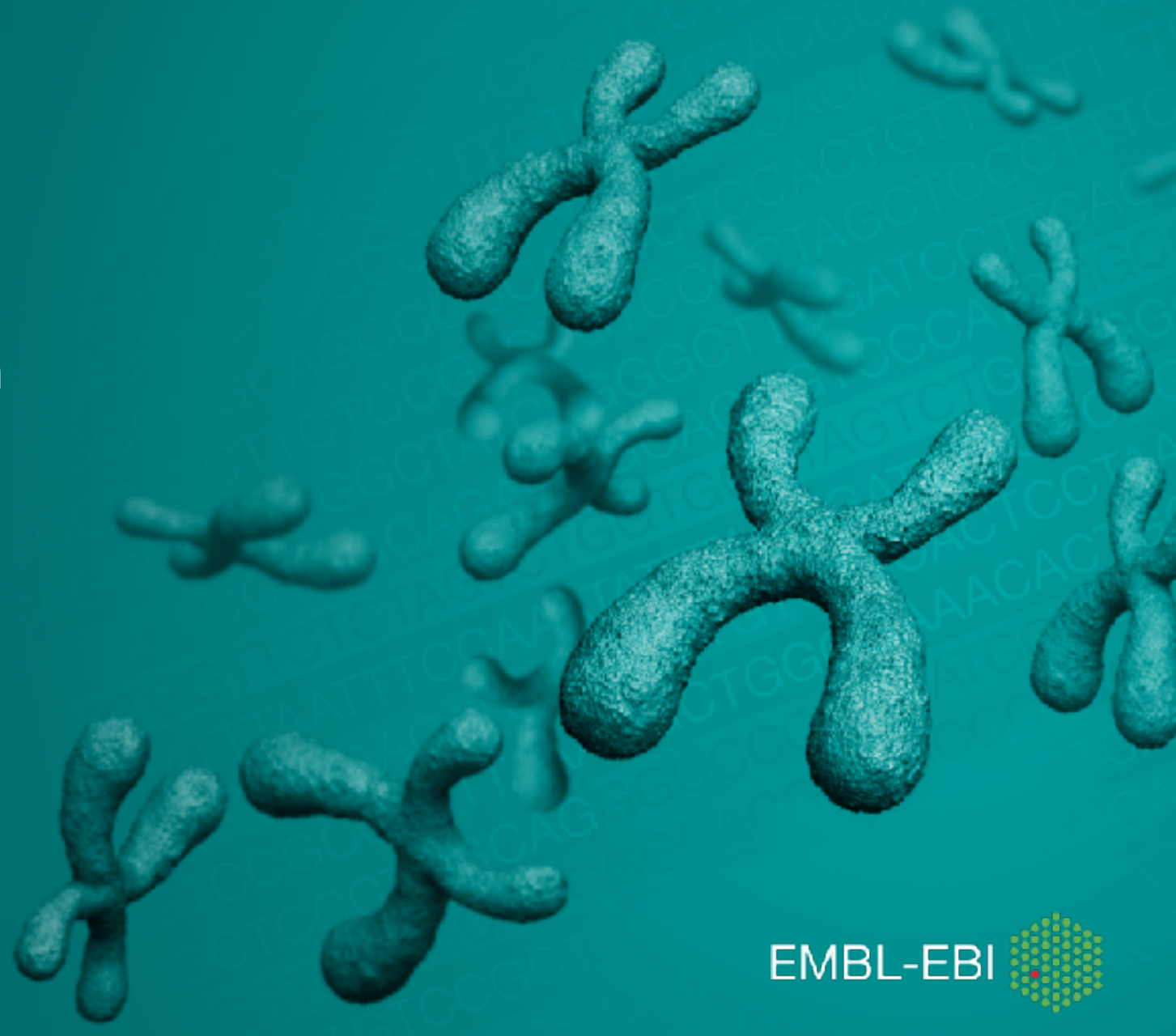
EBI metagenomics: building analysis workflows for *all* metagenomics studies

Rob Finn (rdf@ebi.ac.uk), @robdfinn

Sequence Families Team Leader

19th June 2017



EMBL-EBI

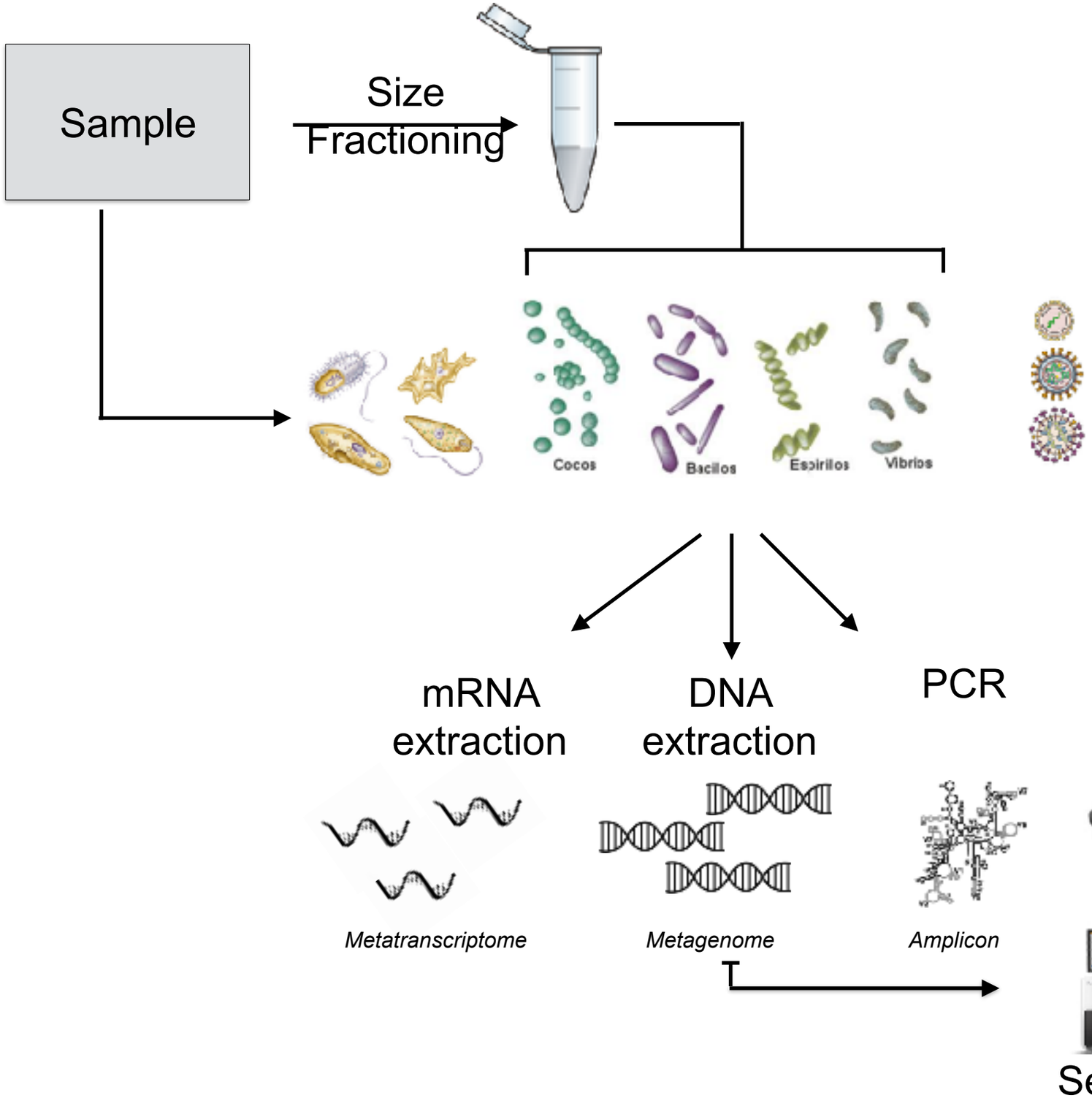
‘Metagenomics’: a broad range of environments

‘Metagenomics’: a broad range of environments

Change in microbiome composition?

Enzymes involved in a pathway?

Different experimental design



EBI Metagenomics

http://www.ebi.ac.uk/metagenomics

A ***free*** resource for the analysis, archiving & browsing of amplicon, metagenomic and metatranscriptomic data

Data submission process

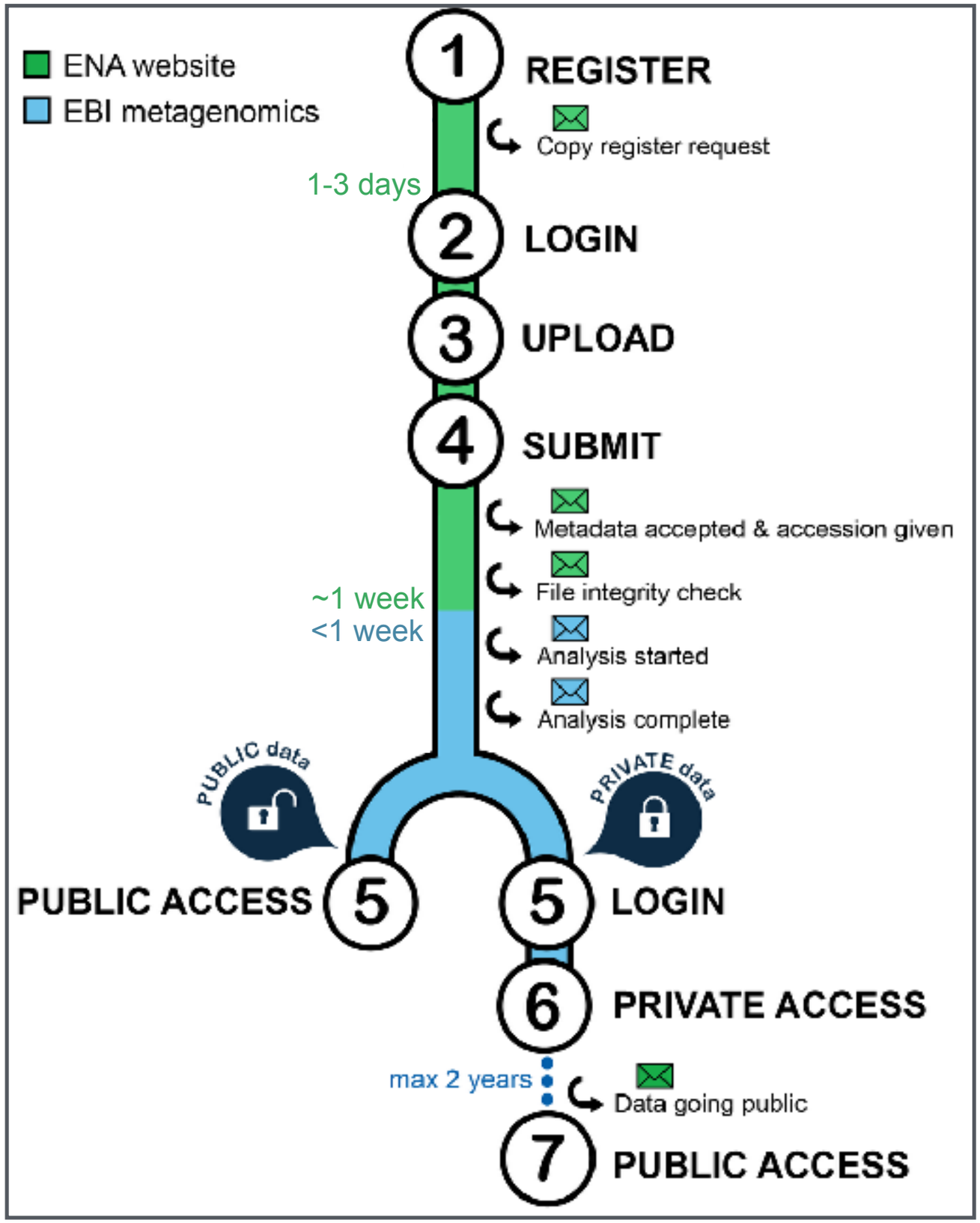
(1) Register for an account

(2) Upload sequence data and metadata

(3) Sequence data is archived in ENA and accessioned

(4) Sequence data is analysed by the metagenomics pipeline

(5) Projects, metadata and results are made available on the website for private or public browsing / download



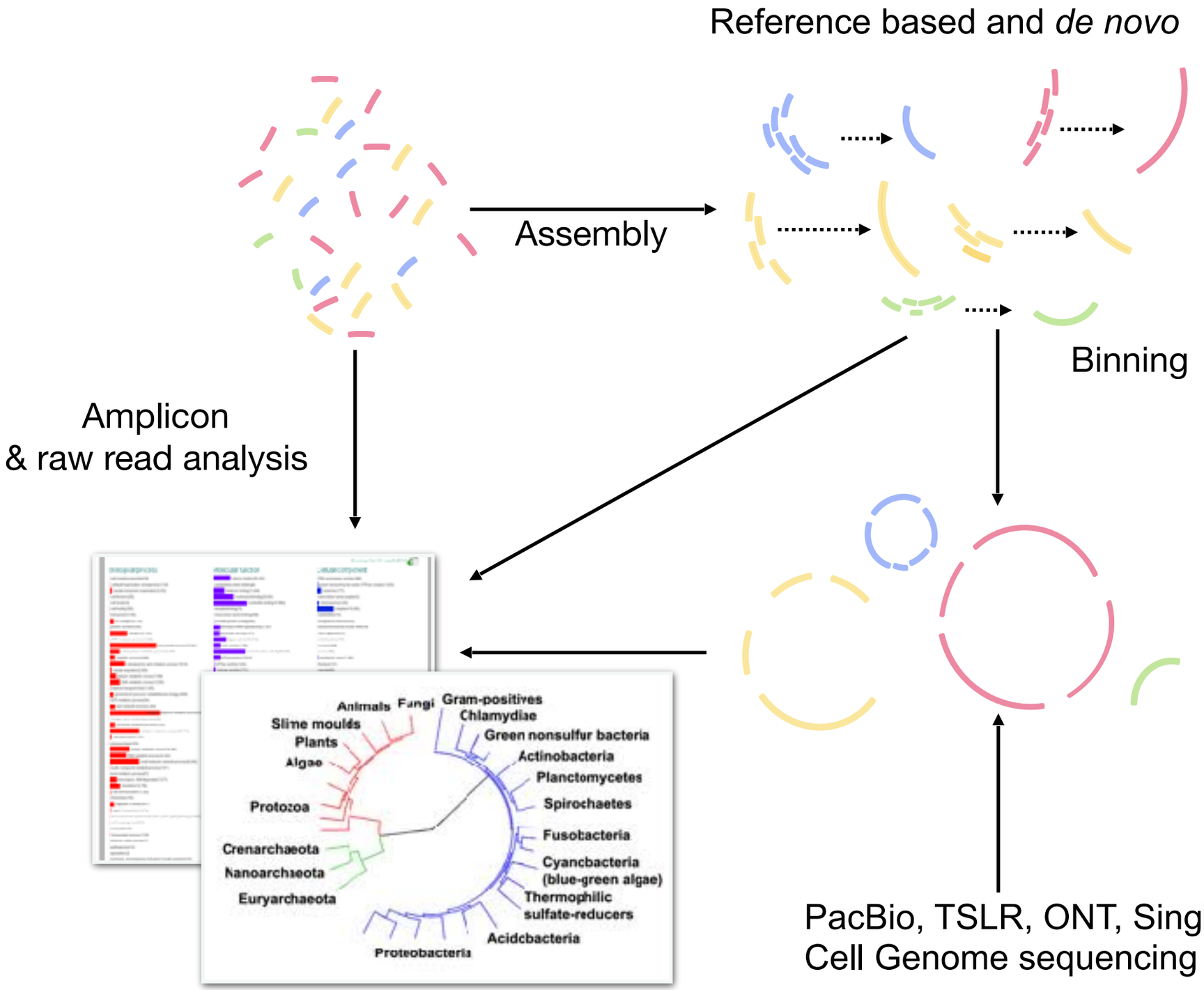
Longevity of data

For data to have longevity and be useful to the scientific community, sequences need to be archived with contextual metadata

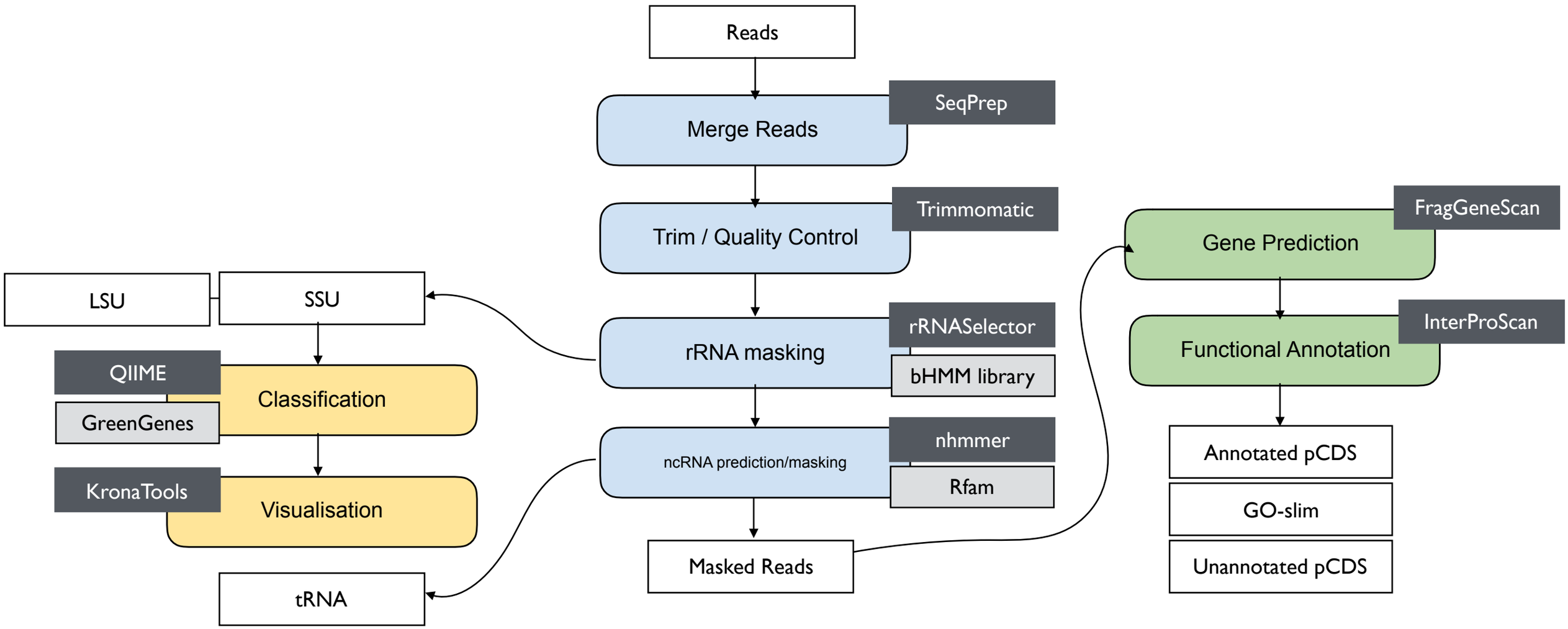
- How was it sampled? How was it extracted? How was it stored? What sequencing platform was used?

- Where did it come from? What were the environmental conditions (lat/long, depth, pH, salinity, temperature…) or clinical observations?

Assembly and (meta-)genome annotation

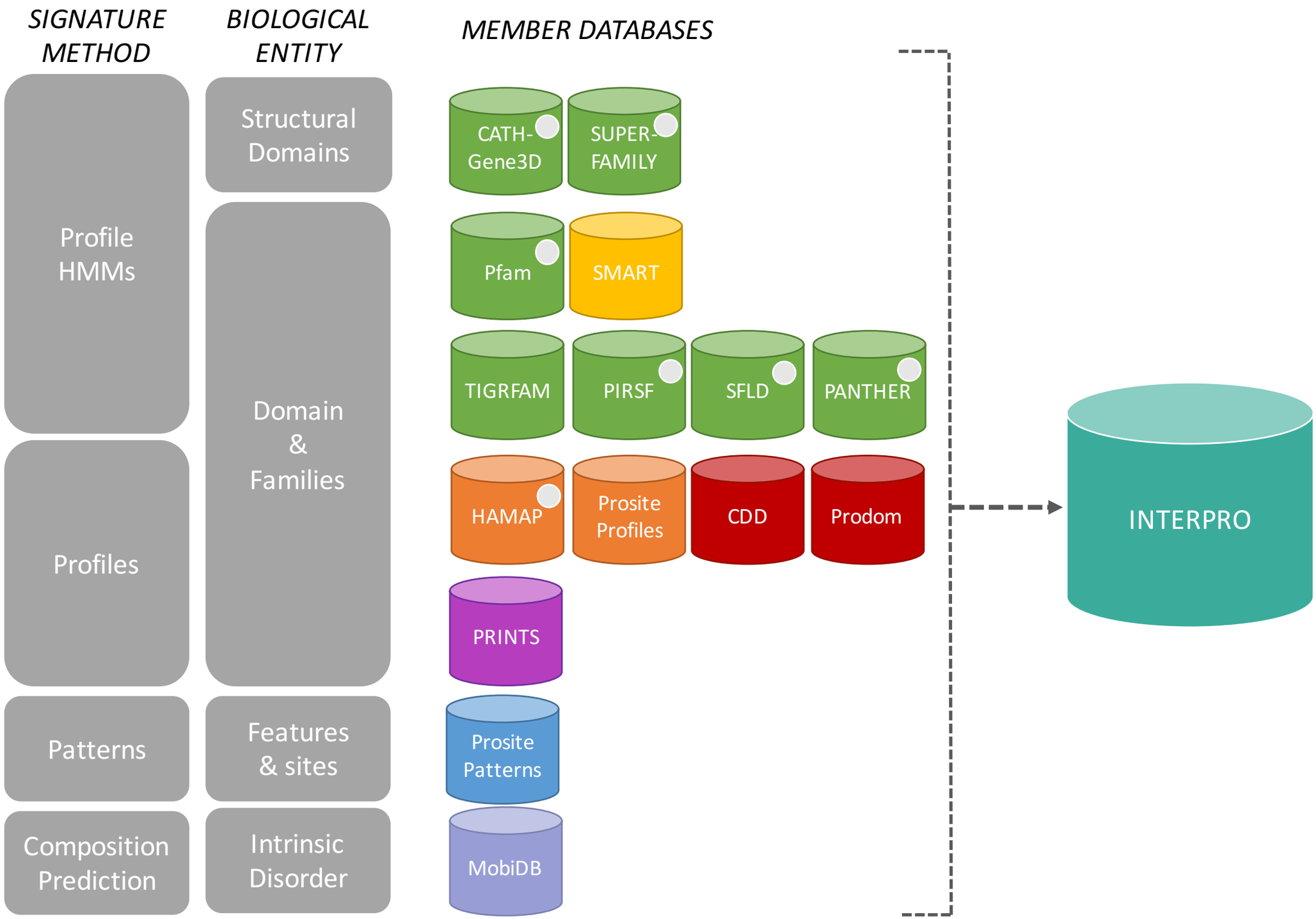


Pipeline overview, version 3.0



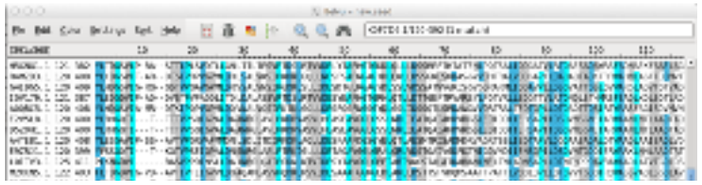
Current pipeline used for all analysis

InterPro - A integrated classification of protein families



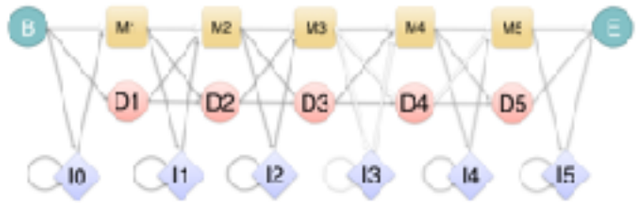
Pfam Entries

Seed Alignment
Representative Members



hmmbuild

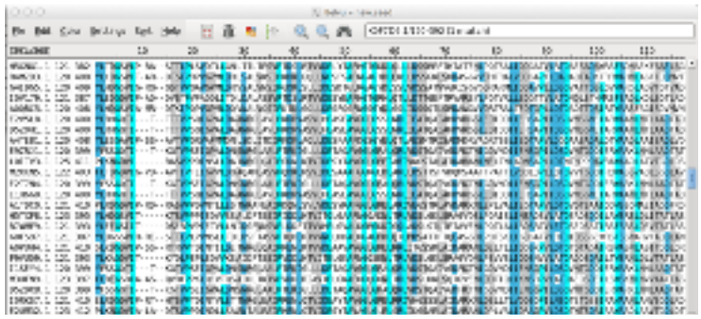
Profile HMM
Seed weighted probabilities



hmmsearch against UniProtKB

Iterate

Full Alignment
All members



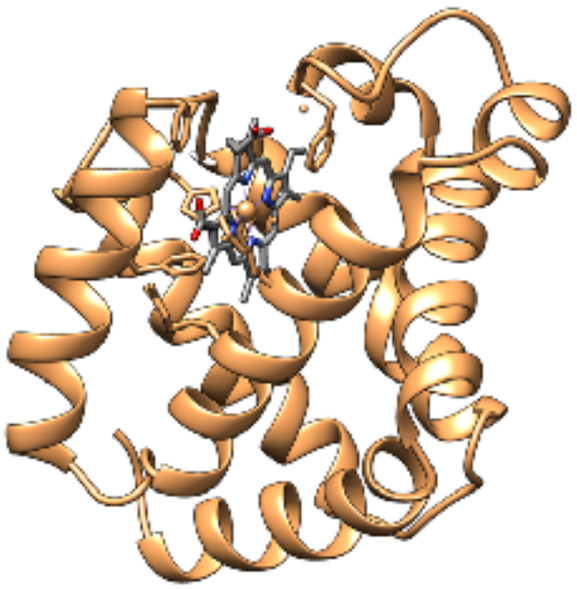
Quality Assurance

DESC-ription File
Annotations and curated thresholds

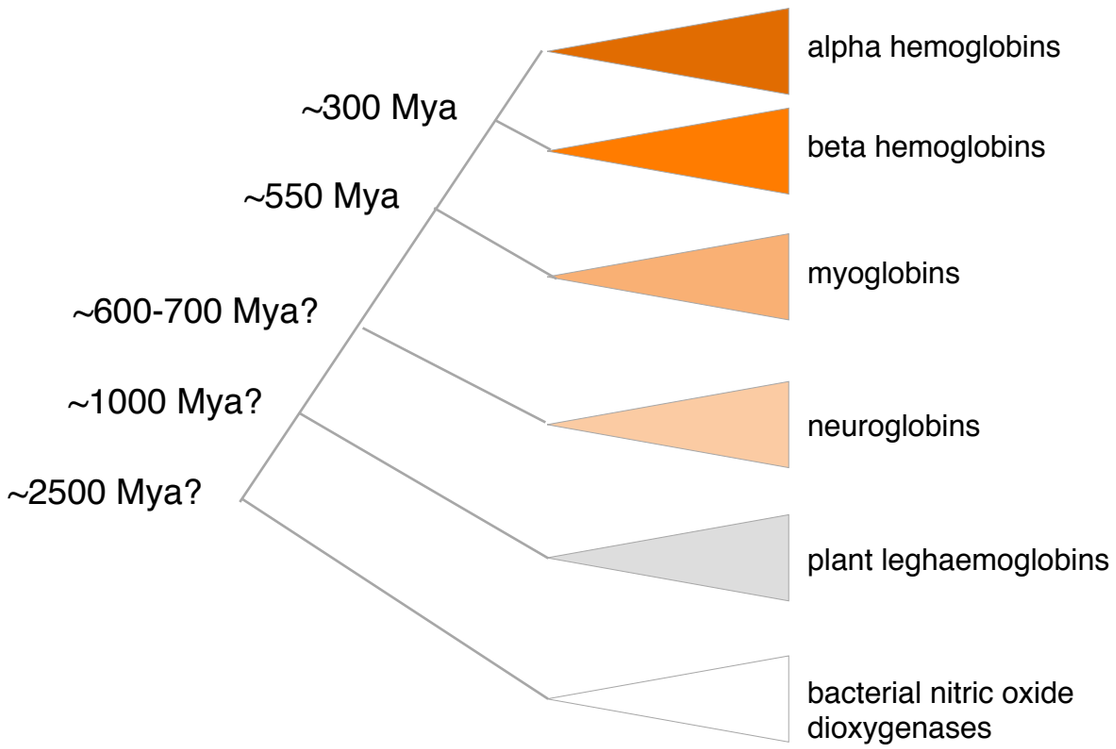
```
ID   dCache_1
AC   PF02743
DE   Cache domain
GA   45.00 45.00;
```

EMBL-EBI

Anecdotal search example: globin superfamily



Aplysia myoglobin (PDB 1mba)



	PSI-BLAST	HMMER
HBA_HUMAN	4e-46	9e-62
HBA_MOUSE	3e-42	4e-55
HBB_HUMAN	2e-57	4e-64
HBB1_MOUSE	9e-50	2e-57
MYG_HUMAN	1e-45	2e-58
MYG_MOUSE	2e-41	6e-54
NGB_HUMAN	-	1e-7
NGB_MOUSE	-	2e-7
LGB1_PEA	1.1	5e-5
LGB2_PEA	0.45	5e-6
HMP_VIBCH	-	0.004
HMP_ECOLI	-	-

E-values(statistical significance)

query: alignment of three vertebrate hemoglobins and one myoglobin

target db: UniProt 7.0 (207K seqs) (contains about 1060 known globins)

at E <= 0.01:
PSI-BLAST sees: 915 globins (9 sec)
HMMER3 sees: 1002 globins (8sec)

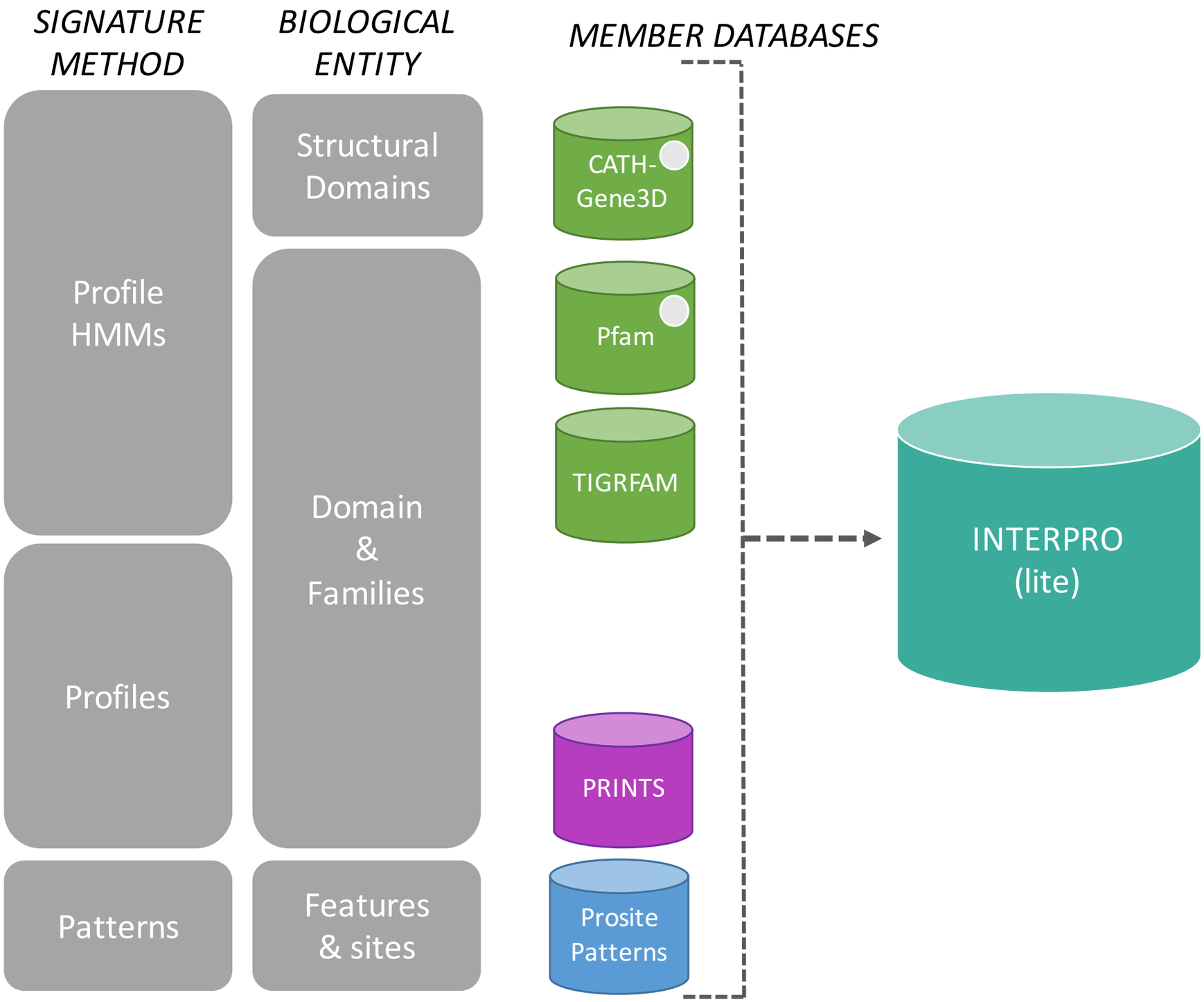
EMBL-EBI

Scalability of profile HMMs

	HMM Reference database (10,000)	*Sequence Reference database (10,000,000)*
1 million	10^{10}	10^{13}
10 million	10^{11}	10^{14}
100 million	10^{12}	10^{15}
1 billion	10^{13}	10^{16}

Consequence - inferior sequence similarity methods, e.g. BLAT, are being used to increase speed

InterPro - A integrated classification of protein families



Anecdotal analysis of OSD data

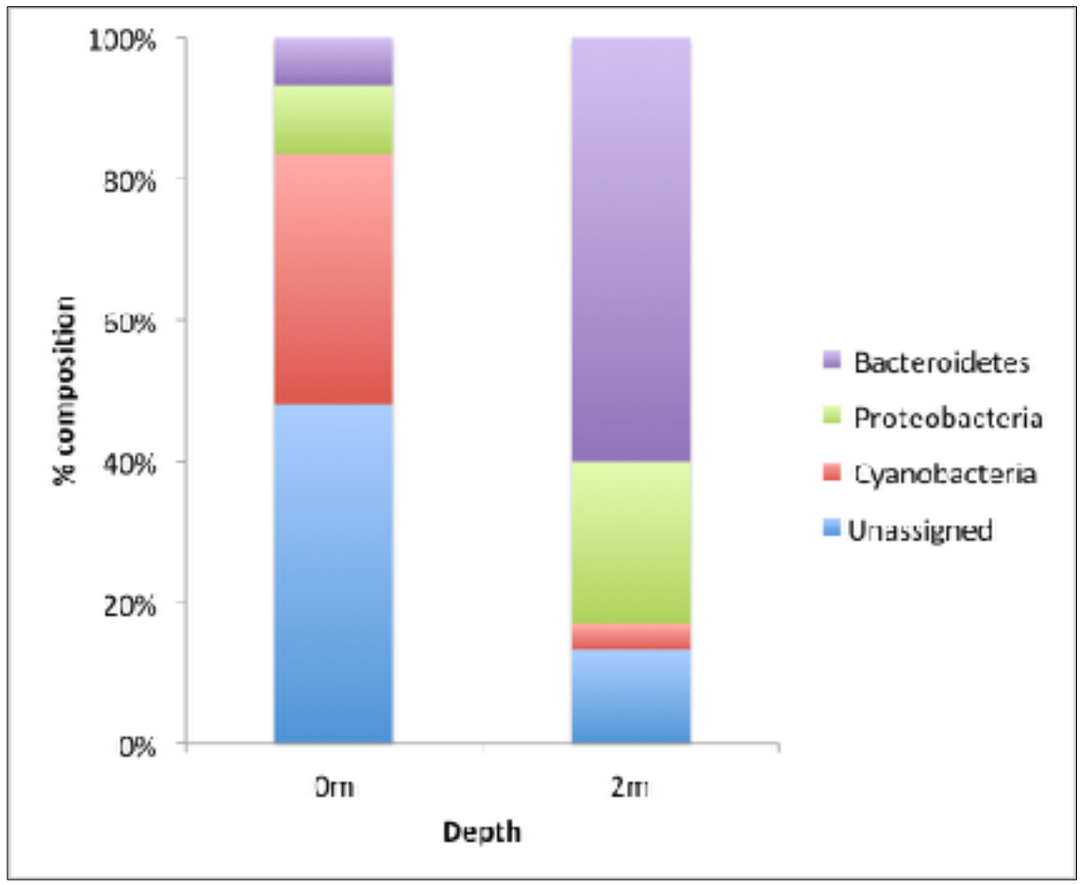
Young Sound, Greenland

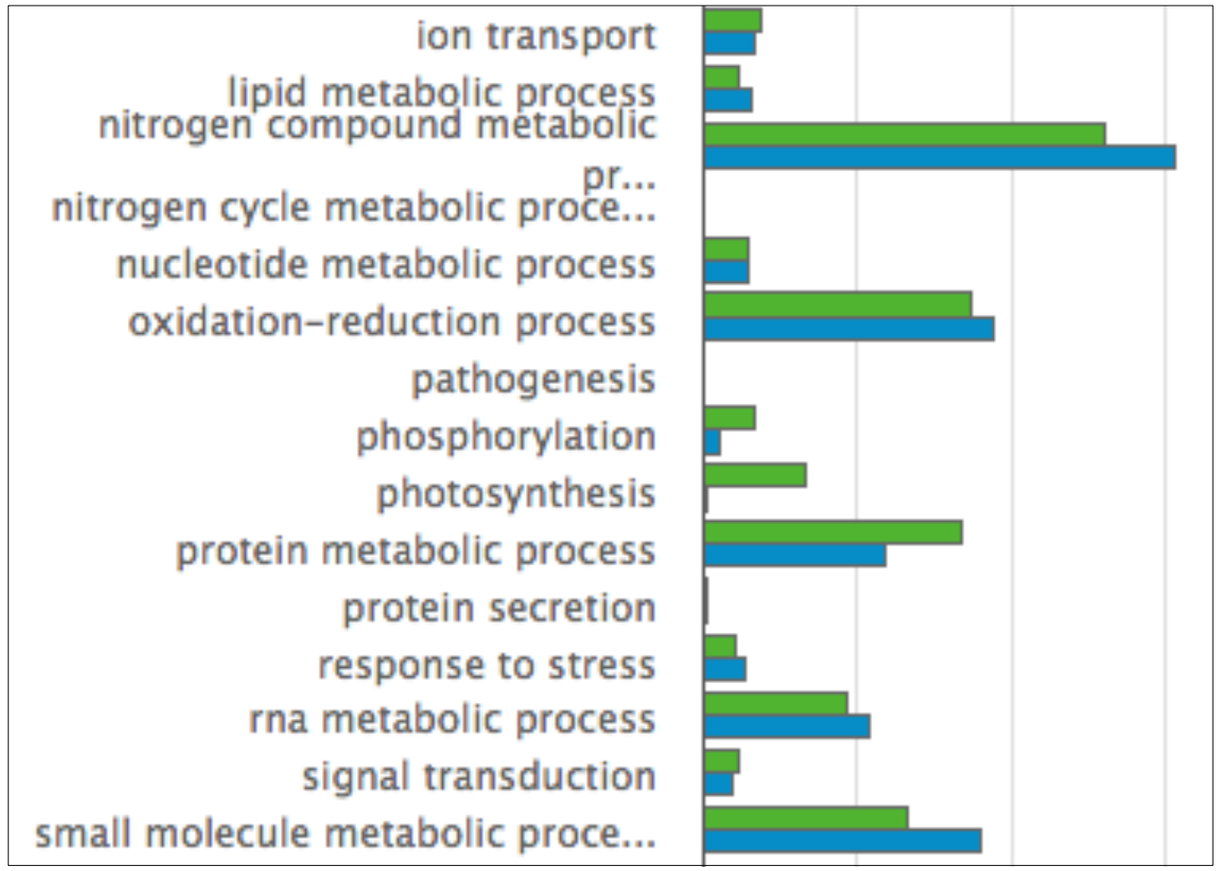
ERR770971

Temperature	-0.1 °C
Project name	Micro B3
Geographic location (depth)	0 m
Environmental package	Water
Salinity	5 psu

ERR770970

Temperature	-1.6 °C
Project name	Micro B3
Geographic location (depth)	2 m
Environmental package	Water
Salinity	32 psu

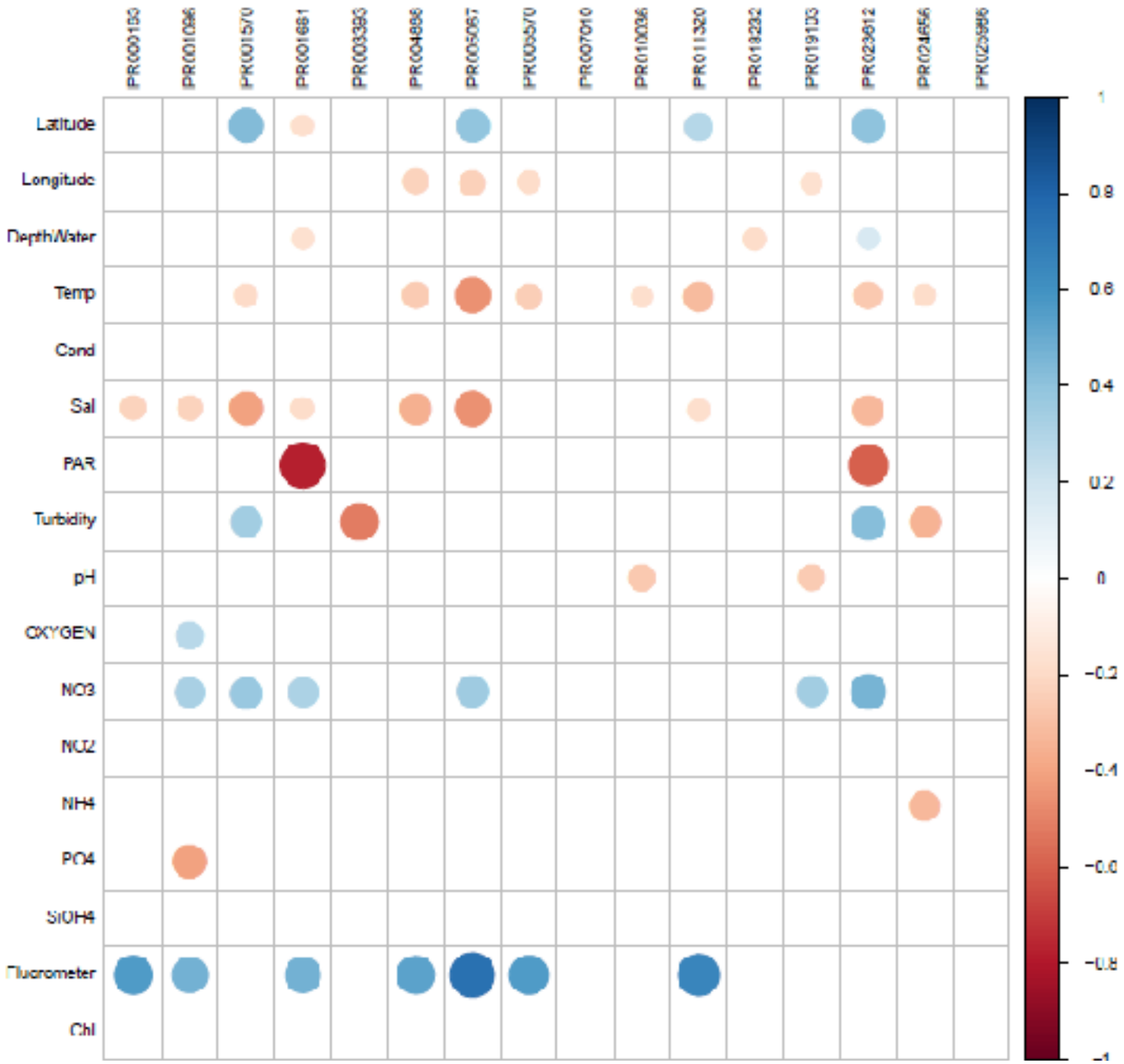




EMBL-EBI

In depth analysis of OSD data

EBI Metagenomics



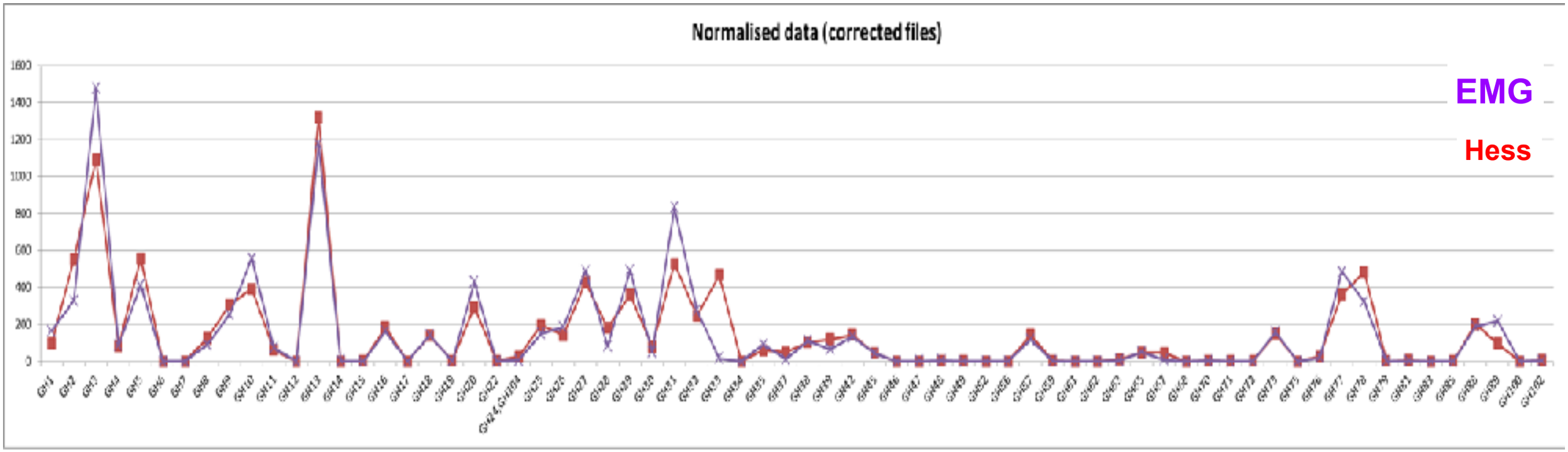
Courtesy of Bernardo Duarte, João Canning-Clôde,Catarina Magalhães, Luís Torgo, Isabel Caçador. Manuscript in preparation.

Heatmap of significant Spearman correlations between protein families and environmental conditions across 150 sites

Annotations with/without assembly

Re-analysis of Biomass-Degrading Genes from Cow Rumen

Comparison of the normalised number of genes / reads corresponding to CAZy Glycoside Hydrolase Family from the Hess et al paper and from the EMG pipeline.



Hess et al: genome assembly then gene prediction using a subset of Pfam.
EMG pipeline: no assembly and gene prediction using InterPro.

Discrepancies are due to the different ways in which significance cut-off are calculated.

EBI Metagenomics

- Number of different projects: **1,314**
- Number of different samples: **78,271**
- Number of runs: **100,179**

- Predicted rRNAs**: >4 billion**
- Predicted CDS: **>200 billion**
- Total InterProScan matches: **>70 billion**

Powerful analysis

Data archiving

http://www.ebi.ac.uk/metagenomics

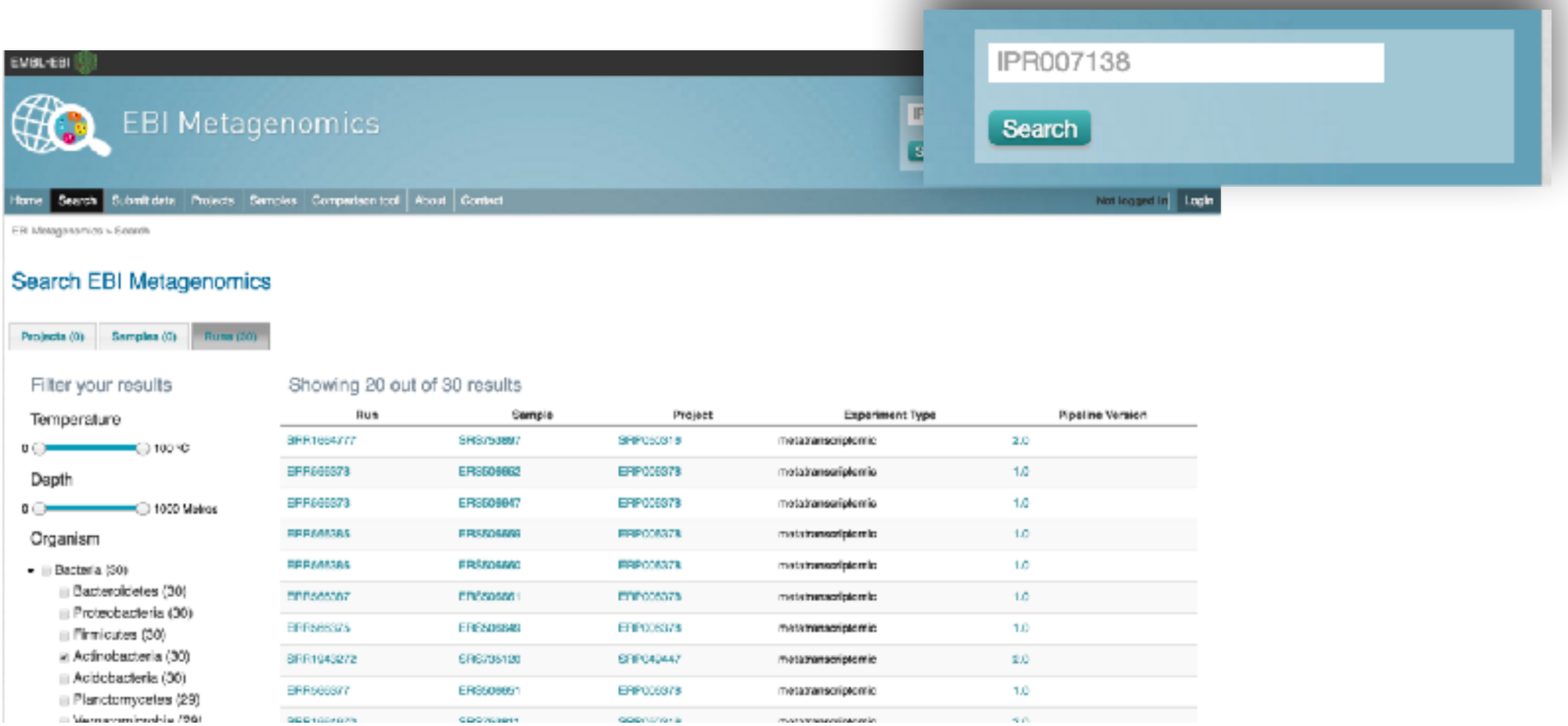
EMBL-EBI

Discoverability

- EBI-search underpins search interface
  - e.g. Give me all antibiotic biosynthesis monooxygenases in soil, where Actinobacteria are found, determined using metatranscriptomics

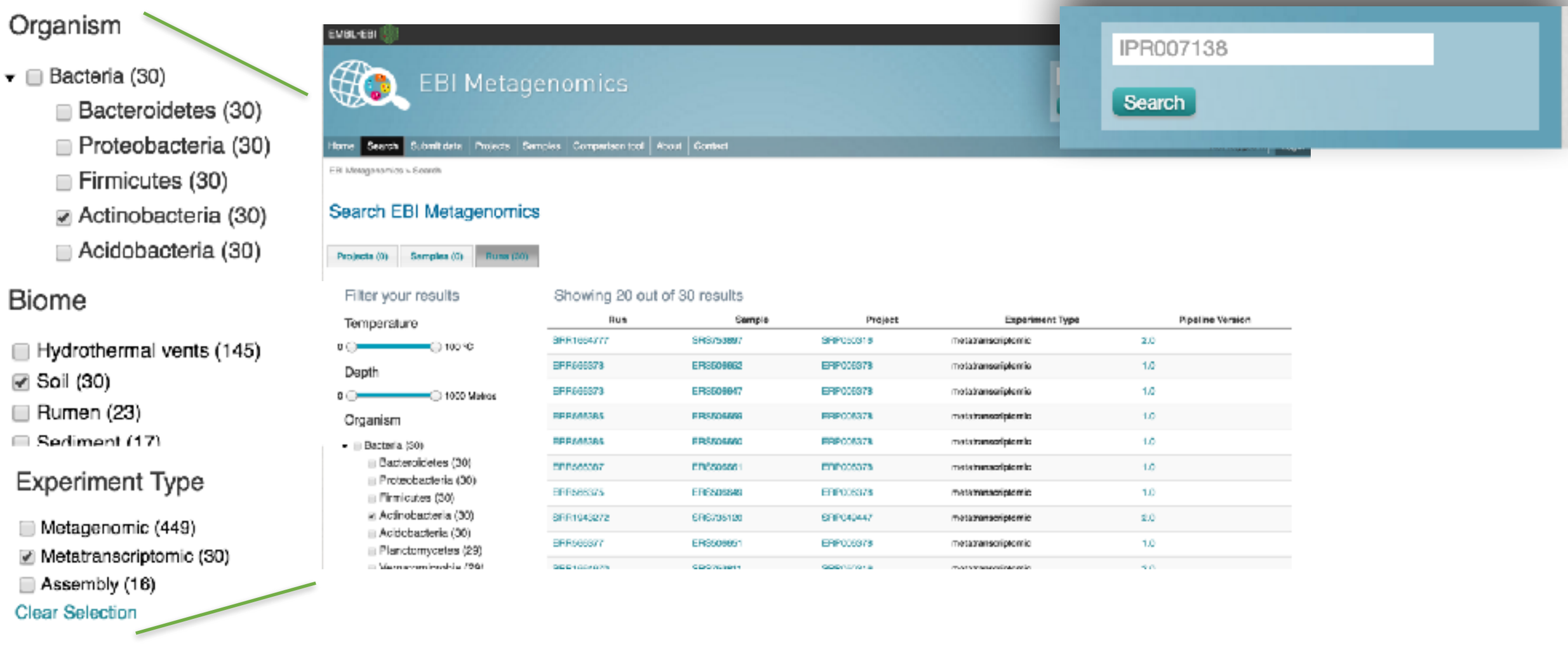
Discoverability

- EBI-search underpins search interface
  - e.g. Give me all antibiotic biosynthesis monooxygenases in soil, where Actinobacteria are found, determined using metatranscriptomics



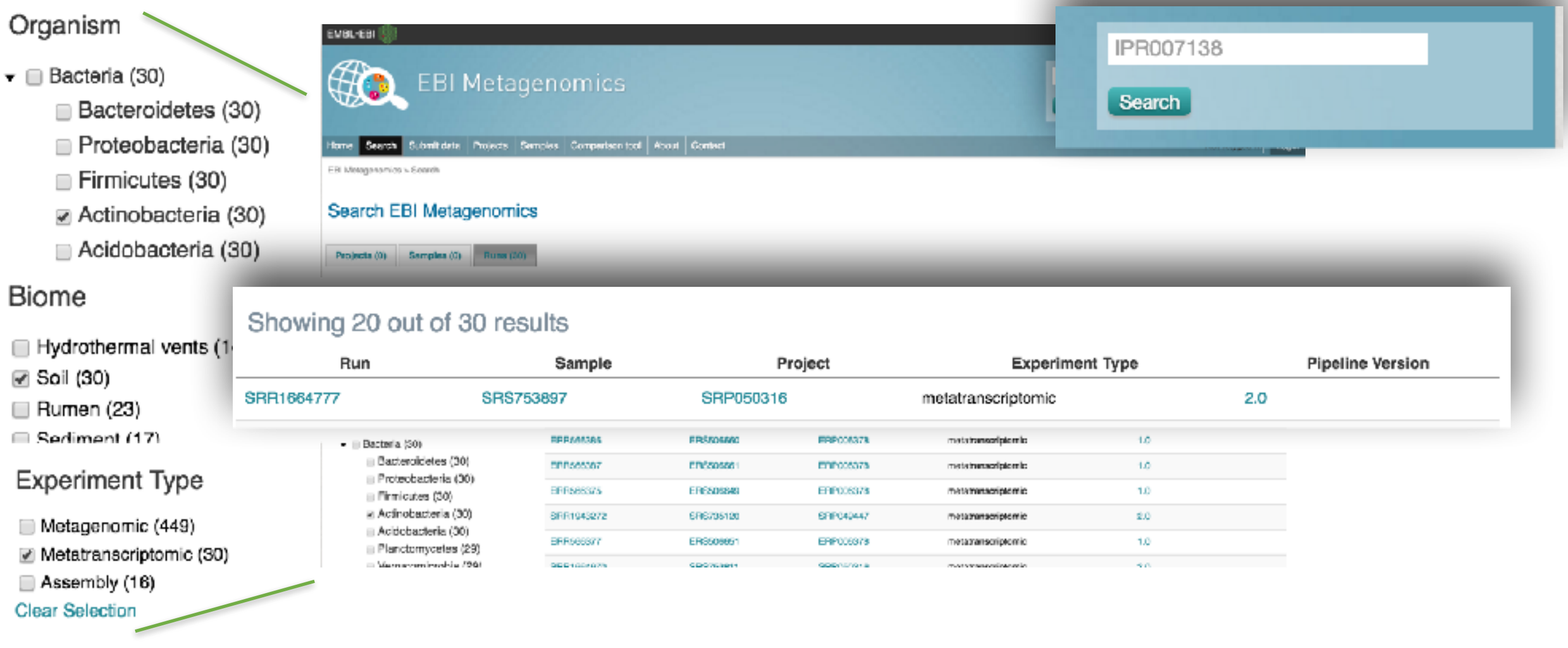
Discoverability

- EBI-search underpins search interface
  - e.g. Give me all antibiotic biosynthesis monooxygenases in soil, where Actinobacteria are found, determined using metatranscriptomics



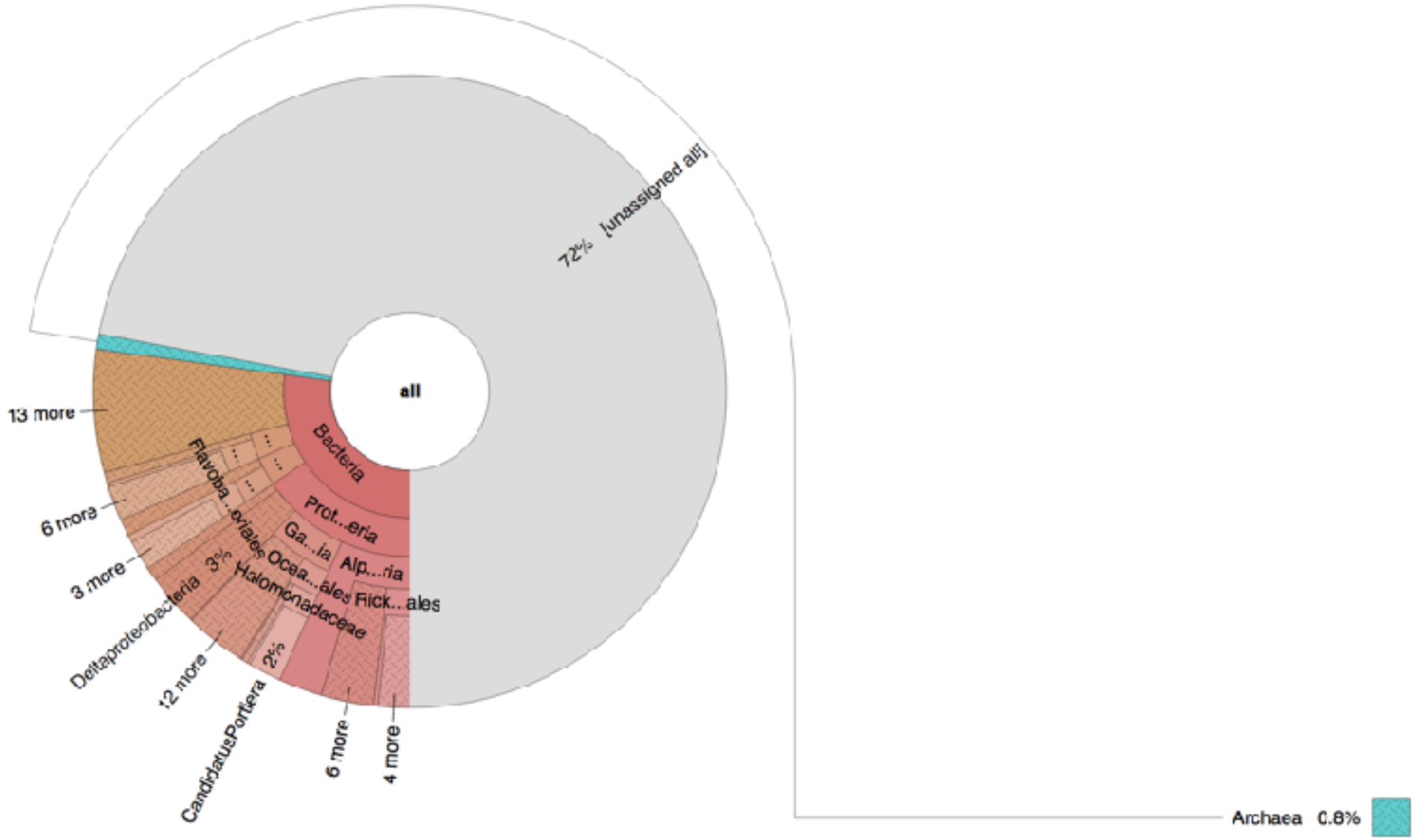
Discoverability

- EBI-search underpins search interface
  - e.g. Give me all antibiotic biosynthesis monooxygenases in soil, where Actinobacteria are found, determined using metatranscriptomics



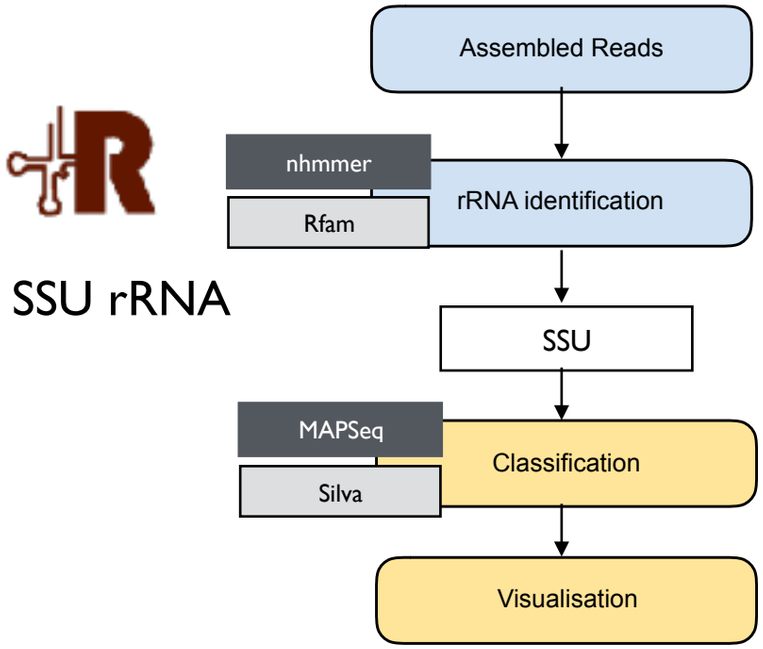
Ensuring broad taxonomic profiling

Out of the 7872 identified SSU rRNA,only 28% classified



Assembly of metagenome from Gulf of Aqaba in the Rea Sea: ERZ376968 (GCA_900157355)

Workflow to validate SSU rRNA

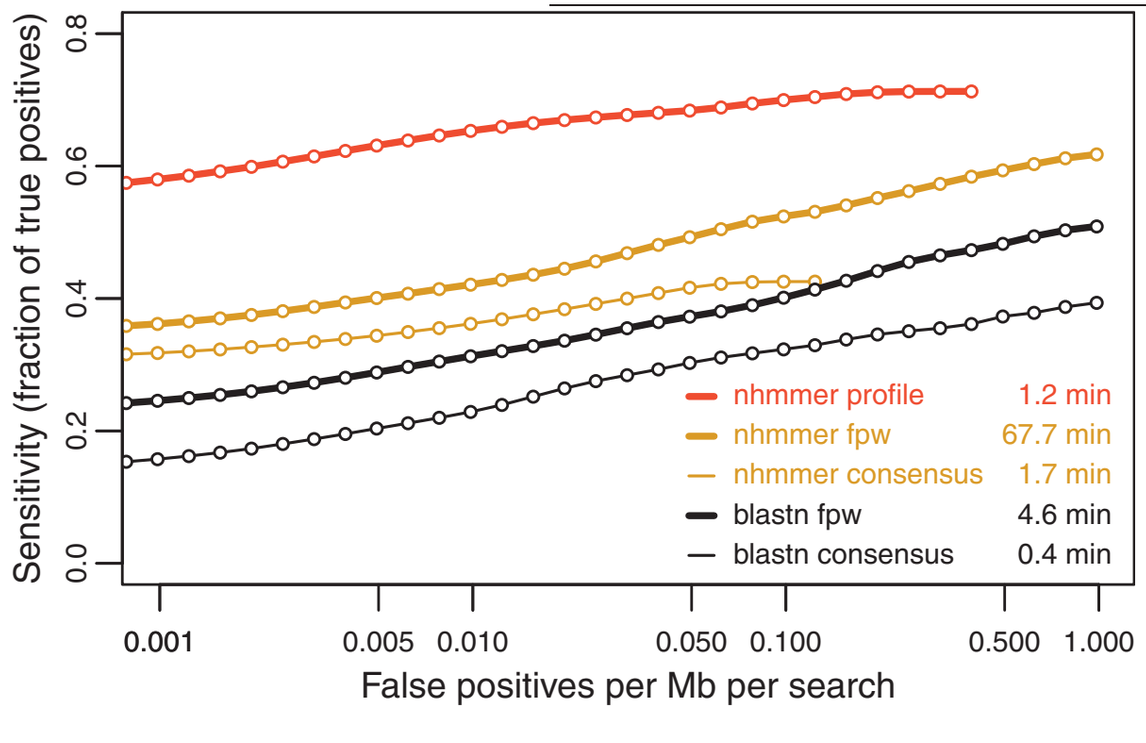


BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 19 2013, pages 2487–248
doi:10.1093/bioinformatics/btt40
Sequence analysis
Advance Access publication July 9, 201

nhmmer: DNA homology search with profile HMMs

Travis J. Wheeler* and Sean R. Eddy
HHMI Janelia Farm Research Campus, Ashburn, VA 20147, USA

Associate Editor: Alfonso Valencia



Amplicon data (V1V3)

>read1.1
ATGCATGCATGC...
>read1.2
ATGCATGCATGC...

Amplicon data (V3V5)

>read2.1
ATGCATGCATGC...
>read2.2
ATGCATGCATGC...

WGS data

>read3.1
ATGCATGCATGC...
>read3.2
ATGCATGCATGC...

MAPseq Algorithm
- word search + full alignment
- confidence estimation
- multiple taxonomy and OTU classification
- fast and accurate

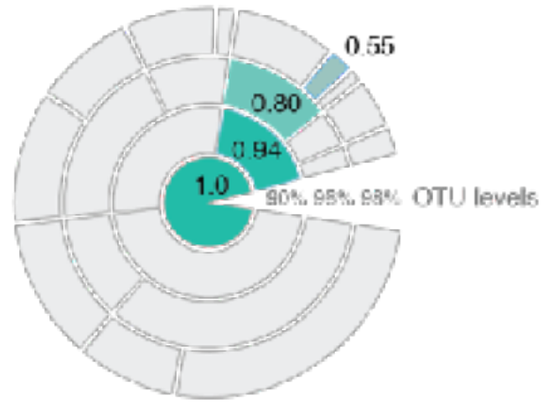
MAPseq Reference DB
- Full-length rRNA sequences
- Preprocessed taxonomies (NCBI, SILVA, custom)
- Preclustered hierarchical OTUs

Taxonomic classification

Bacteria (1.0); *Firmicutes* (0.90); ...; *Enterococcus sp.* (0.5)

Hierarchical OTU Classification

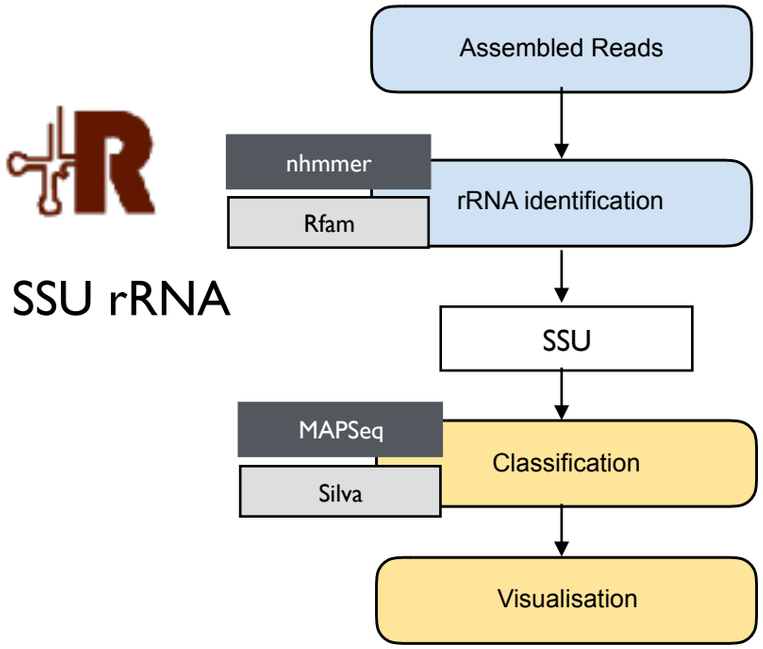
Bacteria (1.0)
Γ132 (0.94)
G432 (0.80)
S85 (0.55)



MAPseq from von Mering lab

EMBL-EBI

Workflow to validate SSU rRNA

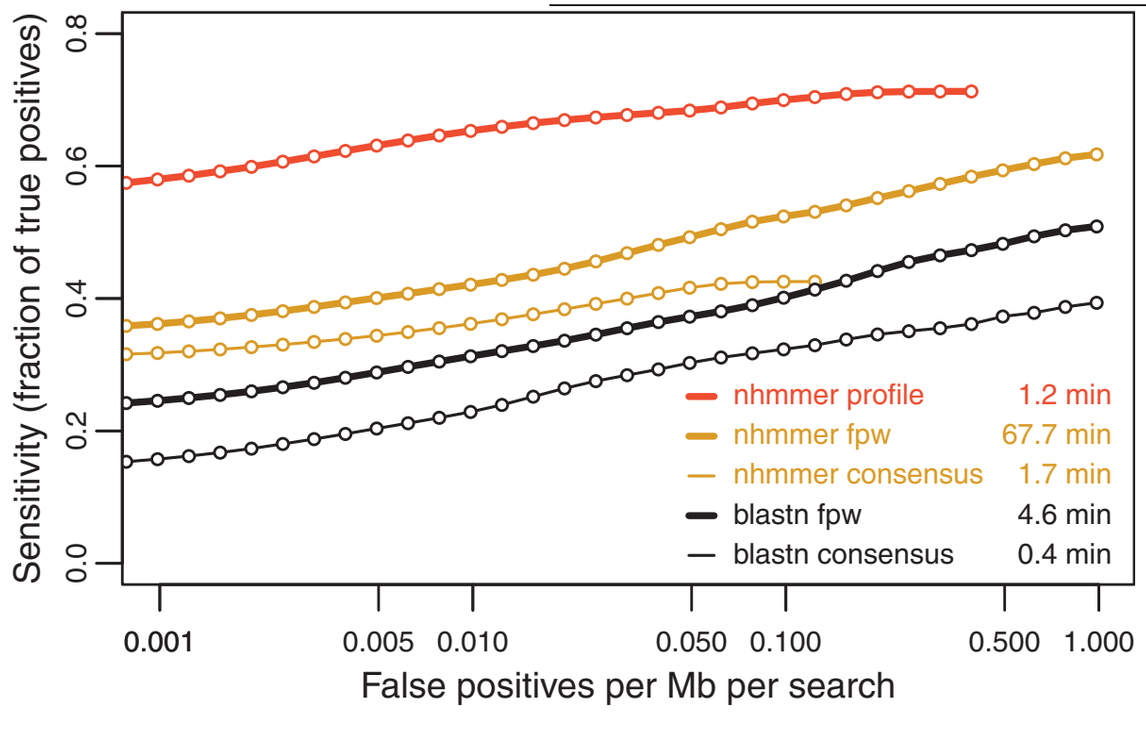


BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 19 2013, pages 2487–248
doi:10.1093/bioinformatics/btt40
Sequence analysis
Advance Access publication July 9, 201

nhmmer: DNA homology search with profile HMMs

Travis J. Wheeler* and Sean R. Eddy
HHMI Janelia Farm Research Campus, Ashburn, VA 20147, USA

Associate Editor: Alfonso Valencia



Amplicon data (V1V3)

>read1.1
ATGCATGCATGC...
>read1.2
ATGCATGCATGC...

Amplicon data (V3V5)

>read2.1
ATGCATGCATGC...
>read2.2
ATGCATGCATGC...

WGS data

>read3.1
ATGCATGCATGC...
>read3.2
ATGCATGCATGC...

MAPseq Algorithm
- word search + full alignment
- confidence estimation
- multiple taxonomy and OTU classification
- fast and accurate

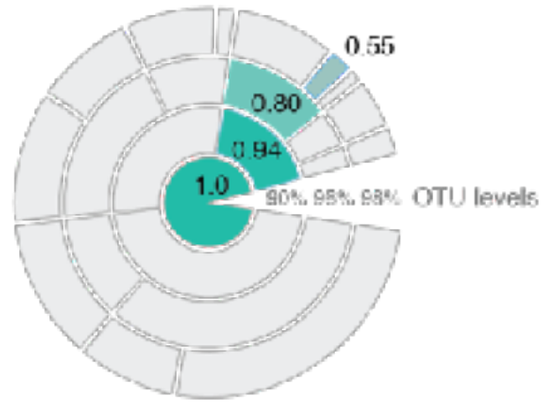
MAPseq Reference DB
- Full-length rRNA sequences
- Preprocessed taxonomies (NCBI, SILVA, custom)
- Preclustered hierarchical OTUs

Taxonomic classification

Bacteria (1.0); *Firmicutes* (0.90); ...; *Enterococcus sp.* (0.5)

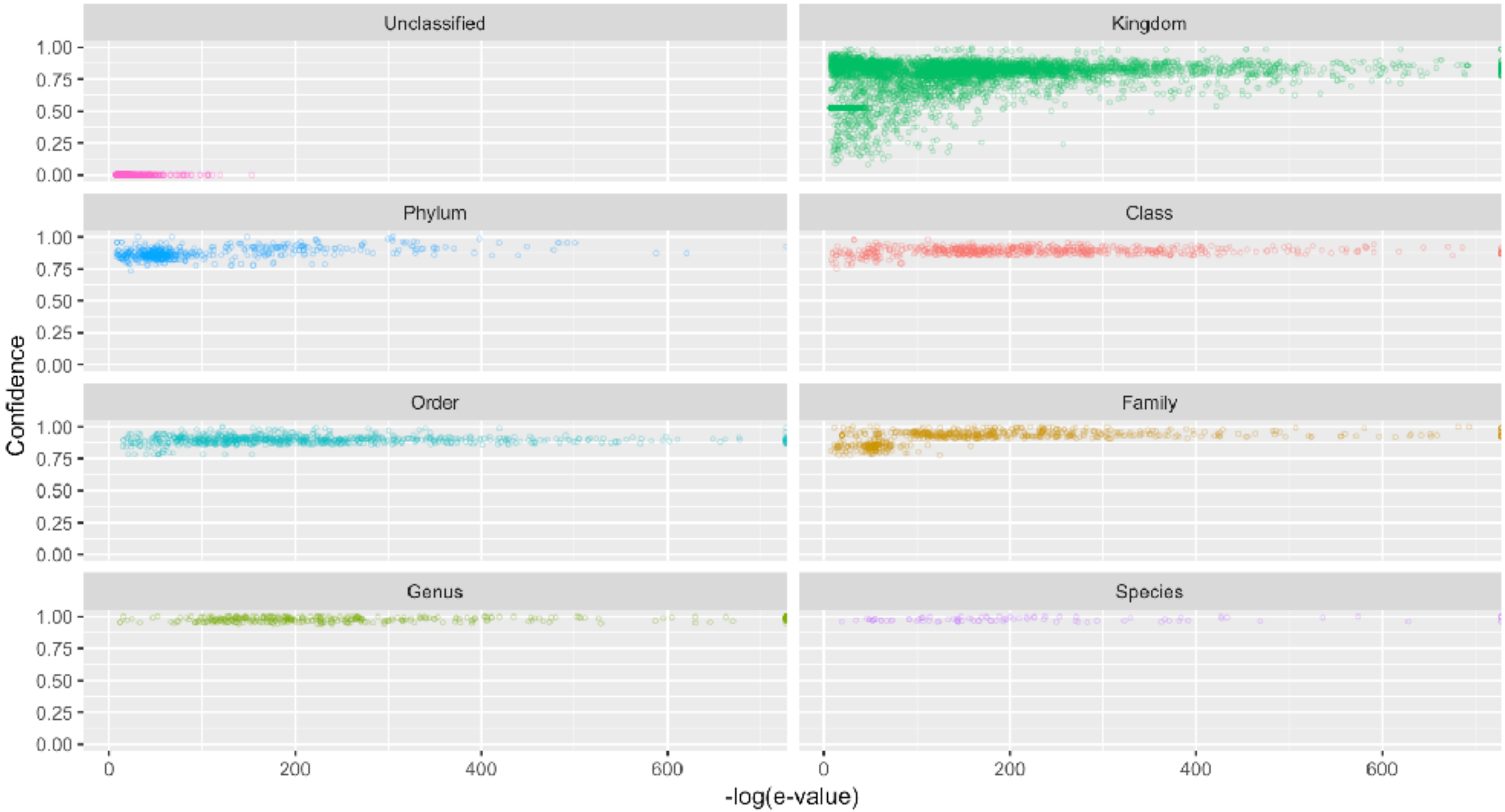
Hierarchical OTU Classification

Bacteria (1.0)
Γ132 (0.94)
G432 (0.80)
S85 (0.55)

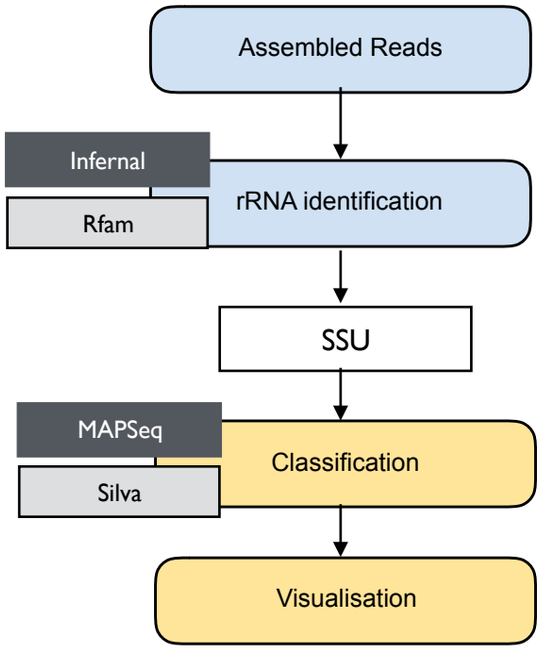


MAPseq from von Mering lab

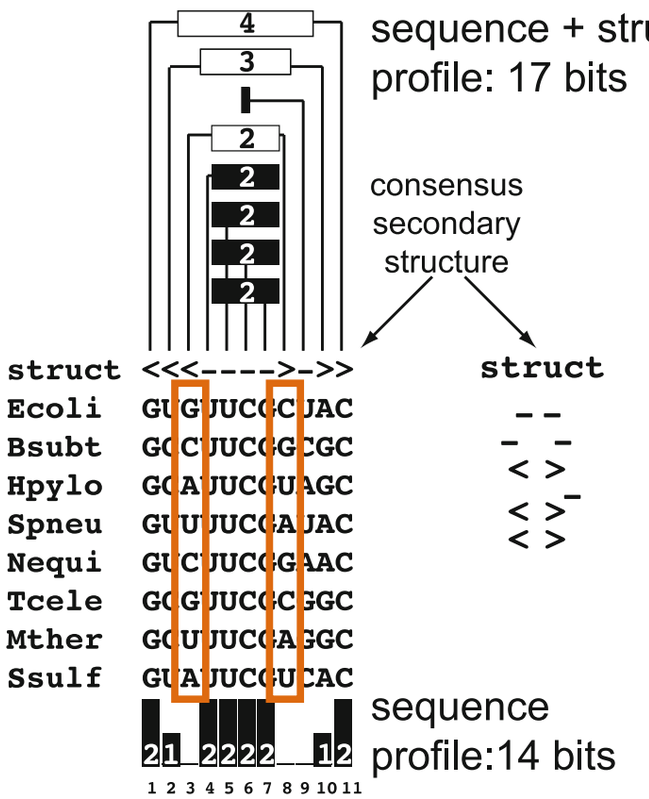
EMBL-EBI

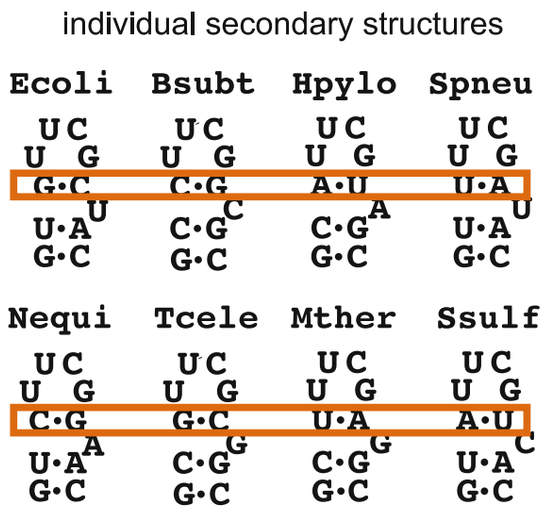


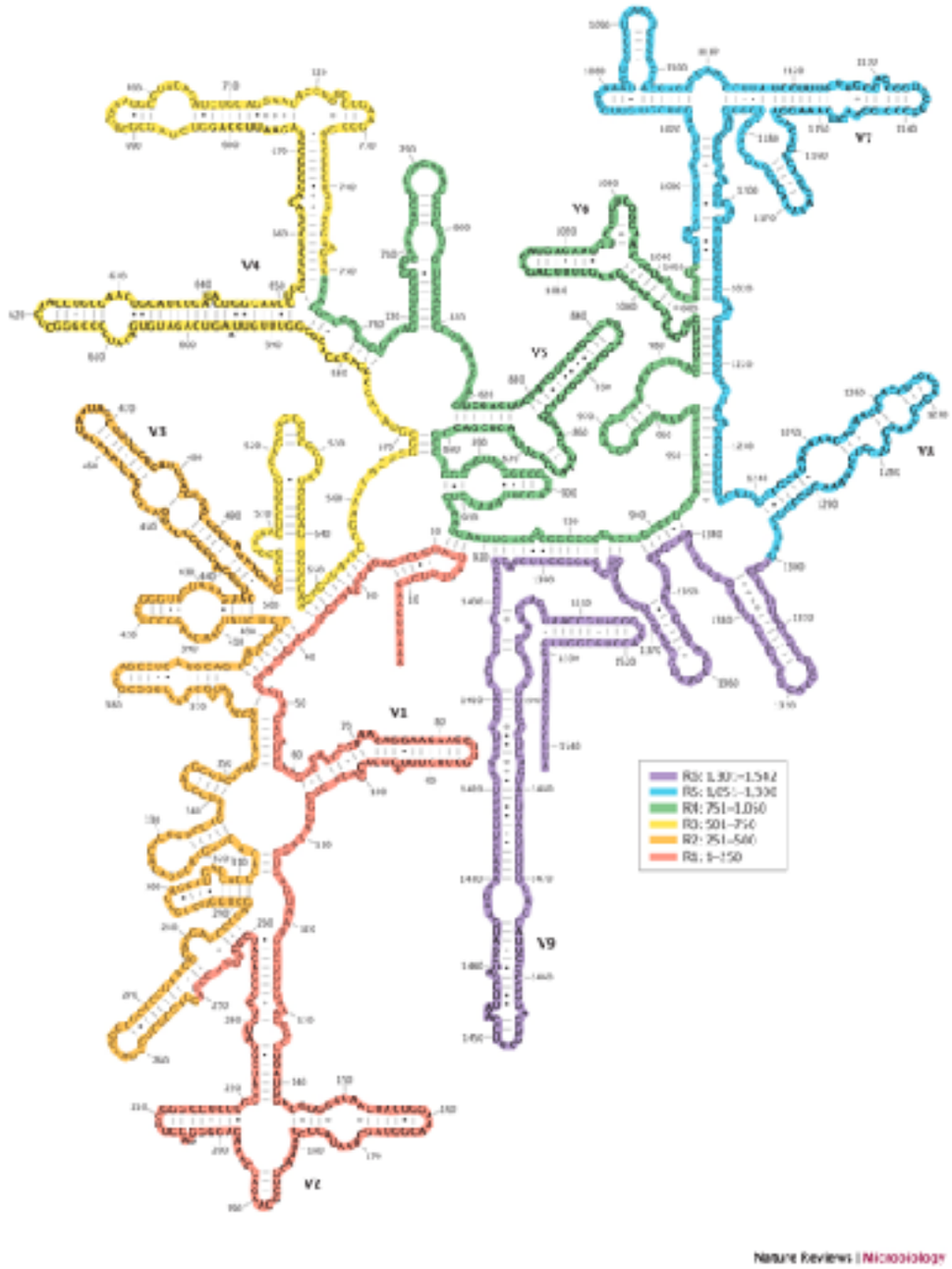
More classified, but majority ambiguous



Using full co-variance model

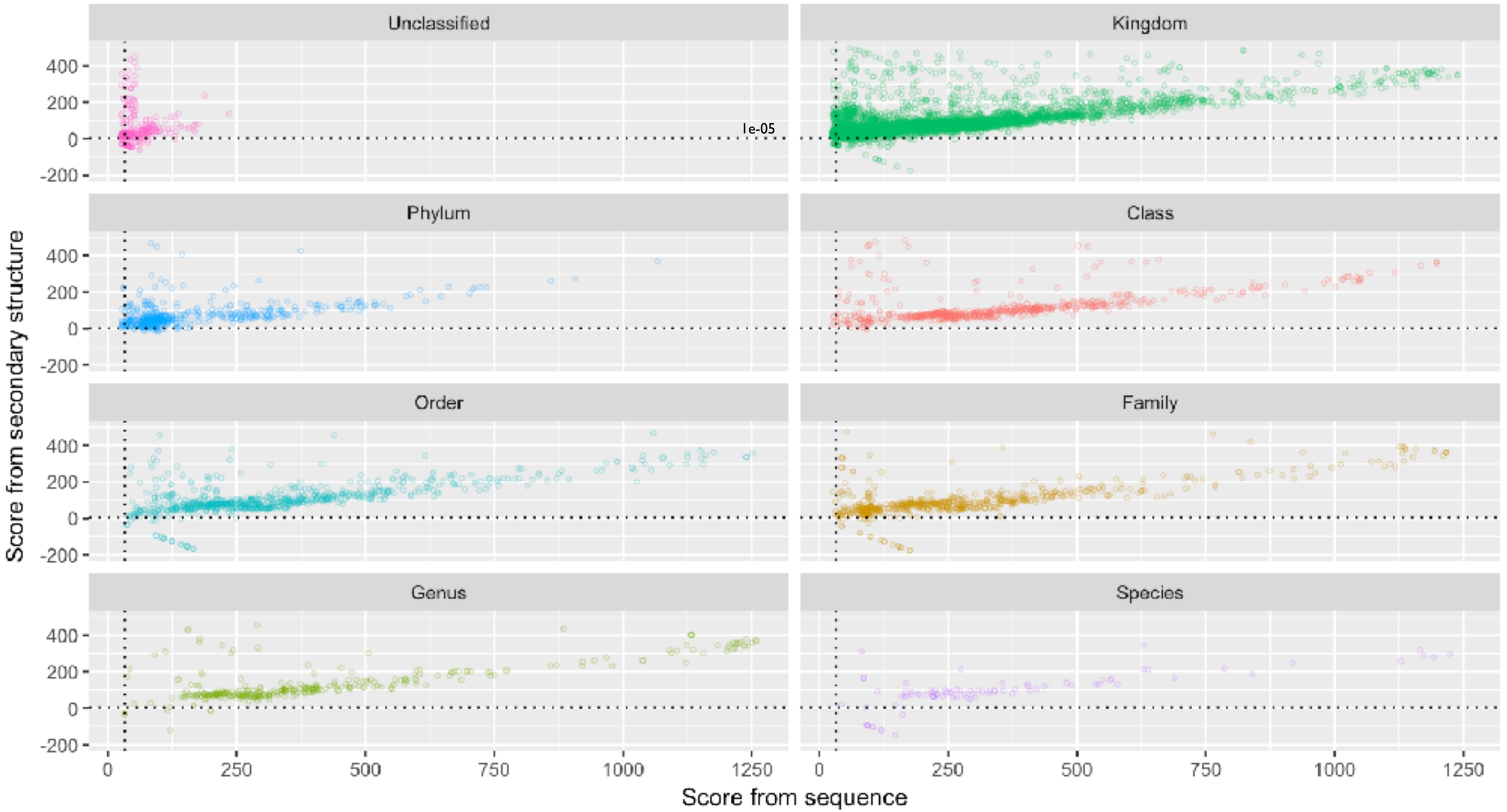






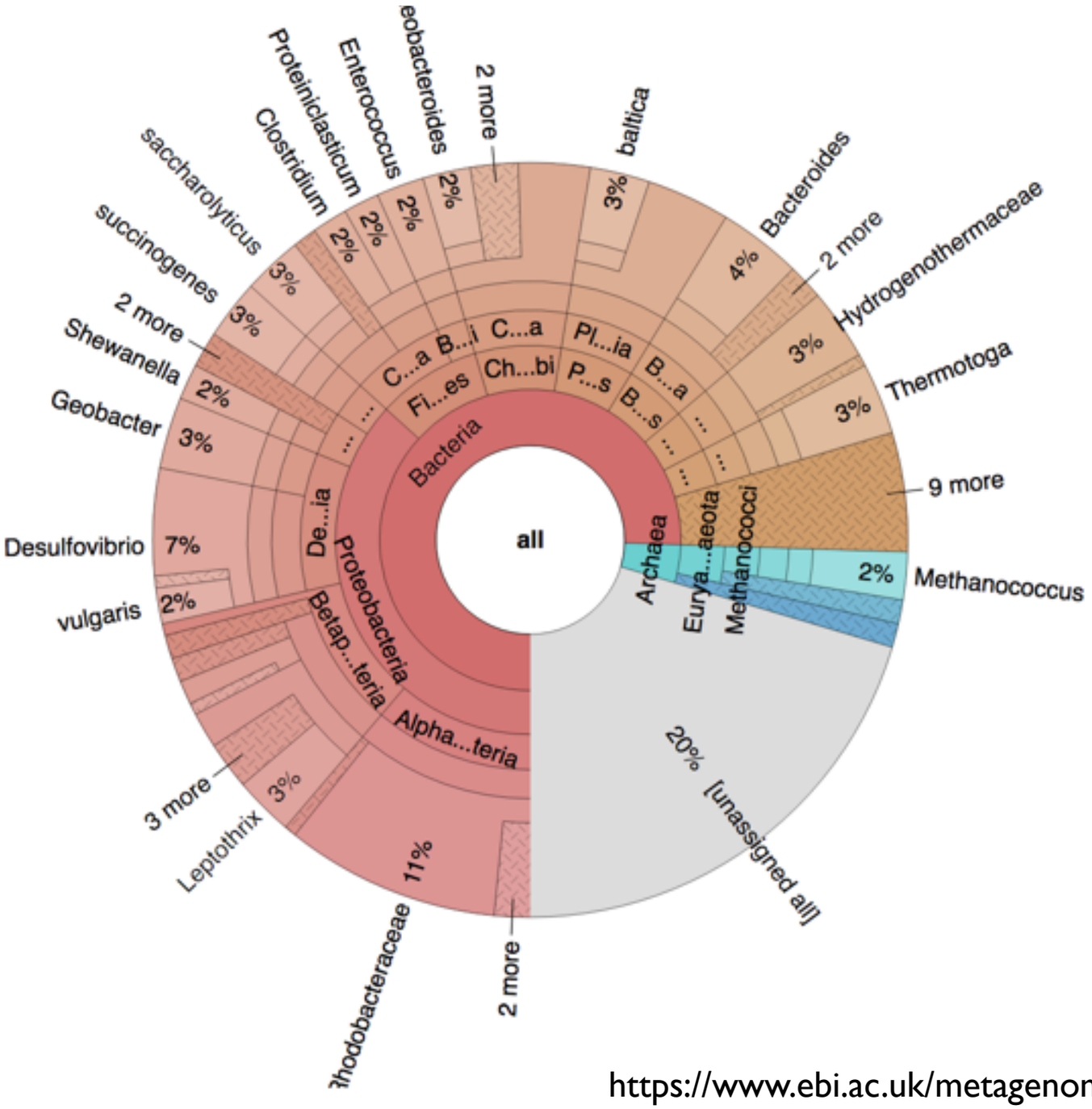
EMBL-EBI

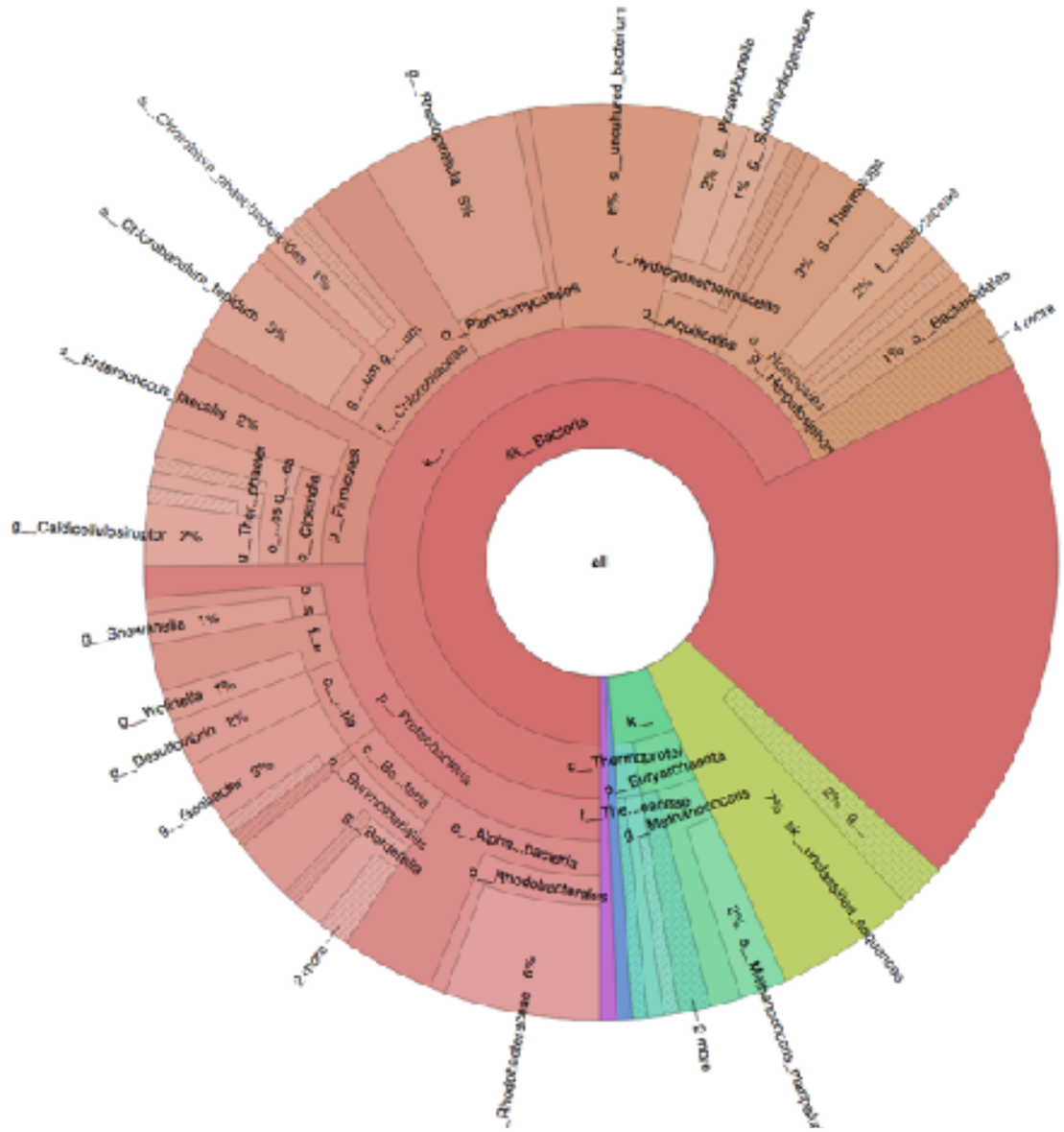
Valid SSU - incomplete reference databases



EMBL-EBI

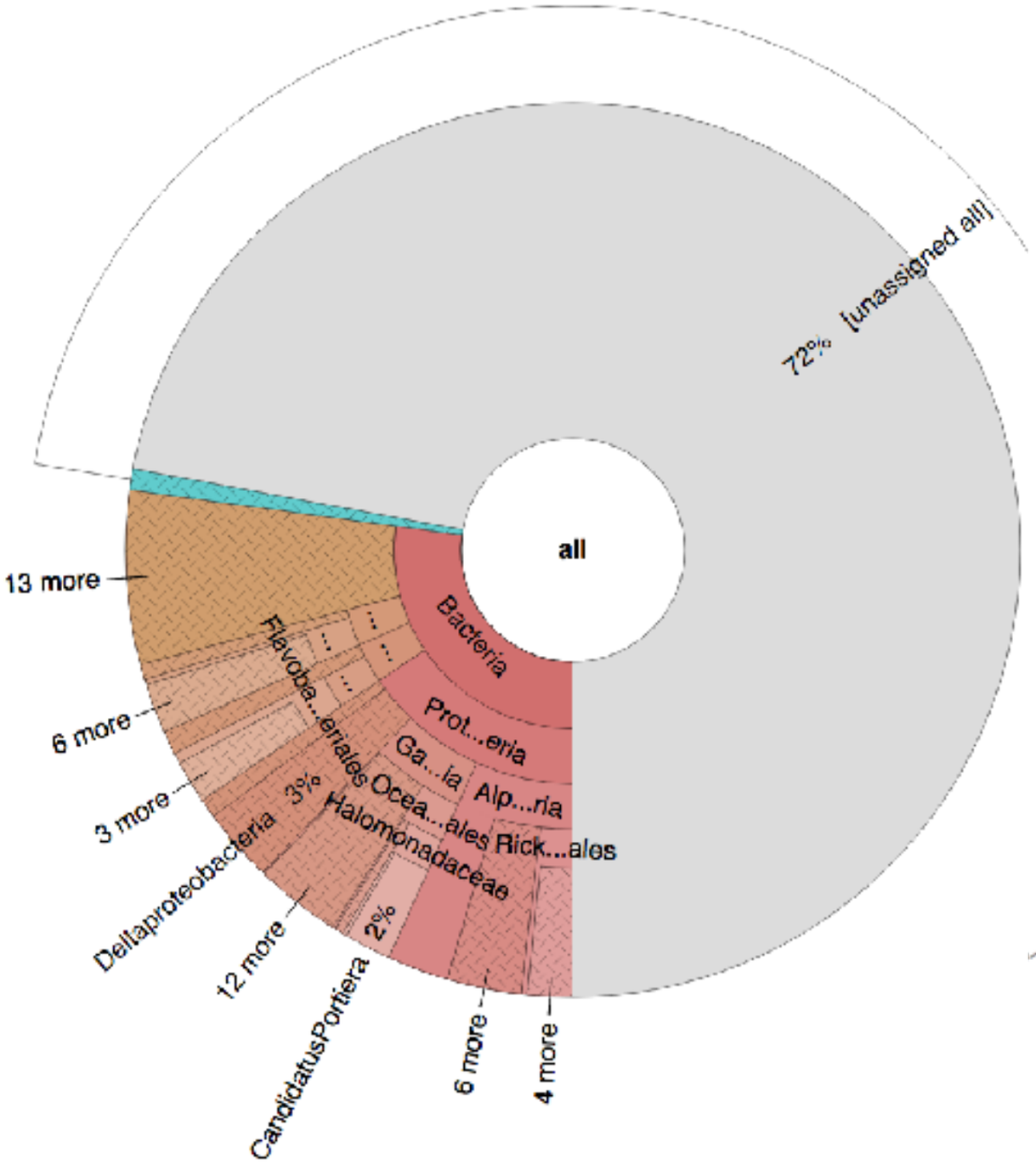
Mock Communities - Bacteria and Archaea

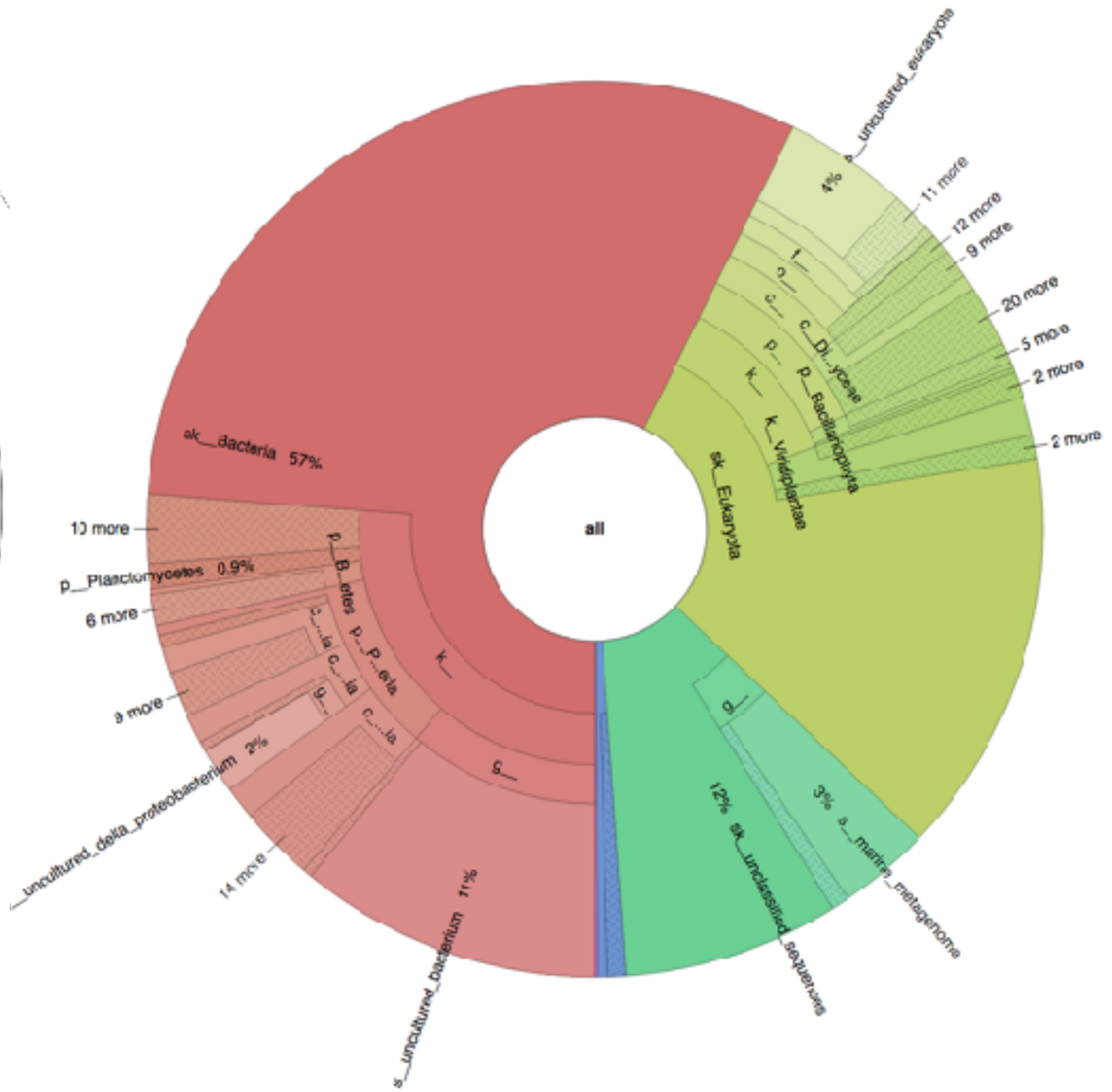




https://www.ebi.ac.uk/metagenomics/projects/SRP016523/samples/SRS372410/runs/SRR606245

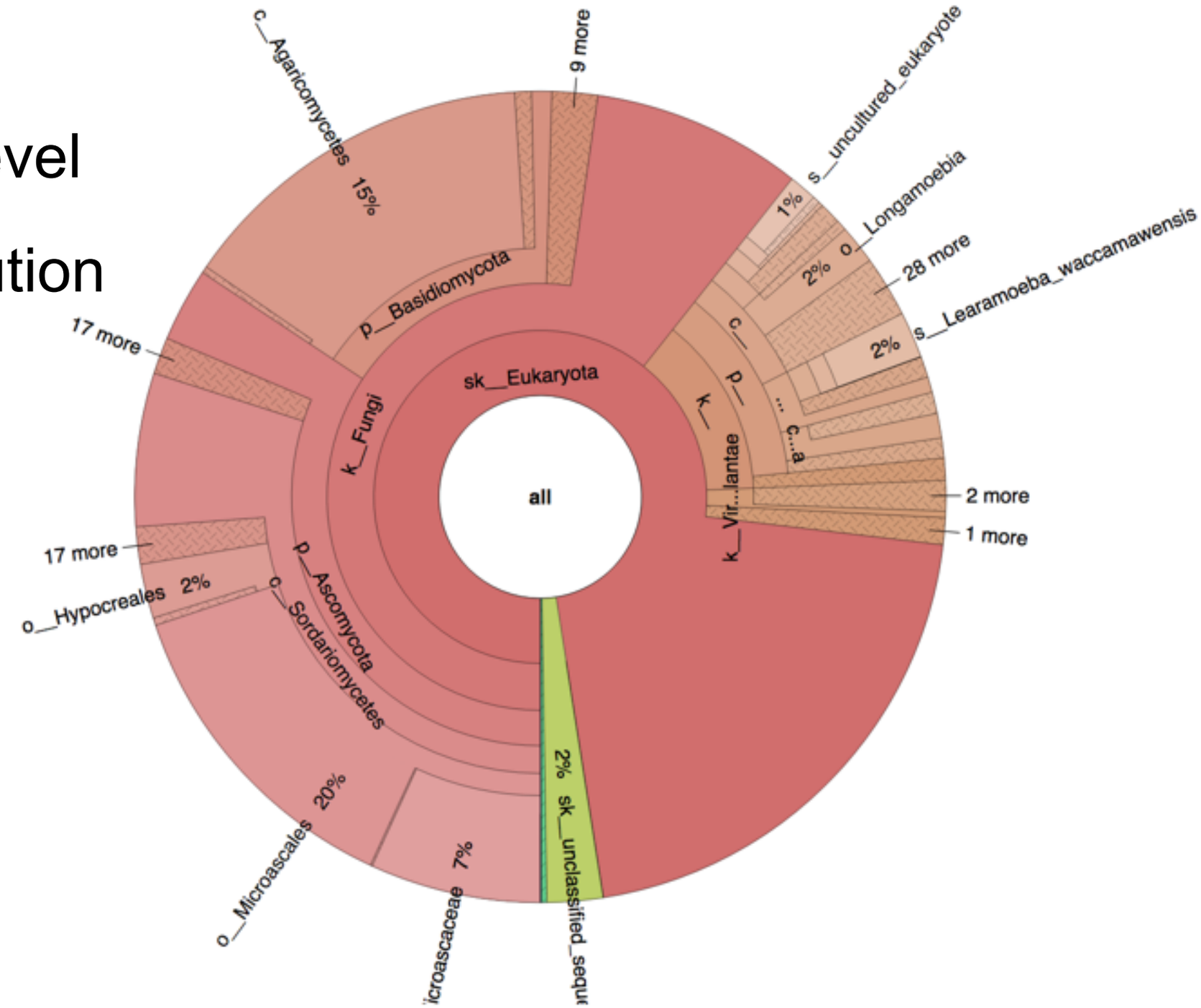
Improvement on Red Sea



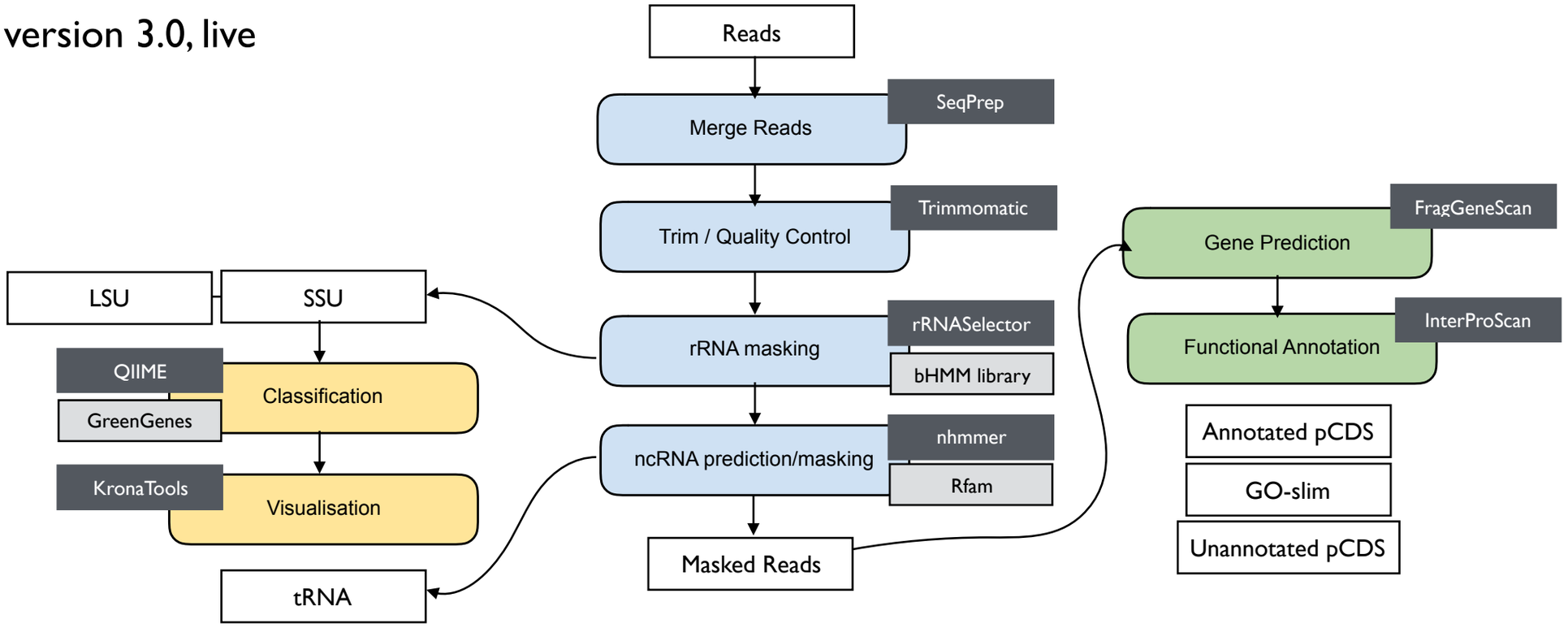


Complex soil communities

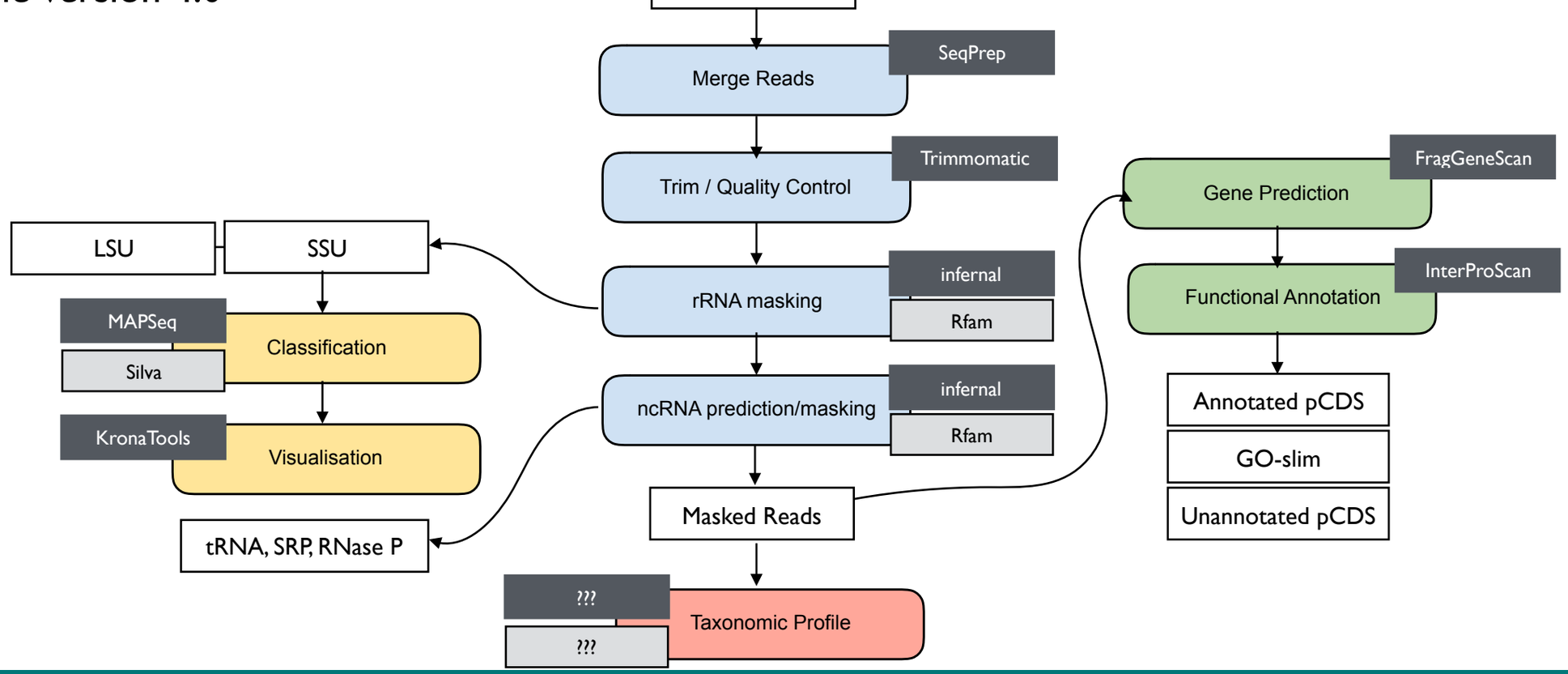
- 18S amplicon
- Class/order level
- Limit of resolution



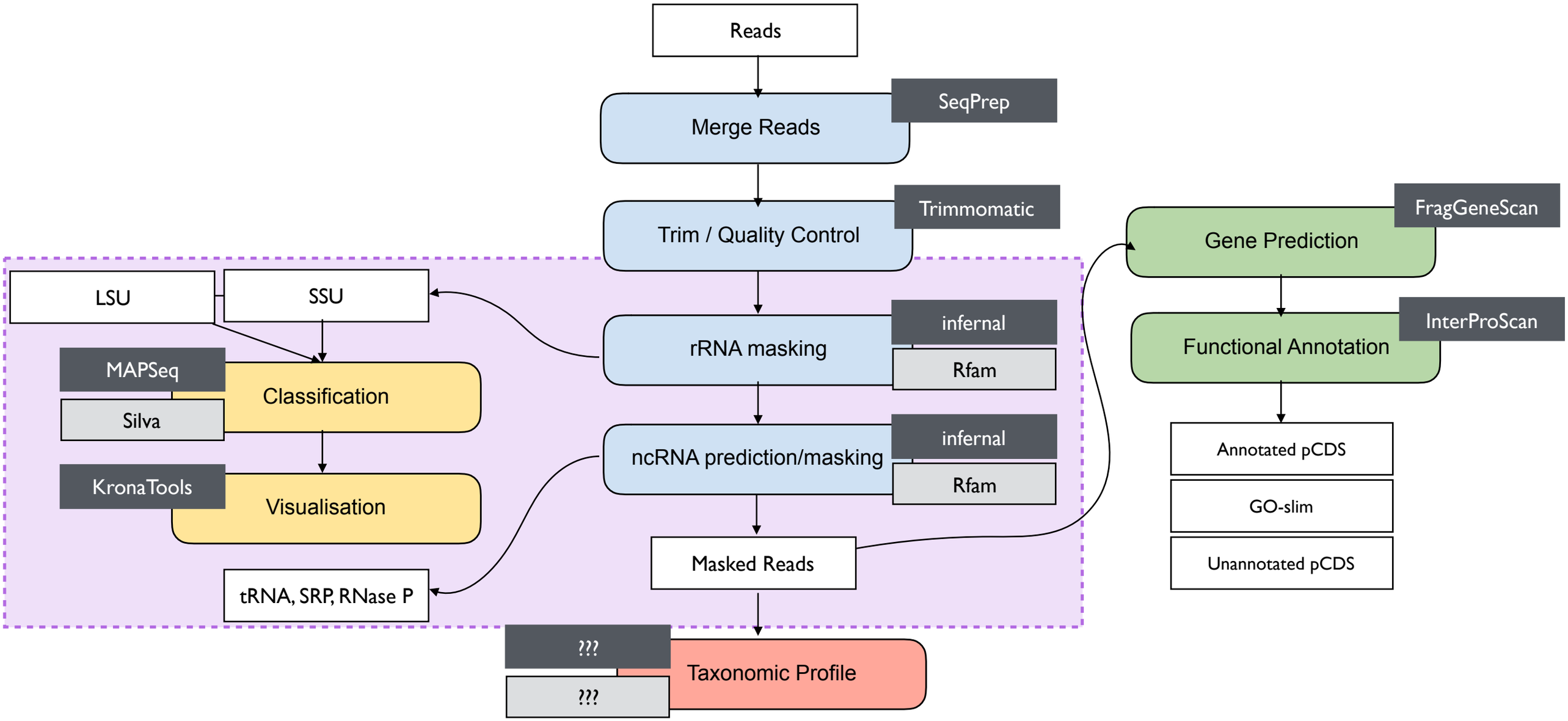
EMG, pipeline version 3.0, live



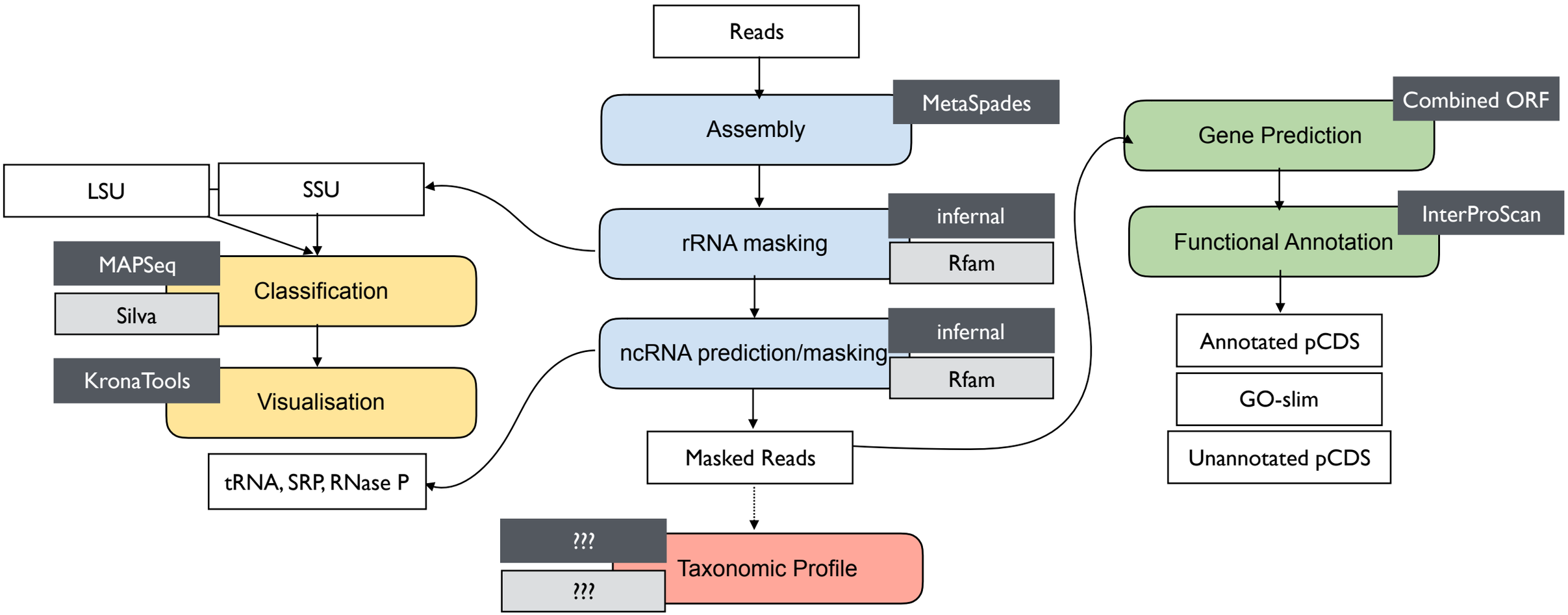
EMG, pipeline version 4.0



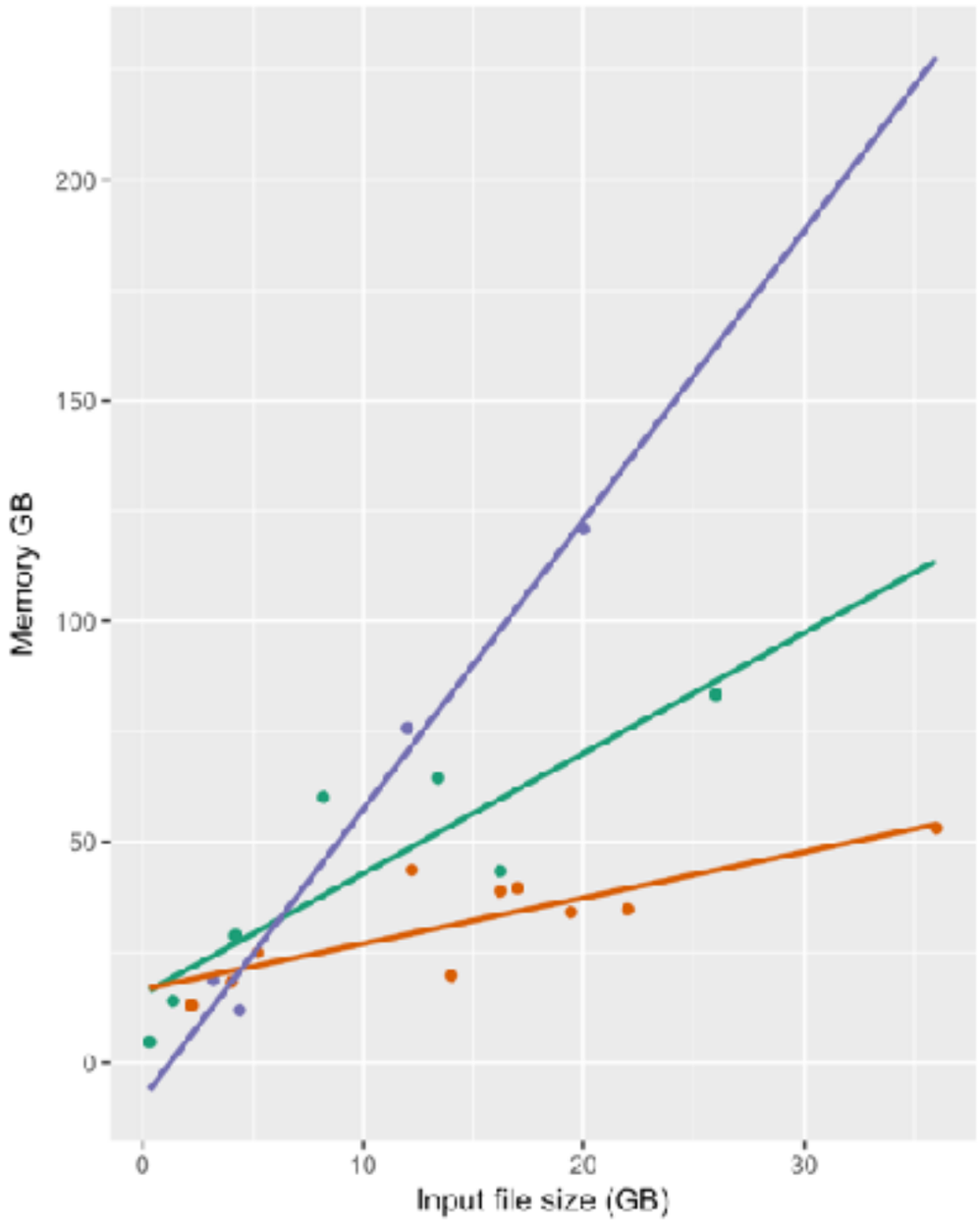
EMG, pipeline version 4.0

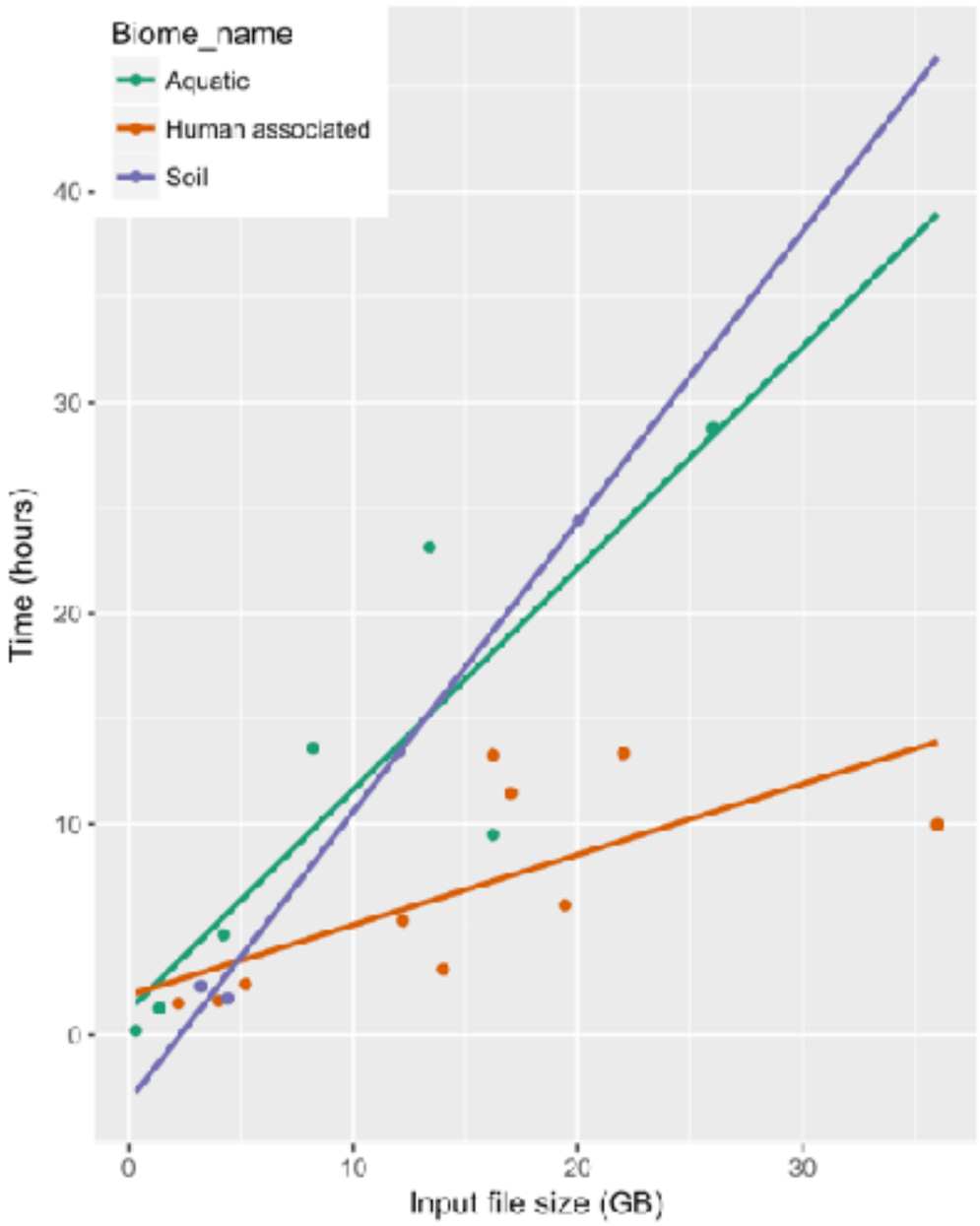


Assembly Pipeline

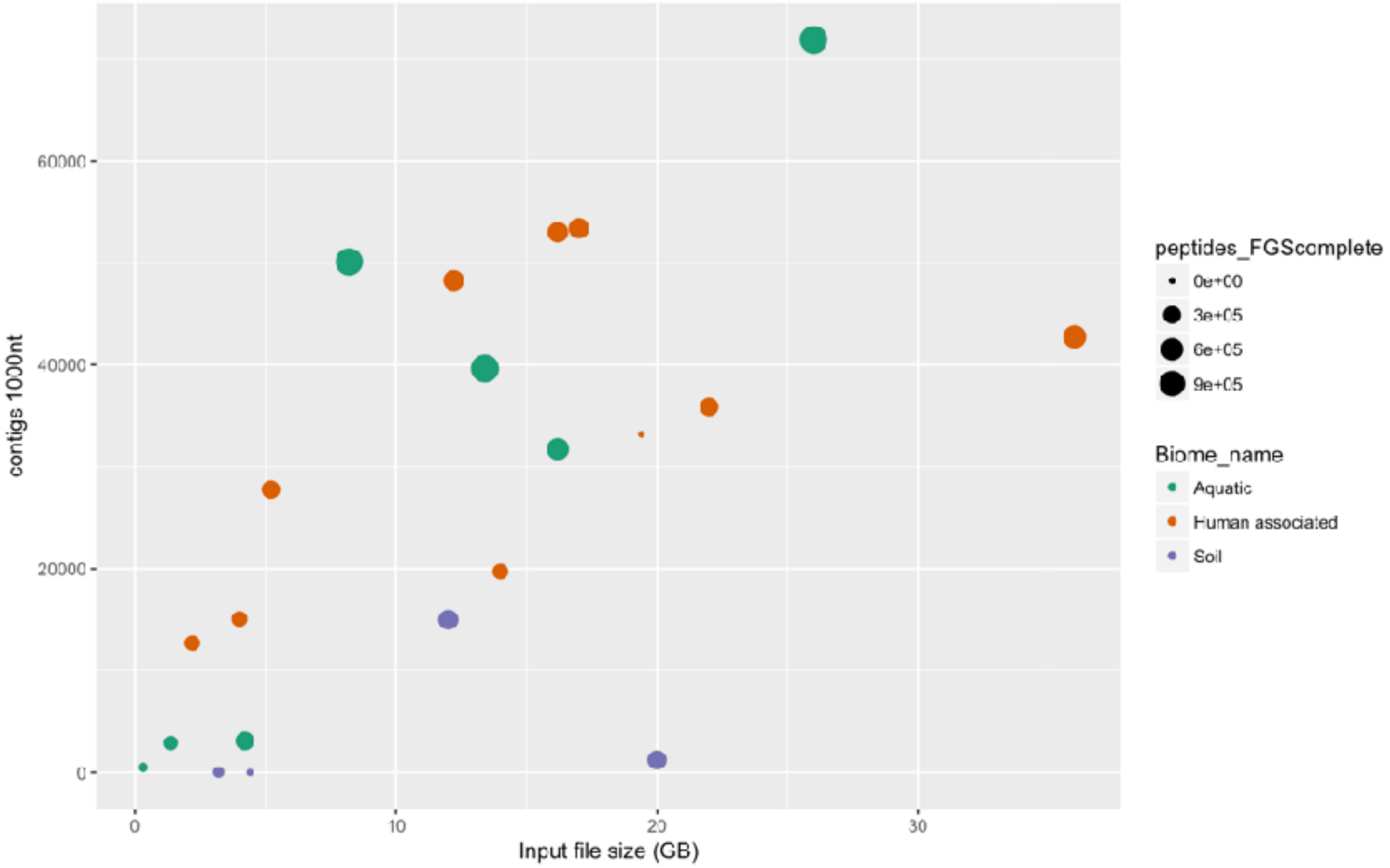


Assembly Profiling

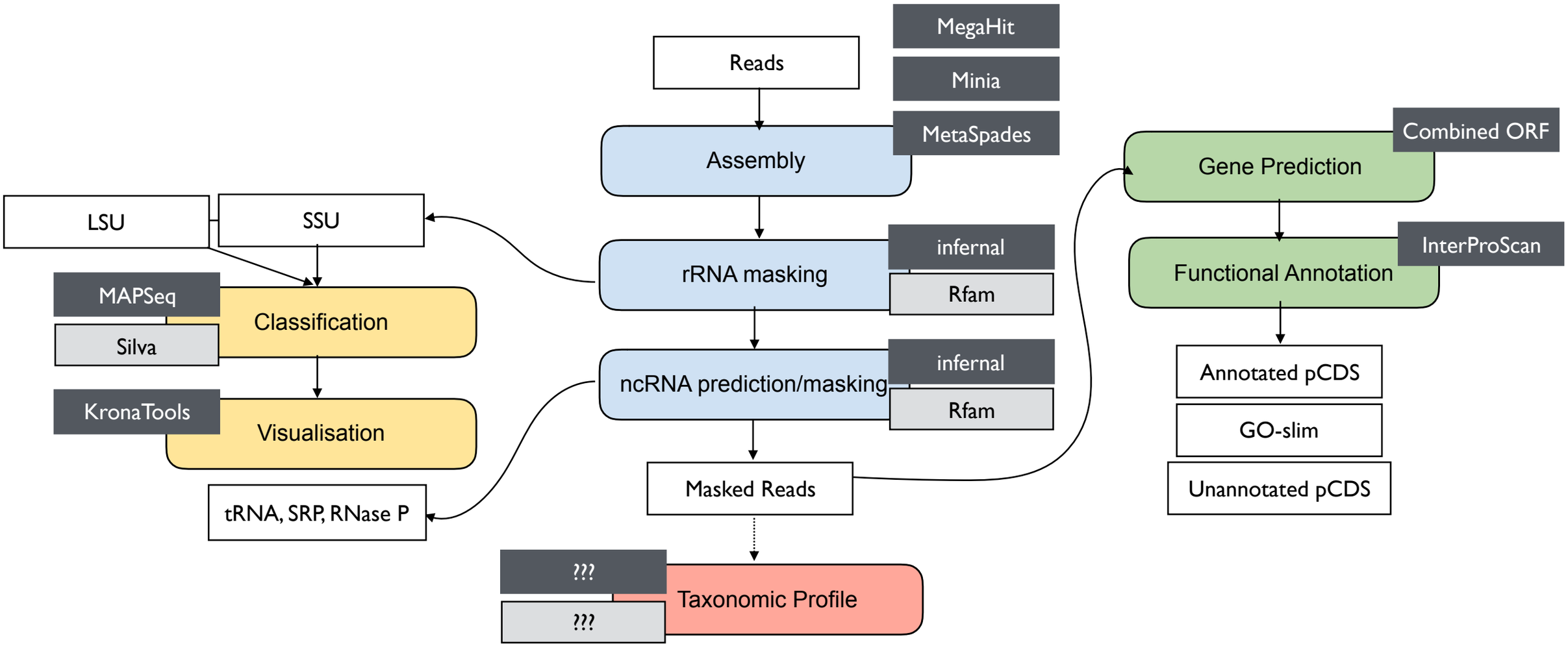




Peptides from some of our assemblies



Assembly Pipeline - multiple assemblers



Pipeline reproducibility and extension

- Formal description of each tool, inputs, parameters

```
#!/usr/bin/env cwl-runner
cwlVersion: v1.0
class: CommandLineTool

requirements:
  - class: InlineJavascriptRequirement
inputs:

  biomMethod:
    type: string
    inputBinding:
      position: 1
 ……..
  otuFormat:
    type: string
    inputBinding:
      prefix: "--"
      separate: false
      position: 4
    default: “to-json"

baseCommand: [ biom ]

outputs:
  otuBiom:
    type: File
    outputBinding:
      glob: otu_table.json
```

Tool

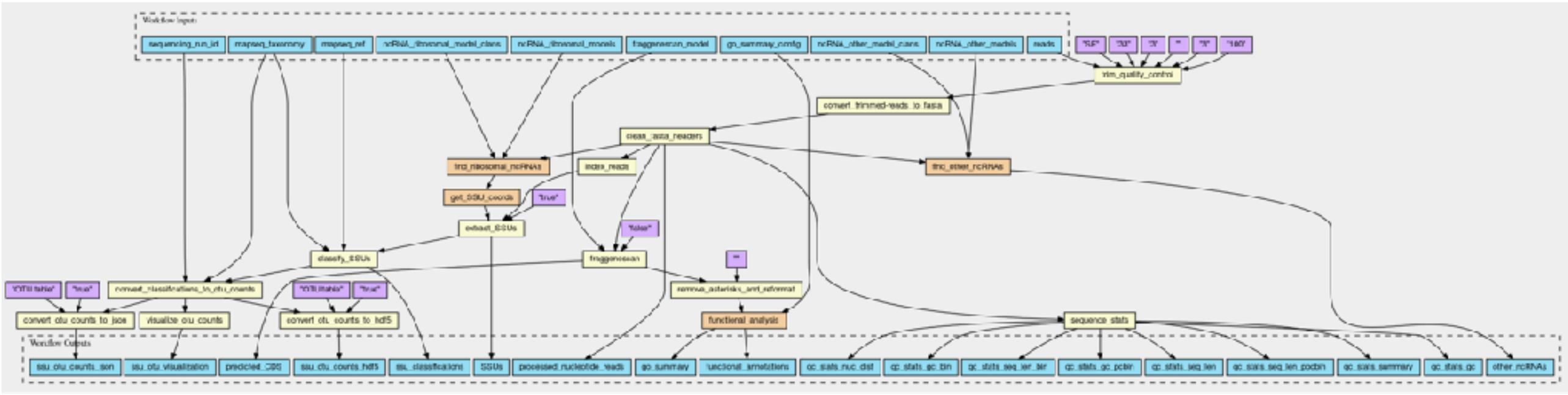
```
biomMethod: "convert"
otuTable:
  class: File
  path:  Other_otus.txt
otuFormat: "to-json"
otuOutputFile: "otu_table.biom"
```

Parameters

Pipeline reproducibility and extension

CWL

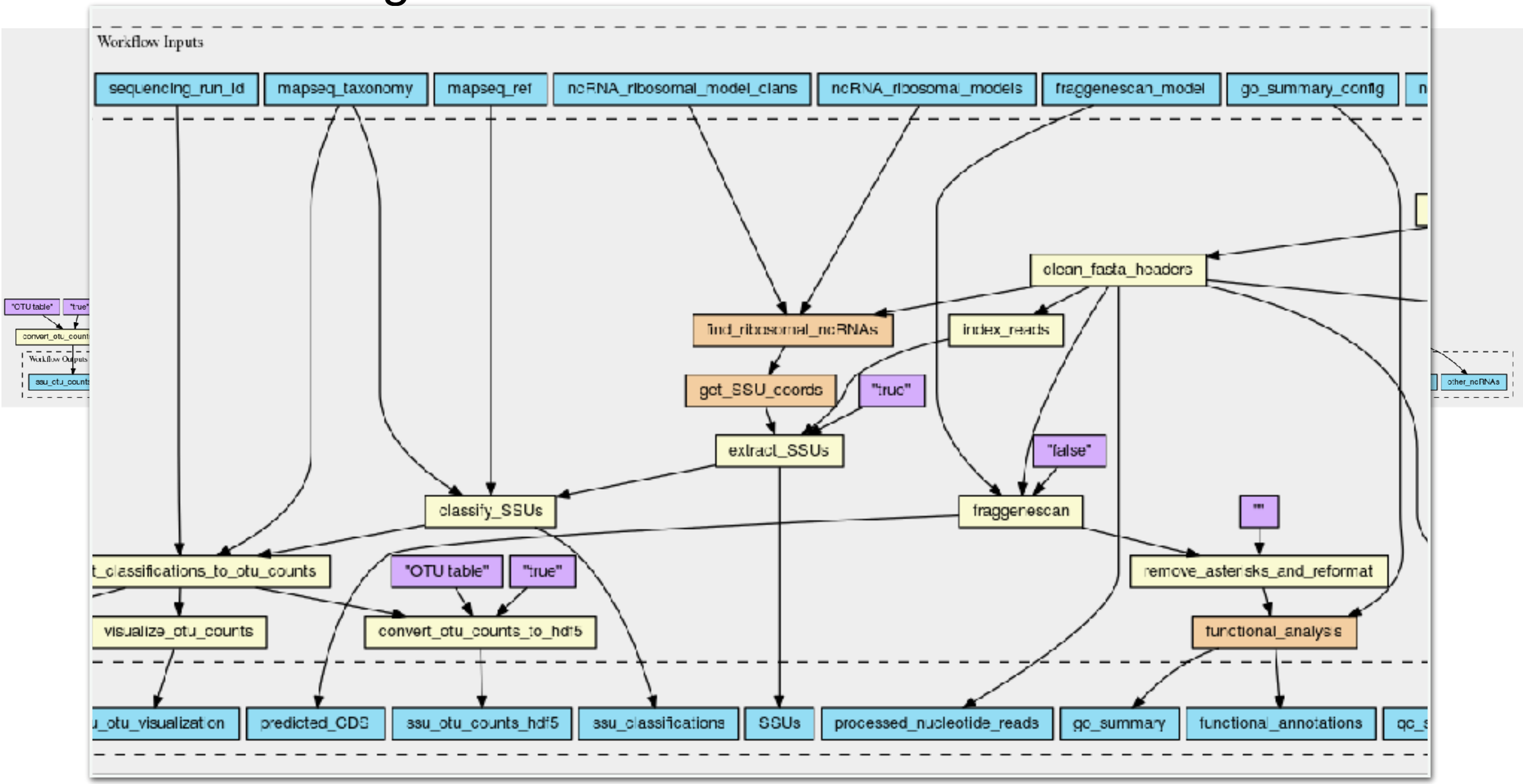
- Combining tools into workflows



EMBL-EBI

Pipeline reproducibility and extension

- Combining tools into workflows



Acknowledgements

EMBL-EBI

The Sequence Families team:
Joanna Argasinska
Matthias Blum
Boris Burkov
Hsin-Yu Chang
Hubert Denise
Sara El-Gebali
Matthew Fraser
Ioanna Kalvari
Aurelien Luciani
Jaina Mistry
Alex Mitchell
Gift Nuka
Typhaine Paysan-Lafosse
Sebastien Pesseat
Anton Petrov
Simon Potter
Matloob Qureshi
Neil Rawlings
Lorna Richardson
Gustavo Salazar-Orejuela
Amaia Sangrador
Maxim Scheremetjew
Blake Sweeney
Siew-Yit Yong

The ENA team:
Guy Cochrane
Petra ten Hoopen + many others in ENA team

External Collaborators:
InterPro Consortium

Harvard
Sean Eddy

University of Montana
Travis Wheeler

BioCatalysts Ltd
Mark Blight

CWL community developer
Michael Crusoe

Newcastle University
Tom Curtis
Darren Wilkinson

WTSI
Trevor Lawley
Sam Forster

University Tromso
Nils-Peder Willassen

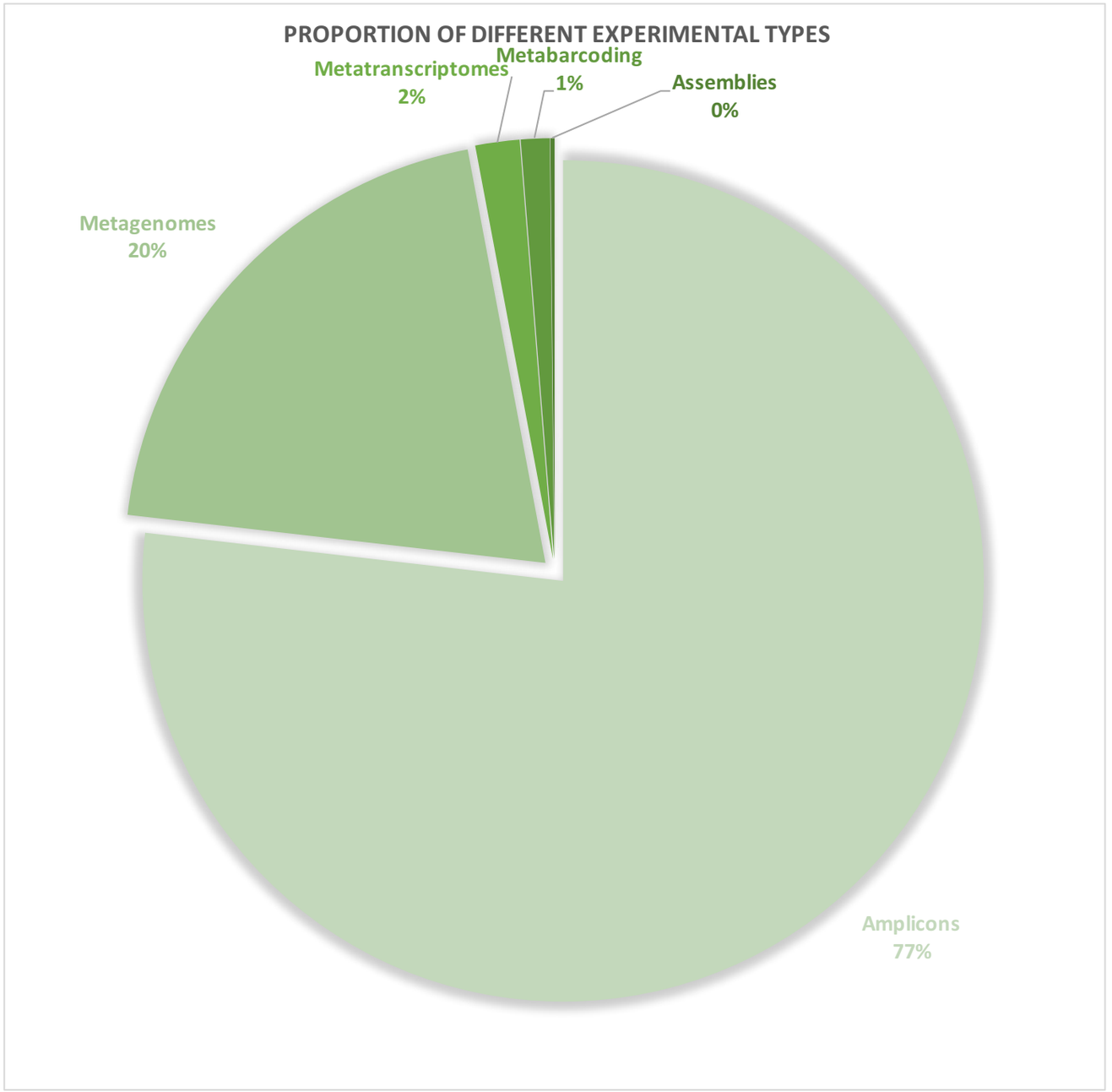
Ghent University
Bart Mesuere
Peter Dawyndt

EMBL-EBI

EBI Metagenomics

- Nucleotide sequence reads: >**250 billion**
- Average length per sequence: **120 nt**

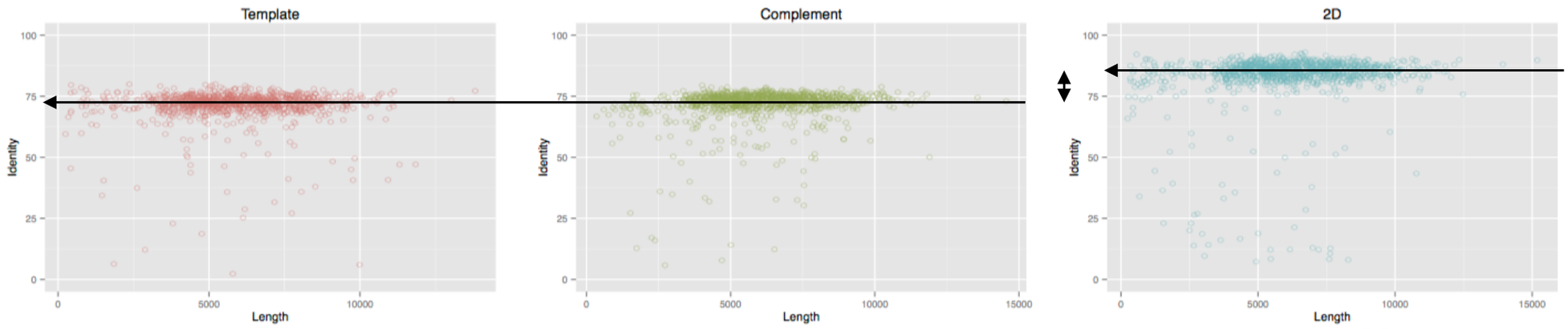
- Predicted rRNAs**: 3.6 billion**
- Predicted CDS: **150 billion**
- Total InterProScan matches: **50 billion**



Emerging trends - ONT

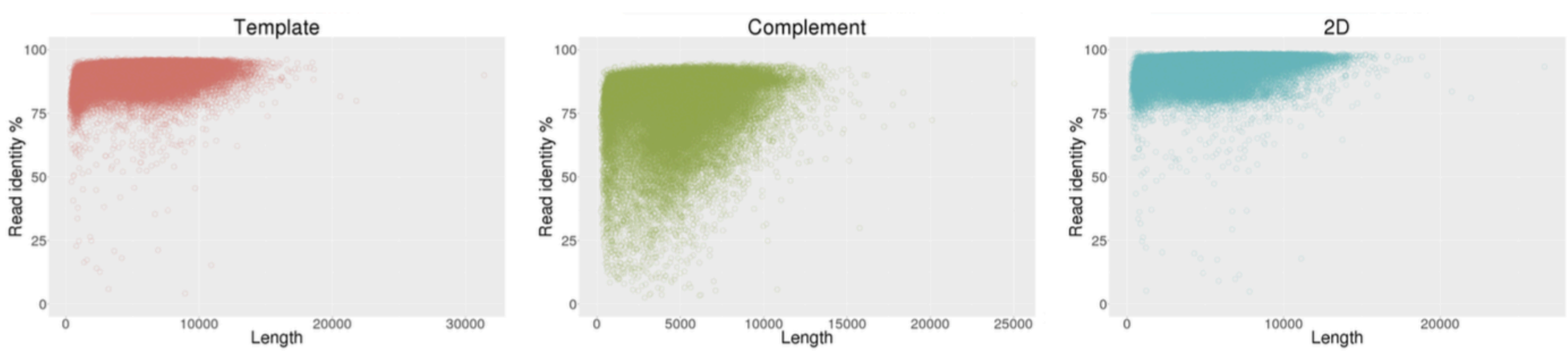
Compare different chemistry accuracies

- R7.3



Mean ~85%

- R9



Mean ~95%

www.earlham.ac.uk

Earlham Institute

EMBL-EBI