Tools for genome annotation
SAPP/GBOL/Empusa

Peter Schaap
Laboratory of Systems and Synthetic Biology
Fairbydesign.nl

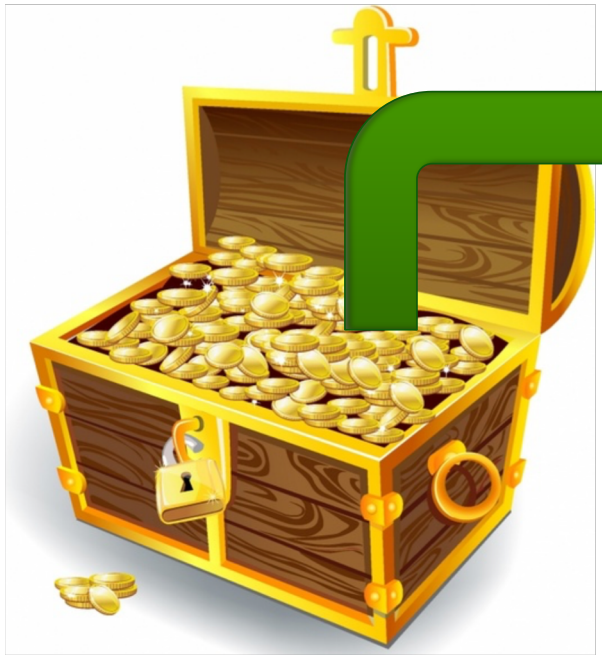
WAGENINGEN
UNIVERSITY & RESEARCH
100years
1918 — 2018

Laboratory of Systems and Synthetic Biology

- Computational Systems Biology –
  - Semantic and Model-based Systems Biology
- Microbial Systems and Synthetic Biology
  - Wet & Dry lab cycle

- How genome information leads to function
- How microbial metabolic processes are regulated and adapt in extant species,
- How microbial organisms and ecosystems respond to (a)biotic environmental cues
- How they can be manipulated to enhance the yield of desired products or to diminish their pathogenicity.

SAPP: Problem definition



Bottom-up
Sequence based pan-genomic analysis - Orthology clustering

~ 50 -100 genomes belonging to a certain taxonomic rank

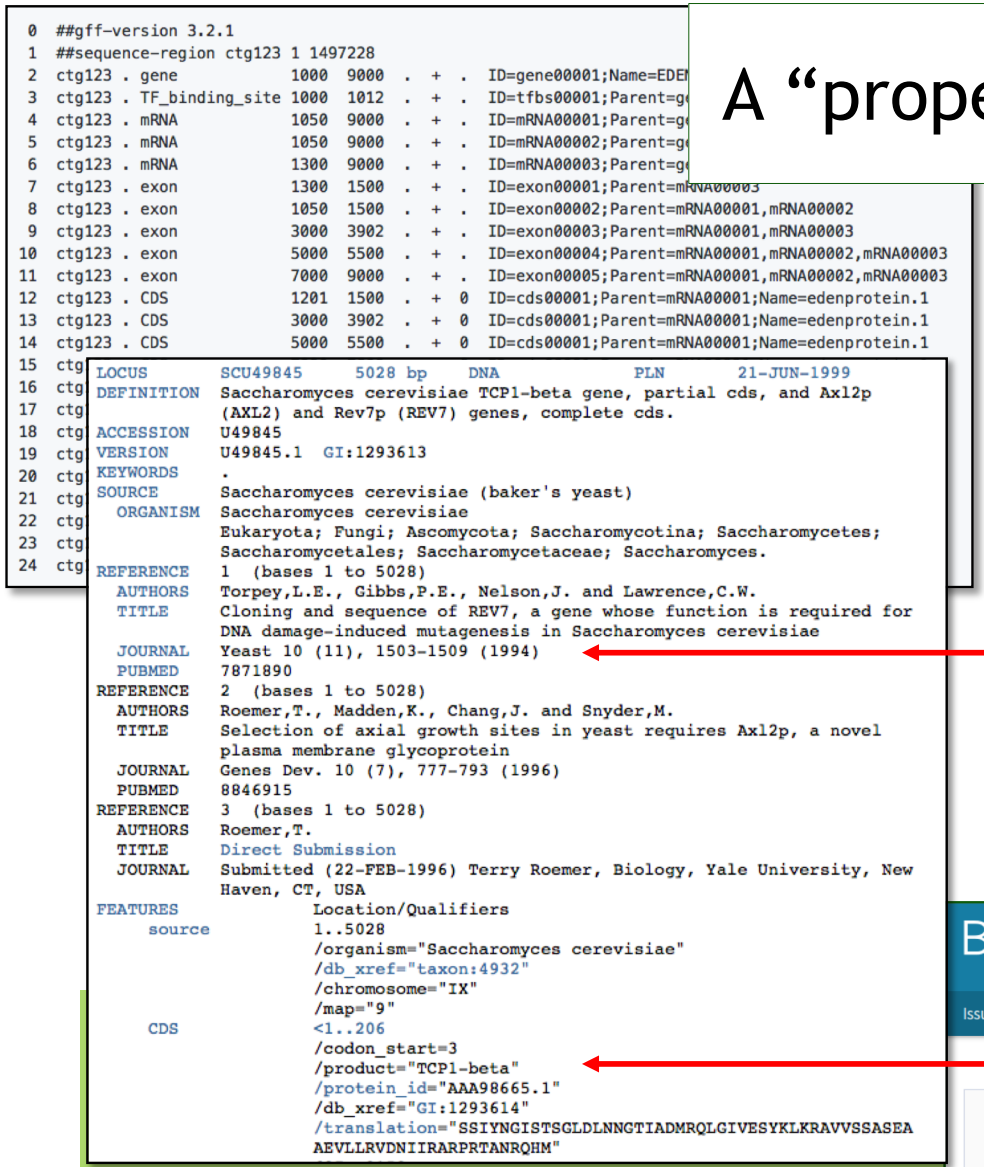
180.000 bacterial genomes and growing exponentially

How to Really Unlock this Treasure Chest?
Key Word: Interoperability

- Bottom-Up: From sequence to function
- Sequences are highly Interoperable
  - Available in an accepted, specified format (Fasta-format and IUPAC code for representation of AA and Nucl.)
- Limitations:
  - Aligning sequences is computational intensive - scales quadraticallly
  - Focus on similarity - what is in common

- Top-Down: From function to sequence
- Both common and unique features
- Very scalable (linear)
- Limitations: Currently not Interoperable
  - Functional descriptions are in a 'free format' Not machine readable
  - Data provenance of derived data not (well) presented in the DDBJ/ENA/GenBank Feature Table

A “proper” format



```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDE
ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=g
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=g
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=g
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=g
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
```

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
```

- Direct linkage of dataset-wise and element-wise provenance with the predictions
  - (What was the tool used and what was the confidence score?)
- Mining enabled
- Query enabled

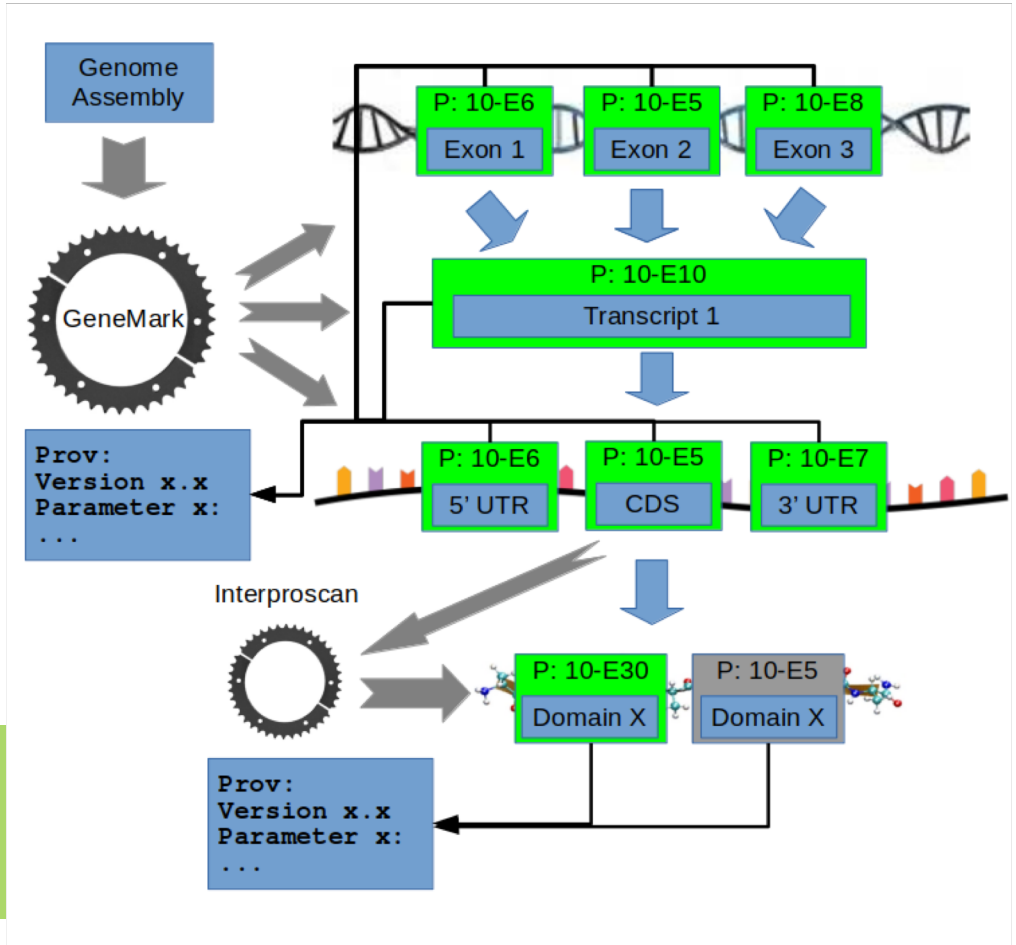
Top-Down; Interoperable Annotation files

computational predictions linked with confidence scores

Example questions: Given the 180.000 sequenced genomes

- What other functions coexist in species that have a desired trait X?
- Which enzymes are available in gram+ bacteria that can catalyze reaction Y (maybe with different cofactors) using a E= $1e^{-50}$ threshold for the domain?

Requires a resource of consistently annotated genomes that can be mined for data and meta-data



FAIR data management: FAIR-ified versus FAIR by design

- FAIR-ified data: deals with making heterogenous data from experiments FAIR.
- This is achieved by using tools such as SEEK and RightField that add Minimal Information for all these types of data in a standardized way/ e.g. the Just Enough Results Model Ontology (JERM) in RDF.

- FAIR by design: works best with computationally derived data. An example of such data is genomic information described in the GBOL ontology. Output = linked (meta) data
- What both approaches have in common is that they use RDF to store the (meta)data and an ontology to standardize the way data and metadata is linked

https://seek4science.org http://gbol.life

Requirements for Interoperable genome mining

- A semantic annotation platform that incorporates common tools and stores the prediction and provenance in “proper” format. **SAPP**
- A graph database that can be mined: **SAGERP**
- A definition of the “proper format”: definitions of biological terms and their relationships: **GBOL ontology**
- Interface to use the ontology: **GBOL stack**
- Tools to develop all of these: **EMPUSA**

- SAPP is the only thing a user would need to use to annotate a genome
- Sager-P is the only thing a user would need to mine the data

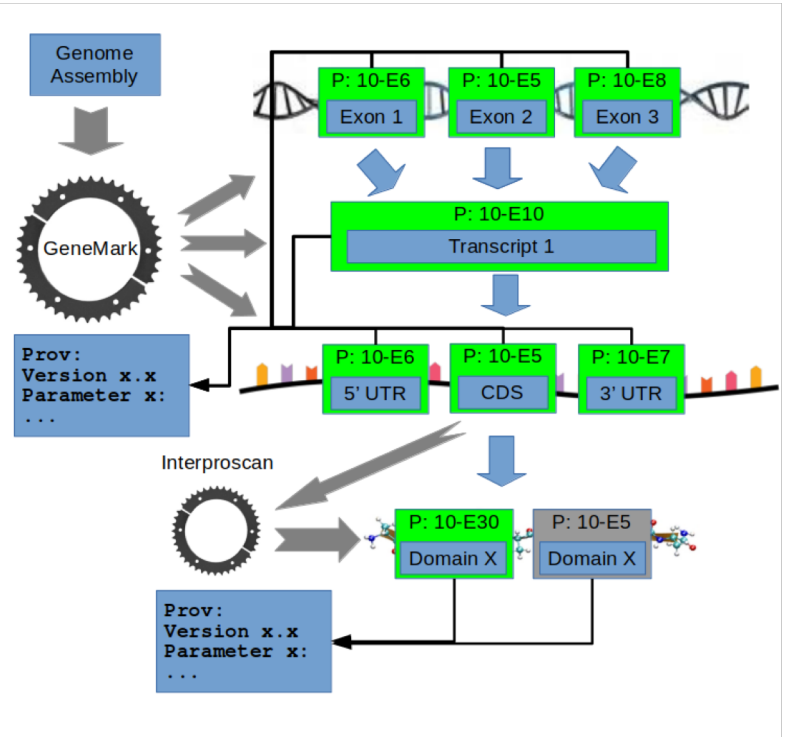
Codebase http://www.gitlab.com/Empusa
Documentation and tutorials http://empusa.org.

SAPP: Annotation information storage

- Wrapper to commonly used annotation tools (prokaryotes and eukaryotes) and generates FAIR-by-design data
- Examples:
  - Uniform annotation of over 100.000 bacterial species.
  - Uniform annotation of salmonoids (fish)

Koehorst et al Bioinformatics 2017
https://gitlab.com/sapp

Documentation:
https://sapp.gitlab.io

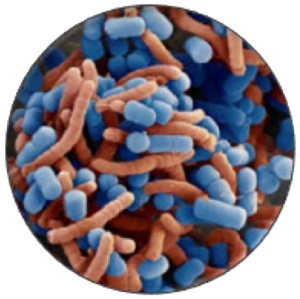
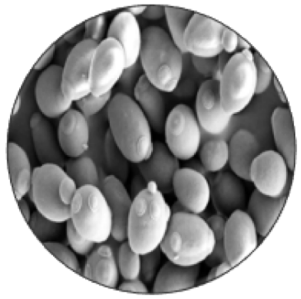


Modular design: Existing tools that query the triple store for input and directly present their output in the RDF data model

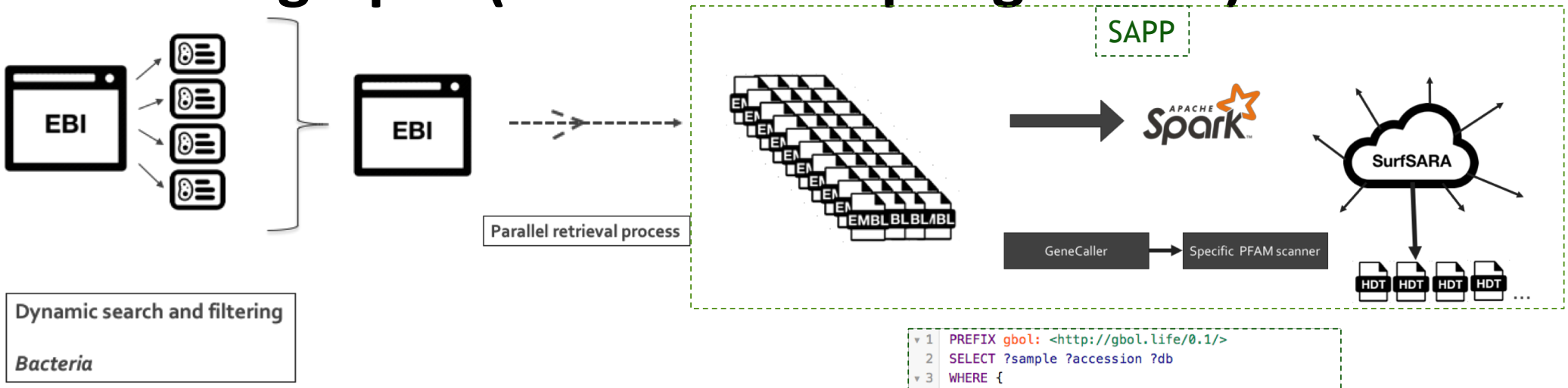
Conversion types
- EMBL / GenBank
- FASTA
- GFF
- QTL
- VCF
- …

Genetic elements
- Gene prediction
- tRNA/rRNA
- Crispr
- …

Functional annotation
- BLAST
- Enzyme predictions
- Domain annotation
- Signal peptides
- Transmembrane
- Localization
- …

High Throughput Re-annotation in SAPP -> 80.000 subgraphs (One HDT file per genome)



Functional domains retrieval

SAGERP

```
PREFIX gbol: <http://gbol.life/0.1/>
SELECT ?sample ?accession ?db
WHERE {
    ?sample a gbol:Sample .
    ?dnaobject gbol:sample ?sample .
    ?dnaobject gbol:feature ?gene .
    ?gene gbol:transcript ?transcript .
    ?transcript gbol:feature ?cds .
    ?cds gbol:protein ?protein .
    ?protein gbol:feature ?domain .
    ?domain gbol:xref ?xref .
    ?xref gbol:db <http://gbol.life/0.1/db/pfam> .
    ?xref gbol:accession ?accession .
} LIMIT 10
```

A HDT file = a turtle (ttl) file in binary format

WAGENINGEN UNIVERSITY & RESEARCH

Use cases:

Computational genomics:
In silico Bioprospecting

&

Organizing QTL data

1, 3-propanediol candidate species

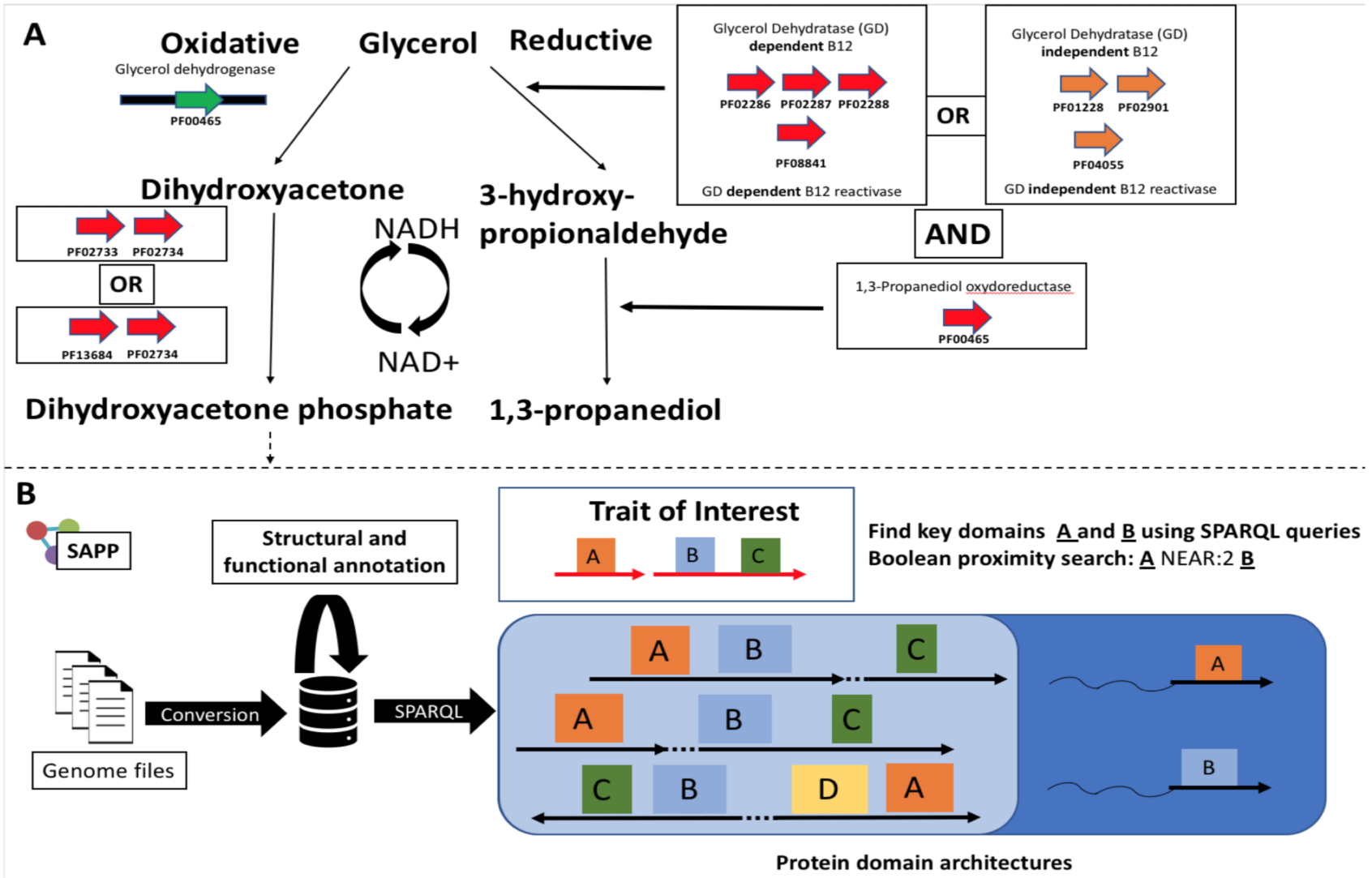


Table 7.2: Properties of key domains involved in glycerol dissimilation in 1,3 PD producers

Domain	Mean Copy Number*	Proximity Search query (compounds and distance)
	Oxidative pathway	
PF02733 (DAK1) PF13684 (DAK1_2) PF02734 (DAK2)	2.2 1.6 2.7	DAK1 AND DAK2 OR DAK1_2 AND DAK2 (immediately adjacent)
	B12-dependent reductive Pathway	
PF00465 PF02286 PF02287 PF02288 PF08841	16.6 1.7 1.7 3.7 1.7	All, within 20.000 up or downstream of the B12 dependent dehydratase domains
	B12-independent reductive Pathway	
PF01228 PF02901 PF04055	13.8 14.4 32.4	All, within 20.000 up or downstream of the B12 independent dehydratase domains

Gene fusion and fission event are frequently observed

A) Operonic structure of the training set

SPARQL Query

B) Operonic structure of the selection used for wet-lab verification

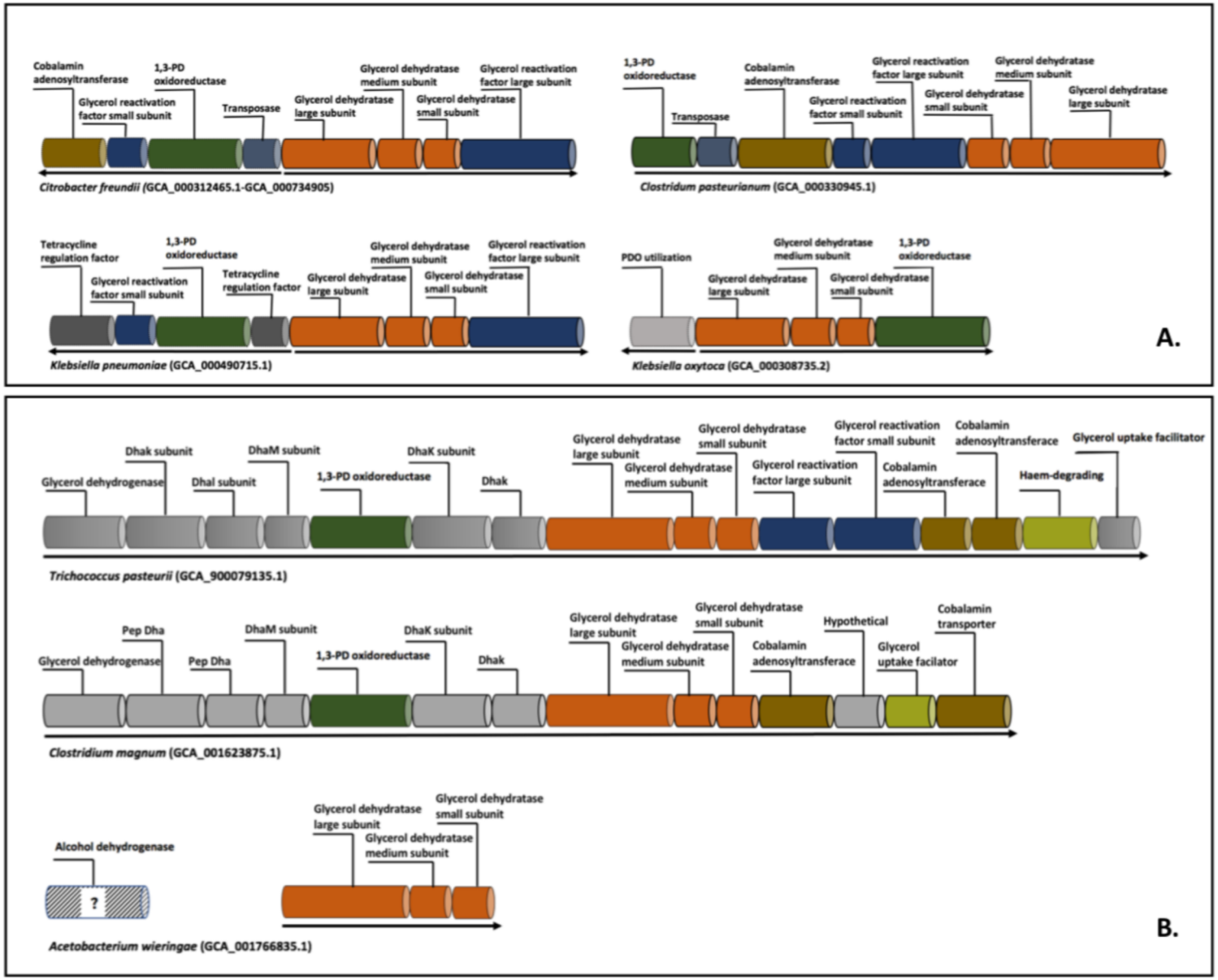
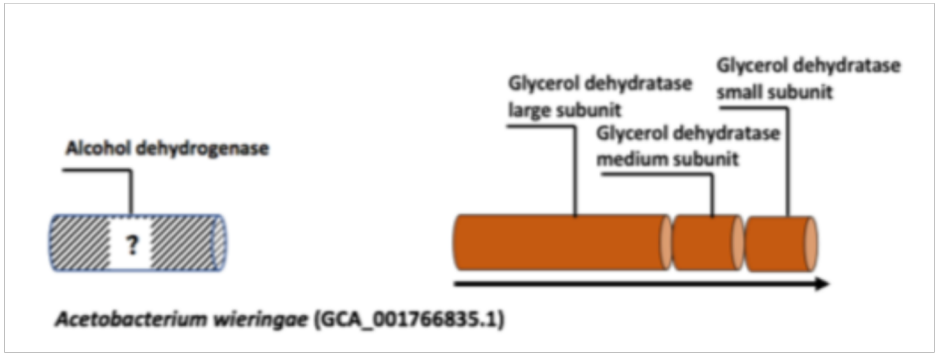


Table 7.3: 1,3-propanediol and acetate yields from glycerol fermentations of selected strains.

Organism	Genome assembly ID	1,3-PD (mol/mol)	Acetate (mol/mol)	OD
Acetobacterium wieringae DSM 1911	GCA_001766835.1	0.18	0.94	0.179
Carnobacterium funditum DSM 5970	GCA_000744185.1	0.33	0.07	0.266
Clostridium magnum DSM 2767	GCA_001623875	0.56	0.03	0.180
Trichococcus pasteurii DSM 2381	GCA_900079135.1	0.66	0.12	0.325



Conclusions

- Semantic Systems Biology and model-based Systems Biology are data integration and analysis approaches that strive to achieve complementary goals.
- Model-based Systems Biology uses mathematical modelling to analyse biological data.
- Integration and sharing of data, information and knowledge is in the realm of Semantic Systems Biology.
- The deliberate exploitation of Semantic Web technologies for integration and sharing of heterogeneous bio-data sources with computational predictions and associated meta-data will lead to:
  - the development of new, testable hypotheses
  - the ability to directly link data and data provenance (FAIR by design)
  - new ways for computational support in quality checking of computationally inferred annotations. (meta-analysis of element-wise provenance)

Acknowledgments

- Jesse van Dam -> Thesis defence 23 January 2019
- Jasper Koehorst -> Thesis defence 25 January 2019
- Maria Suarez Diez, Bart Nijsse, Niels Zondervan
- MycoSynVac: CRG Luis Serrano's group
- Digital Salmon: Jon Olav Vik
- SynBioChem: Carole Goble, Rainer Breitling and Paul Mulherin
- Manchester University (FAIRdom): Natalie Stanford

WAGENINGEN
UNIVERSITY & RESEARCH

IBISBA 1.0

UNLOCK

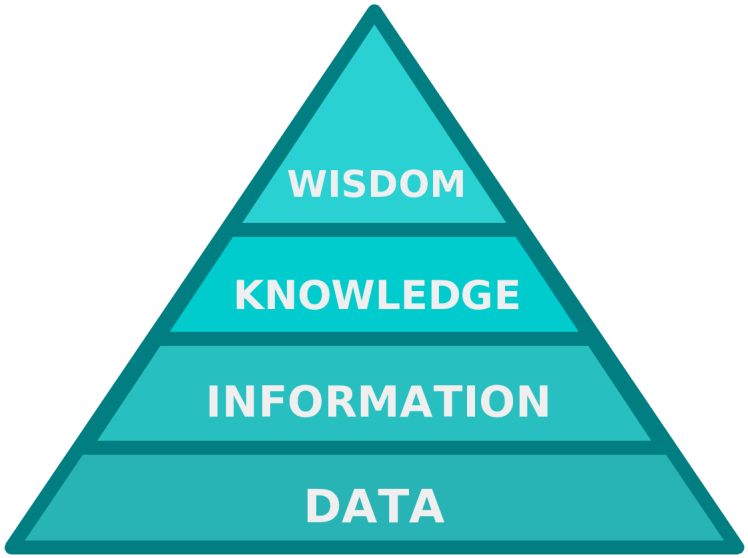
MYCOSYNVAC

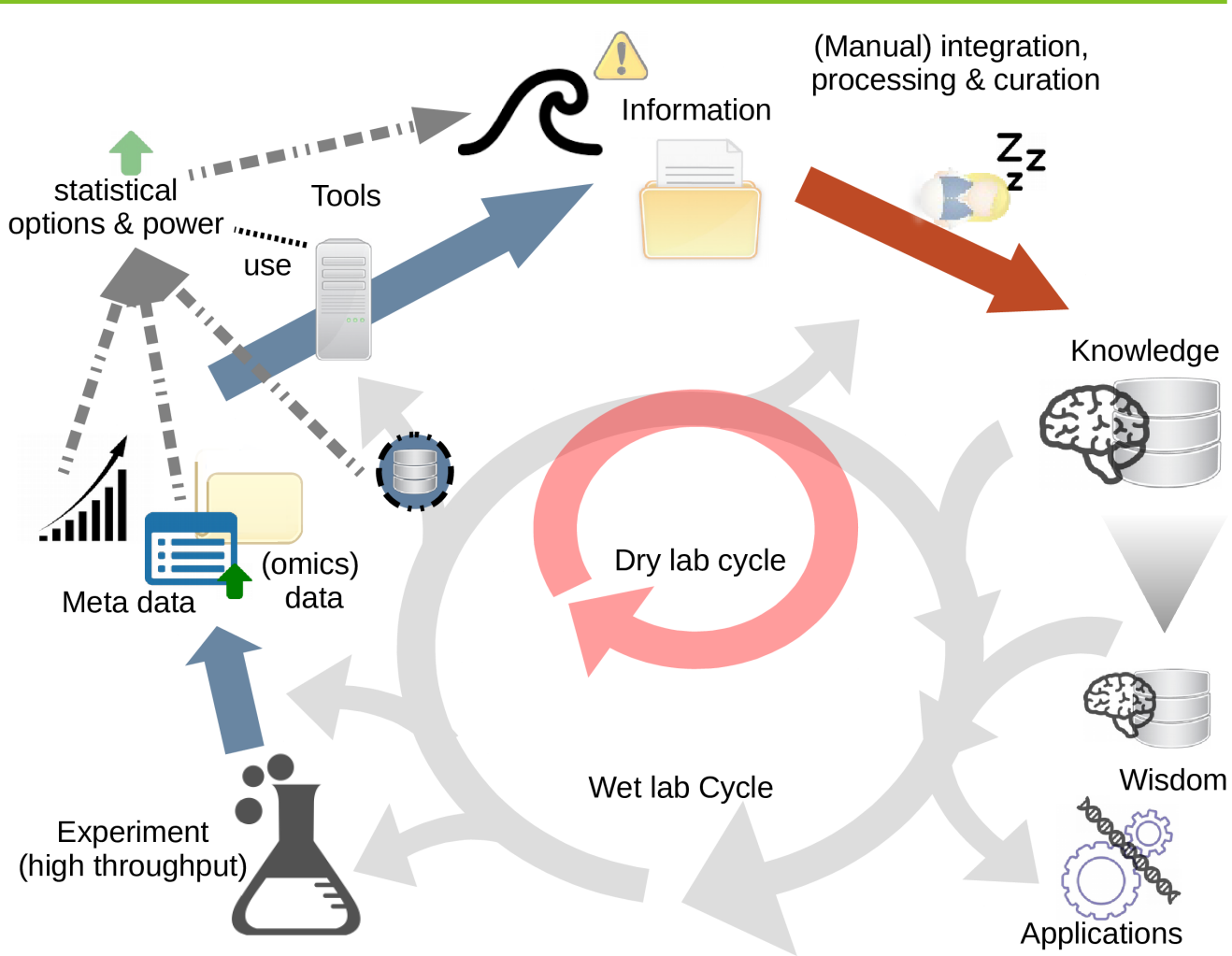
1000100 1010011
DIGITAL SALMON

SYNBIOCHEM
Manchester Synthetic Biology Research Centre
for Fine and Speciality Chemicals

FAIRDOM

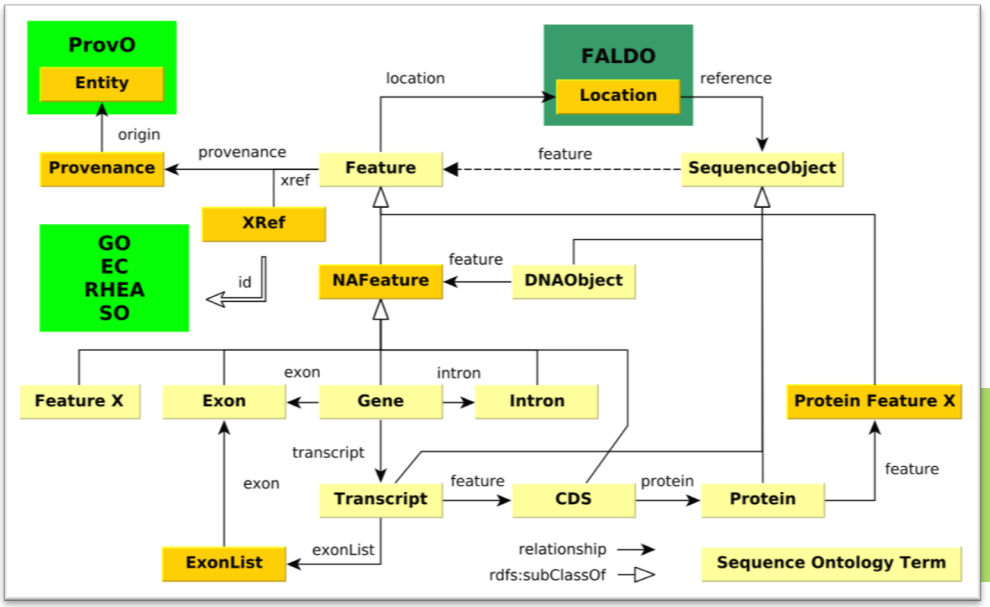
Big/FAIR data at SSB: tool development

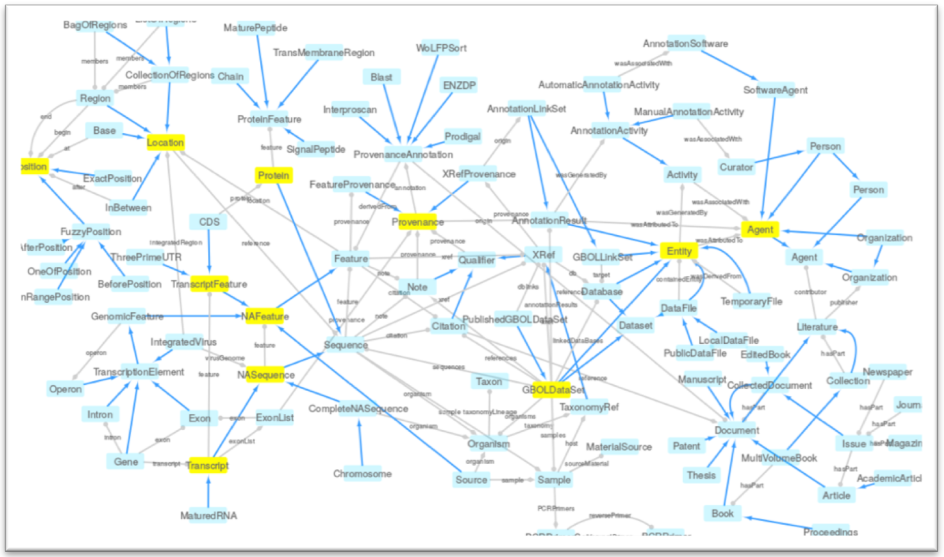




GBOL: Genome Biology Ontology Language

Sub domain	Classes	Properties
Genomic locations	16	17
Genes		
transcripts and features	114	133
Document structure	27	107
Dataset-wise provenance	22	54
Element-wise provenance	5	9
BIBO	59	90





Embedded with existing ontologies.

Van Dam et al. Journal of biomedical semantics 2015

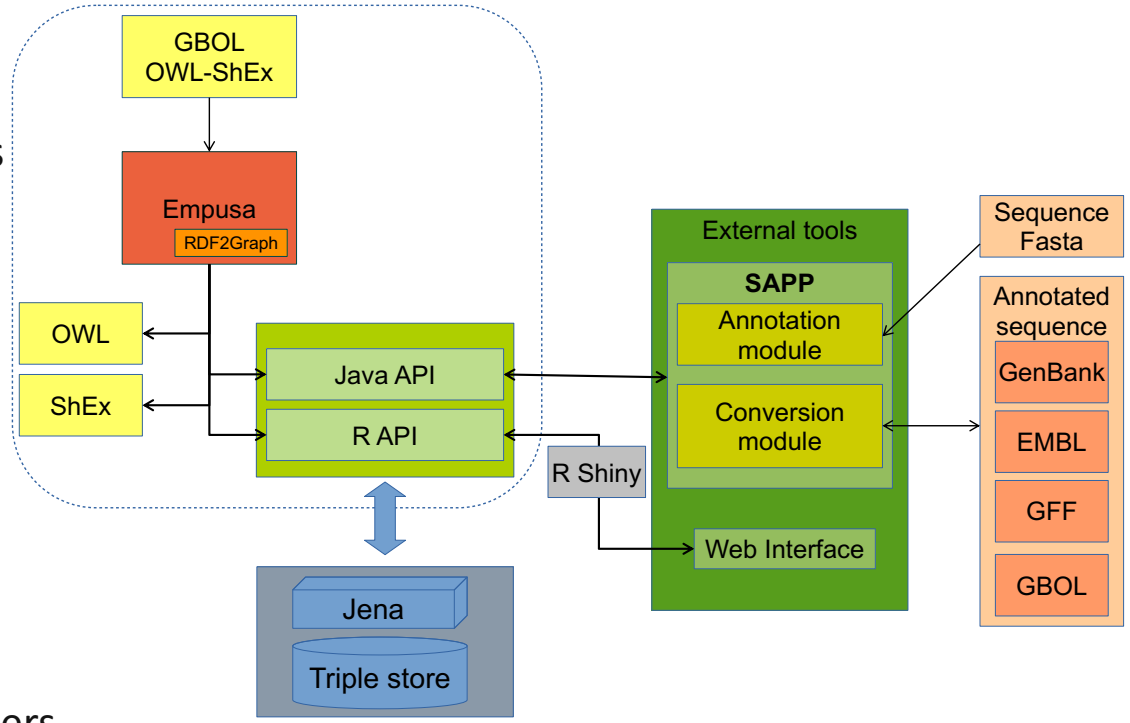
WAGENINGEN UNIVERSITY & RESEARCH

Tool development for FAIR genome annotation

- **SAPP**: an annotation platform
- **SAGERP**: resource with annotated genomes

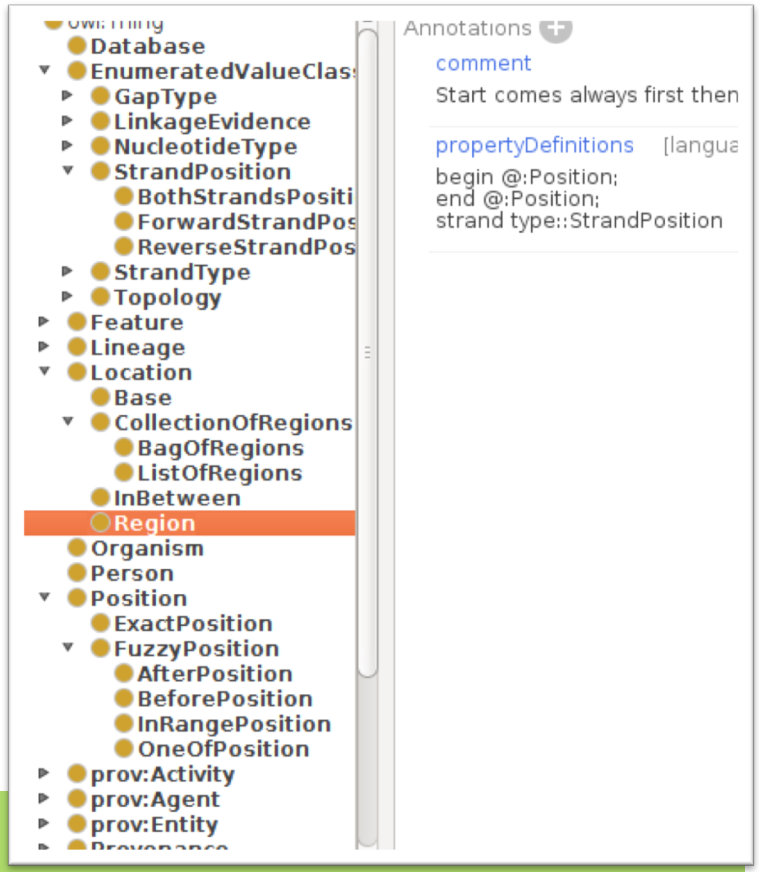
Developer:

- **GBOL stack:**
  - GBOL ontology (backbone)
  - Java/R Api
  - Owl/ShEx
  - Interface gate keeper **Empusa**
- Code generator: **Empusa** useful for developers



Code generation: EMPUSA

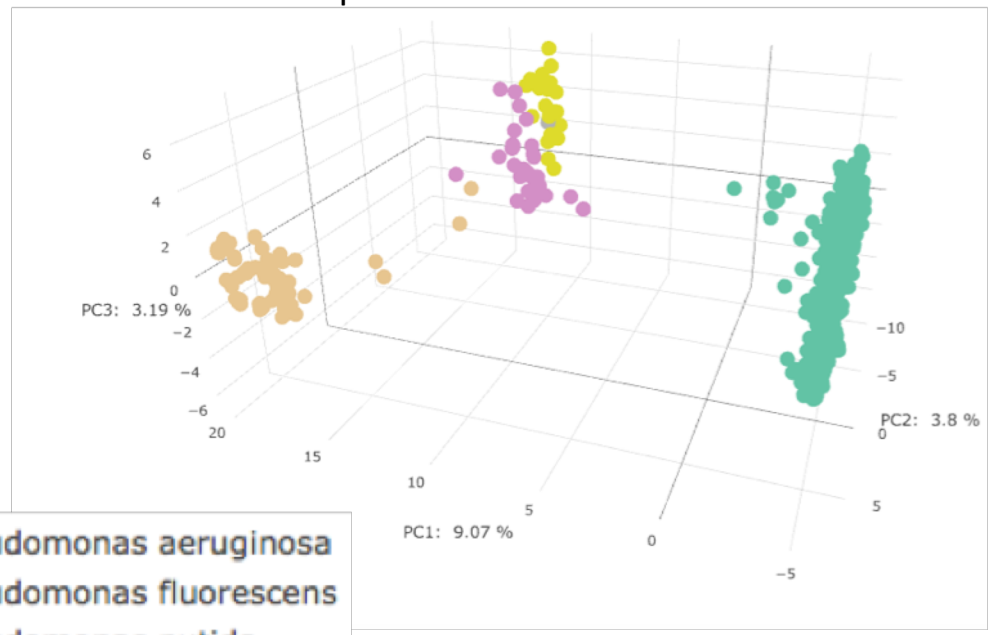
- Linked data graph is free format: Ontology defines structure but does not enforce it.
  - **NEED TO MANTAIN CONSISTENCY ->**
  - **Gatekeeper tool**
- From Ontology (protégé file)
  - OWL + ShEx
- API: Java + R
  - Instance validation included
- > 80.000 lines of code generated



WAGENINGEN
UNIVERSITY & RESEARCH

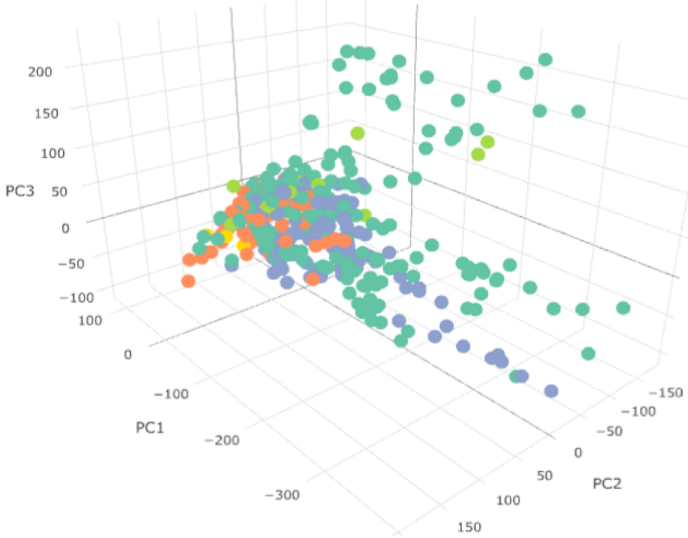
Functional variation

Phylogeny and phenotype relationships with the functional landscape



- Pseudomonas aeruginosa
- Pseudomonas fluorescens
- Pseudomonas putida
- Pseudomonas syringae

Scored phenotypes



- Aerobe
- Anaerobe
- Facultative
- Microaerophilic
- Obligate aerobe
- Obligate anaerobe

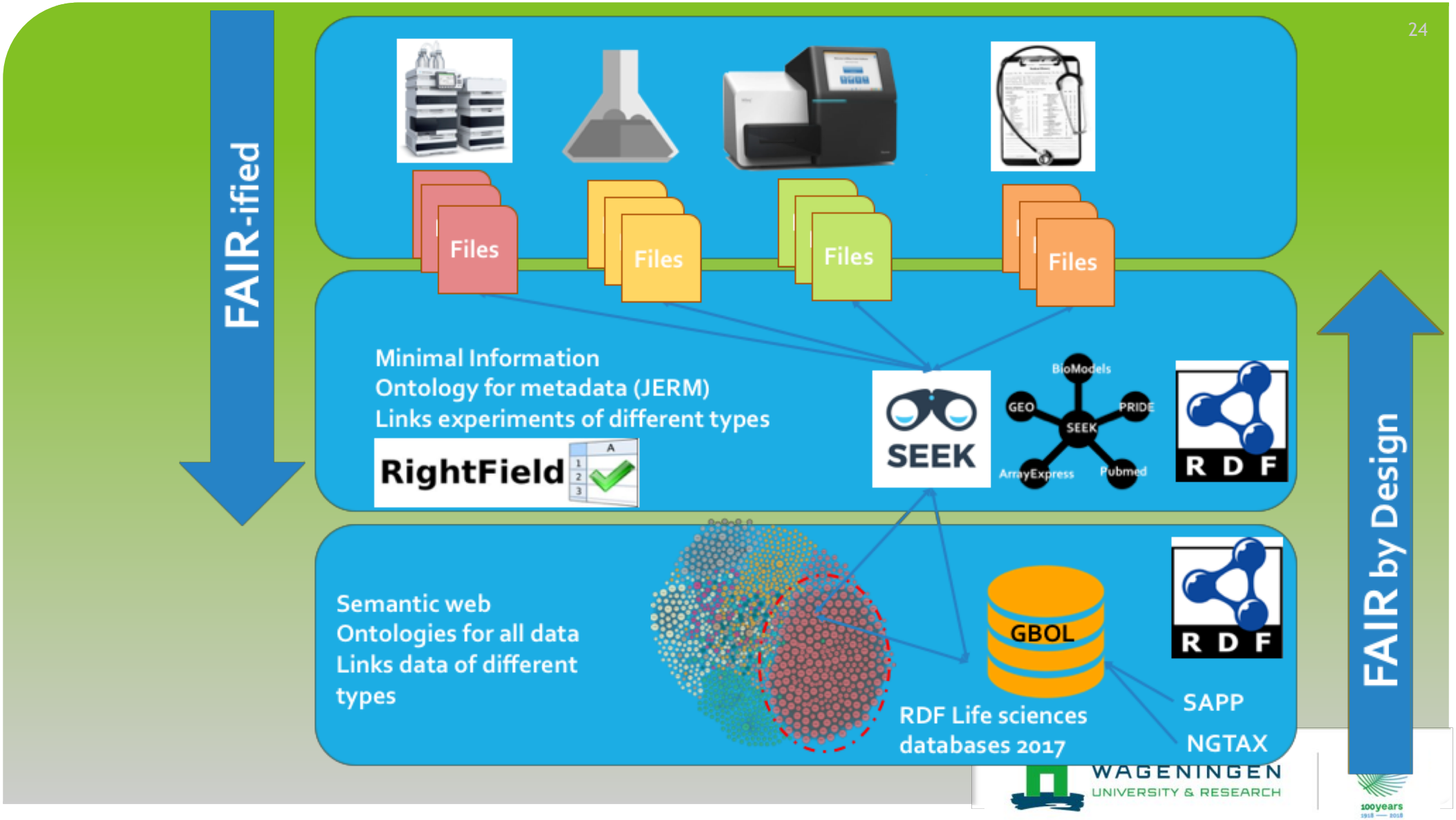
BacDive
The Bacterial Diversity Metadatabase

Pfam

Koehorst, Jasper J., et al. *Scientific reports* 6 (2016):

WAGENINGEN
UNIVERSITY & RESEARCH

100years


FAIR-ified
Files
Files
Files
Files
Minimal Information
Ontology for metadata (JERM)
Links experiments of different types
RightField
SEEK
BioModels
GEO
SEEK
PRIDE
ArrayExpress
Pubmed
R D F
Semantic web
Ontologies for all data
Links data of different types
GBOL
RDF Life sciences databases 2017
R D F
SAPP
NGTAX
FAIR by Design