Standards for omics data – A personal overview

June 18th 2020

X-Omics workshop on data integration and standards

Juan Antonio Vizcaíno

EMBL-European Bioinformatics Institute (EMBL-EBI)

Hinxton, Cambridge

EMBL-EBI

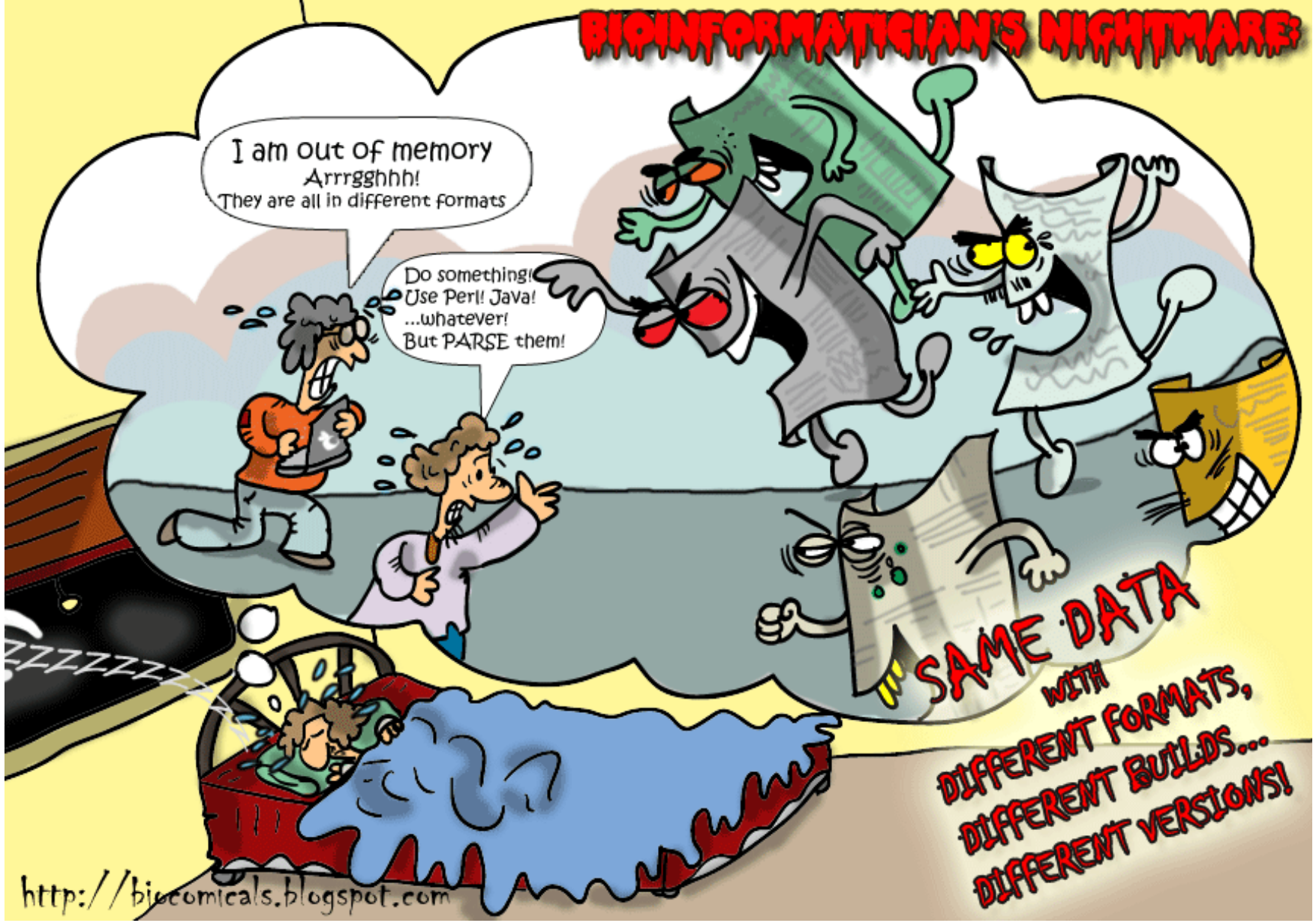
Overview

- A couple of slides about the need of data standards
- Proteomics data standards as an example: The Proteomics Standards Initiative and ProteomeXchange
- DNA/RNA Sequencing standards: Introduction to GAG4H standards
- Data integration using data standards

Data standards are needed

Standards are needed in everyday life: also in bioinformatics…

With a small number
of standards,
converters are feasible



Taken from Biocomicals, http://biocomicals.blogspot.com

The typical dilemma

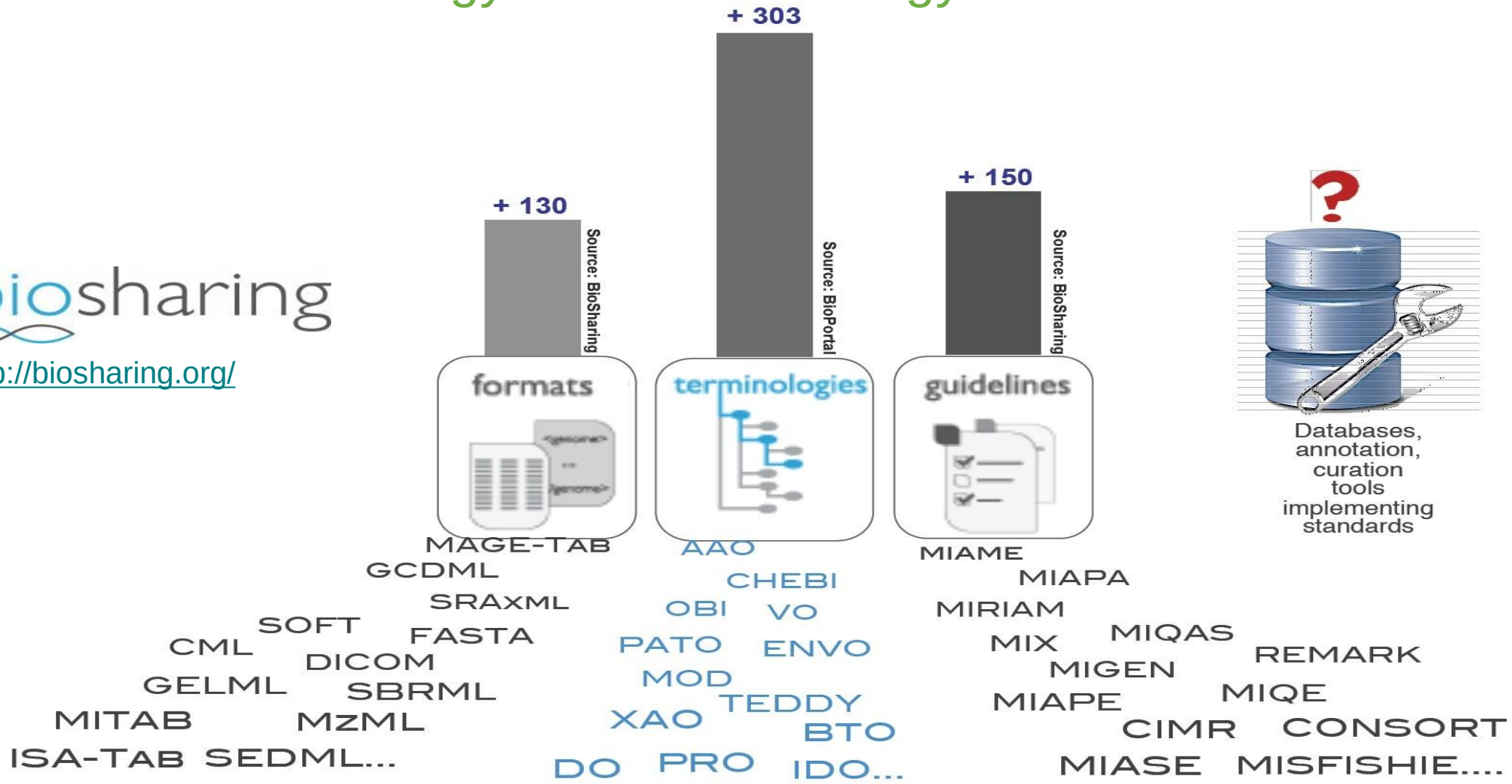
- Data standards **need to be stable** to promote adoption

- Very often data standards for **omics data need to evolve very rapidly:**
  - Data is inherently **very complex**
  - Experimental techniques **are evolving** all the time

Data standards in biology and biotechnology

biosharing

http://biosharing.org/

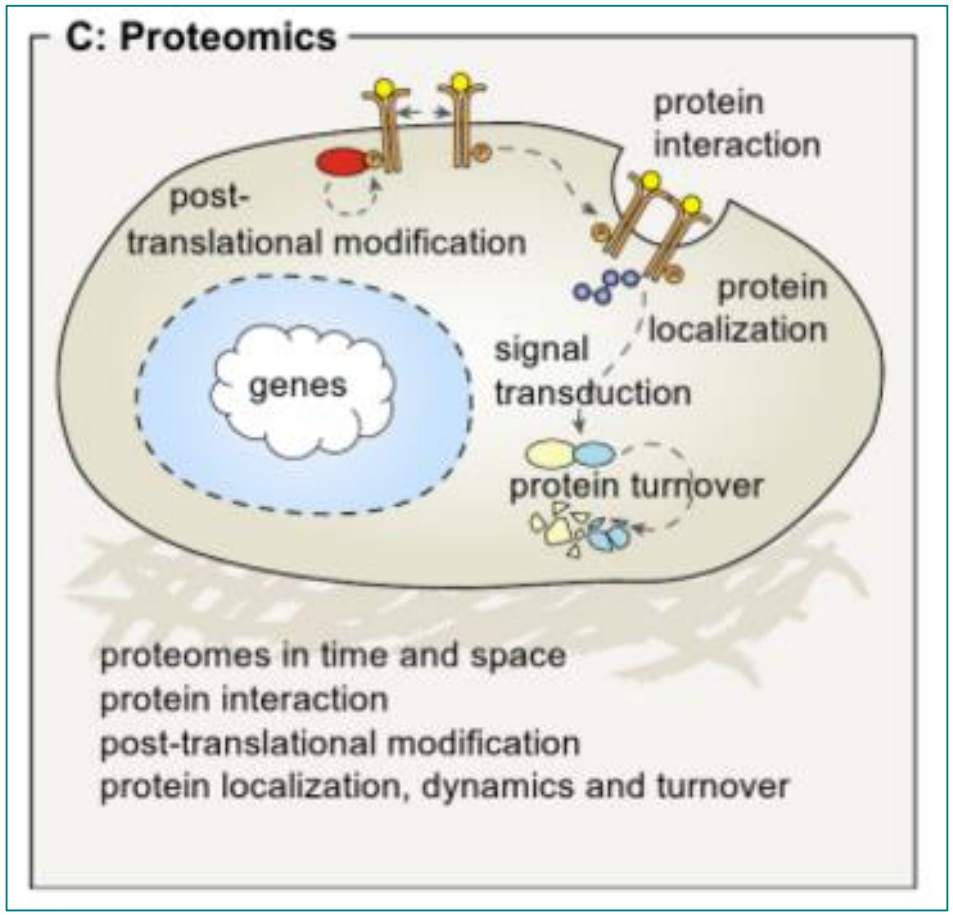


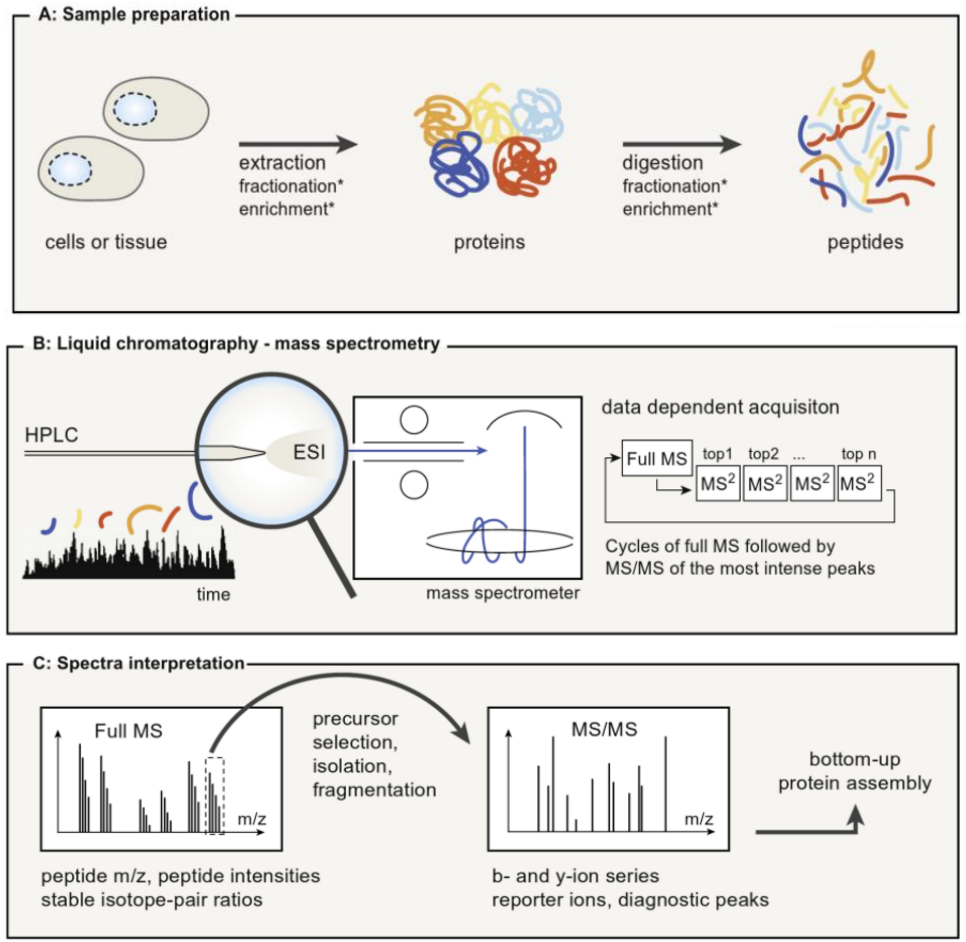
Source: Susanna-Assunta Sansone (University of Oxford, UK)

Overview

- A couple of slides about the need of data standards
- **Proteomics standards: The Proteomics Standards Initiative and ProteomeXchange**
- DNA/RNA Sequencing standards: Introduction to GAG4H standards
- Data integration using data standards

One slide intro to MS based proteomics





Hein *et al.*, *Handbook of Systems Biology*, 2012

HUPO Proteomics Standards Initiative

- Develops data standards for proteomics.
- Both **data *representation*** and ***annotation*** standards.
- Involves data producers, database providers, software producers, publishers, everyone who wants to be involved…
- Active Workgroups: MI, MS, PI, Mod and the new QC.
- Inter-group activities: MIAPE and **Controlled Vocabularies**.
- **Started in 2002**, so some experience already…
- **One annual meeting** in March-April, **regular phone calls**.
- Close interaction with the **metabolomics community (MSI)**.

http://www.psidev.info

PSI Deliverables – As an example

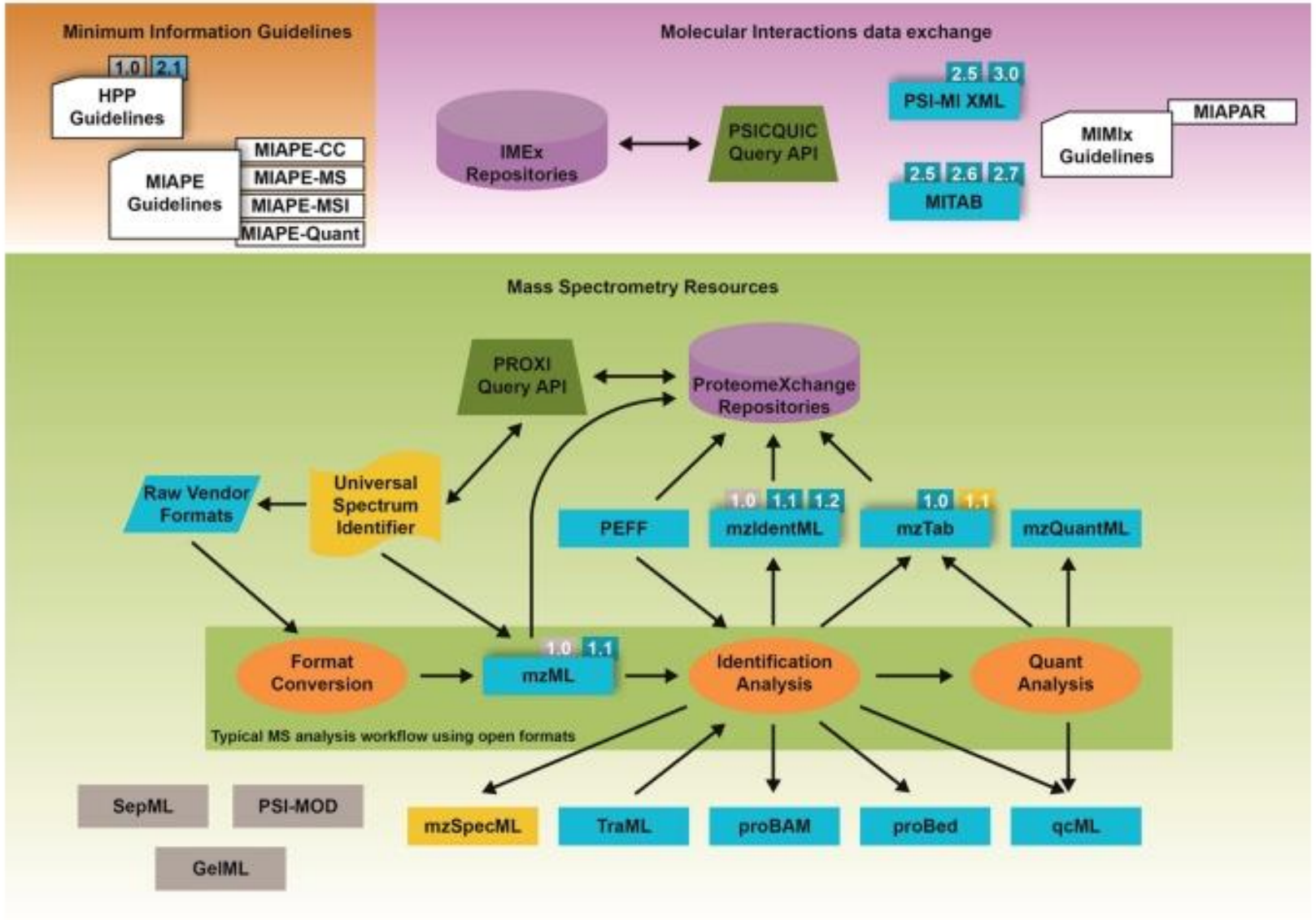
- Formats: Usually **XML-based** (but also **tab-delimited files**), capable of representing the relevant Minimum Information, plus additional detailed data for the domain.

- Controlled vocabularies: Usually an OBO-style hierarchical controlled vocabulary precisely defining the **metadata** that are encoded in the formats.

- Databases and Tools: **Foster open software implementations** to make the standards truly useful.

- Community interaction to ensure **adoption of the standards** and **public deposition** of data in proteomics repositories.

Summary slide



Deutsch *et al.*, *JPR*, *2017*

PSI Deliverables – As an example

- Formats: Usually **XML-based** (but also **tab-delimited files**), capable of representing the relevant Minimum Information, plus additional detailed data for the domain.

- Controlled vocabularies: Usually an OBO-style hierarchical controlled vocabulary precisely defining the **metadata** that are encoded in the formats.

- Databases and Tools: **Foster open software implementations** to make the standards truly useful.

- Community interaction to ensure **adoption of the standards** and **public deposition** of data in proteomics repositories.

Current PSI Standard File Formats for MS

MS data	• **mzML**
Identification	• mzIdentML
Quantitation	• mzQuantML
Final Results	• mzTab
SRM	• TraML

Data formats for mass spectra data

Binary data

Agilent Technologies

Waters
THE SCIENCE OF WHAT'S POSSIBLE.™

SCIEX

ThermoFisher SCIENTIFIC

XML-based open files

mzData

mzXML

mzML

Peak lists

.dta, .pkl, .mgf, .ms2

An example of success story: mzML

- A data format for the **storage and exchange of MS output files**
  - Designed by merging the best aspects of both mzData and mzXML
  - Developed with full participation of academic researchers, hardware and software vendors
  - Expected to replace mzXML and mzData, but not expected to completely replace vendor binary formats
  - Captures spectra (raw data or peak lists), chromatograms and related metadata

  - Version 1.0 released in June 2008, **v1.1** released in June 2009
  - **Many implementations already exist**
  - Version 1.2 with enhanced compression considered for the near future.
  - Also used for **MS metabolomics data**.

Martens *et al.*, MCP, 2011

An example of success story: mzML

Product	Source	Contact	Support comments
ProteoWizard	USC	Parag Mallick	Full mzML support today
TPP	ISB	Eric Deutsch	Full mzML support today (including embedded X!Tandem)
Insilicos Viewer	Insilicos	Erik Nilsson	Full mzML support today
X!Tandem	GPM	Ron Beavis	Full mzML support today
Myrimatch	Vanderbilt	Matt Chambers	Full mzML support today
InSilicoSpectro	SIB	Alex Masselot	Full mzML support today
Proteios SE	Univ Lund	Fredrik Levander	Full mzML support today
NCBI C++ toolkit	NCBI	Douglas Slotta	available in next release
OpenMS/TOPP	Univ Tübingen	Marc Sturm	Full mzML support today

http://www.psidev.info/mzml_1_0_0

The most popular search engines support mzML

Many parser libraries available

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
Home » Bioconductor 3.0 » Software Packages » n
mzR

GitHub This repository Search
pymzml / pymzML

jmzml
java implementation of the PSI-MS mzML specification v1.1

proteowizard

Conversion from raw files into mzML and other formats

Current PSI Standard File Formats for MS

MS data	• mzML
Identification	• **mzIdentML**
Quantitation	• mzQuantML
Final Results	• mzTab
SRM	• TraML

mzIdentML -> Data standard for peptide and protein IDs

- XML-based data standard for peptide and protein identifications e.g. following database search and protein inference
- Sections for all PSMs, proteins/protein groups, protocols/parameters etc.

- **Timeline:**
  - Original 1.0 version in Aug 2009
  - Version 1.1 stable (Aug 2011); Original manuscript published in MCP in 2012*
  - Well supported in lots of open source and commercial software
  - Fully supported by ProteomeXchange resources
  - 2012 onwards (mzIdentML 1.2): extended use cases
    - Better support for protein grouping. 2017 mzIdentML 1.2 release; manuscript published at MCP**

* Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O., *et al.*, The mzIdentML data standard for mass spectrometry-based proteomics results. *Molecular & Cellular Proteomics* 2012, *11*, M111.014381.
** Vizcaíno, J. A., Mayer G., Perkins S., Barsnes H., *et al.,* The mzIdentML Data Standard Version 1.2, Supporting advances in Proteome Informatics. *Molecular & Cellular Proteomics* 2017, *16*, 1275-1285.

2011-2012

mzIdentML 1.1

Data standard for peptide and protein identification data

Increasingly supported by the most-used proteomics software and databases

MATRIX SCIENCE

OPENMS

ProteoAnnotator

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Scaffold

proteowizard

mzid Library

jmzIdentML

PEAKS® Complete Software for Proteomics

PIA

CRUX

MyriMatch

ProteinPilot

ms-data-core-api

CPTAC

National Cancer Institute at the National Institutes of Health

ProCon

ProteomeXchange

2017

mzIdentML 1.2

New support for:

- Cross-linking approaches
- Peptide level scores
- Modification localization scores
- Proteogenomics approaches

Improved support for:

- Protein inference
- Pre-fractionation
- *de novo* sequencing
- Spectral library searches

Current PSI Standard File Formats for MS

MS data	• mzML
Identification	• mzIdentML
Quantitation	• mzQuantML
Final Results	• **mzTab**
SRM	• TraML

mzTab – Aims and concept

- To provide a simple and efficient way of exchanging results from MS approaches.
  - **Simpler summary** report of the experimental results
  - Peptides and proteins identified in a given experimental setting
  - **Small molecules** identified
  - Reported **quantification** values
  - Technical and biological metadata

- **Easier to parse** and use by the research community, systems biologists as well as providers of knowledge bases.
- It can be used by non-experts in bioinformatics.
- It does not aim to replace mzIdentML and mzQuantML

mzTab - Sections

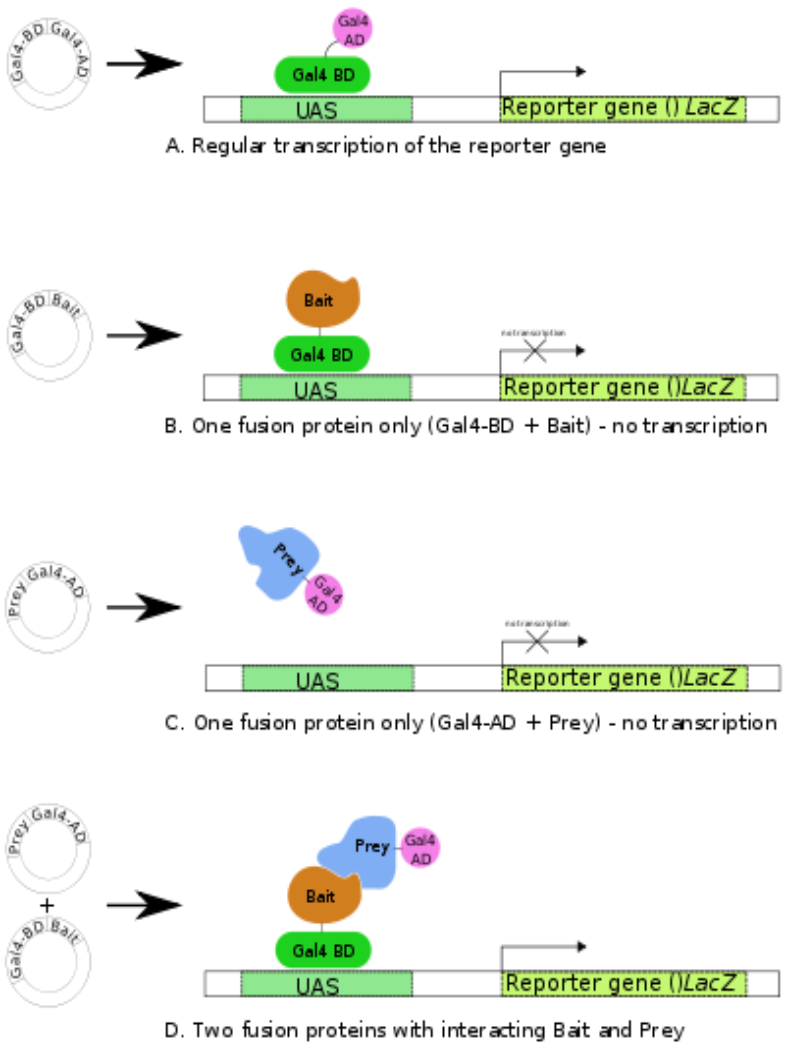
Metadata	• Basic information about experiment and sample • Key-Value pairs
Protein	• Basic information about protein identifications • Table-based
Peptide	• Information about quantified peptides • Table-based
PSM	• Information about identified spectra • Table-based
Small Molecule	• Basic information about identified small molecules • Table-based

Griss *et al.*, MCP, 2014

Metadata section - Example

```
MTD    mzTab-version                 1.0
MTD    mzTab-mode                    Complete
MTD    mzTab-type                    Identification
MTD    mzTab-ID                      PRIDE assay metadata example
MTD    title                         COFRADIC N-terminal proteome of unstimulated human
MTD    instrument[1]-name            [PRIDE, PRIDE:0000131, Instrument model, Micromass
MTD    instrument[1]-source          [MS, MS:1000008, Ionization Type, ESI]
MTD    instrument[1]-analyzer          [MS, MS:1000010, Analyzer Type, Quadrupole-TOF]
MTD    instrument[1]-detector        [MS, MS:1000026, Detector Type, MultiChannelPlate]
MTD    software[1]                    [MS, MS:1001456, analysis software, MassLynx v3.5]
MTD    protein_search_engine_score[1][MS, MS:1002367, probability for proteins,]
MTD    publication[1]                pubmed:16038019|pubmed:12665801|pubmed:16518876
MTD    contact[1]-name               Kristian Flikka
MTD    contact[1]-affiliation        Computational Biology Unit, University of Bergen
MTD    contact[1]-email              flikka@ii.uib.no
MTD    ms_run[1]-format              [MS, MS:1000564, PSI mzData file, ]
MTD    ms_run[1]-location  …
       ftp://ftp.ebi.ac.uk/pub/databases/pride/PRIDE_Exp_Complete_Ac_1643.xml
```

And also… protein-protein interactions



PSI-XML: XML-based format

MITAB: tab-delimited format

PSI Deliverables – As an example

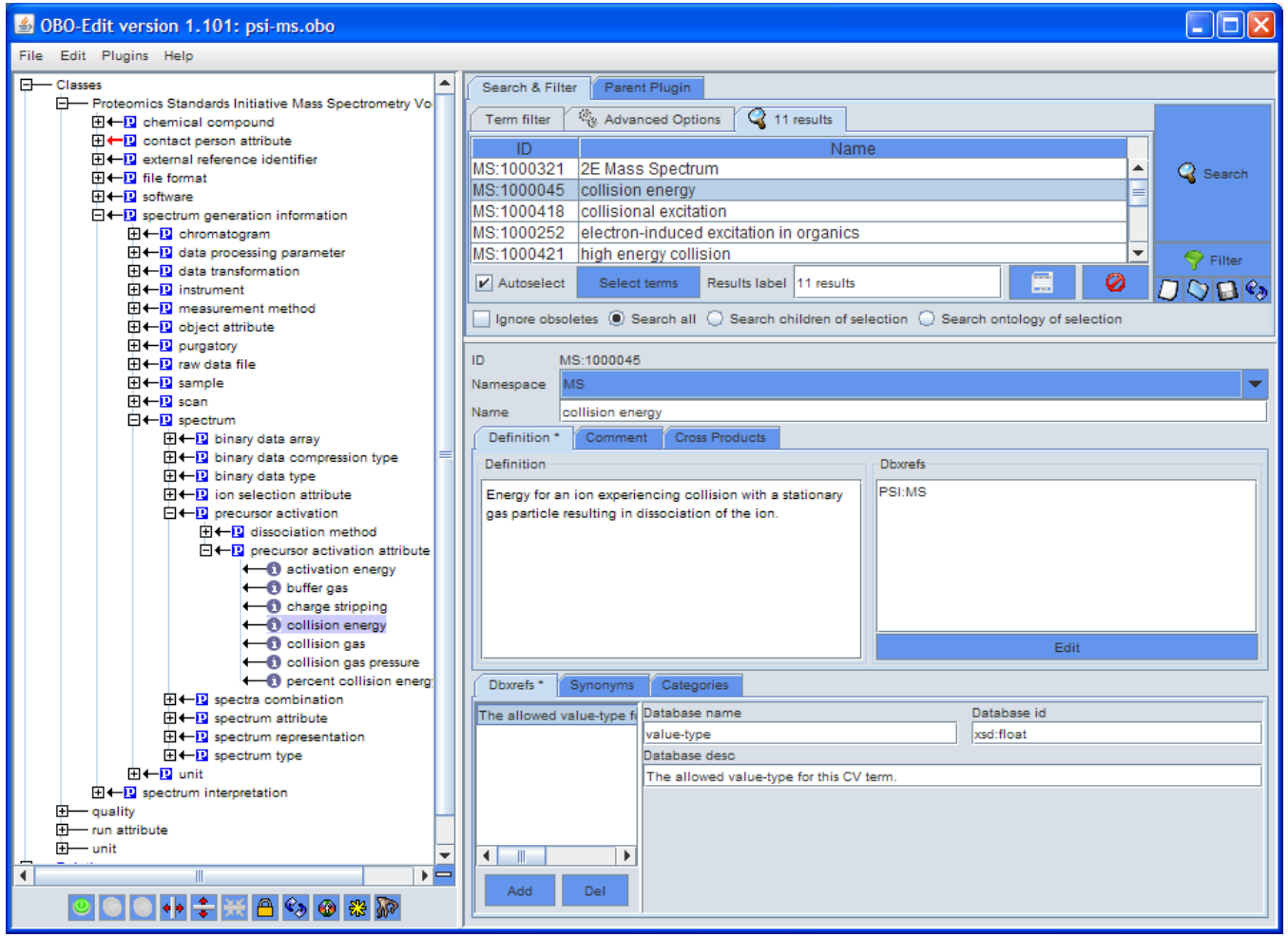
- Formats: Usually **XML-based** (but also **tab-delimited files**), capable of representing the relevant Minimum Information, plus additional detailed data for the domain.

- Controlled vocabularies: Usually an OBO-style hierarchical controlled vocabulary precisely defining the **metadata** that are encoded in the formats.

- Databases and Tools: **Foster open software implementations** to make the standards truly useful.

- Community interaction to ensure **adoption of the standards** and **public deposition** of data in proteomics repositories.

PSI MS Controlled Vocabulary



>3,000 terms

Mayer *et al.*, Database, 2013

The Ontology Lookup Service (OLS)

http://www.ebi.ac.uk/ontology-lookup/

PSI Deliverables – As an example

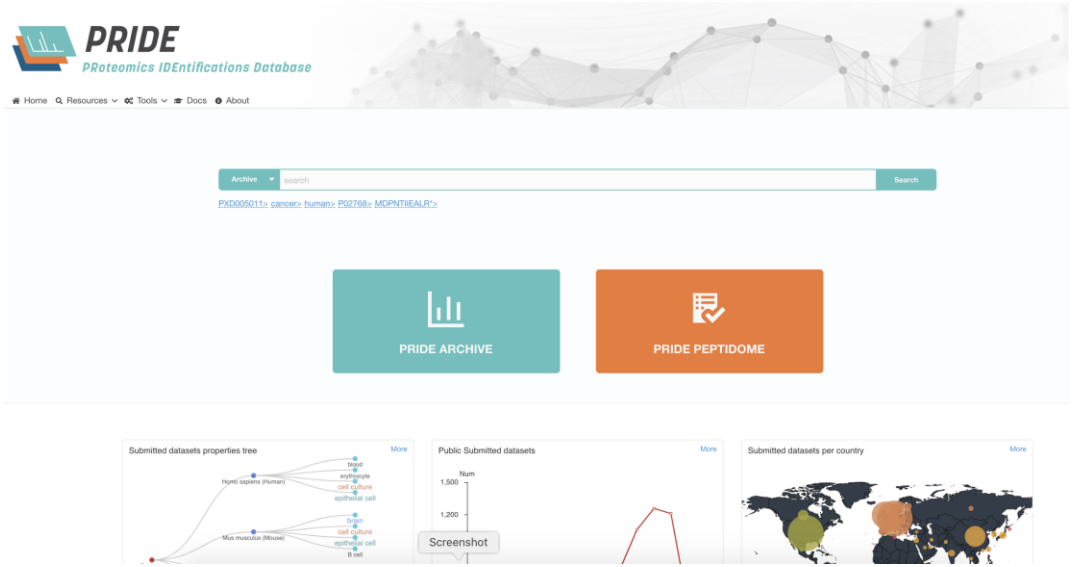
- Formats: Usually **XML-based** (but also **tab-delimited files**), capable of representing the relevant Minimum Information, plus additional detailed data for the domain.

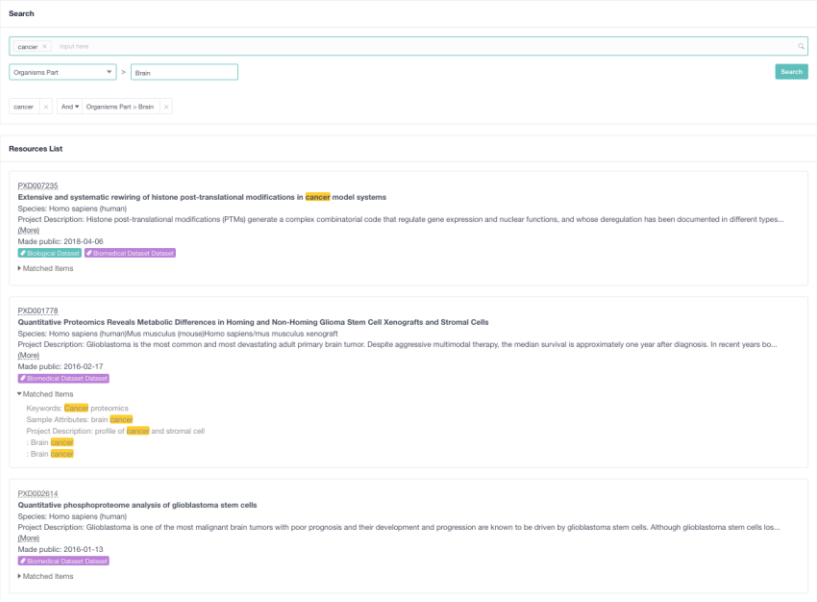
- Controlled vocabularies: Usually an OBO-style hierarchical controlled vocabulary precisely defining the **metadata** that are encoded in the formats.

- Databases and Tools: **Foster open software implementations** to make the standards truly useful.

- Community interaction to ensure **adoption of the standards** and **public deposition** of data in proteomics repositories.

The PRIDE database





PRIDE is the world-leading resource storing mass spectrometry-based proteomics datasets:

- Mass spectra (raw data, peak lists), peptide and protein expression data
- **All proteomics approaches are supported**
- **>17,000 datasets (June 2020).**

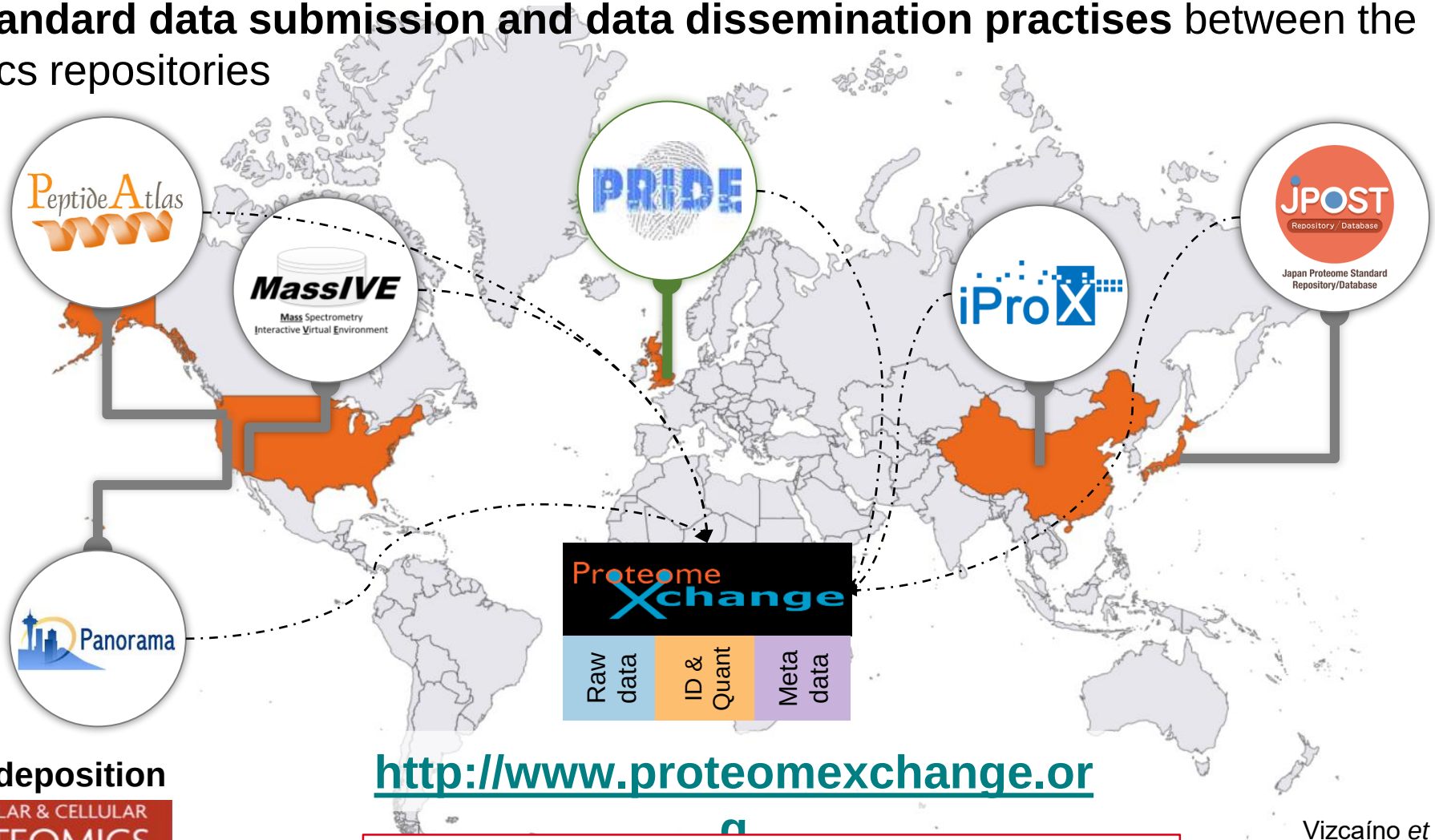
http://www.ebi.ac.uk/pride/archive/

- **ELIXIR Core data resource and data deposition database**

Vizcaíno *et al.*, *NAR*, 2016
Perez-Riverol *et al.*, *NAR*, 2019

ProteomeXchange: A global, distributed proteomics infrastructure

Implements **standard data submission and data dissemination practises** between the main proteomics repositories



Mandatory data deposition

MCP MOLECULAR & CELLULAR PROTEOMICS

Journal of proteome research

nature

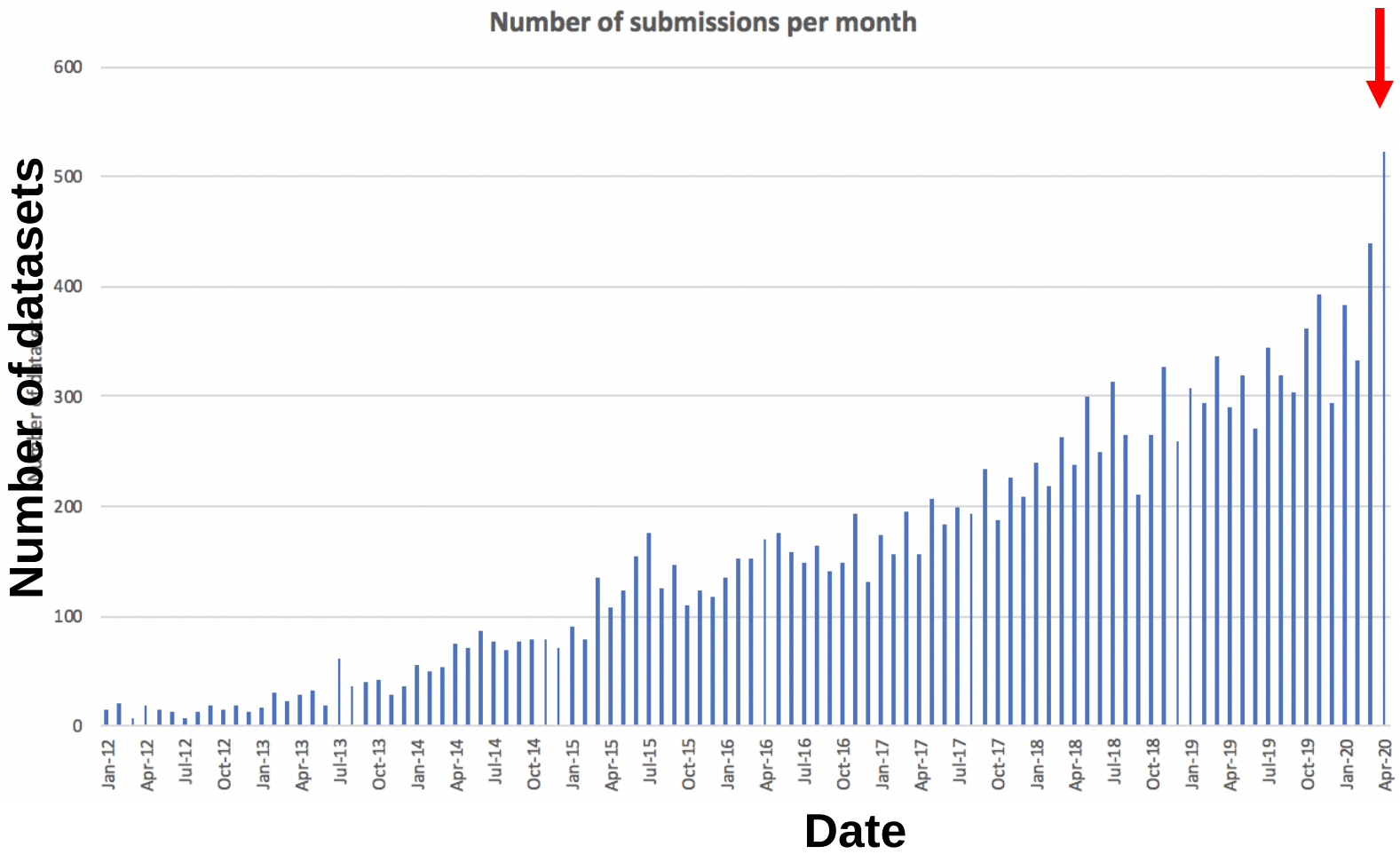
PLOS

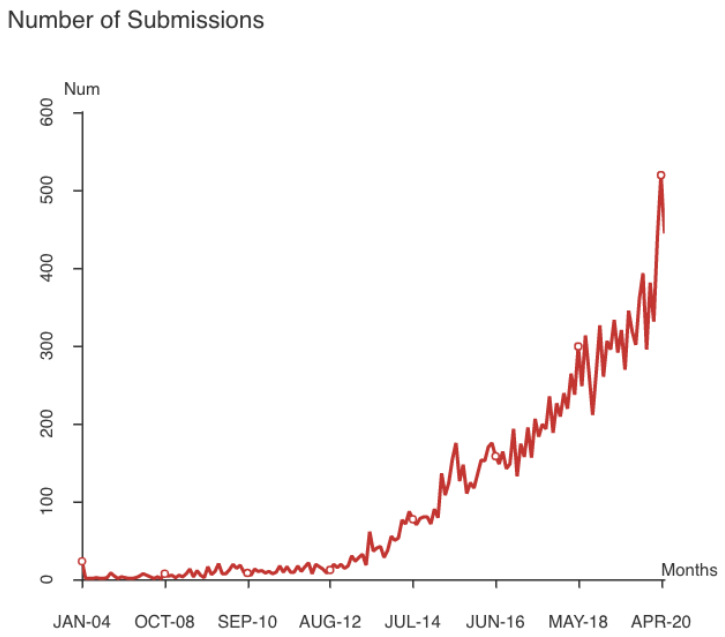
http://www.proteomexchange.org

Data sharing has generalized in the field

Vizcaíno *et al.*, *Nat Biotechnol*, 2014
Deutsch *et al.*, *NAR, 2017*
Deutsch *et al.*, *NAR, 2020*

Dataset submissions keep increasing in number and volume





PRIDE: More than **500 datasets** in a single month (April 2020)

PSI Deliverables – As an example

- Formats: Usually **XML-based** (but also **tab-delimited files**), capable of representing the relevant Minimum Information, plus additional detailed data for the domain.

- Controlled vocabularies: Usually an OBO-style hierarchical controlled vocabulary precisely defining the **metadata** that are encoded in the formats.

- Databases and Tools: **Foster open software implementations** to make the standards truly useful.

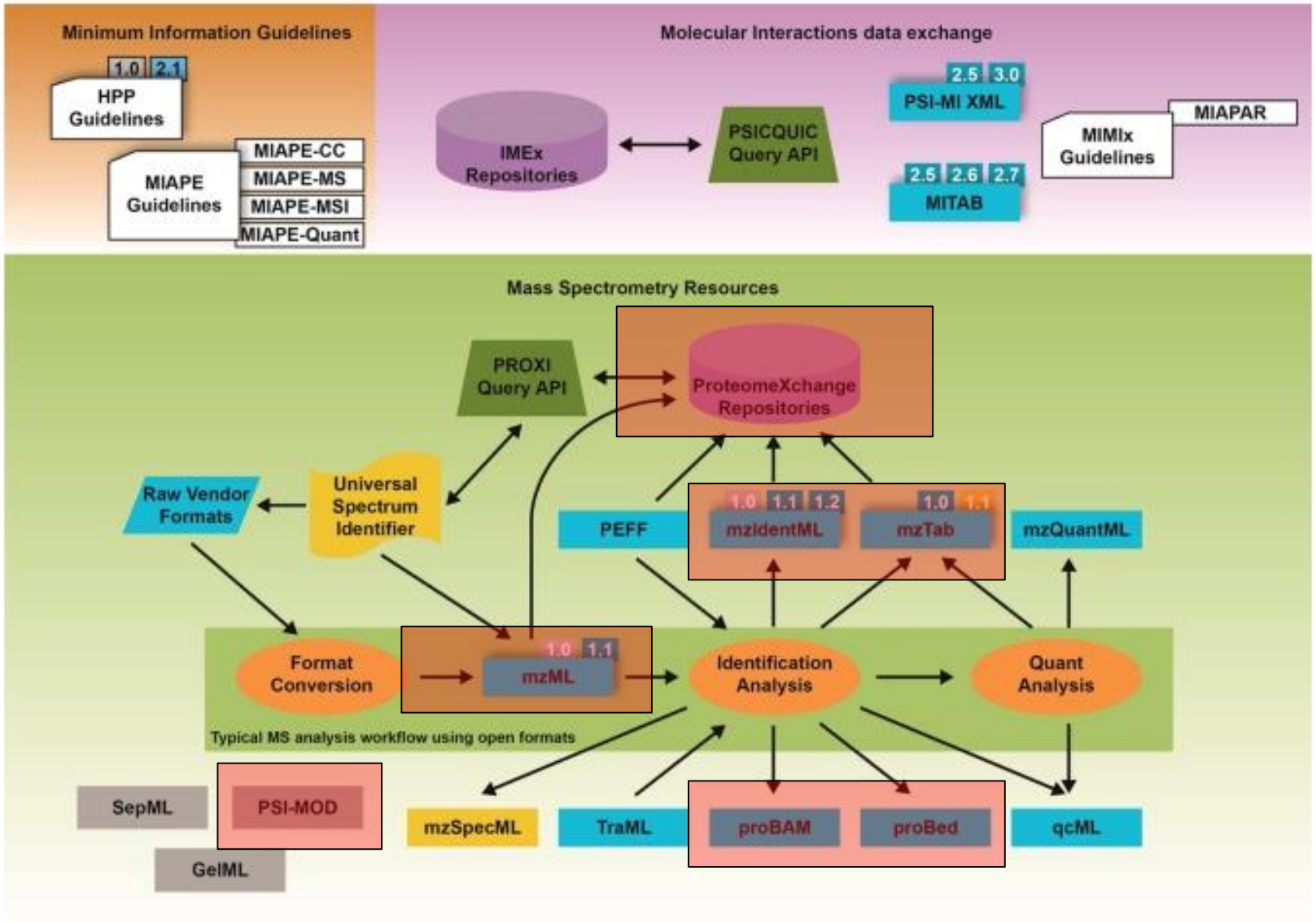
- Community interaction to ensure **adoption of the standards** and **public deposition** of data in proteomics repositories.

Importance of making software available

PSI promotes implementations. The reference libraries are always open source and can be used by anyone!

jmzML (https://github.com/PRIDE-Utilities/jmzml)	Cote *et al.*, *Proteomics*, 2009
jmzIdentML (https://github.com/PRIDE-Utilities/jmzidentML)	Reisinger *et al.*, *Proteomics*, 2012
jmzReader (https://github.com/PRIDE-Utilities/jmzReader)	Griss *et al.*, *Proteomics*, 2012
jmzQuantML (https://github.com/UKQIDA/jmzquantml)	Qi *et al., Proteomics*, 2014
jmzTab (https://github.com/PRIDE-Utilities/jmzTab)	Xu *et al., Proteomics*, 2014
ms-data-core-api (https://github.com/PRIDE-Utilities/ms-data-core-api)	Perez-Riverol *et al.*, *Bioinformatics*, 2015

Summary slide



Deutsch *et al.*, *JPR*, *2017*

Overview

- A couple of slides about the need of data standards
- Proteomics standards: The Proteomics Standards Initiative and ProteomeXchange
- **DNA/RNA Sequencing standards: Introduction to GAG4H standards**
- Data integration using data standards

The GA4GH Mission...

Global Alliance for Genomics & Health

LAUNCH of GA4GH in 2013

The Global Alliance for Genomics and Health aims to accelerate progress in genomic science and human health by developing standards and framing policy for responsible genomic and health-related data sharing.

GA4GH 2019 Strategic Roadmap

Global Alliance for Genomics & Health

- Approved
- Ongoing
- In Development

Clin/Pheno Data Capture
- Phenopackets
- Pedigree Representation
- Health Open Terminology

Cloud
- Workflow Execution Service API
- Cloud Testbed Interoperability
- Data Repository Service API
- Tool Registry Service
- Task Execution Service

Discovery
- Beacon API
- Service Info
- Service Registry
- Search API

Data Security
- Policy Initiatives
- Authentication & Authorization Infrastructure
- Data Security and Infrastructure Policy
- Breach Response Protocol

Data Use & Researcher Identity
- Data Use Ontology
- GA4GH Passports

Genomic Knowledge Standards
- Variant Representation
- Variant Annotation
- Gene Annotation

Large Scale Genomics
- htsget API
- refget API
- Read file formats
- Variation File Formats
- Variation Benchmarking
- Crypt4GH
- rnaget API

Regulatory & Ethics
- Framework
- Policy Initiatives
  - Ethics Review Recognition, Accountability, Consent, Privacy & Security Technical Standards & IP, Return of Results
- Responsible Data Sharing Resource Documents
- GDPR Forum

Discovery Work Stream

Global Alliance for Genomics & Health

Create unified data discovery platform for genomic and clinical data

Approved and available for implementation

- **Beacon API V1**: discovery of variant information by remote researcher
- **Service Info/Registry API**: will allow for dynamic registration and on-demand discovery of online GA4GH APIs (data, tools, services) to enable their real time discovery and use.

In development

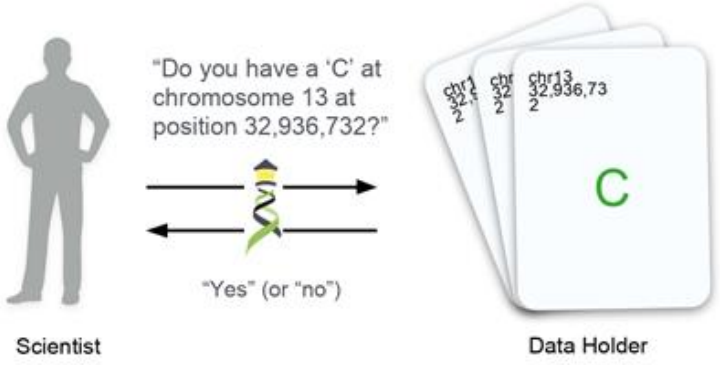
- **Search API**: specification for query language across genomic, phenotypic, and clinical data

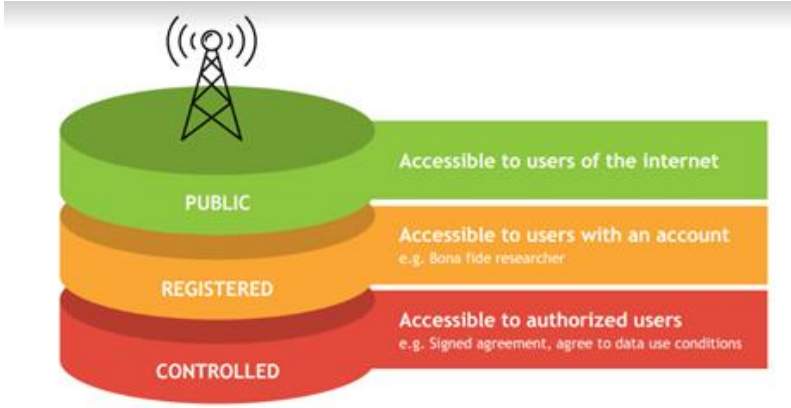
Beacon API v1

Global Alliance for Genomics & Health

The Beacon API can be implemented as a web-accessible service that users may query for information about a specific allele.

Approved: October 3, 2018





Example Users

EMBL-EBI · Australian Genomics Health Alliance · SciLifeLab · elixir · University of California Santa Cruz · Broad Institute · European Genome-Phenome Archive · International Cancer Genome Consortium

Large Scale Genomics Work Stream

Global Alliance for Genomics & Health

Standardize methods for accessing large-scale genomic data

Approved and available for implementation

- **htsget Streaming API V1**: secure standard interface for slicing and streaming sequencing data that decouples the assumption of a file at the remote location
- **refget API V1**: framework to retrieve 'reference sequences' by a unique checksum
- **CRAM V3, SAM V1, & BAM V1**: standard file formats for storing read data
- **VCF V4 & BCF V2**: standard format to represent genomic variation
- **Crypt4GH:** An encrypted container file format suitable for genomic data

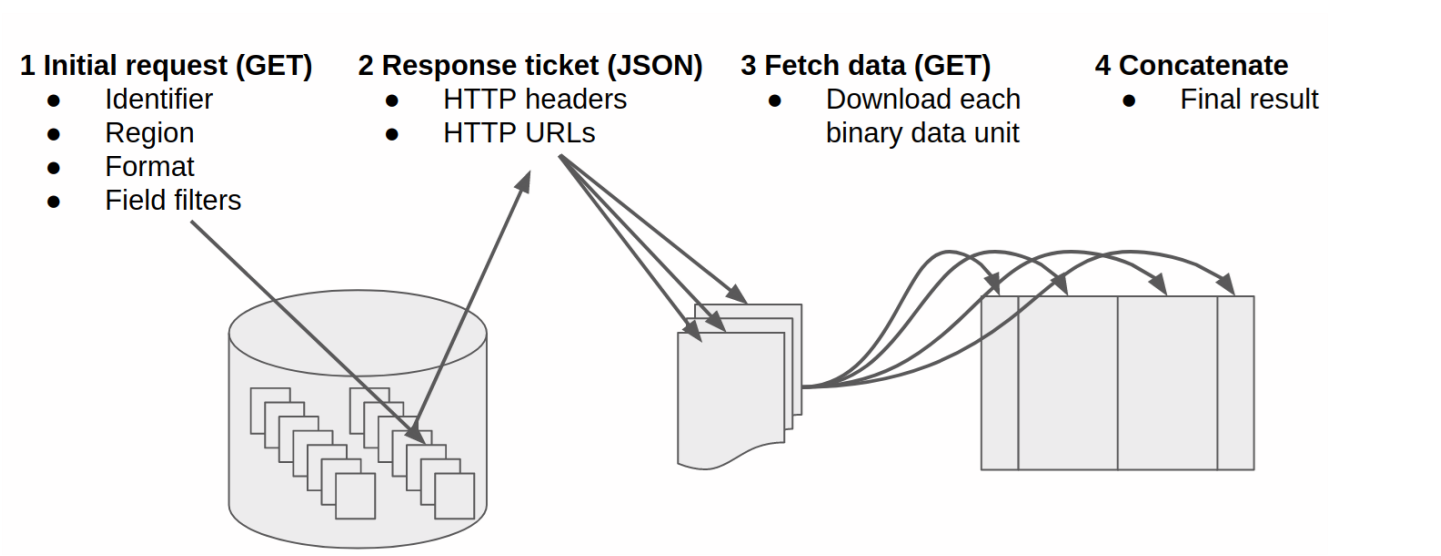
In development

- **rnaget API:** request a URL for a required RNASeq Expression Matrix

htsget API v1

htsget is a genomic data retrieval specification that allows users to download read data for subsections of the genome in which they are interested.

Approved: October 7, 2017



Example Users

Australian Genomics Health Alliance
Genomics england
BROAD INSTITUTE
ENA European Nucleotide Archive
DNAnexus
EUROPEAN GENOME-PHENOME ARCHIVE
Google Cloud Platform
UNIVERSITY OF OXFORD
wellcome sanger institute

refget API v1

Global Alliance for Genomics & Health

The GA4GH refget API enables access to reference genomic sequences using a checksum identifier based on the sequence content itself.

Approved: October 3, 2018

Reference sequence

Chromosome 1
GRCh38 (hg38)

Normalise sequence

A-Z only
Uppercase
No whitespace

Calculate checksum (md5)

6681ac2f62509cfc220d78751b8dc524

Calculate checksum (TRUNC512)

959cb1883fc1ca9ae1394ceb475a356ead1ecceff5824ae7

Example Users

BROAD INSTITUTE

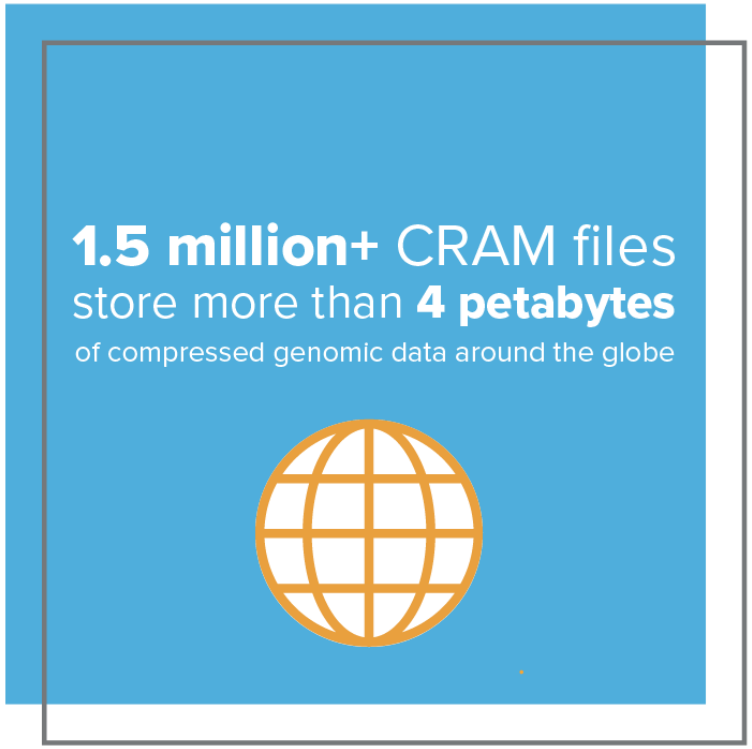
EUROPEAN GENOME-PHENOME ARCHIVE

aws

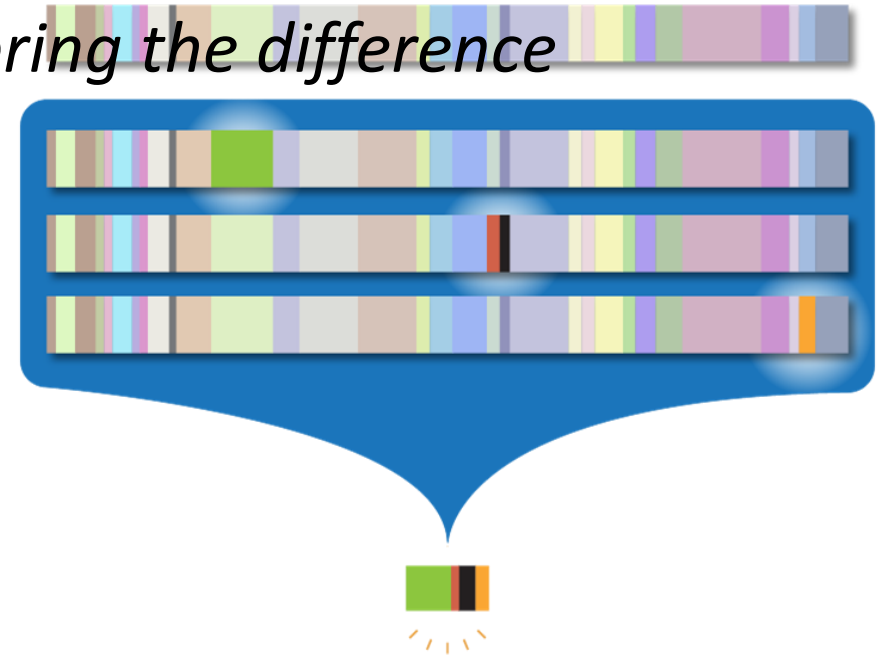
CRAM

Global Alliance for Genomics & Health

CRAM is a file format for storing compressed genomic data. To make files small and efficient, the algorithm compresses information by only storing the parts that are different from the reference human genome.



CRAM compresses data by only storing the difference



CRAM Implementation

Global Alliance for Genomics & Health

CRAM is supported by the following libraries and tools:

Software Libraries: htslib | htsjdk | PySam | Bio::DB::HTS | RustBio

Tools: Samtools | GATK | Picard | IGV | Crumble

Data Archives: European Nucleotide Archive (ENA) | European Genome-phenome Archive (EGA)

Genome Browsers: ENSEMBL | JBrowse | UCSC Genome Browser

Example Users

EMBL-EBI | BROAD INSTITUTE | wellcome sanger institute | illumina | NHLBI TOPMed | Australian Genomics Health Alliance | H3Africa Human Heredity & Health in Africa | Genomics england | SciLifeLab

Crypt4GH v1

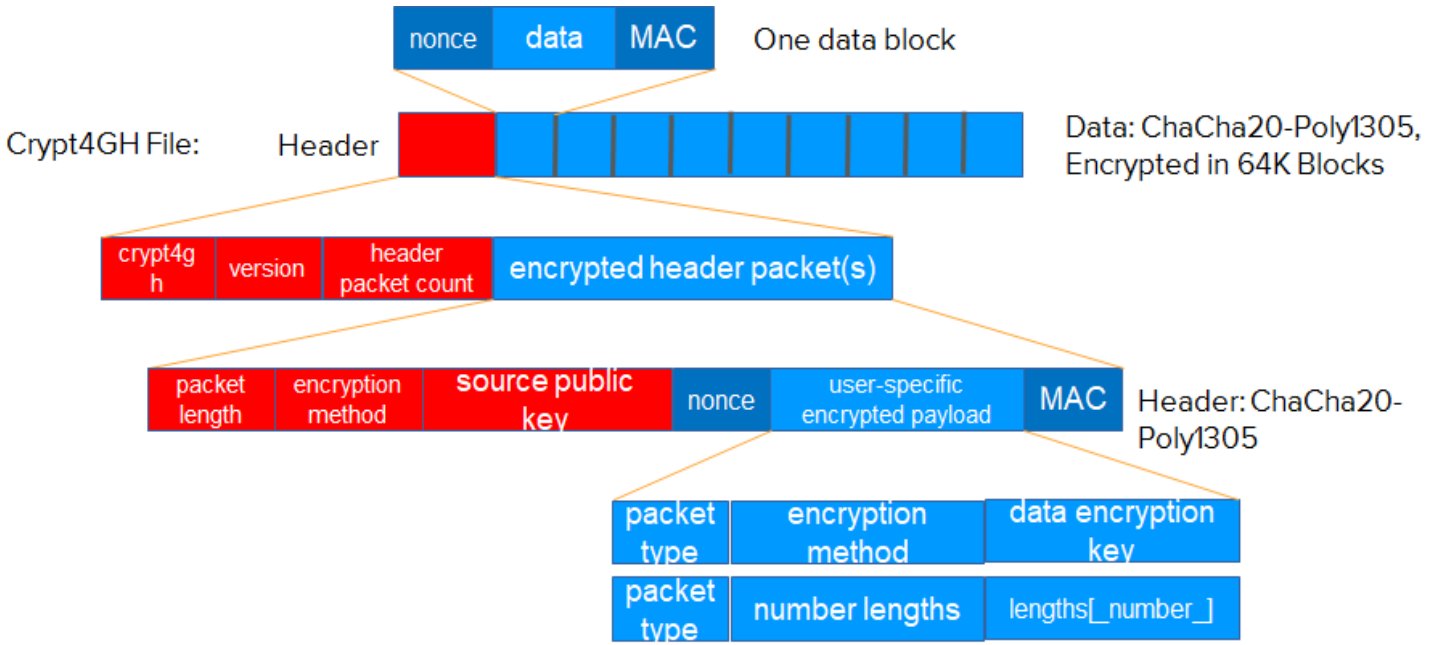
Global Alliance for Genomics & Health

Crypt4GH is a random-access encrypted file container format for securely sharing large-scale genomic data

Approved: September 3, 2019

Example Users

EUROPEAN GENOME-PHENOME ARCHIVE



VCF/BCF

Global Alliance for Genomics & Health

The Variant Call Format (VCF) specifies the format of a text file used in bioinformatics for storing gene sequence variations. The Binary Call Format (BCF) is the Binary equivalent, smaller and more efficient to process.

Software Libraries: htslib | htsjdk

Tools: Samtools | BCFtools

Databases: European Variation Archive (EVA) | dbGAP | dbSNP | 1000 Genomes Projects / IGSR

Genome Browsers: ENSEMBL | JBrowse | UCSC Genome Browser

Example Users

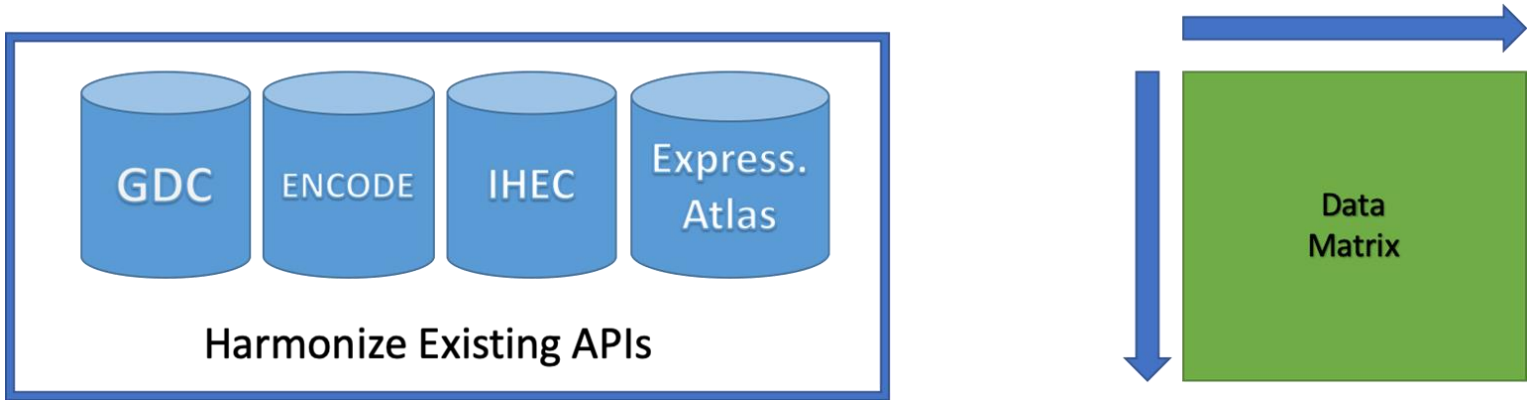
All of Us RESEARCH PROGRAM | BROAD INSTITUTE | elixir | EMBL-EBI | Genomics england

RNAget v1

Global Alliance for Genomics & Health

RNAget API enables search and retrieval of RNA data at scale.

Approved: September 3, 2019



Example Users

CanDIG

NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

epi share

ENCODE

HUMAN CELL ATLAS

Expression Atlas

Genomic Knowledge Standards Work Stream

Global Alliance for Genomics & Health

Create standards-based components for exchange of genomic information

Approved and available for implementation

- **Variant Representation:** an extensible data model and message schema specification for the representation of variants

In development

- **Variant Annotation: Data Model:** guide the linkage of annotations and structured clinical interpretations to variant data

Variation Representation V1

Global Alliance for Genomics & Health

The Variant Representation Specification is a standard way of exchanging genetic variation data with precision and consistency

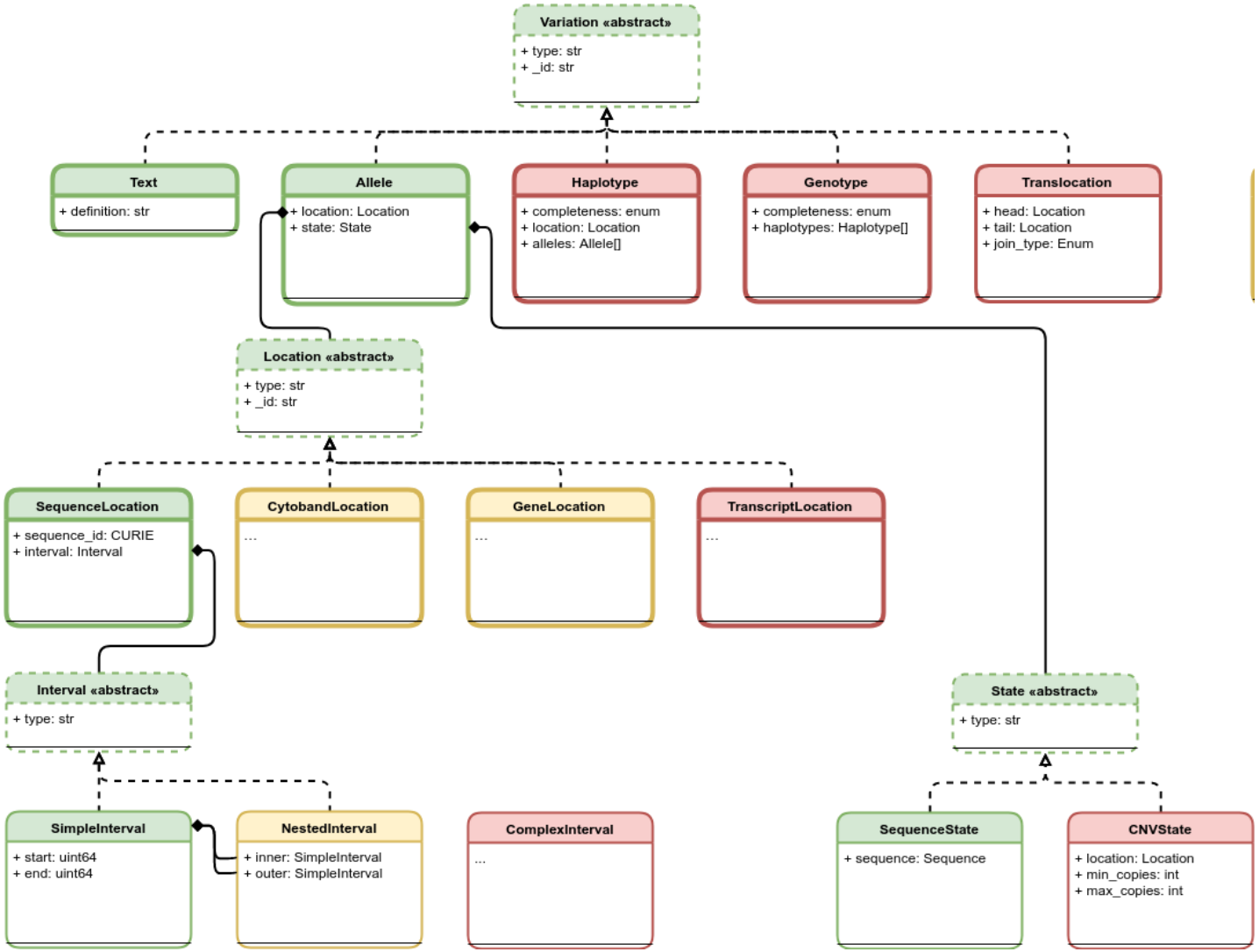
Approved: September 3, 2019

Example Users

VICC Variant Interpretation for Cancer Consortium

ClinGen Clinical Genome Resource

BRCA CHALLENGE



Cloud Work Stream

Global Alliance for Genomics & Health

Bring algorithms to the data' by creating standards for portable workflows

Approved and available for implementation

- **Workflow Execution Service (WES) V1**: execute the same scientific tools and workflows in a variety of environments without modification
- **Tool Registry Service (TRS)**: portable exchange of tools and workflows
- **Data Repository Service API (DRS)**: create a common way to refer to data and access it regardless of cloud or platform, making it easier to do work across projects and environments

In development

- **Task Execution Service (TES):** common interface for batching execution of tasks in multiple systems
- **Testbed Interoperability Demonstration**: use GA4GH cloud APIs to demonstrate that workflows can be exchanged between Driver Project sites and used reproducibly

Workflow Execution Service API v1

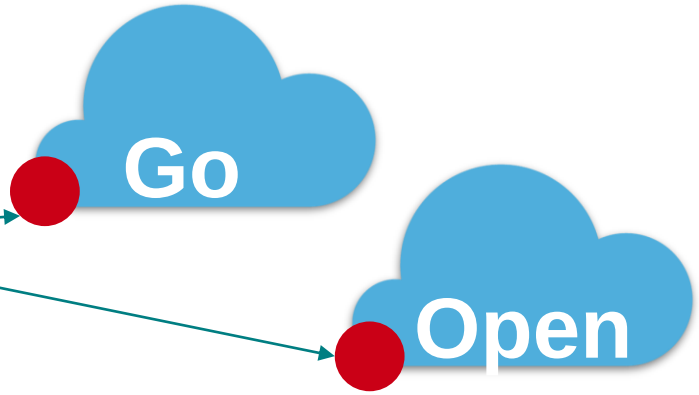
Global Alliance for Genomics & Health

Workflow Execution Service (WES) allows complex workflows written in CWL or WDL (two common workflow languages) to run on disparate cloud environments (e.g. Google, Cloud, AWS, OpenStack) with identical flow control and execution.

Approved: January 28, 2019

Cancer Workflow
ICGC Argo 1.x

ARGO Accelerating Research in Genomic Oncology
International Cancer Genome Consortium



Example Users

BROAD INSTITUTE
NHLBI TOPMed
HUMAN CELL ATLAS
All of Us RESEARCH PROGRAM
Australian Genomics Health Alliance
Genomics england
elixir

Regulatory and Ethics Work Stream

Global Alliance for Genomics & Health

Technology standards and best practices for protecting data

Ongoing initiatives

- **International Participant Values Survey:** "Your DNA, Your Say" explores how people around the world feel about the collection, use, and sharing of genetic and health data for research
- **GDPR & International Health Data Sharing Forum**: a primer followed by monthly briefs that answer questions about the GDPR's impact on various aspects of international health research

In development

- **Return of Results Policy**: what to consider when deciding whether to tell participants about genomic findings relevant to their health

GA4GH 2019 Strategic Roadmap

Global Alliance for Genomics & Health

- Approved
- Ongoing
- In Development

Clin/Pheno Data Capture
- Phenopackets
- Pedigree Representation
- Health Open Terminology

Cloud
- Workflow Execution Service API
- Cloud Testbed Interoperability
- Data Repository Service API
- Tool Registry Service
- Task Execution Service

Discovery
- Beacon API
- Service Info
- Service Registry
- Search API

Data Security
- Policy Initiatives
- Authentication & Authorization Infrastructure
- Data Security and Infrastructure Policy
- Breach Response Protocol

Data Use & Researcher Identity
- Data Use Ontology
- GA4GH Passports

Genomic Knowledge Standards
- Variant Representation
- Variant Annotation
- Gene Annotation

Large Scale Genomics
- htsget API
- refget API
- Read file formats
- Variation File Formats
- Variation Benchmarking
- Crypt4GH
- rnaget API

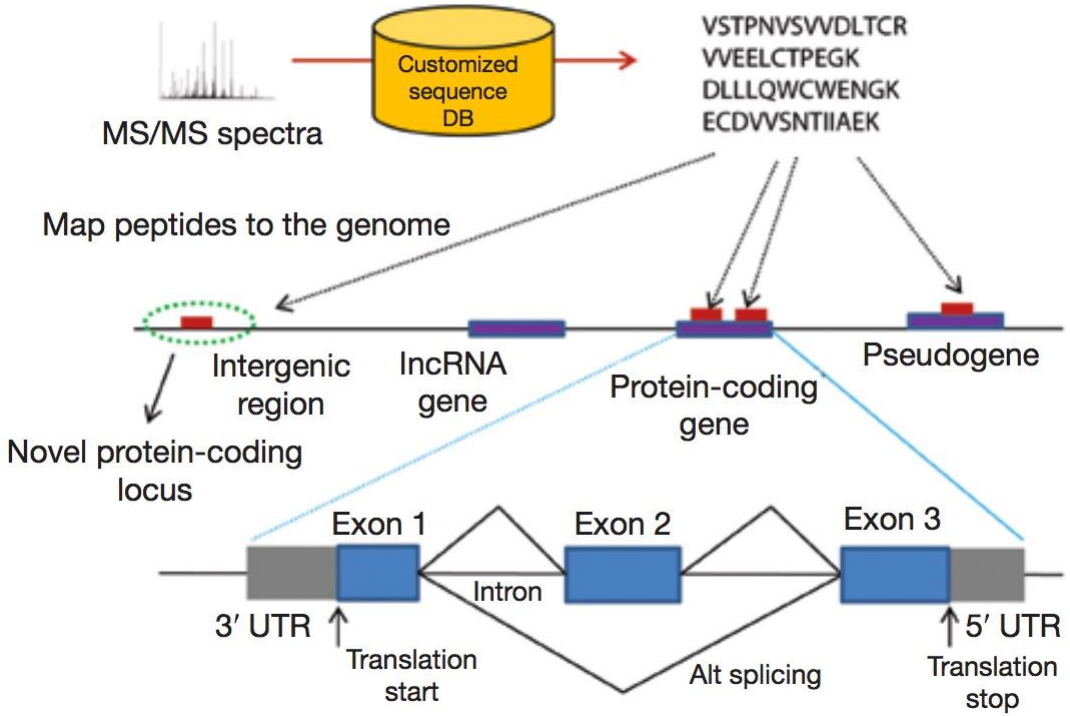
Regulatory & Ethics
- Framework
- Policy Initiatives
  - Ethics Review Recognition, Accountability, Consent, Privacy & Security Technical Standards & IP, Return of Results
- Responsible Data Sharing Resource Documents
- GDPR Forum

Overview

- A couple of slides about the need of data standards
- Proteomics standards: The Proteomics Standards Initiative and ProteomeXchange
- DNA/RNA Sequencing standards: Introduction to GAG4H standards
- **Data integration using data standards**

Across-omics -> Proteogenomics approaches

- Proteomics data is **combined with genomics and/or transcriptomics information**, typically by using sequence databases generated from **DNA sequencing efforts, RNA-Seq experiments, Ribo-Seq approaches, and long-non-coding RNAs**.



Historically: Used for genome annotation purposes

Nesvizhskii, *Nat Methods*, *2014*

Proteogenomics related data formats

- Two ongoing formats are being developed: **proBed** (version 1 available) and **proBAM** (still under review).

- Same overall objective: to map identified peptides to genome coordinates.

- Different level of detail:
  - **proBed** is tab-delimited and simpler, **based on the original BED format**. Less level of detail.
  - **proBAM** is based in the original **SAM/BAM formats**, widely used in genomics. Much higher level of detail.

Proteogenomics related data formats

Menschaert *et al. Genome Biology*
DOI 10.1186/s13059-017-1377-x

Genome Biology

OPEN LETTER Open Access

The proBAM and proBed standard formats: enabling a seamless integration of genomics and proteomics data

Gerben Menschaert[1*†], Xiaojing Wang[2,3*†], Andrew R. Jones[4], Fawaz Ghali[4,5], David Fenyö[6,7], Volodimir Olexiouk[1], Bing Zhang[8,9], Eric W. Deutsch[10], Tobias Ternent[11] and Juan Antonio Vizcaíno[11*]

Abstract

On behalf of The Human Proteome Organization (HUPO) Proteomics Standards Initiative, we introduce here two novel standard data formats, proBAM and proBed, that have been developed to address the current challenges of integrating mass spectrometry-based proteomics data with genomics and transcriptomics information in proteogenomics studies. proBAM and proBed are adaptations of the well-defined, widely used file formats SAM/BAM and BED, respectively, and both have been extended to meet the specific requirements entailed by proteomics data. Therefore, existing popular genomics tools such as SAMtools and Bedtools, and several widely used genome browsers, can already be used to manipulate and visualize these formats "out-of-the-box." We also highlight that a number of specific additional software tools, properly supporting the proteomics information available in these formats, are now available providing functionalities such as file generation, file conversion, and data analysis. All the related documentation, including the detailed file format specifications and example files, are accessible at http://www.psidev.info/probam and at http://www.psidev.info/probed.

proBAM	Description	Example
QNAME	Spectrum name	index=7096_PXD001524
FLAG	Bitwise FLAG	16
RNAME	Reference sequence NAME	chr21
POS	1-based leftmost mapping POSition	33907431
MAPQ	-	255
CIGAR	CIGAR string	23M1628N28M
RNEXT	-	*
PNEXT	-	0
TLEN	-	0
SEQ	Coding sequence	TCGACCATTTTCAGCAAG CAAATTGATCAGATTGGT AGTGAGGGGAGAGAA
QUAL	-	*
XL	Number of peptides to which the spectrum maps	XL:i:1
XM	Modification(s): semicolon-separated list of modifications	XM:Z:*
XB	Mass difference (exp - calcul); experimental mass; calculated mass	XB:Z:0.0002109709;;
XQ	PSM FDR (i.e. q-value or 1-PEP)	XQ:f:1.06E-04
XS	PSM score	XS:f: 79.78288685
NH	Number of genomic locations to which the peptide sequence maps	NH:i:1
XO	Peptide uniqueness (1…5)	XO:Z:unique
XC	Peptide Charge	XC:i:2
XI	Peptide intensity	XI:f:-1
XP	Peptide sequence from the original search result	XP:Z:FSPLTTNLINLLAENGR
XR	Reference peptide sequence	XR:Z:FSPLTTNLINLLAENGR
XF	Reading frame of the peptide (0, 1, 2)	XF:Z:0,1
XA	Whether the peptide is well annotated (0,1,2)	XA:i:0
XG	Peptide type (N, V, W, J, A, M, C, E, B, O, T, R, I, G, D, U, X)	XG:A:N
YP	Protein accession ID from the original search	YP:Z:ENSP00000290299
XE	Enzyme used in the experiment	XE:i:1
XN	Number of missed cleavages in the peptide	XN:i:0
XT	Enzyme specificity (0, 1, 2, 3)	XT:i:3
YA	Following amino acids (2 AA)	YA:Z:LS
YB	Preceding amino acids (2 AA)	YB:Z:ER
XU	Uniform Resource Identifier	*
Z?	Custom fields	.

proBed	Description	Example
chrom	Reference sequence chromosome	chr21
chromStart	Start position of the first DNA base	33907430
chromEnd	End position of the last DNA base	33909107
name	Unique name	ENSP00000290299_3845
score	Score	276
strand	+ or - for strand	-
thickStart	Coding region start	33907430
thickEnd	Coding region end	33909107
reserved	Always 0	0
blockCount	Number of blocks	2
blockSizes	Block sizes	25,26
chromStarts	Block starts	0,1651
psmScore	PSM score	79.78288685
fdr	Estimated global false discovery rate	1.06E-04
modifications	Post-translational modifications	15-UNIMOD:7
expMassToCharge	Experimental mass to charge value	936.499
calcMassToCharge	Calculated mass to charge value	936.497
psmRank	Peptide-Spectrum Match rank.	1
charge	Charge value	2
peptideSequence	Peptide sequence	FSPLTTNLINLLAENGR
uniqueness	Peptide uniqueness	unique
proteinAccession	Protein accession number	ENSP00000290299
genomeReferenceVersion	Genome reference version number	Homo_sapiens.GRCh38.77
datasetID	Dataset Identifier	PXD001524_reprocessed
uri	Uniform Resource Identifier	.

Color legend
Genomic locations
Mapping details
Nucleotide sequence
PSM information
Peptide information
Protein information
Enzyme information
Data source

Provide your own data to genome browsers

e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Using this website | Annotation and prediction | Data access | API & software | About us

In this section

- Retrieving sequences
- Gene Expression
- Compare genes across species
- Variants for my gene
- Diseases and Phenotypes
- ENCODE data in Ensembl
- Use my own data

Search documentation. Go

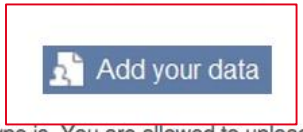
Help & Documentation > Using this website > Tutorials > Use my own data

Use my own data in Ensembl

NGS reads and more

Ensembl supports a number of different filetypes for upload and visualisation along the genome.

The most popular page in which to view your data is the Location tab, Region in Detail. Use the 'Add your data' button in the left hand menu (N.B. the button will change to 'Manage your data' once you've added at least one file).



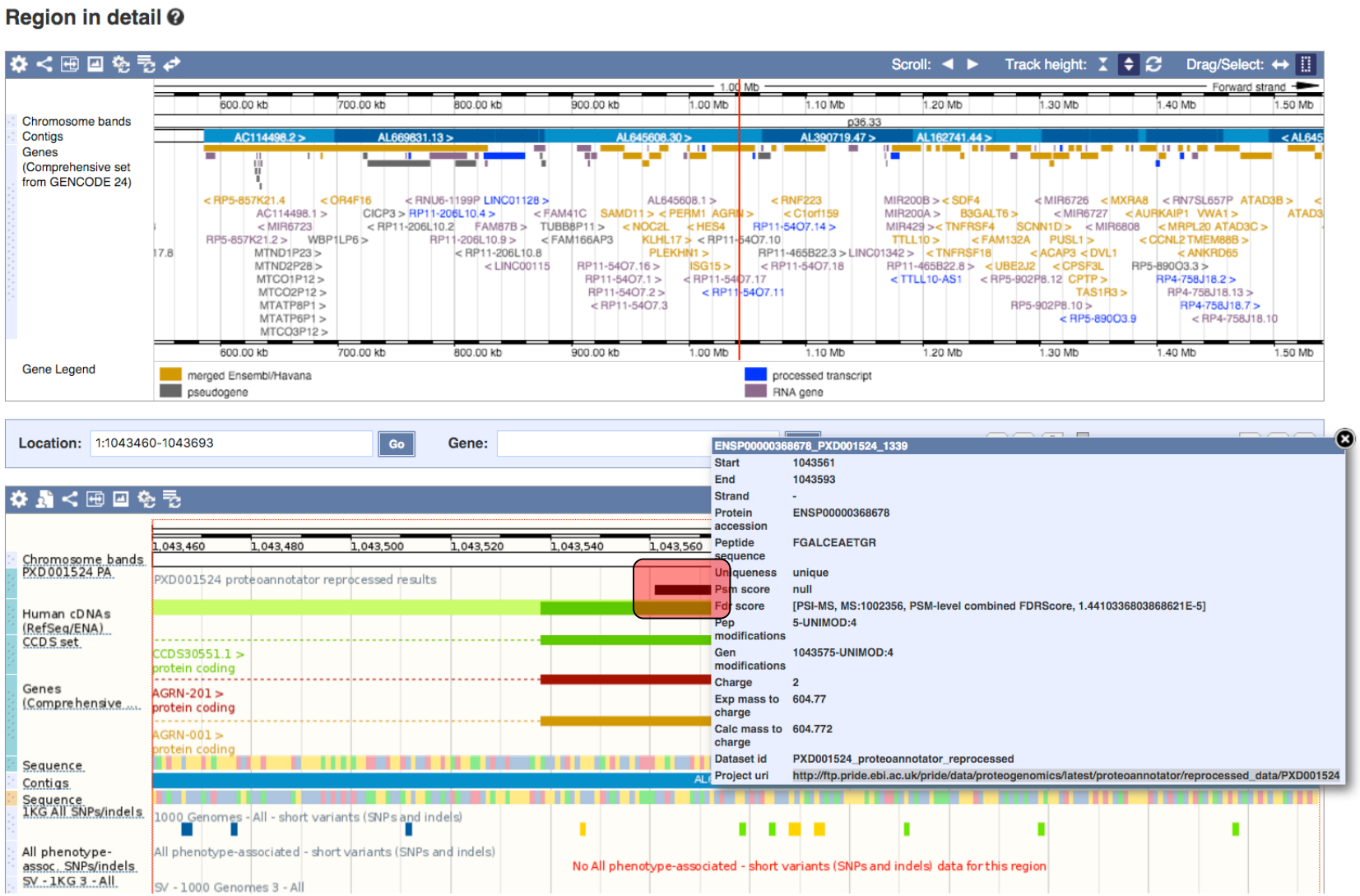
A dialogue will ask you what your file type is. You are allowed to upload smaller files. In the case of large data files, only the url attachment option is available.

[Click to enlarge]

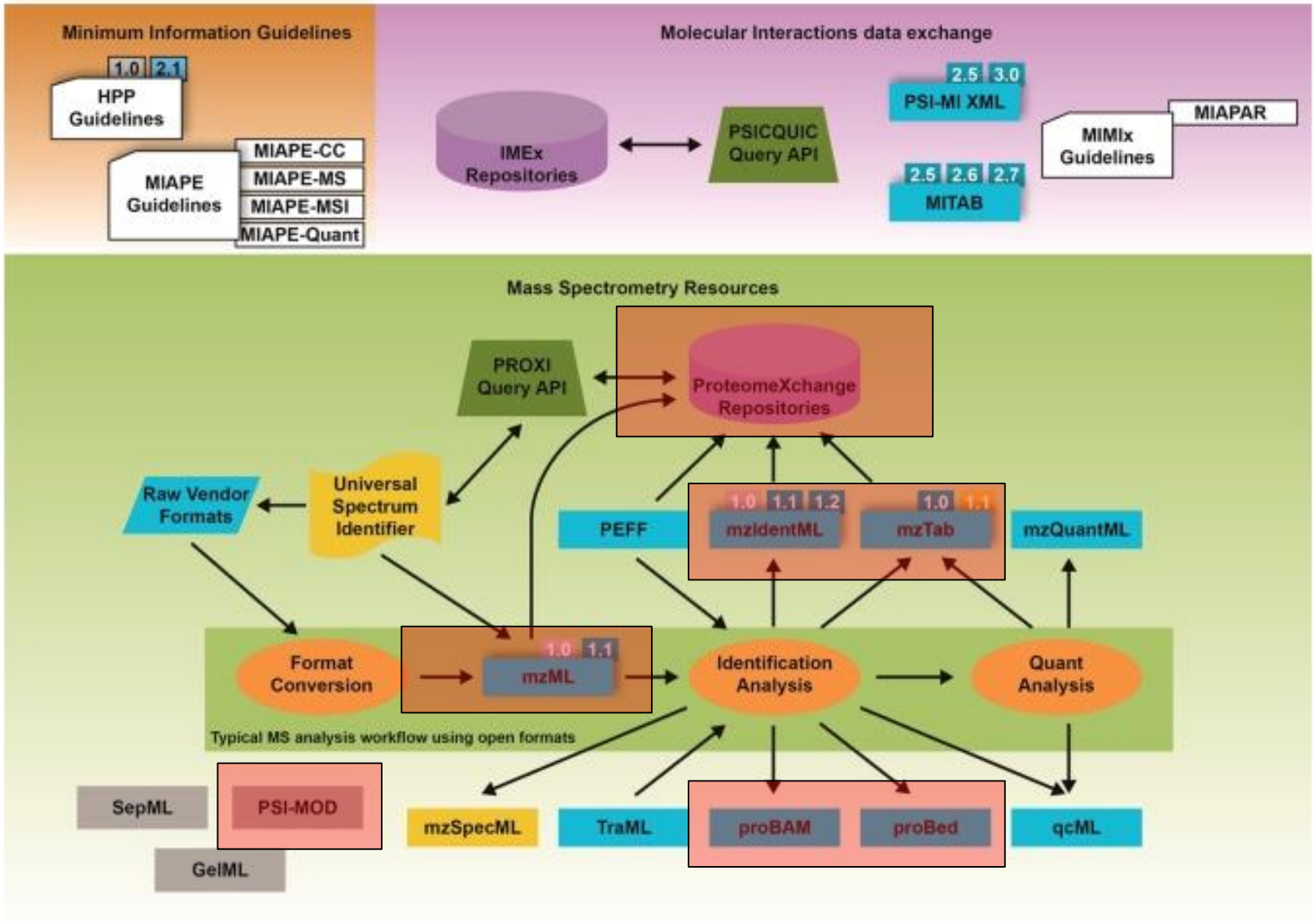
Here is an example of sequence reads attached as a BAM file, along chromosome 20.

[Click to enlarge]

TrackHubs in Genome Browsers



Summary slide



Deutsch *et al.*, *JPR*, *2017*

Protein sequences/proteomics data and GA4GH standards

It definitely makes sense that protein sequence/proteomics information is incorporated into GA4GH projects

- Reference for proteins: **Refget**
- **Variation standards** and **Beacons**
- Others?

refget API v1

Global Alliance for Genomics & Health

The GA4GH refget API enables access to reference genomic sequences using a checksum identifier based on the sequence content itself.

Approved: October 3, 2018

Reference sequence

Chromosome 1
GRCh38 (hg38)

Normalise sequence

A-Z only
Uppercase
No whitespace

Calculate checksum (md5)

6681ac2f62509cfc220d78751b8dc524

Calculate checksum (TRUNC512)

959cb1883fc1ca9ae1394ceb475a356ead1ecceff5824ae7

Example Users

BROAD INSTITUTE

EUROPEAN GENOME-PHENOME ARCHIVE

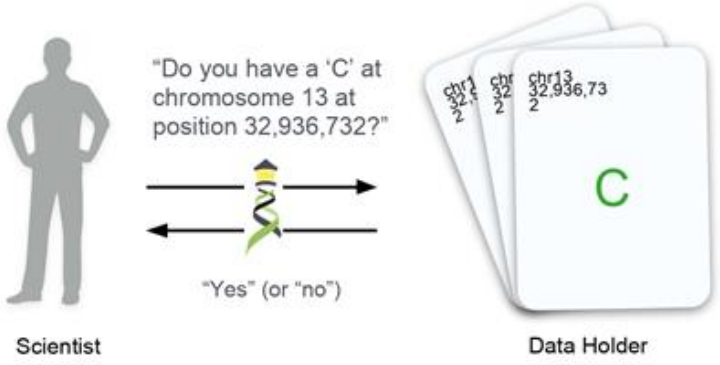
aws

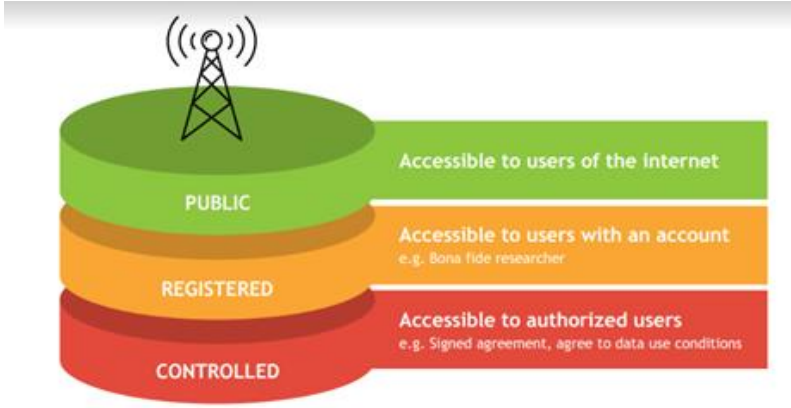
Beacon API v1

Global Alliance for Genomics & Health

The Beacon API can be implemented as a web-accessible service that users may query for information about a specific allele.

Approved: October 3, 2018





Example Users

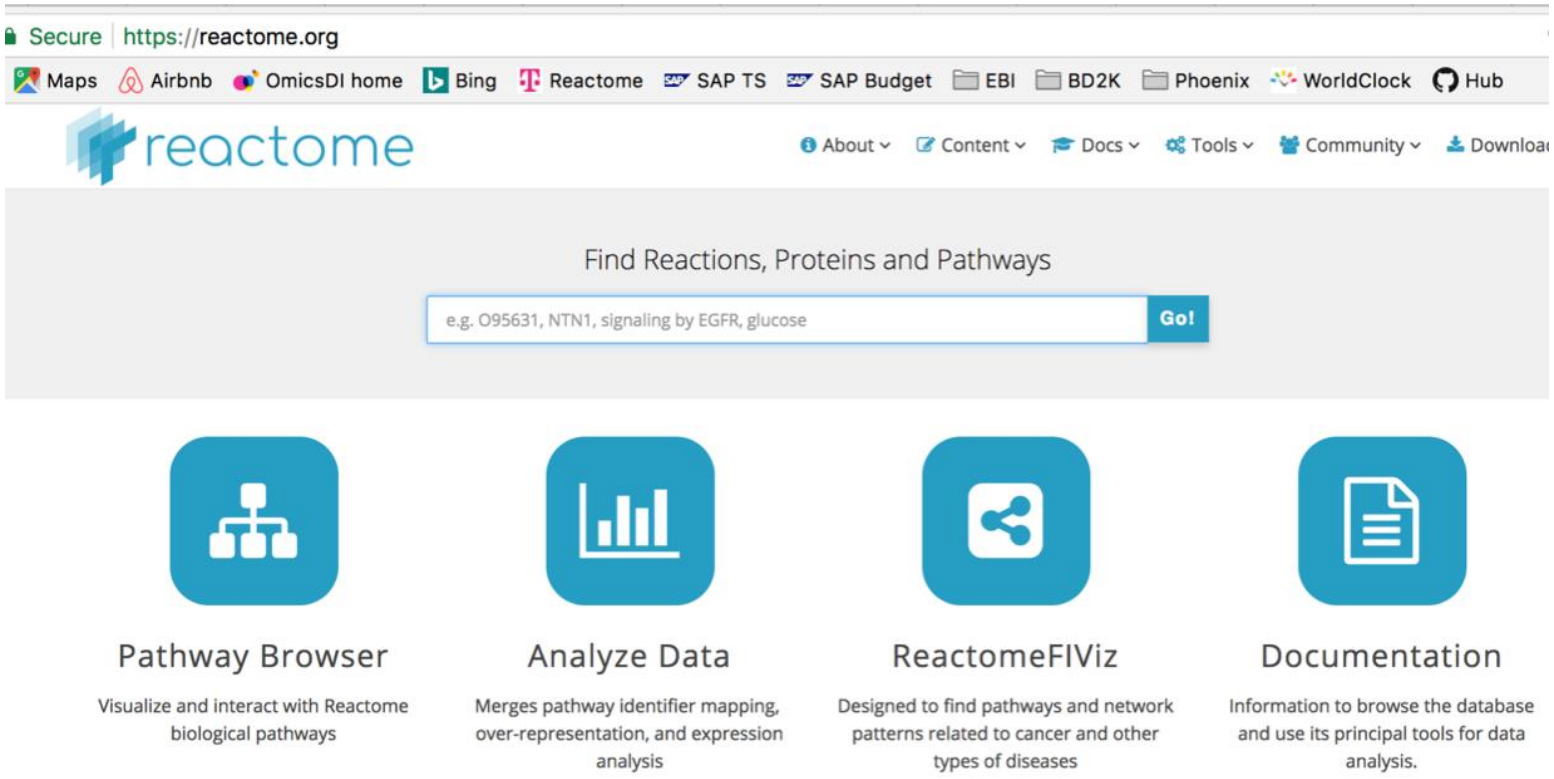
EMBL-EBI · Australian Genomics Health Alliance · SciLifeLab · elixir · University of California Santa Cruz · Broad Institute · European Genome-Phenome Archive · International Cancer Genome Consortium

Two main use cases for Beacons

- **Discoverability** of **experimentally confirmed variants** at the proteome level (those ones that are actually expressed).
  - For completely **open data** – any species
  - **UniProt** – Curated protein sequence data
  - **PRIDE** – Experimental proteomics data coming (mainly) from proteogenomics studies

- For **sensitive human proteomics** data.
  - Controlled-Access datasets (in the future)
  - **Same use case** that for sequencing data, but at the proteome level
- Representation of **more complex data** would be possible as well (PTMs, expression levels, etc)

Reactome – manually curated human pathways



Biological pathways can be used as common reference system to integrate different types of omics data (e.g. proteomics, genomics, small molecules)

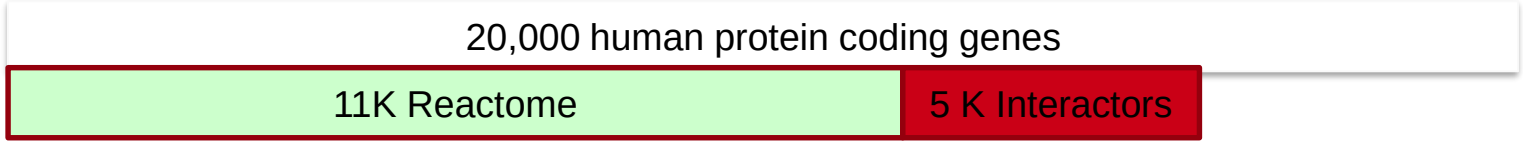
Fabregat A et al. The Reactome Pathway Knowledgebase. NAR 2018

Reactome coverage

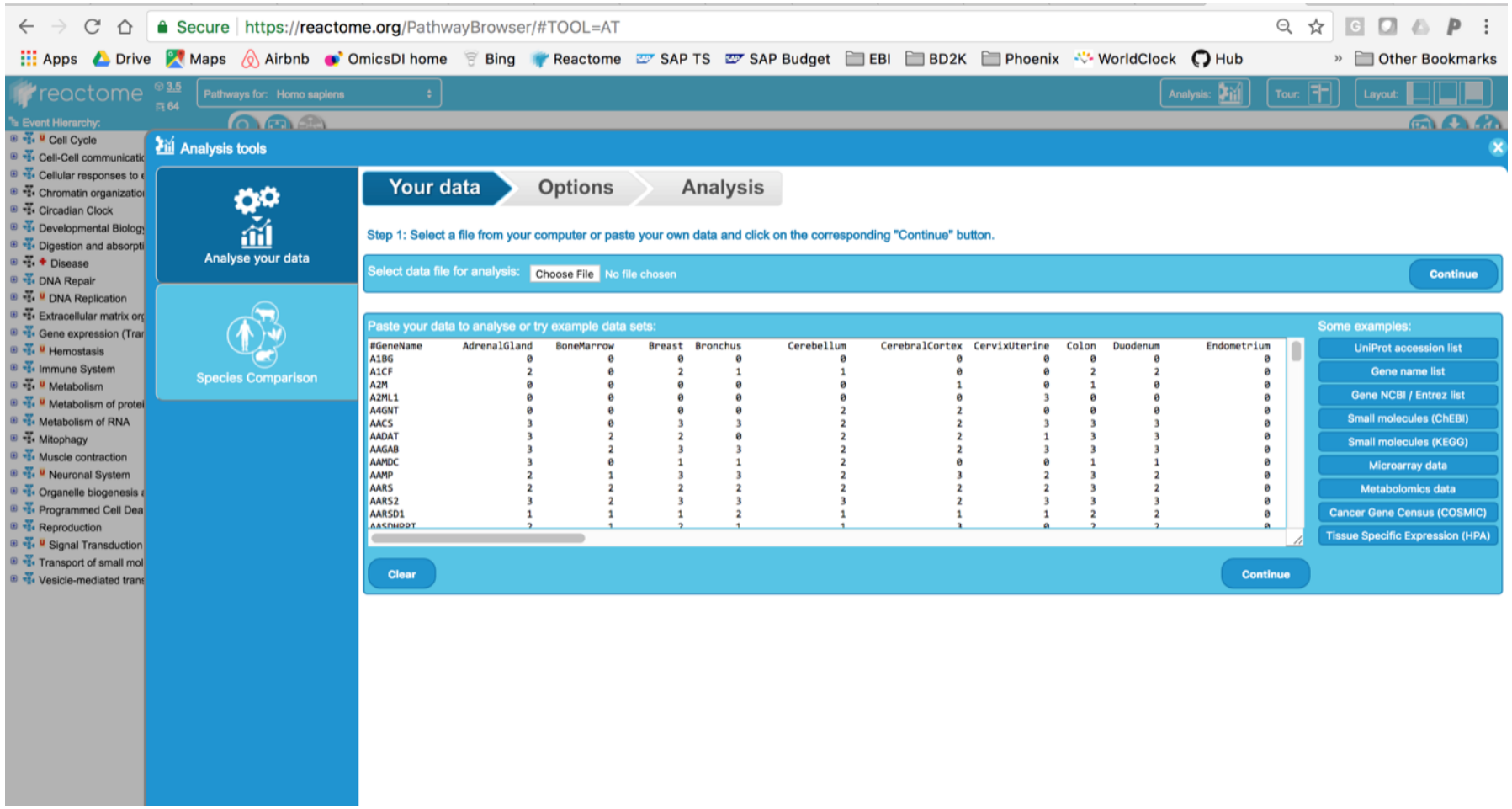
Version 70 released on September 9, 2019

2,287	12,608	10,860	1,856	222	30,398
Human Pathways	Reactions	Proteins	Small Molecules	Drugs	Literature References

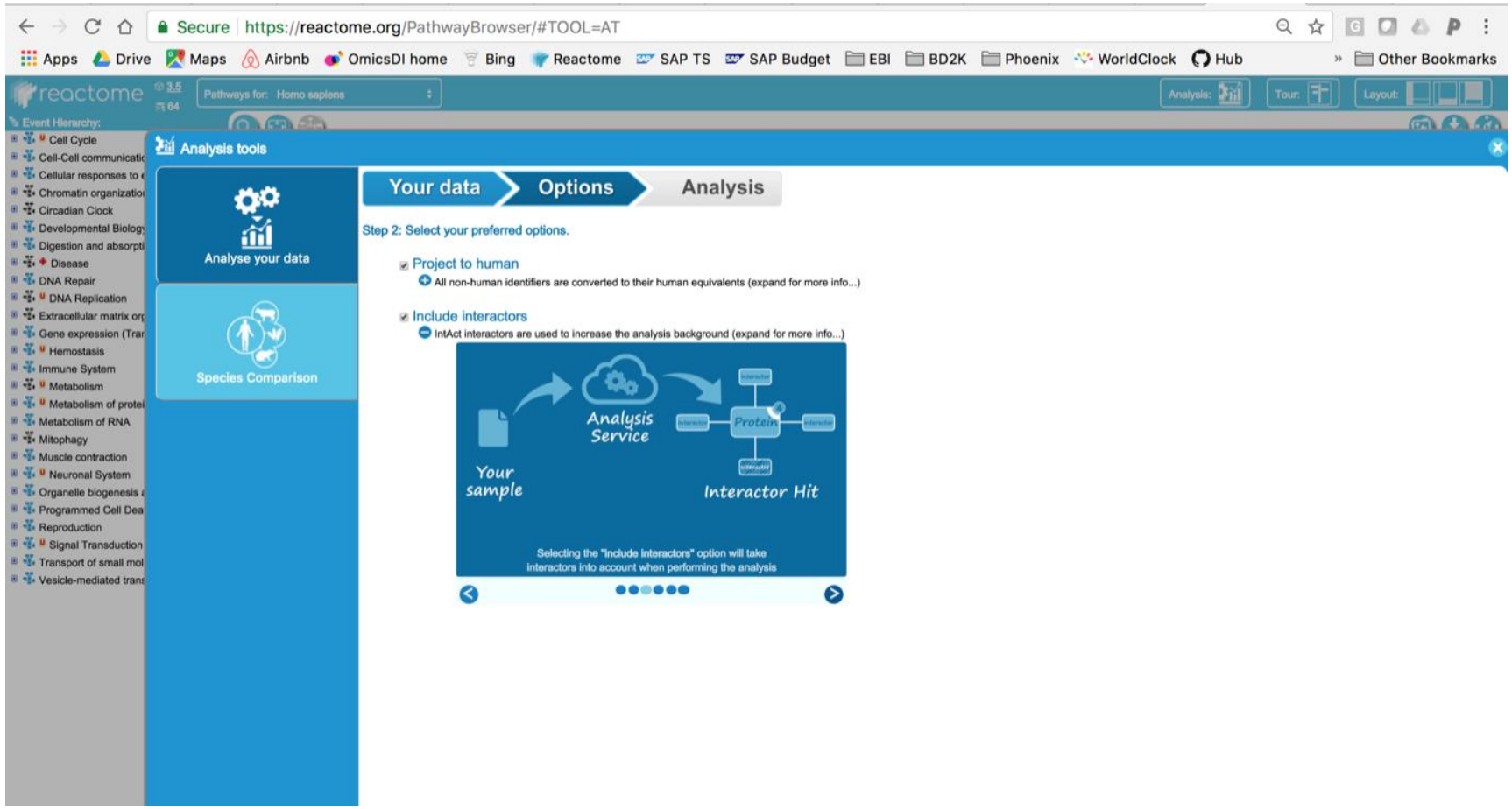
- Ca. 5,000 high confidence interactors from IntAct



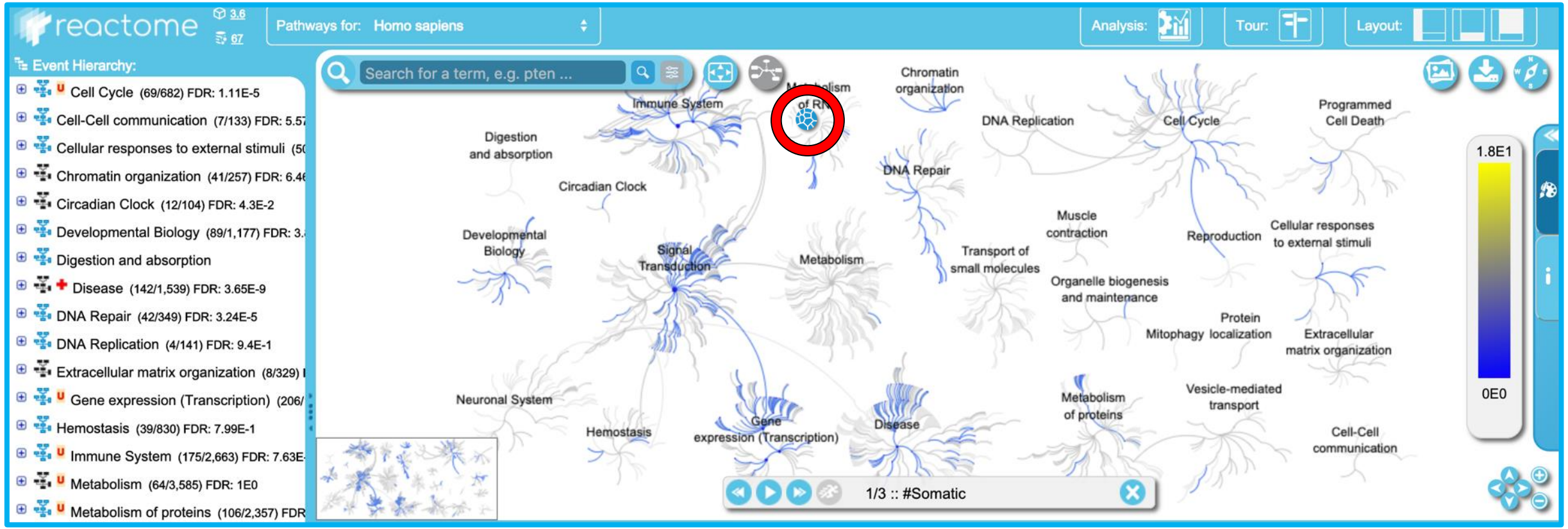
Pathway Analysis



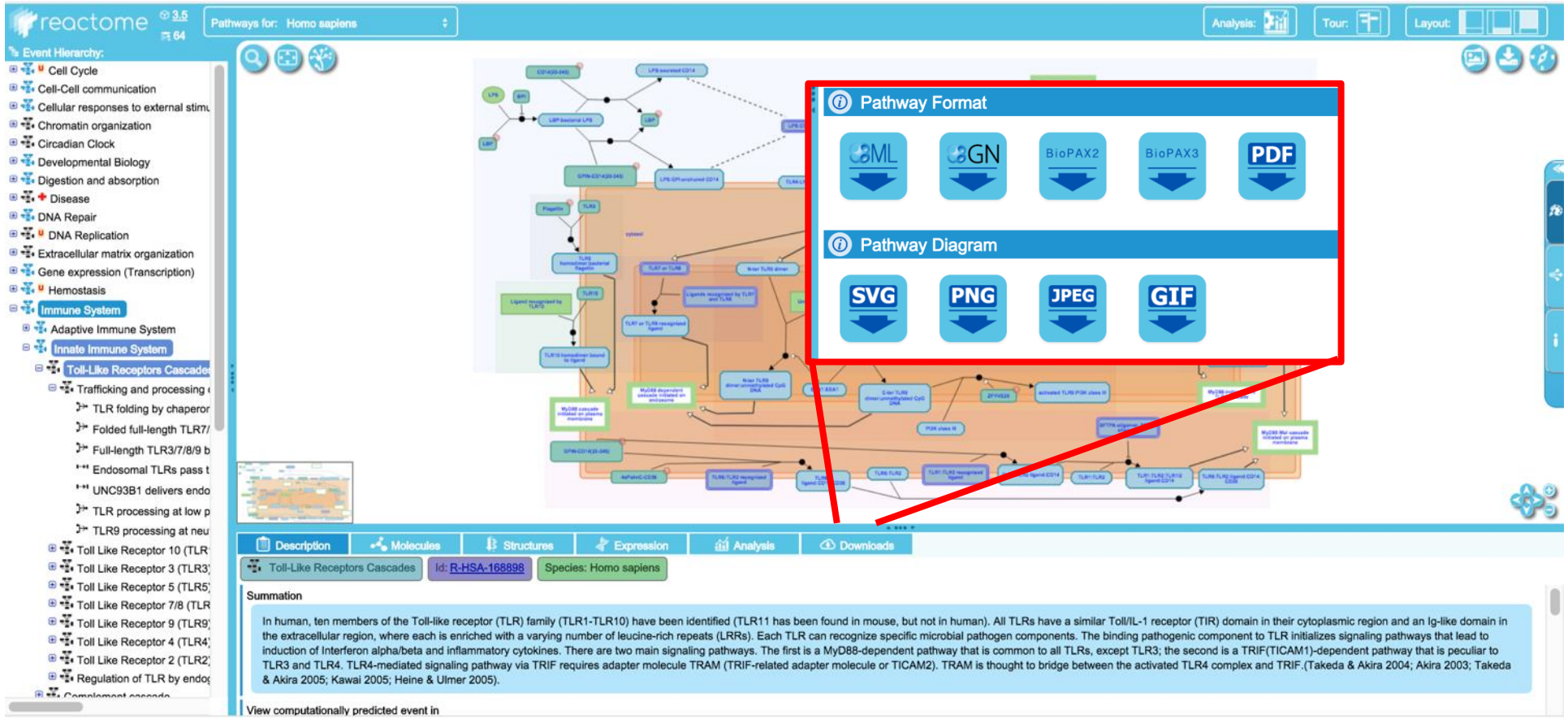
Pathway Analysis



Pathway Analysis: Results overview



Multiple download formats



Overview of standards for pathway related information

Core Standards

Standards for Knowledge Representation

BioPAX

SBOL

Standards for Visual Representation

SBGN

SBOL VISUAL

Standards for Models and their Analyses

SBML

CellML

NeuroML

SED ML

Associated Standards

Used by core standards

Projects

Infrastructure

BioModels.net qualifiers

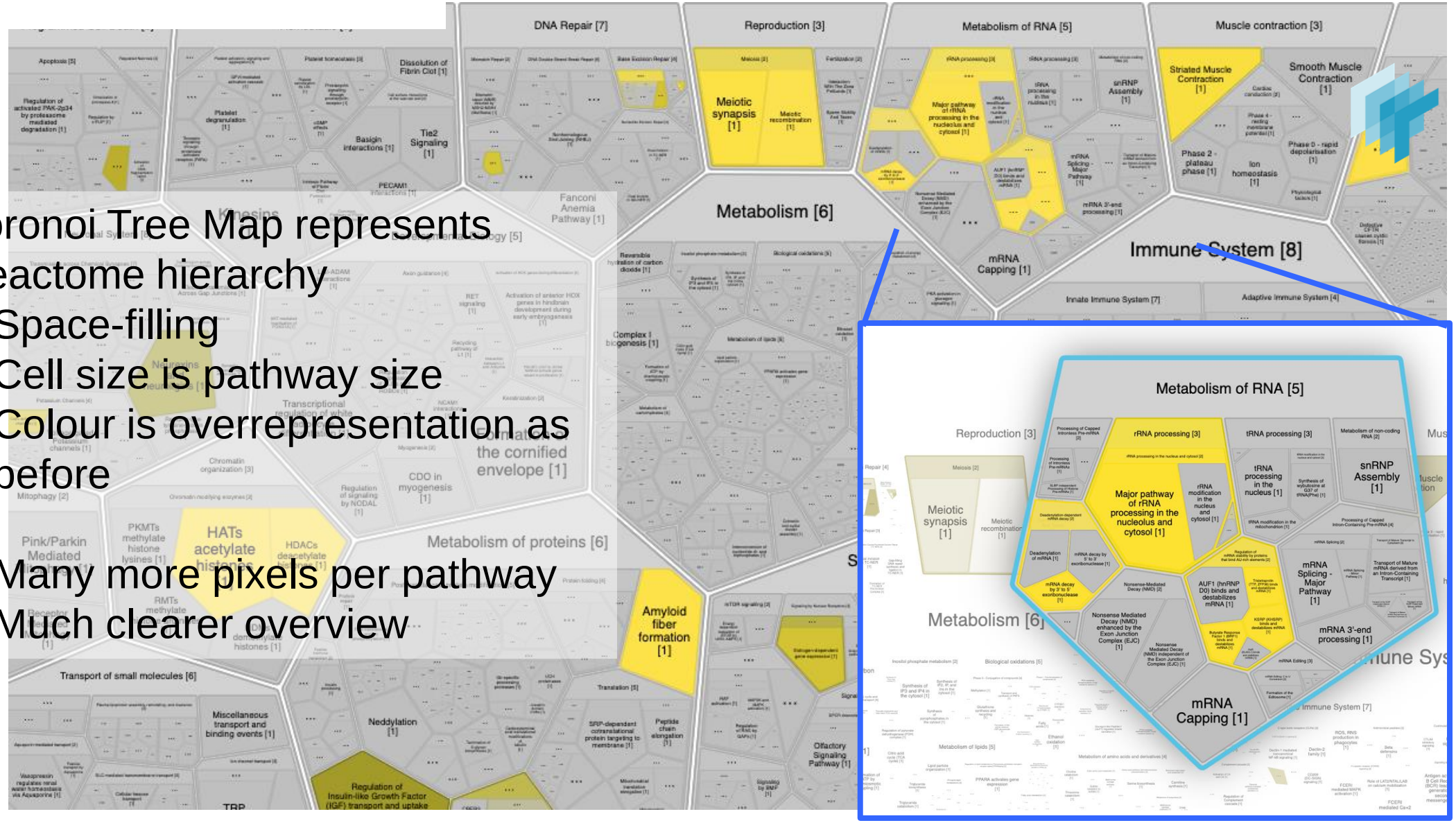
Controlled Vocabularies

SBO

KiSAO

adapted from:
Schreiber F, Bader GD, Gleeson P, Golebiewski M, Hucka M, Le Novère N, Myers C, Nickerson D, Sommer B, Walthemath D:
Specifications of Standards in Systems and Synthetic Biology: Status and Developments in 2016
J Integr Bioinform. (2016) 13:289. doi: 10.2390/biecoll-jib-2016-289

Results overview



Voronoi Tree Map represents Reactome hierarchy

- Space-filling
- Cell size is pathway size
- Colour is overrepresentation as before
- Many more pixels per pathway
- Much clearer overview

Summary

- We have covered the activities of the PSI and the main data formats that it has developed over the years
- Another model: development of GA4GH standards
- Data integration ideas
- Take home message: the development of data standards requires a lot of time and effort.

Acknowledgements

Journal of proteome research

Perspective

pubs.acs.org/jpr

Proteomics Standards Initiative: Fifteen Years of Progress and Future Work

Eric W. Deutsch,[*,†] Sandra Orchard,[‡] Pierre-Alain Binz,[§] Wout Bittremieux,[∥] Martin Eisenacher,[⊥] Henning Hermjakob,[‡,○] Shin Kawano,[◆] Henry Lam,[□▼] Gerhard Mayer,[⊥] Gerben Menschaert,[#] Yasset Perez-Riverol,[‡] Reza M. Salek,[‡] David L. Tabb,[+] Stefan Tenzer,[¶] Juan Antonio Vizcaíno,[‡] Mathias Walzer,[‡] and Andrew R. Jones[▽]

[†]Institute for Systems Biology, Seattle, Washington 98109, United States

[‡]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

[§]CHUV Centre Hospitalier Universitaire Vaudois, 1011 Lausanne, Switzerland

[∥]Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, 2020 Antwerp, Belgium

[⊥]Medizinisches Proteom Center (MPC), Ruhr-Universität Bochum, D-44801 Bochum, Germany

[#]Lab of Bioinformatics and Computational Genomics (BioBix), Faculty of Bioscience Engineering, Ghent University, 9000 Ghent, Belgium

[▽]Institute of Integrative Biology, University of Liverpool, South Wirral L64 4AY, United Kingdom

[○]State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, National Center for Protein Sciences, Beijing, Beijing 102206, China

[◆]Database Center for Life Science, Joint Support Center for Data Science Research, Research Organization of Information and Systems, Kashiwa, Chiba 277-0871, Japan

[¶]Institute for Immunology, University Medical Center of the Johannes-Gutenberg University Mainz, 55131 Mainz, Germany

[+]SA MRC Centre for TB Research, DST/NRF Centre of Excellence for Biomedical TB Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

[□]Division of Biomedical Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, P. R. China

[▼]Department of Chemical and Biomolecular Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, P. R. China

and many others….

Andy Yates (EMBL-EBI, GA4GH)
H. Hermjakob (EMBL-EBI, Reactome)

Nicolas Rodriguez (Babraham)