Tools for genome annotation
SAPP/GBOL/Empusa

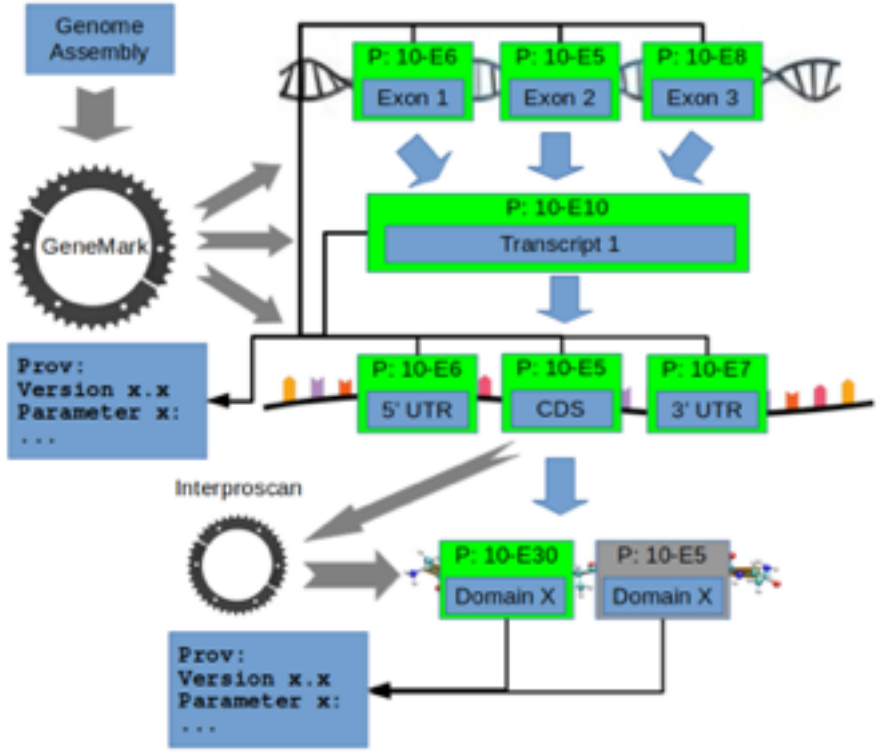
Jasper Koehorst
Laboratory of Systems and Synthetic Biology
Fairbydesign.nl

Annotation information storage

Example questions e.g:

- What (genes) distinguish species that have a desired trait?
- Which enzymes are there that can catalyze reaction X (maybe with different cofactors?

Requires a resource of consistently annotated genomes that can be mined



WAGENINGEN
UNIVERSITY & RESEARCH

Requirements for genome mining

- A semantic annotation platform that incorporates common tools and stores the results 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**: code generator

- 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

A “proper” format

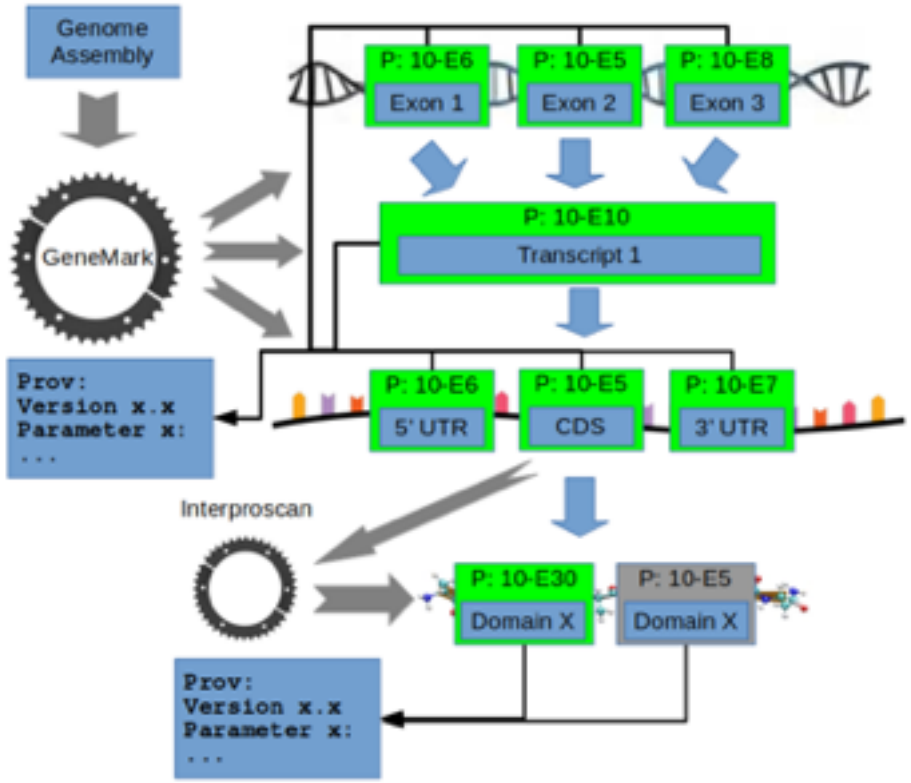
```
0  ##gff-version 3.2.1
1  ##sequence-region ctg123 1 1497228
2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=g
4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=g
5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=g
6  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=g
7  ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=
8  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
9  ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
10 ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11 ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12 ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13 ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14 ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15 ctg
16 ctg
17 ctg
18 ctg
19 ctg
20 ctg
21 ctg
22 ctg
23 ctg
24 ctg
```

```
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"
```

- Dataset-wise and element-wise provenance
- Mining enabled
- Query enabled

SAPP: Annotation information storage

- Wrapper to commonly used annotation tools (prokaryotes and eukaryotes) that generates FAIR data
- Examples:
  - Uniform annotation of over 10 000 bacterial species.
  - Uniform annotation of salmonoids (fish)



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

Documentation:
https://sapp.gitlab.io

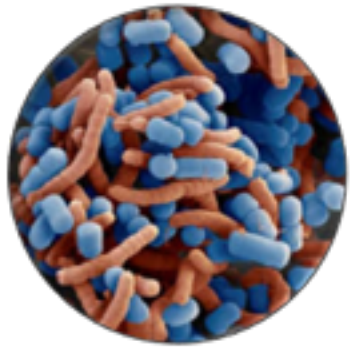
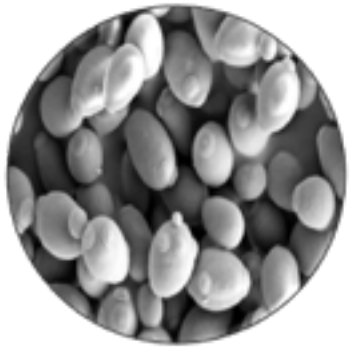
WAGENINGEN
UNIVERSITY & RESEARCH

Modular design

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
- …

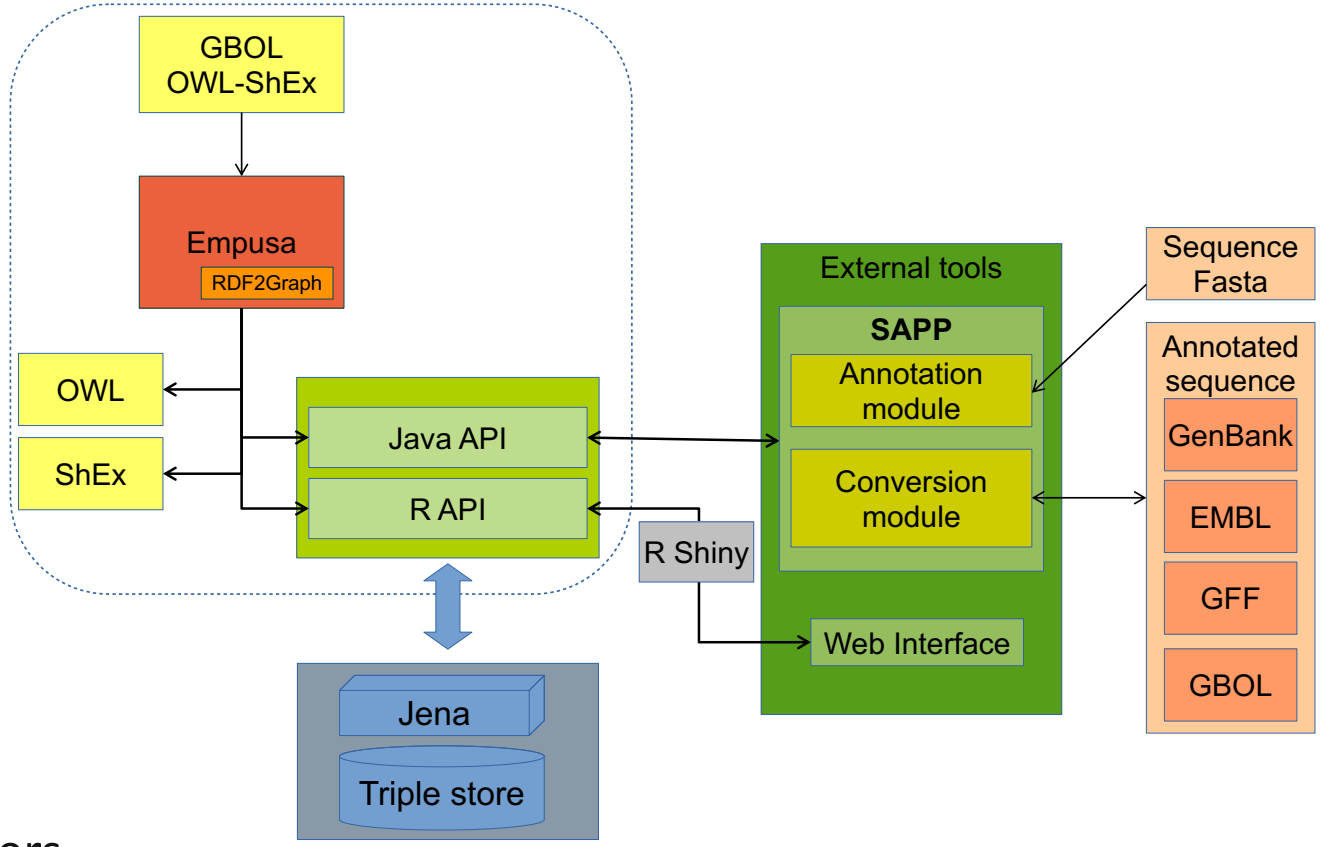



Tool development for FAIR genome annotation

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

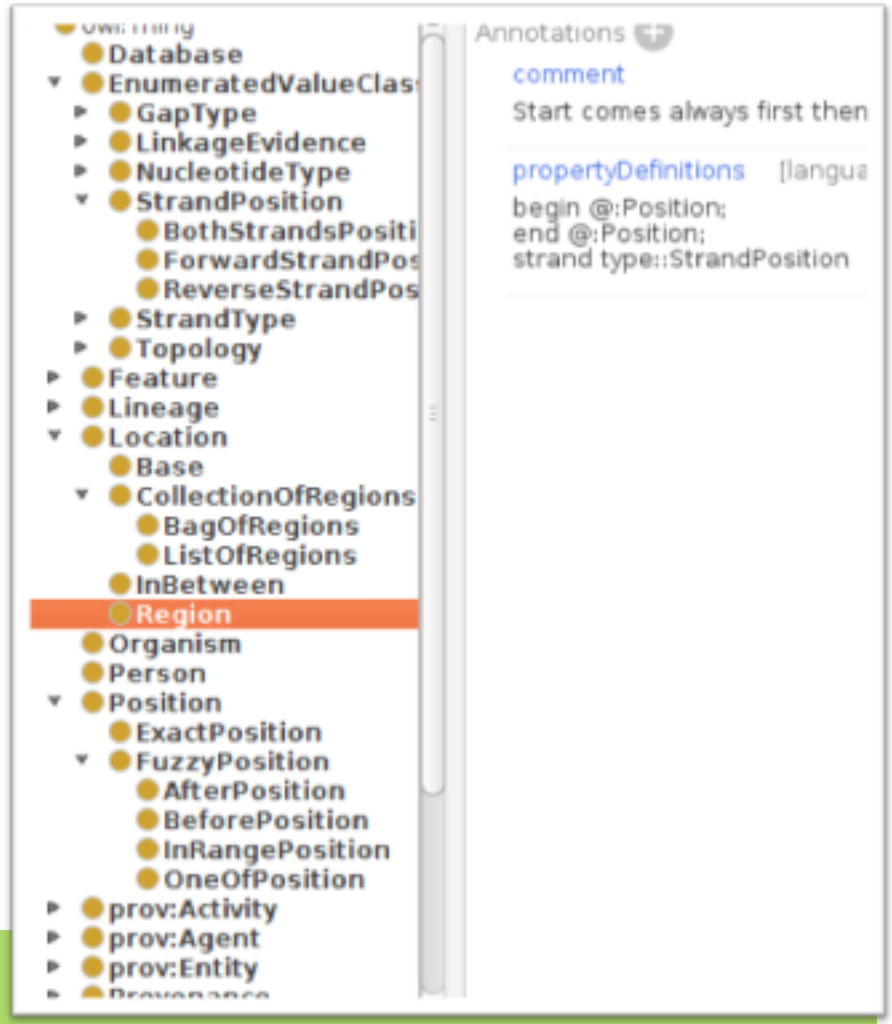
Developer:

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



Code generation: EMPUSA

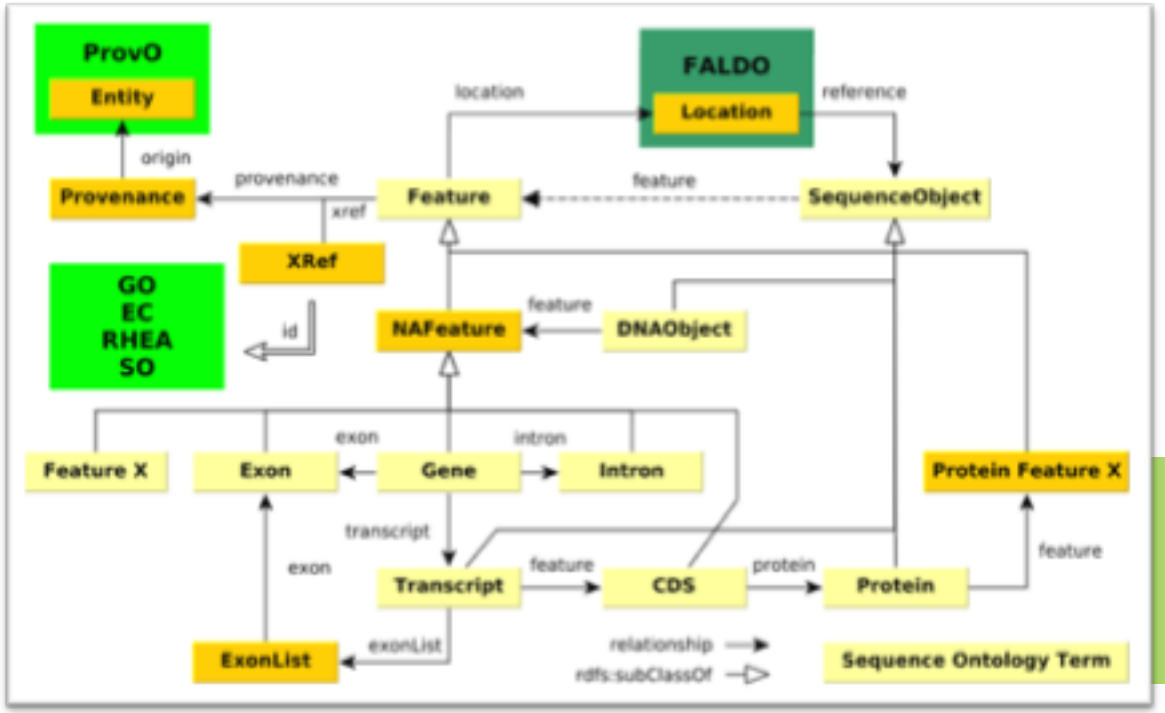
- Linked data graph is free format: Ontology defines structure but does not enforce it.
  - **NEED TO MANTAIN CONSISTENCY**
- From Ontology (protégé file)
  - OWL + ShEx
- API: Java + R
  - Instance validation included
- > 80.000 lines of code generated



WAGENINGEN
UNIVERSITY & RESEARCH

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

100years
1918 — 2018

Use cases:

Computational genomics

&

Organizing QTL data

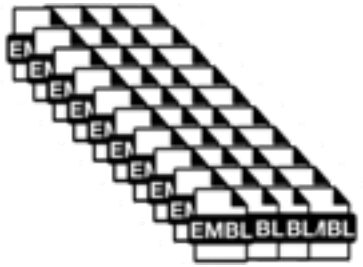
High Throughput annotation

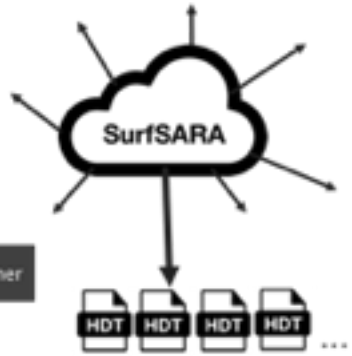
Parallel retrieval process

Dynamic search and filtering

Bacteria

SAPP





GeneCaller → Specific PFAM scanner

Functional domains retrieval

SAGERP

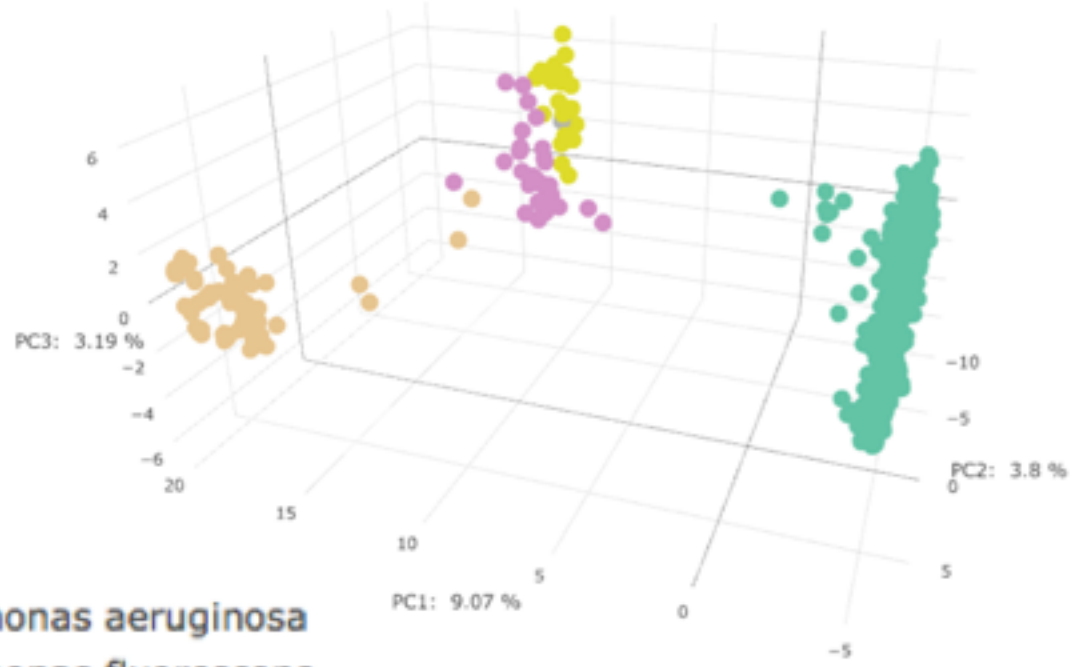
```
PREFIX gbol: <http://gbol.life/0.1/>
SELECT ?sample ?accession ?db
WHERE {
    ?sample a gbol:Sample .
    ?dnaobject gbol:sample ?sample .
    ?dnaobject gbol:feature ?gene .
    ?gene gbol:transcript ?transcript .
    ?transcript gbol:feature ?cds .
    ?cds gbol:protein ?protein .
    ?protein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

WAGENINGEN
UNIVERSITY & RESEARCH

100years
1918 — 2018

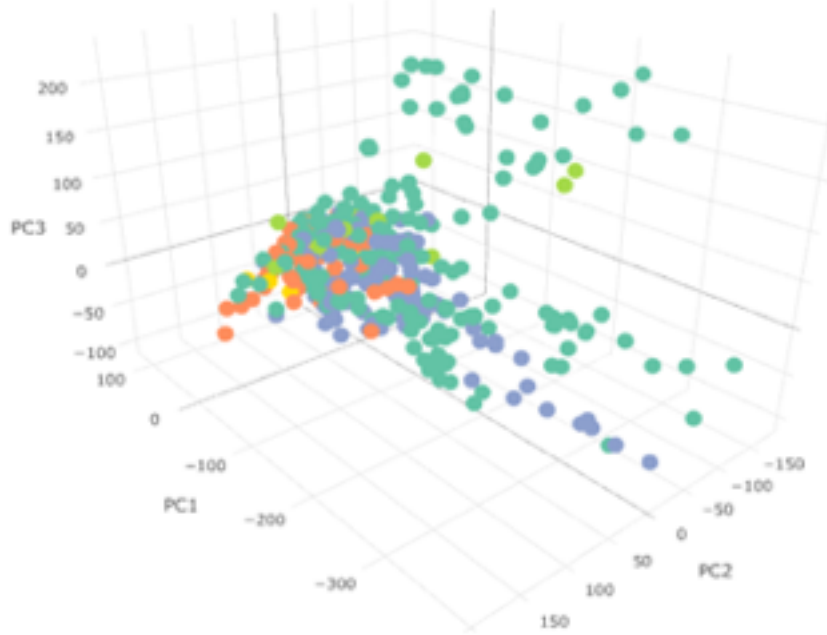
Functional variation

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

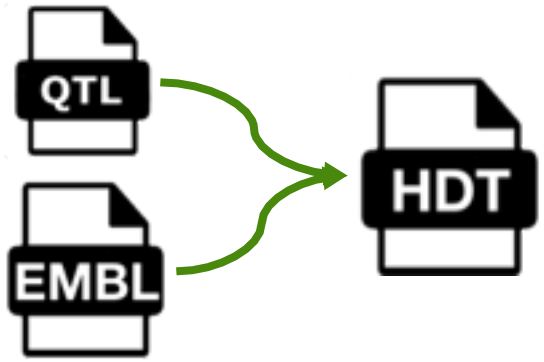
100years

QTL incorporation

QTL data of 3 species

- *Oryza sativa* (Rice)
- *Sorghum bicolor* (Sorghum)
- *Glycine max* (Soybean)

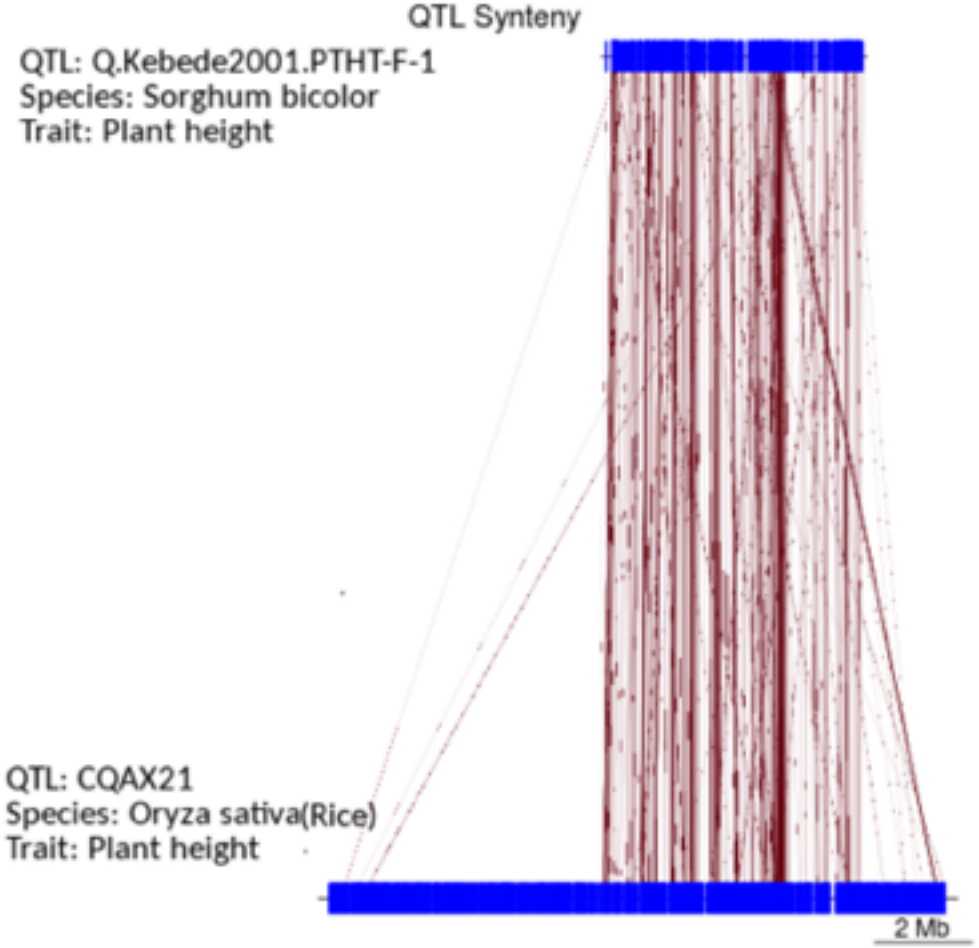
Conversion and comparison

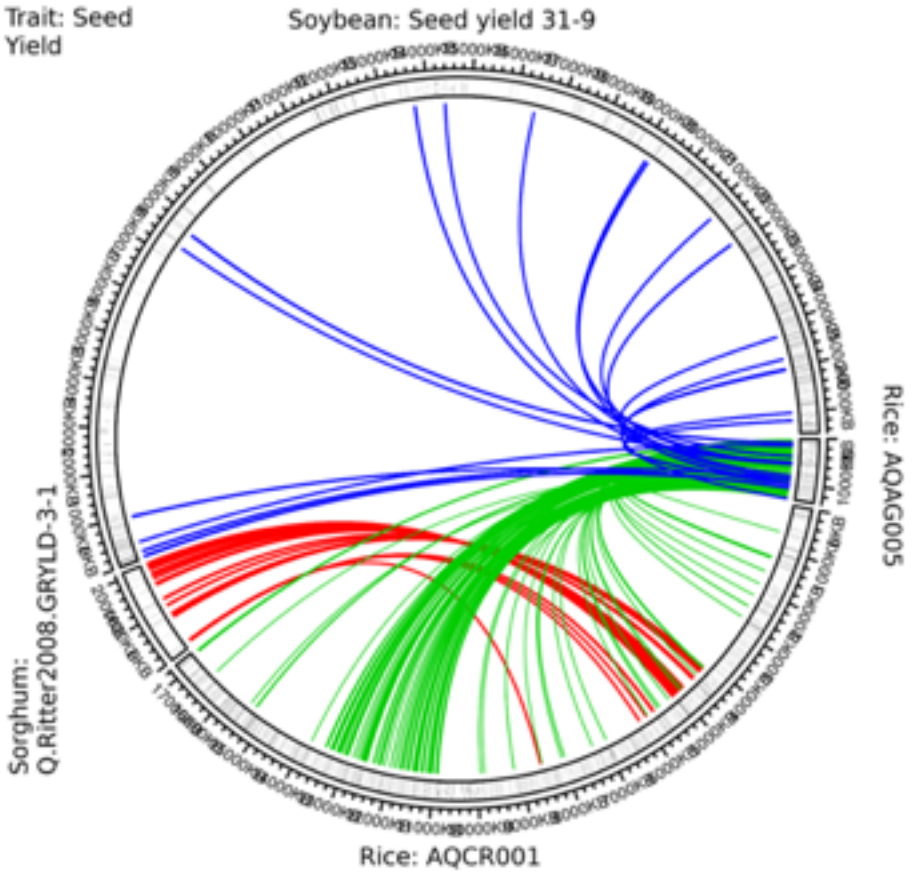


Pairwise protein sequence comparison

Organism	Number of QTLs	QTLs with < 450 genes	Number of Traits
Oryza sativa	8203	2958	230
Sorghum bicolor	604	277	72
Glycine max	691	451	88

Synteny comparison of QTL regions





Current publications using SAPP

1. Comparative genomics highlights symbiotic capacities and high metabolic flexibility of the marine genus Pseudovibrio
   D Versluis, B Nijsse, MA Naim, JJ Koehorst, J Wiese, JF Imhoff, ... **Genome biology and evolution** 10 (1), 125-142

2. Concurrent haloalkanoate degradation and chlorate reduction by Pseudomonas chloritidismutans AW-1T
   P Peng, Y Zheng, JJ Koehorst, PJ Schaap, AJM Stams, H Smidt, ... **Applied and environmental microbiology** 83 (12), e00325-17

3. Persistence of Functional Protein Domains in Mycoplasma Species and their Role in Host Specificity and Synthetic Minimal Life
   T Kamminga, JJ Koehorst, P Vermeij, SJ Slagman, ... **Frontiers in cellular and infection microbiology** 7, 31

4. Complete Genome Sequence of Akkermansia glycaniphila Strain PytT, a Mucin-Degrading Specialist of the Reticulated Python Gut
   JP Ouwerkerk, JJ Koehorst, PJ Schaap, J Ritari, L Paulin, C Belzer, ... **Genome announcements** 5 (1), e01098-16

5. Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining
   JCJ van Dam, JJ Koehorst, JO Vik, PJ Schaap, M Suarez-Diez **bioRxiv**, 184747

6. Reverse methanogenesis and respiration in methanotrophic archaea
   PHA Timmers, CU Welte, JJ Koehorst, CM Plugge, MSM Jetten, ... **Archaea** 2017

7. Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data
   JJ Koehorst, JCJ Van Dam, RGA Van Heck, E Saccenti, ... **Scientific reports** 6, 38699

8. Complete genome sequence of thermophilic Bacillus smithii type strain DSM 4216 T
   EF Bosma, JJ Koehorst, SAFT van Hijum, B Renckens, B Vriesendorp, ... **Standards in genomic sciences** 11 (1), 52

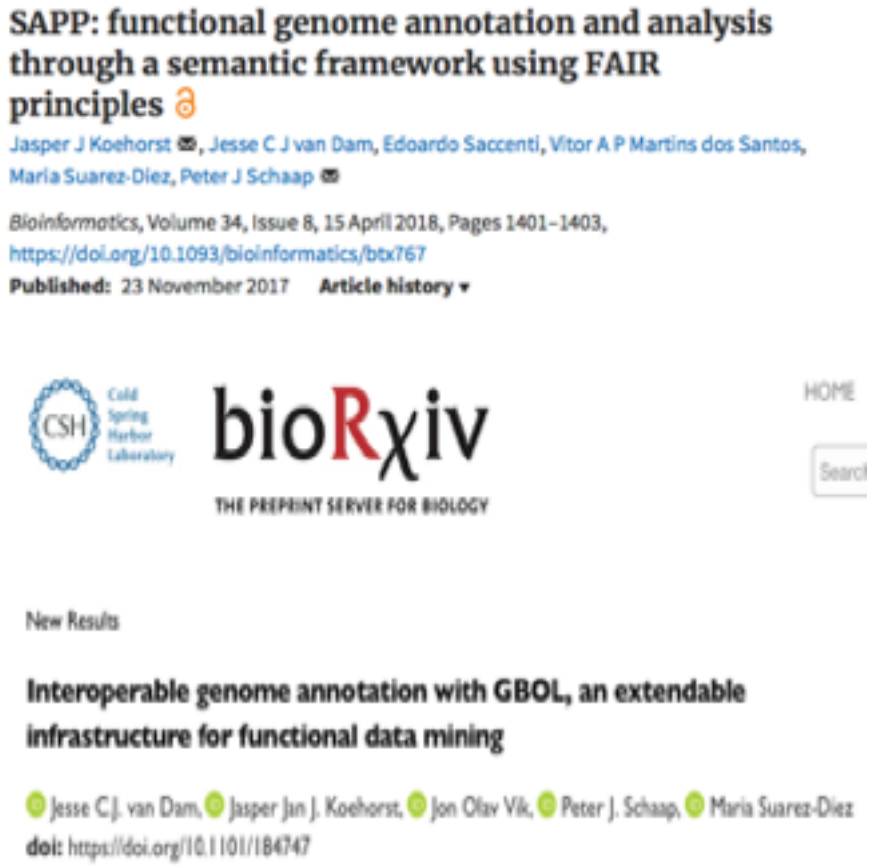
9. Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics
   JJ Koehorst, E Saccenti, PJ Schaap, VAPM dos Santos, M Suarez-Diez **F1000Research**

10. Assessing the metabolic diversity of streptococcus from a protein domain point of view
    E Saccenti, D Nieuwenhuijse, JJ Koehorst, VAPM dos Santos, PJ Schaap **PloS one** 10 (9), e0137908

11. A genomic view on syntrophic versus non-syntrophic lifestyle in anaerobic fatty acid degrading communities
    P Worm, JJ Koehorst, M Visser, VT Sedano-Núñez, PJ Schaap, ... Biochimica et Biophysica Acta **(BBA)-Bioenergetics** 1837 (12), 2004-2016

Availability

- **SAPP** Koehorst et al Bioinformatics 2017 https://sapp.gitlab.io
- **Empusa**: https://gitlab.com/Empusa
- **GBOL:** Documentation & namespace: http://gbol.life/0.1/



SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles

Jasper J Koehorst, Jesse C J van Dam, Edoardo Saccenti, Vitor A P Martins dos Santos, Maria Suarez-Diez, Peter J Schaap

Bioinformatics, Volume 34, Issue 8, 15 April 2018, Pages 1401–1403, https://doi.org/10.1093/bioinformatics/btx767
Published: 23 November 2017 **Article history**

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining

Jesse C.J. van Dam, Jasper Jan J. Koehorst, Jon Olav Vik, Peter J. Schaap, Maria Suarez-Diez

doi: https://doi.org/10.1101/184747

Acknowledgements

Laboratory Systems and Synthetic Biology

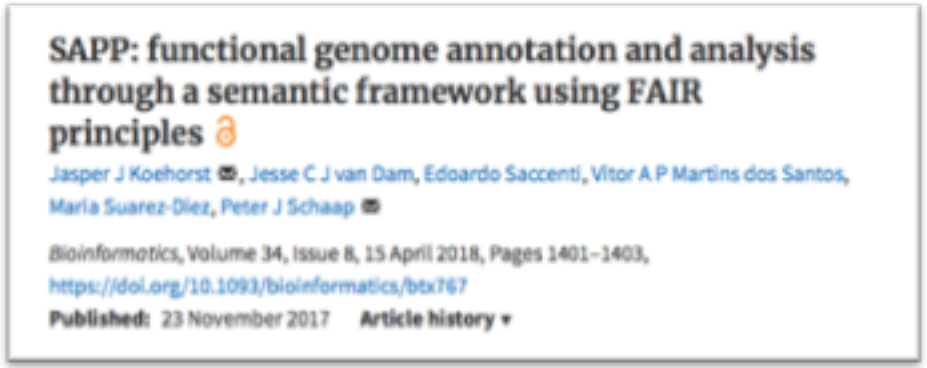
Jesse van Dam
Benoit Carreres
Maria Suarez-Diez
Edoardo Saccenti
Peter Schaap
Vitor Martins dos Santos

Norwegian university of Life Sciences

Jon Olav Vik
Fabian Grammes
Arne Bjørke Gjuvsland

Availability

- **SAPP** Koehorst et al Bioinformatics 2017 https://sapp.gitlab.io
- **Empusa**: https://gitlab.com/Empusa
- **GBOL:** Documentation & namespace: http://gbol.life/0.1/



SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles

Jasper J Koehorst, Jesse C J van Dam, Edoardo Saccenti, Vitor A P Martins dos Santos, Maria Suarez-Diez, Peter J Schaap

Bioinformatics, Volume 34, Issue 8, 15 April 2018, Pages 1401–1403, https://doi.org/10.1093/bioinformatics/btx767

Published: 23 November 2017 Article history

Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME

New Results

Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining

Jesse C.J. van Dam, Jasper Jan J. Koehorst, Jon Olav Vik, Peter J. Schaap, Maria Suarez-Diez

doi: https://doi.org/10.1101/184747