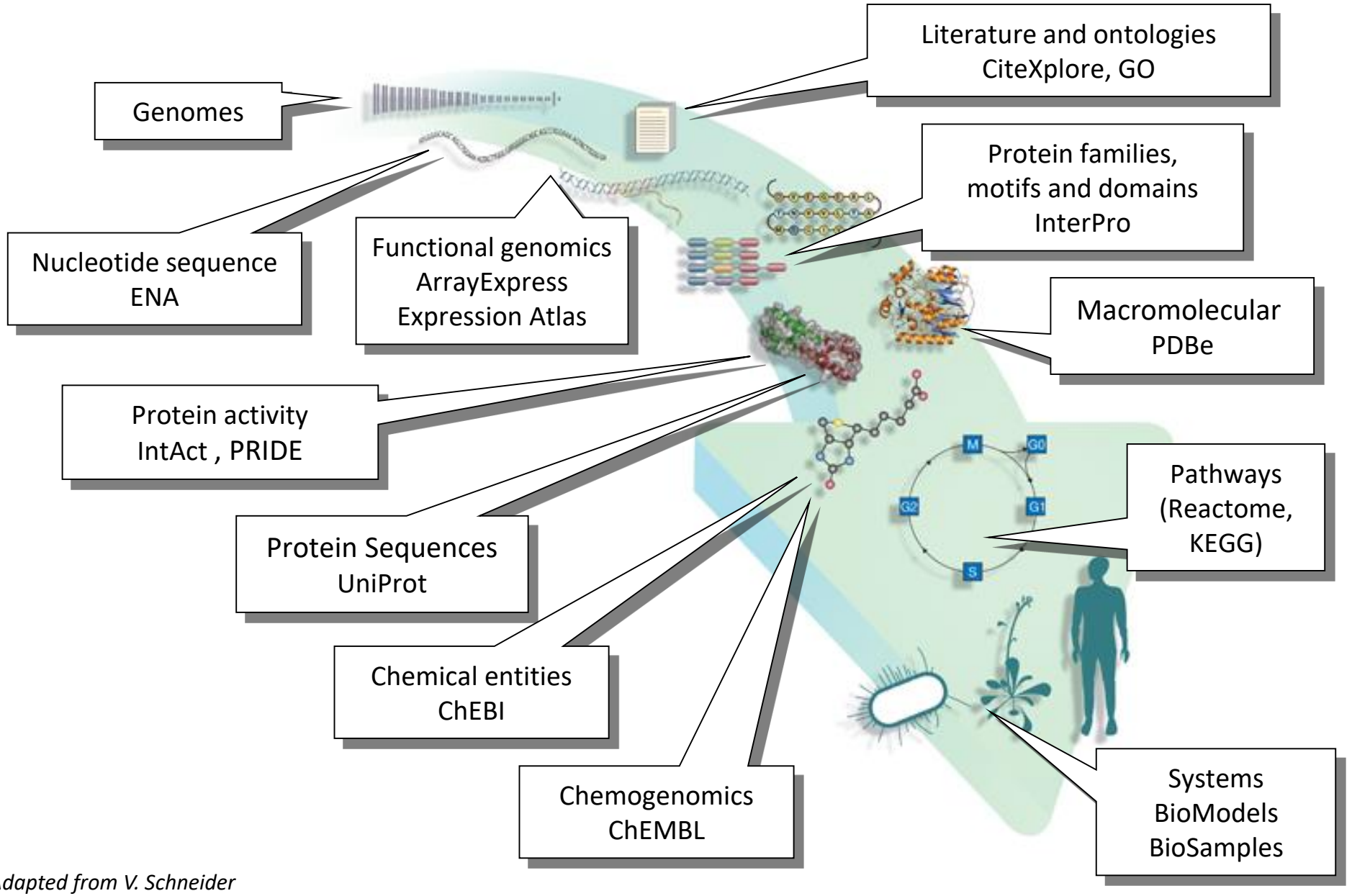
Peptides and fragments m/z, retention time, fragment relative intensity



Targeted data extraction

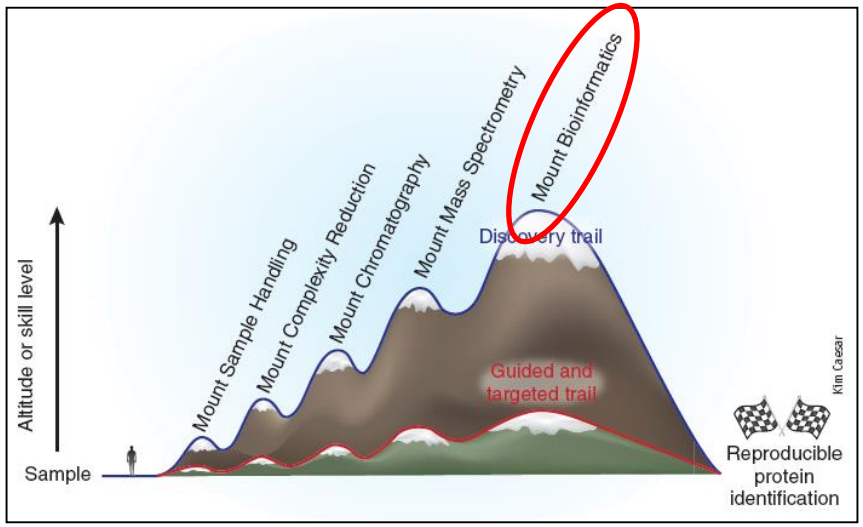
Protein quantification

Databases and tools are required at every step



Adapted from V. Schneider

Conclusion, perspectives, …



*From R. Aebersold, Nature Methods, 6 (6), 411-412, June **2009**.*

Constantly growing computational proteomics / –omics community, visibility and dedicated sessions in all conference programs

Needs are big and various:
-Informaticians for Big Data handling, storing and archiving
-Software developers
-Mathematicians, statisticians for new algorithms, …

… and thanks!

Alexandre Burel, Patrick Guterl, Fabrice Varrier

Jérôme Pansanel, Yannick Patois (IPHC, Strasbourg)
Myriam et Frédéric Bertrand (IRMA, Strasbourg)

WP2 de ProFI: Christophe Bruley
David Bouyssie, Véronique Dupierris, Alexandre Burel, Aymen Romdhani, Jean-Philippe, Julie Poissat, Alexandre Walter, Ibrahim Yapici, Laurent, Yves Vandenbrouck, Thomas Burger et co.
Mais aussi: Emmanuelle Mouton, Anne-Marie Hesse, Magali Rompais