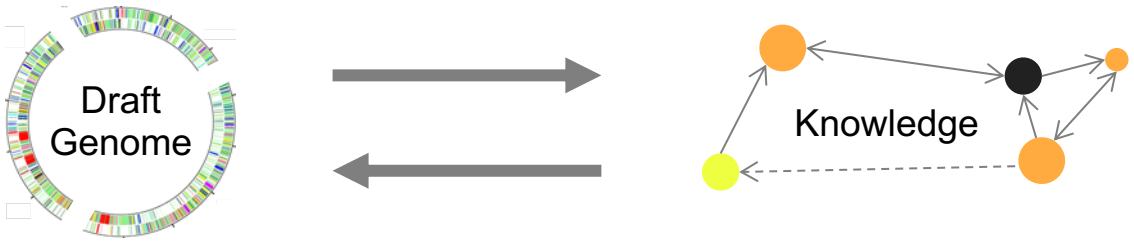
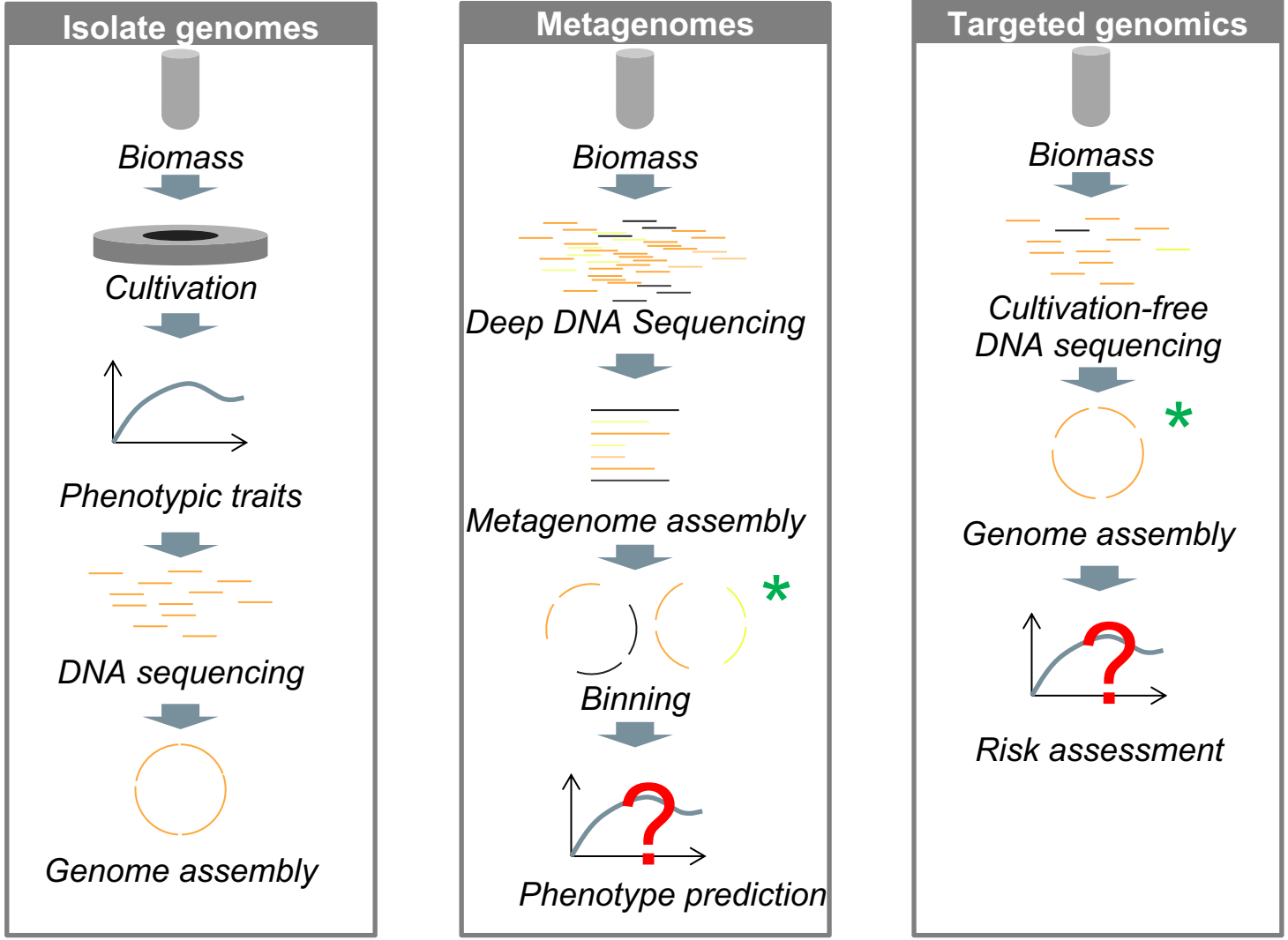


Computational modeling and learning methods for metagenomics of microbial and viral communities

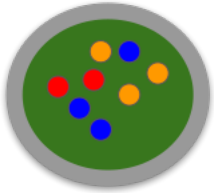
Thomas Rattei
Department of Microbiology and Ecosystem Science
University of Vienna

Microbial communities



* Genome standards for Single-Cell genomes (MISAG) and genomes from metagenomes (MIMAG) of bacteria and archaea.
Bowers et al., Nat Biotechnol 2017

PhenDB



Prediction of bacterial phenotypes

Cube

Division of Computational Systems Biology

Valerie Eichinger

Roman Feldbauer

Javier Geijo

Patrick Hyden

Lukas Lüftinger

Florian Piewald

Peter Peneder

FWF

Der Wissenschaftsfonds.

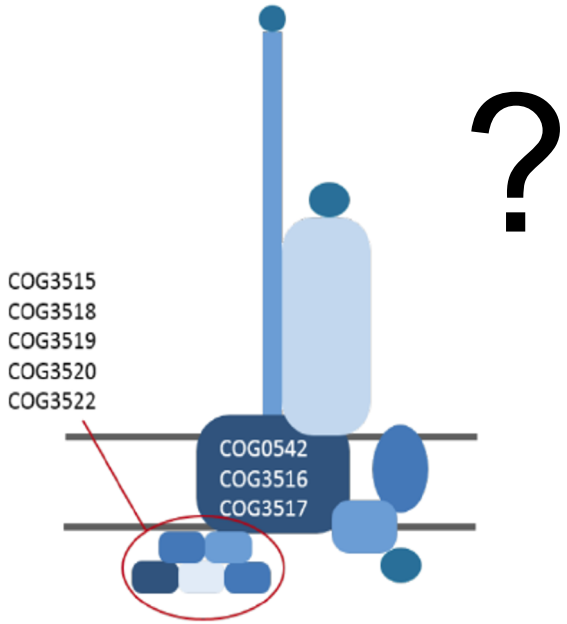
Rationale for PhenDB

Multiple coverage binning

We can recover **nearly complete** genomes.

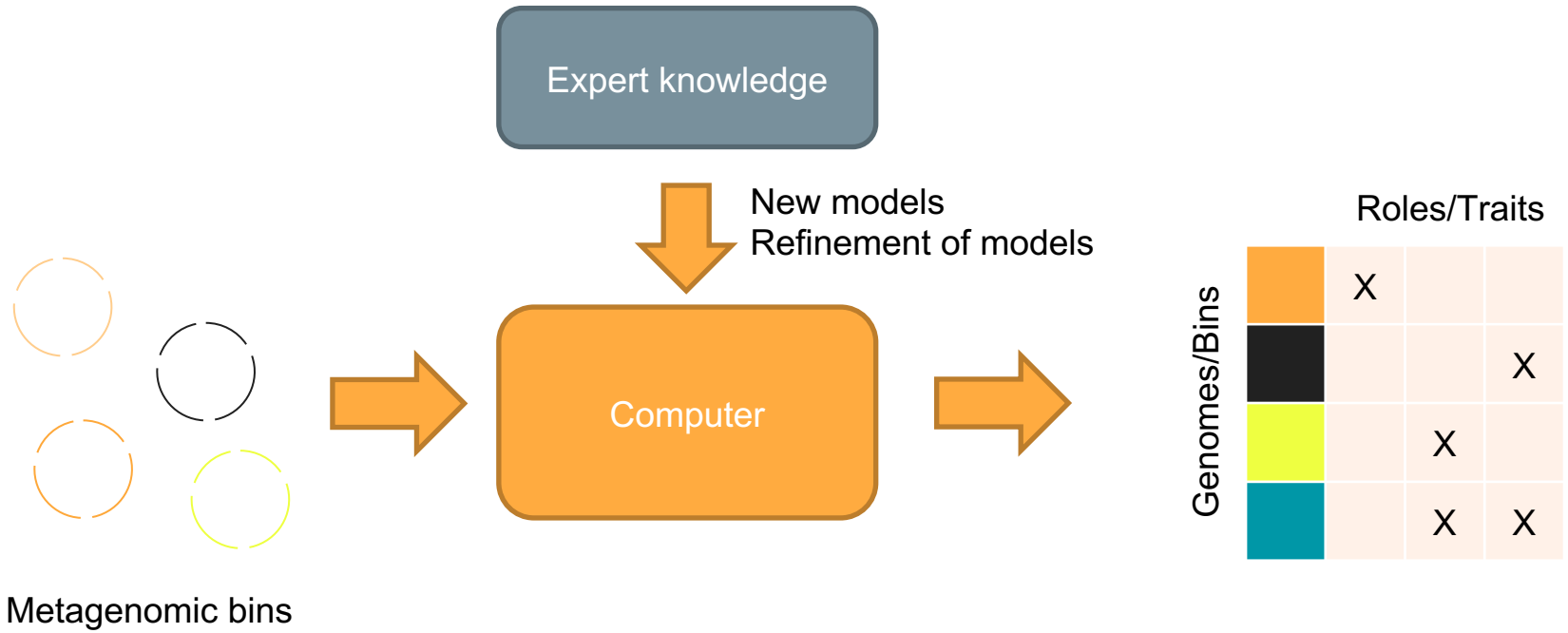
The genome is a **major source of information**.

Lot of genomes are coming from metagenomes. **Speed up data analysis.**

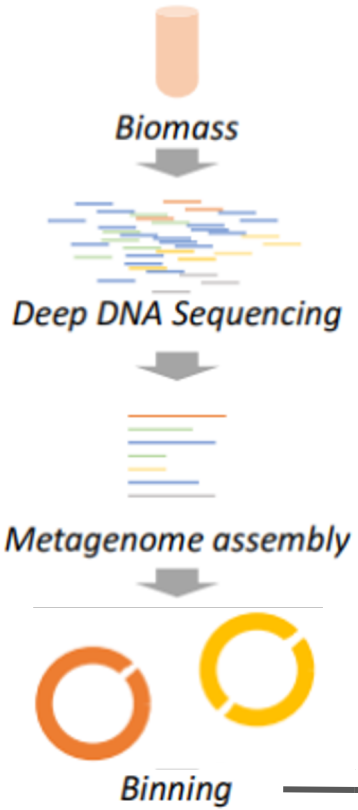


Is there any method able to infer complex traits from nearly complete genomes?

Role and trait prediction in microbial communities



Trait prediction for nearly complete genomes



Method	Features	Predictions	Publication
PICA	Protein family occurrence (COG/eggNOG)	Microbial roles, metabolic traits, protein modules, resistance	Feldbauer et al., BMC Bioinformatics 2015

Computational method based on PICA.

Learns genotype-phenotype association models from completely sequenced genomes of microbes with known traits.

These models can be used to predict the traits in novel genomes.

The whole genome architecture can be used to predict the trait.

Current PICA models

Human pathogens
Penicillin resistance of S. aureus

Symbionts
Obligate intracellular lifestyle
Functional T3 Secretion System
Functional T4 Secretion System
Functional T6 Secretion System

Metabolism
Ammonia-oxidizing bacteria
Nitrite-oxidizing bacteria
Nitrogen fixation
Hydrogen production
Alkane degradation
Arsenic detoxification
Methane utilization as carbon source

Lifestyle
Autotrophic metabolism
Phototroph bacteria

Habitat compatibility
Halophile
Psychrophile
Thermophile
Mesophile
Aerobic respiration
Anaerobic respiration
Facultative anaerobe

Physiology
Motility
Spore formation
Gram negative bacteria

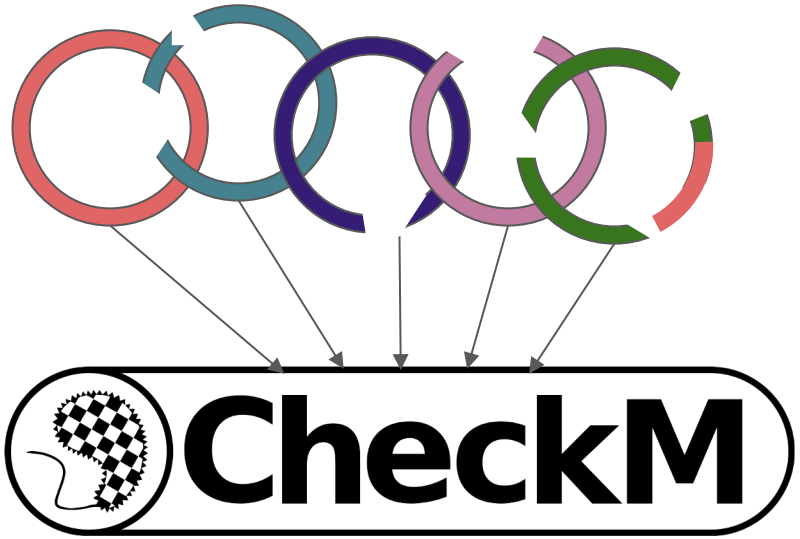
Metabolism
Plant pathogenicity based on AvrE virulence factor
Secondary bile acids production
Benzoate degradation via hydroxylation
Phytate hydrolisis
Glycine and taurine are de-conjugation from bile salts
Butyrate-production
Chitine degradation
Aromatic-ring-hydroxylation
Carbon monoxide assimilation
Trimethylamine production via choline
Bibenzofuran (DF) dibenzo-p-dioxin (DD) degradation
Steroid hormone metabolism
Naphthalene degradation
Fatty acid degradation
Phosphonate hydrolisis
Recycling of organic phosphorus
Capsular polysaccharide biosynthesis
Thiosulfate oxidation
Plant pathogenicity based on Thaxtomins
Ureolytic activity
Reduction of various α,β-unsaturated and nitro compounds

First step: quality assessment

Metagenomic bins

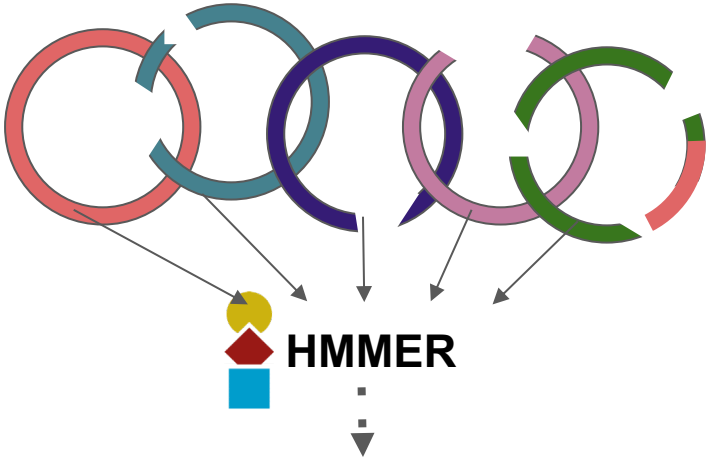
processing ….
(roughly 10 min per bin)
requires 20+ GB RAM

Completeness, contamination

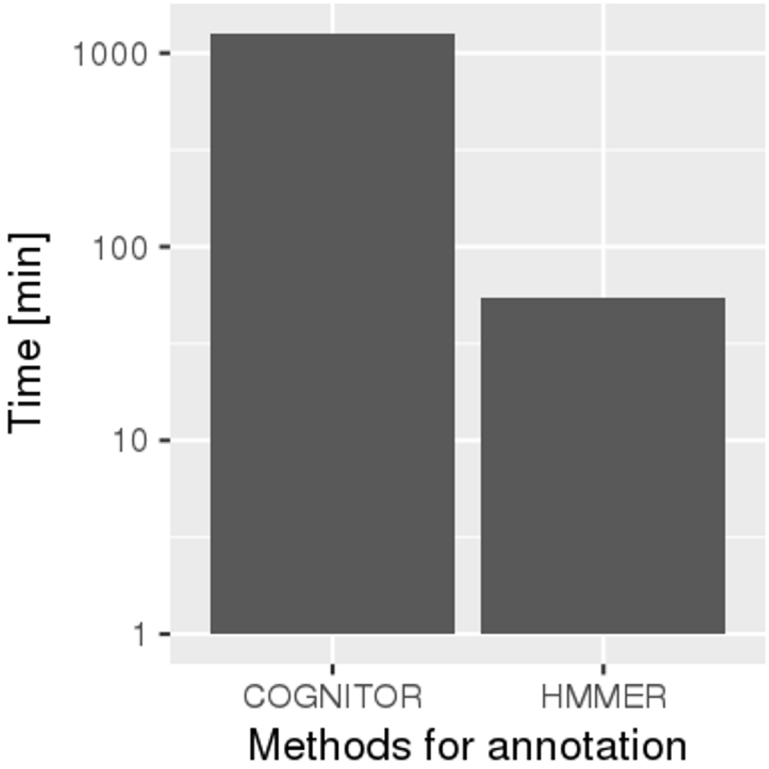


bin_ID	1	2	3	4	5
comp. [%]	99.9	94.5	93.6	91.2	91.3
cont. [%]	0.0	0.0	0.0	0.0	12.2

Second step: annotating genomes

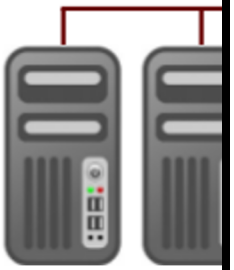


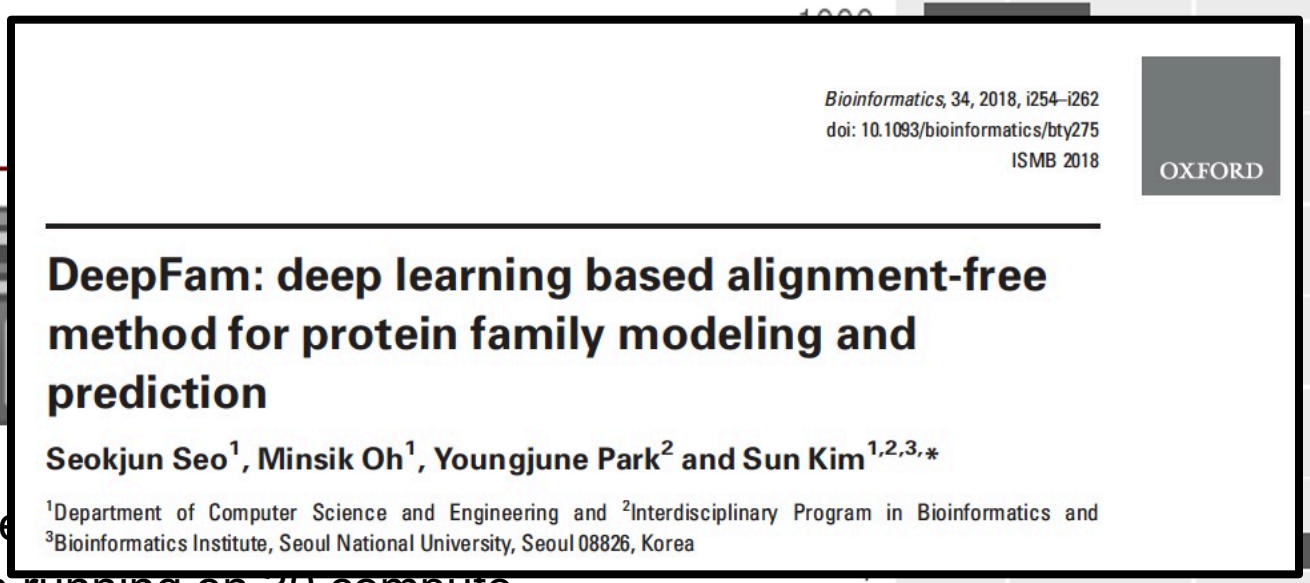
bin_ID	1	2	3	4	5
ENOG1	0	1	0	0	1
ENOG2	1	1	0	0	2
...	...	...	...	...	...



Computational costs for annotation of an *E. coli* genome with EggNOG database

Improvement by parallelization





Bioinformatics, 34, 2018, i254–i262
doi: 10.1093/bioinformatics/bty275
ISMB 2018

OXFORD

DeepFam: deep learning based alignment-free method for protein family modeling and prediction

Seokjun Seo[1], Minsik Oh[1], Youngjune Park[2] and Sun Kim[1,2,3,*]

[1]Department of Computer Science and Engineering and [2]Interdisciplinary Program in Bioinformatics and [3]Bioinformatics Institute, Seoul National University, Seoul 08826, Korea

Further spe

Annotation running on 30 compute nodes of the LiSC

Computational costs for annotation of an *E. coli* genome with EggNOG database

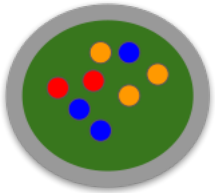
Pipeline - results and summaries

Individual Results file for each uploaded bin

- For each model:
  - **Verdict: Prediction of PhenDB (YES/NO/NA)**
    - "**N/A**" if balanced accuracy below **cut-off** (default 0.7)
  - **PICA probability**
    - Depends on found COGs
  - **Balanced accuracy**
    - Depends on Completeness/Contamination and model

	A	B	C	D
1	Model_name	Verdict	Probability	Balanced_accuracy
2	AEROBE	YES	0.77	0.87
3	ANAEROBE	NO	0.92	0.87
4	AOB	NO	0.91	0.96
5	AUTO	NO	0.84	0.8
6	FACULTATIVE	NO	0.84	0.79
7	GRAMNEGATIVE	YES	0.99	0.99
8	HALO	NO	0.62	0.73
9	METHANOTROPH	NO	0.77	0.84
10	MOTILE	YES	0.84	0.86
11	PEN_180	NO	0.98	0.96
12	PHOTO	NO	0.9	0.92
13	PSYCHRO	N/A	N/A	0.68
14	SPORE	NO	0.97	0.92
15	SYMBIONT	NO	0.98	1

PhenDB



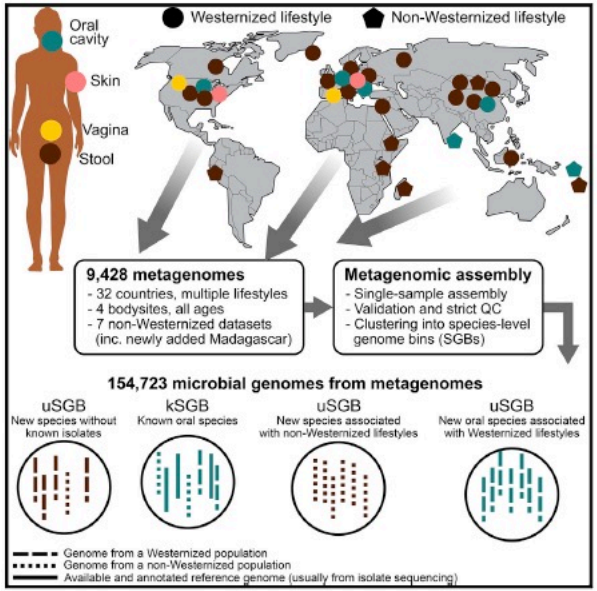
Prediction of bacterial phenotypes

http://phendb.org

Cell

Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle

Graphical Abstract



Authors

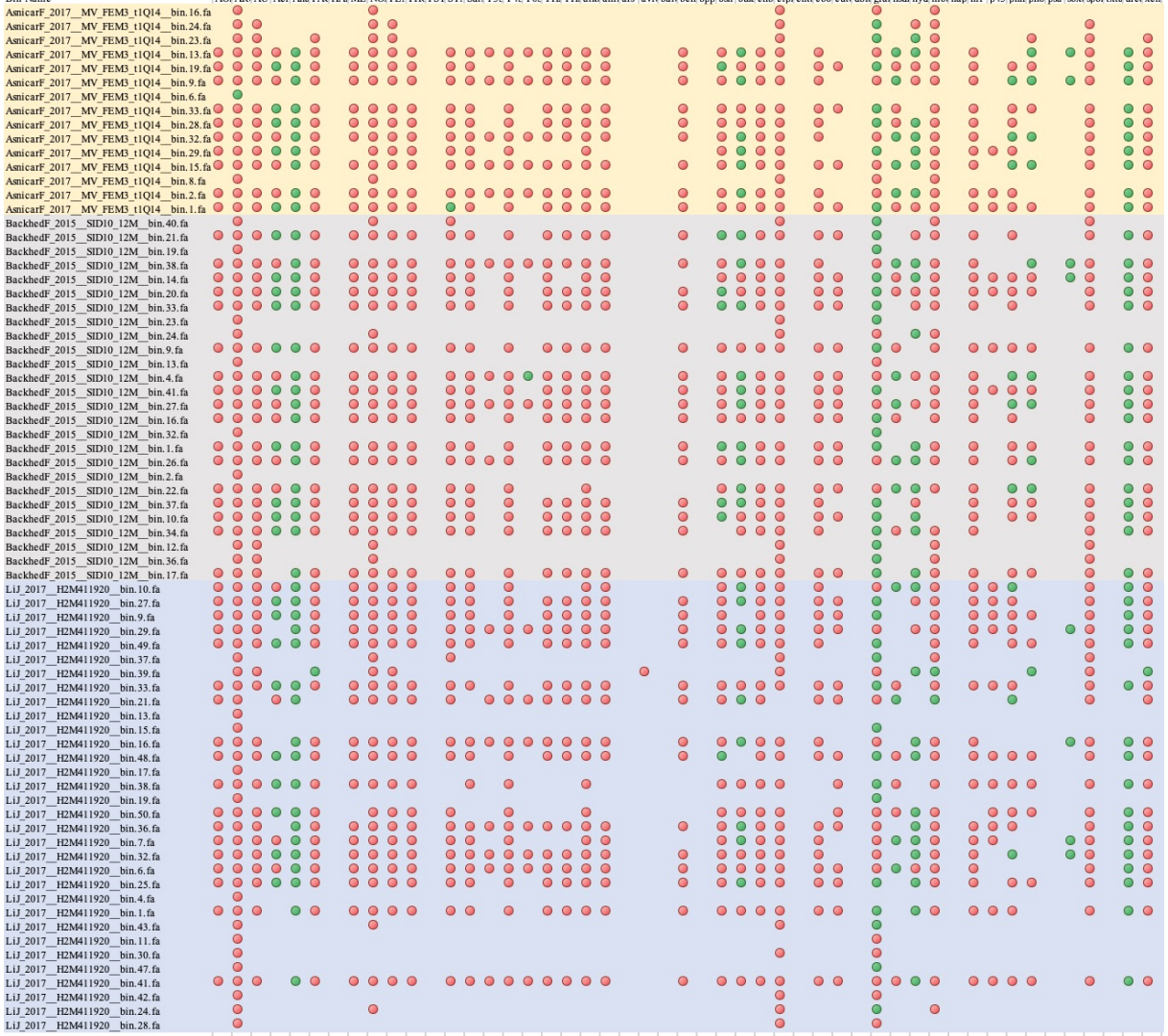
Edoardo Pasolli, Francesco Asnicar, Serena Manara, ..., Christopher Quince, Curtis Huttenhower, Nicola Segata

Correspondence

nicola.segata@unitn.it

In Brief

The human microbiome harbors many unidentified species. By large-scale metagenomic assembly of samples from diverse populations, we uncovered >150,000 microbial genomes that are recapitulated in 4,930 species. Many species (77%) were never described before, increase the mappability of metagenomes, and expand our understanding of global body-wide human microbiomes.


AsnicarF_2017__MV_FEM3_t1Q14__bin.16.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.24.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.23.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.13.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.19.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.9.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.6.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.33.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.28.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.32.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.29.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.15.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.8.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.2.fa
AsnicarF_2017__MV_FEM3_t1Q14__bin.1.fa
BackhedF_2015__SID10_12M__bin.40.fa
BackhedF_2015__SID10_12M__bin.21.fa
BackhedF_2015__SID10_12M__bin.19.fa
BackhedF_2015__SID10_12M__bin.38.fa
BackhedF_2015__SID10_12M__bin.14.fa
BackhedF_2015__SID10_12M__bin.20.fa
BackhedF_2015__SID10_12M__bin.33.fa
BackhedF_2015__SID10_12M__bin.23.fa
BackhedF_2015__SID10_12M__bin.24.fa
BackhedF_2015__SID10_12M__bin.9.fa
BackhedF_2015__SID10_12M__bin.13.fa
BackhedF_2015__SID10_12M__bin.4.fa
BackhedF_2015__SID10_12M__bin.41.fa
BackhedF_2015__SID10_12M__bin.27.fa
BackhedF_2015__SID10_12M__bin.16.fa
BackhedF_2015__SID10_12M__bin.32.fa
BackhedF_2015__SID10_12M__bin.1.fa
BackhedF_2015__SID10_12M__bin.26.fa
BackhedF_2015__SID10_12M__bin.2.fa
BackhedF_2015__SID10_12M__bin.22.fa
BackhedF_2015__SID10_12M__bin.37.fa
BackhedF_2015__SID10_12M__bin.10.fa
BackhedF_2015__SID10_12M__bin.34.fa
BackhedF_2015__SID10_12M__bin.12.fa
BackhedF_2015__SID10_12M__bin.36.fa
BackhedF_2015__SID10_12M__bin.17.fa
LiJ_2017__H2M411920__bin.10.fa
LiJ_2017__H2M411920__bin.27.fa
LiJ_2017__H2M411920__bin.9.fa
LiJ_2017__H2M411920__bin.29.fa
LiJ_2017__H2M411920__bin.49.fa
LiJ_2017__H2M411920__bin.37.fa
LiJ_2017__H2M411920__bin.39.fa
LiJ_2017__H2M411920__bin.33.fa
LiJ_2017__H2M411920__bin.21.fa
LiJ_2017__H2M411920__bin.13.fa
LiJ_2017__H2M411920__bin.15.fa
LiJ_2017__H2M411920__bin.16.fa
LiJ_2017__H2M411920__bin.48.fa
LiJ_2017__H2M411920__bin.17.fa
LiJ_2017__H2M411920__bin.38.fa
LiJ_2017__H2M411920__bin.19.fa
LiJ_2017__H2M411920__bin.50.fa
LiJ_2017__H2M411920__bin.36.fa
LiJ_2017__H2M411920__bin.7.fa
LiJ_2017__H2M411920__bin.32.fa
LiJ_2017__H2M411920__bin.6.fa
LiJ_2017__H2M411920__bin.25.fa
LiJ_2017__H2M411920__bin.4.fa
LiJ_2017__H2M411920__bin.1.fa
LiJ_2017__H2M411920__bin.43.fa
LiJ_2017__H2M411920__bin.11.fa
LiJ_2017__H2M411920__bin.30.fa
LiJ_2017__H2M411920__bin.47.fa
LiJ_2017__H2M411920__bin.41.fa
LiJ_2017__H2M411920__bin.42.fa
LiJ_2017__H2M411920__bin.24.fa
LiJ_2017__H2M411920__bin.28.fa

Take home

- Fast prediction of complex traits for large genome-centered metagenomes.
- PhenDB can not predict completely new mechanisms
- We improve models and expand the web interface
- Only available for bacteria.
- Can we quantitatively compare traits?

VOGDB

Virus Orthologous Groups

Cube

Division of Computational Systems Biology

Hans-Jörg Hellinger

Sabrina Jutz

Jean Mainguy

FWF

Der Wissenschaftsfonds.

AUSTRALIAN INSTITUTE OF MARINE SCIENCE

Nicole Webster

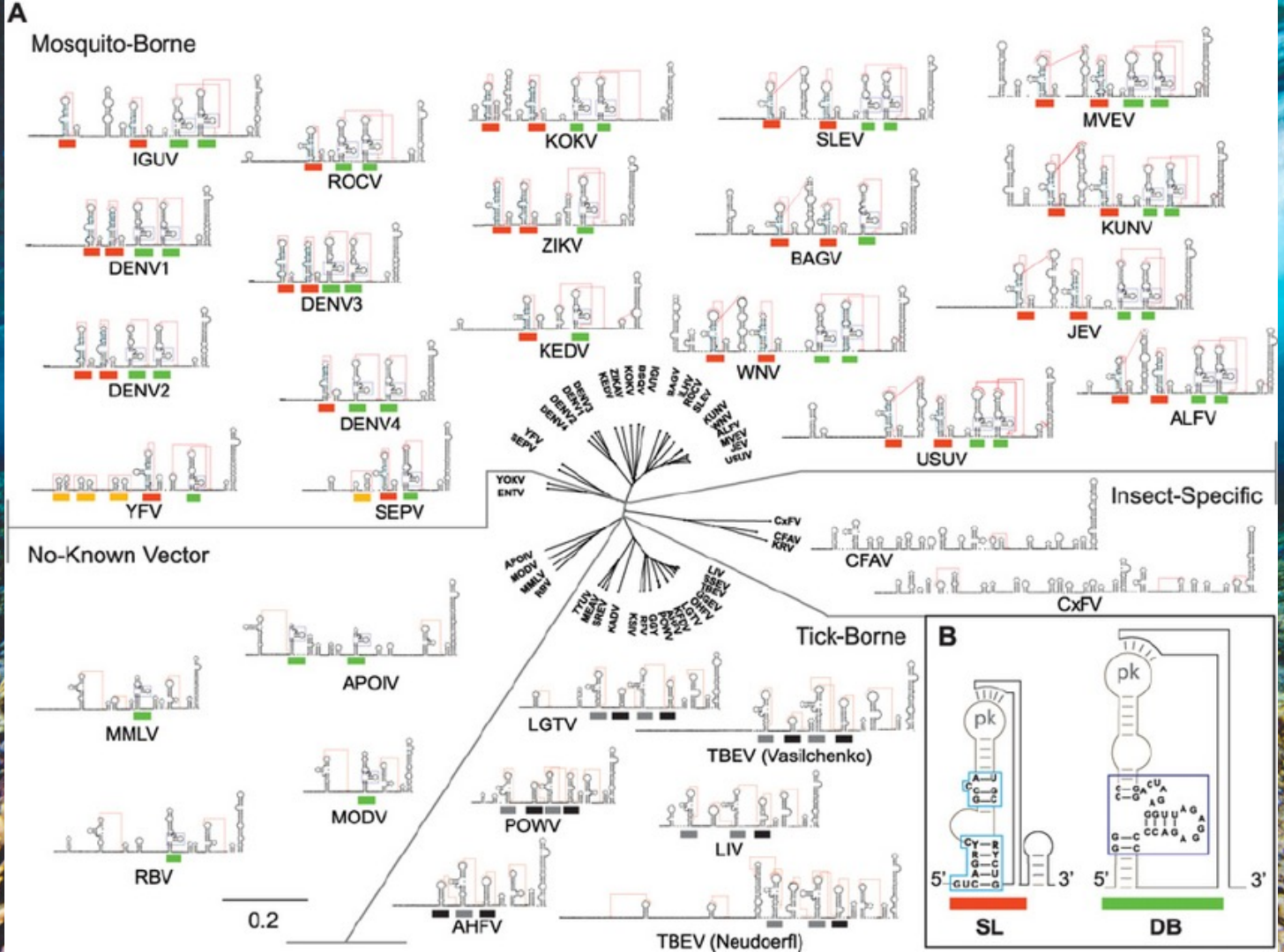
Patrick Laffy

Simon Roux

JGI

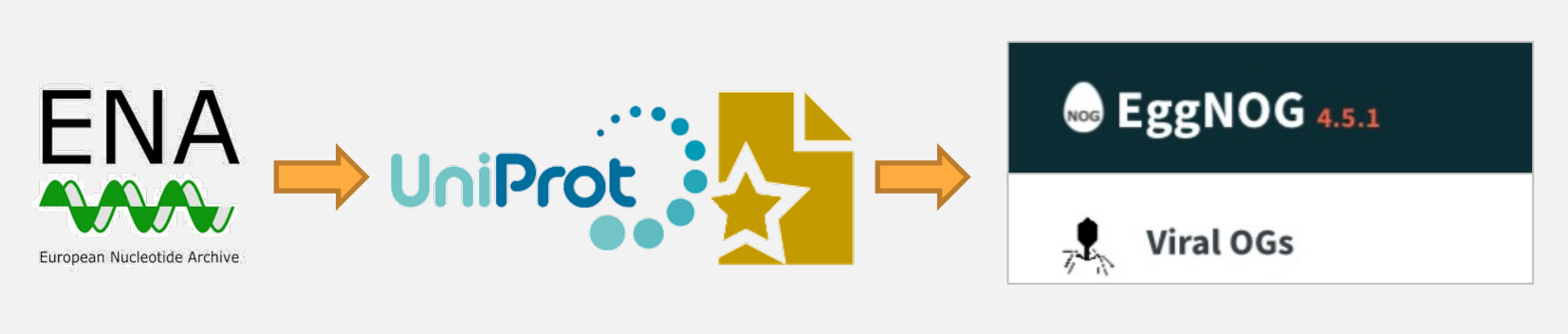
JOINT GENOME INSTITUTE

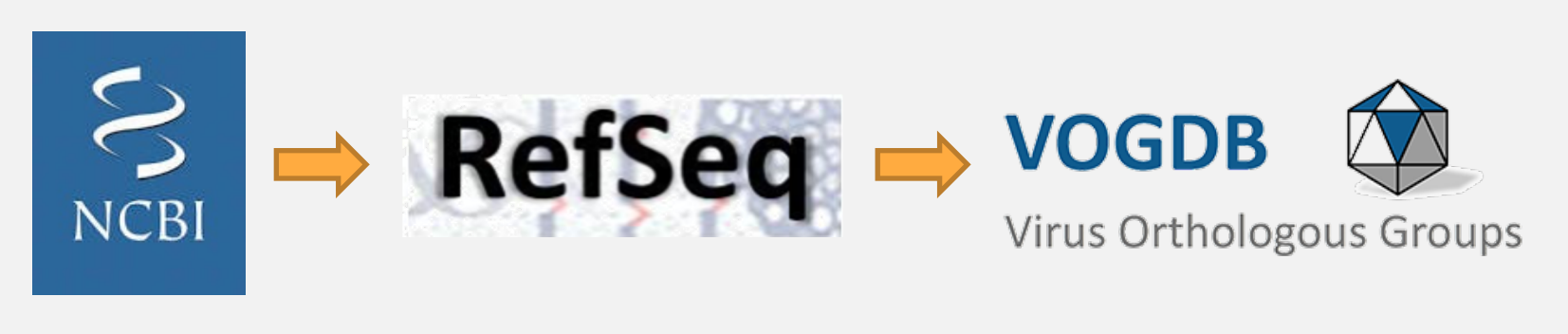
DEPARTMENT OF ENERGY

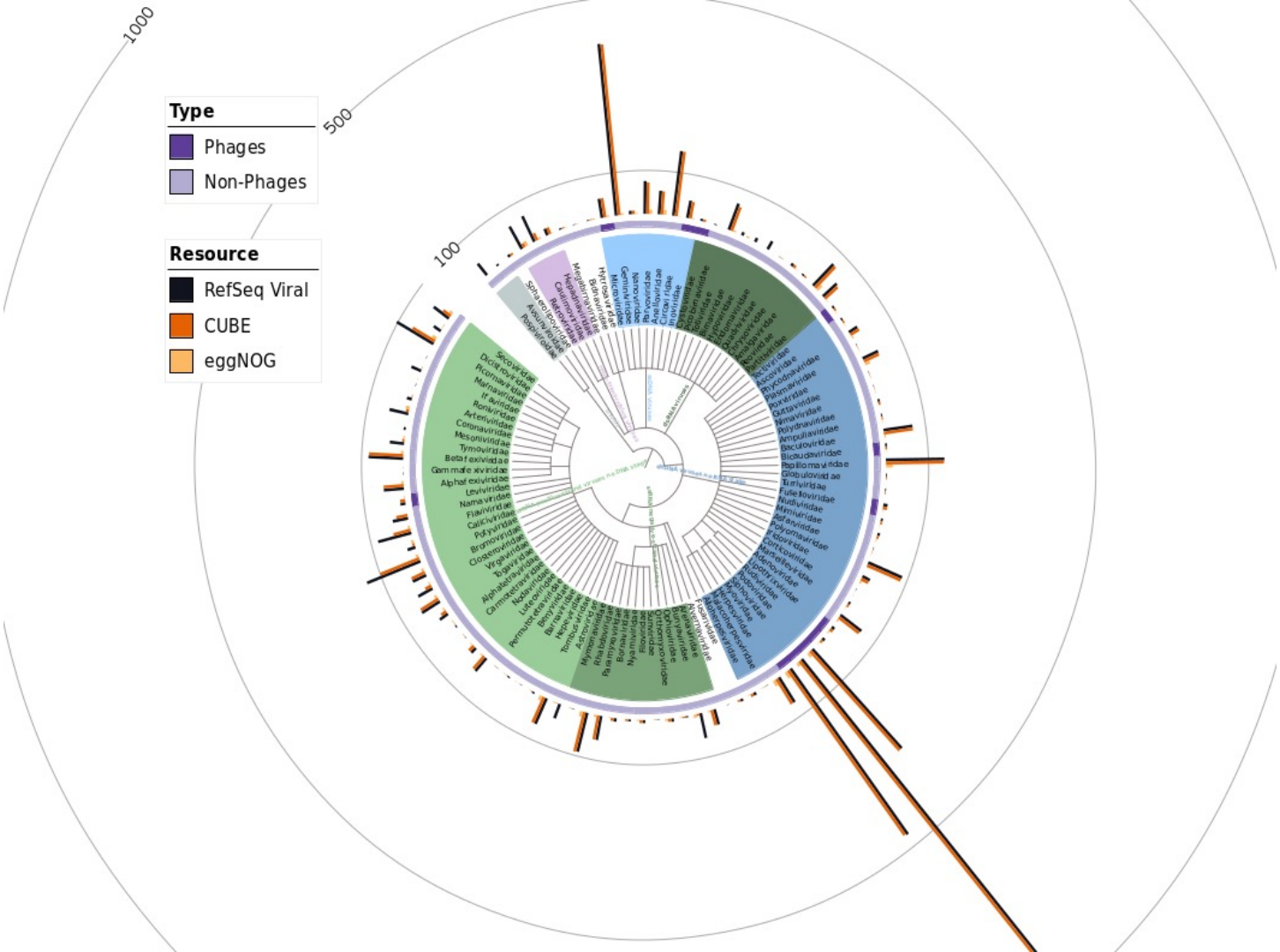

A
Mosquito-Borne
IGUV
ROCV
KOKV
SLEV
MVEV
DENV1
DENV3
ZIKV
BAGV
KUNV
DENV2
KEDV
WNV
JEV
DENV4
ALFV
USUV
YFV
SEPV
YOKV
ENTV
Insect-Specific
CxFV
CFAV
KRV
No-Known Vector
APOIV
MODV
MMLV
RBV
Tick-Borne
LGTV
TBEV (Vasilchenko)
POWV
LIV
AHFV
TBEV (Neudoerfl)
0.2
B
pk
5'
3'
SL
DB

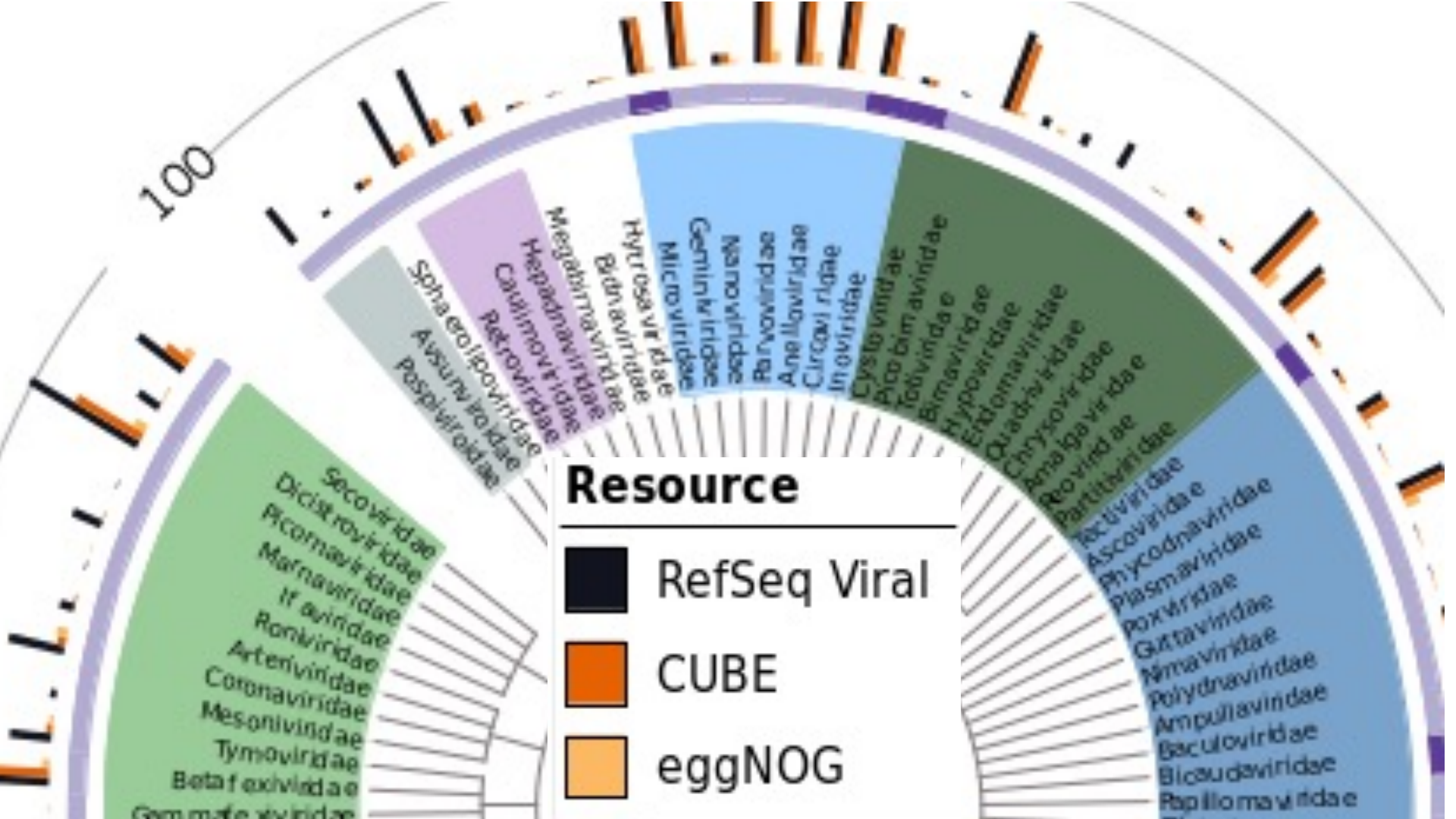
Catalogs of protein families in viruses

All-virus orthologous groups

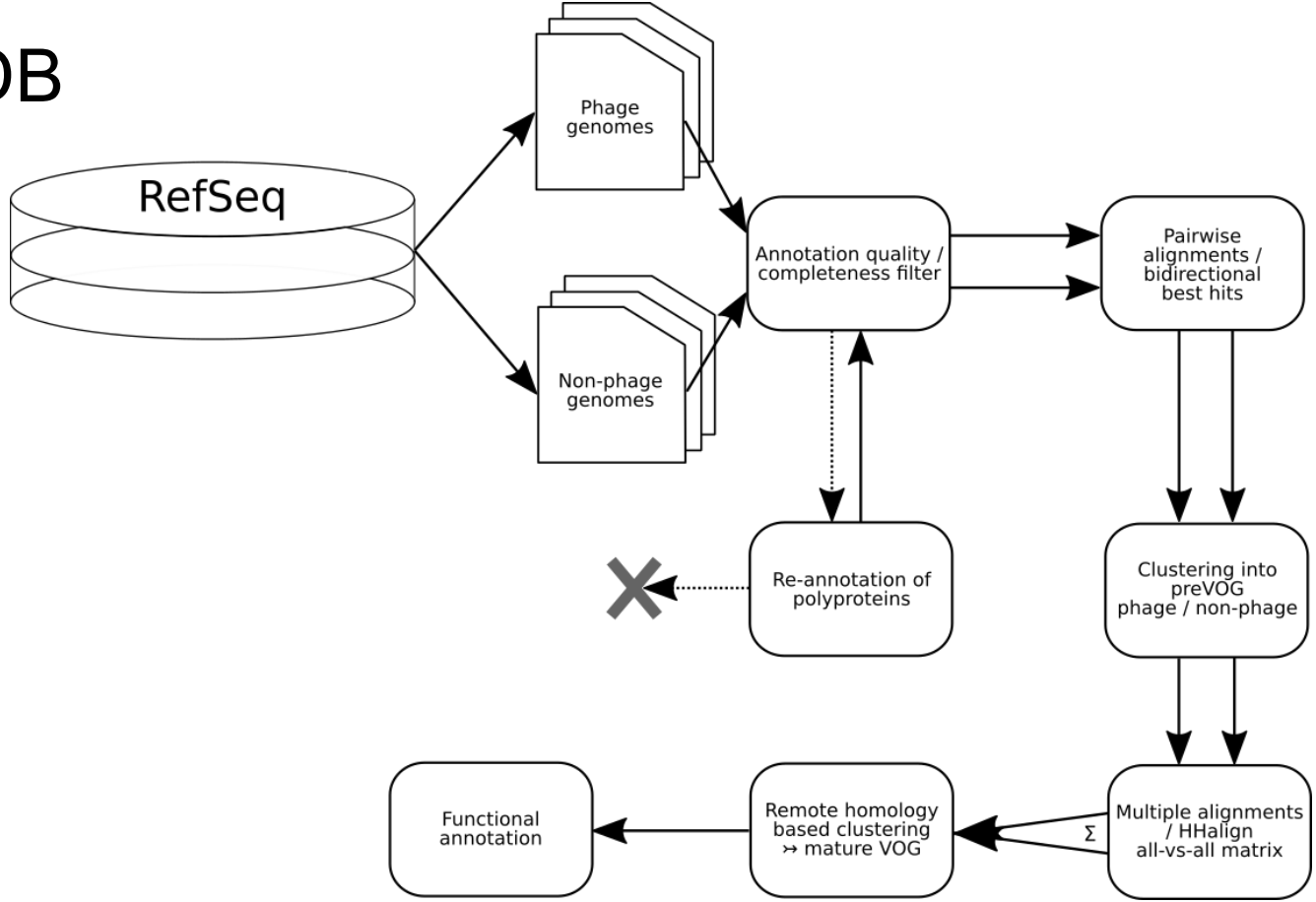





1000
500
100
Type
Phages
Non-Phages
Resource
RefSeq Viral
CUBE
eggNOG


100
Resource
RefSeq Viral
CUBE
eggNOG
Secoviridae
Dicistroviridae
Picornaviridae
Marnaviridae
Iflaviridae
Roniviridae
Arteriviridae
Coronaviridae
Mesoniviridae
Tymoviridae
Betaflexiviridae
Sphaerolipoviridae
Avsunviroidae
Pospiviroidae
Hepadnaviridae
Caulimoviridae
Retroviridae
Megabirnaviridae
Bidnaviridae
Hypovirus
Microviridae
Geminiviridae
Nanoviridae
Parvoviridae
Anelloviridae
Circoviridae
Inoviridae
Cystoviridae
Picobirnaviridae
Totiviridae
Birnaviridae
Hypoviridae
Endornaviridae
Quadriviridae
Chrysoviridae
Amalgaviridae
Reoviridae
Partitiviridae
Ascoviridae
Phycodnaviridae
Plasmaviridae
Poxviridae
Guttaviridae
Nimaviridae
Polydnaviridae
Ampullaviridae
Baculoviridae
Bicaudaviridae
Papillomaviridae

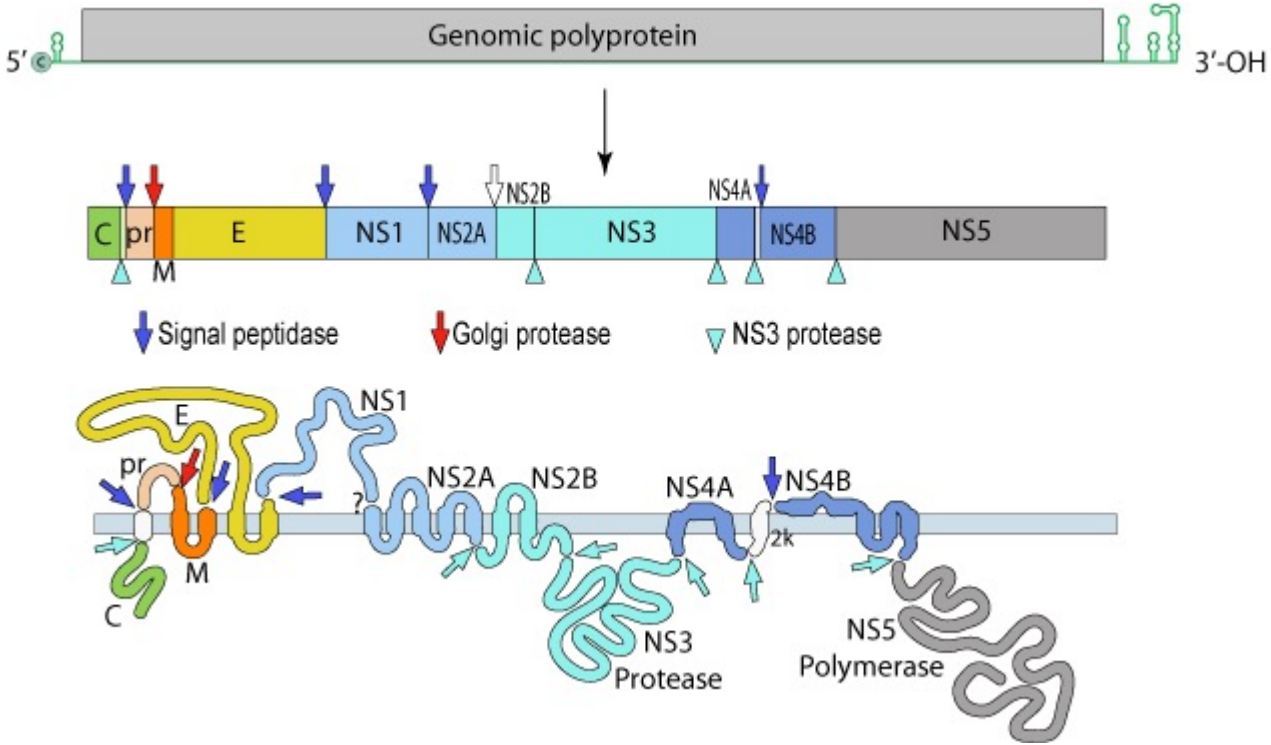
VOGDB


RefSeq
Phage genomes
Non-phage genomes
Annotation quality / completeness filter
Pairwise alignments / bidirectional best hits
Re-annotation of polyproteins
Clustering into preVOG phage / non-phage
Multiple alignments / HHalign all-vs-all matrix
Σ
Remote homology based clustering ↣ mature VOG
Functional annotation

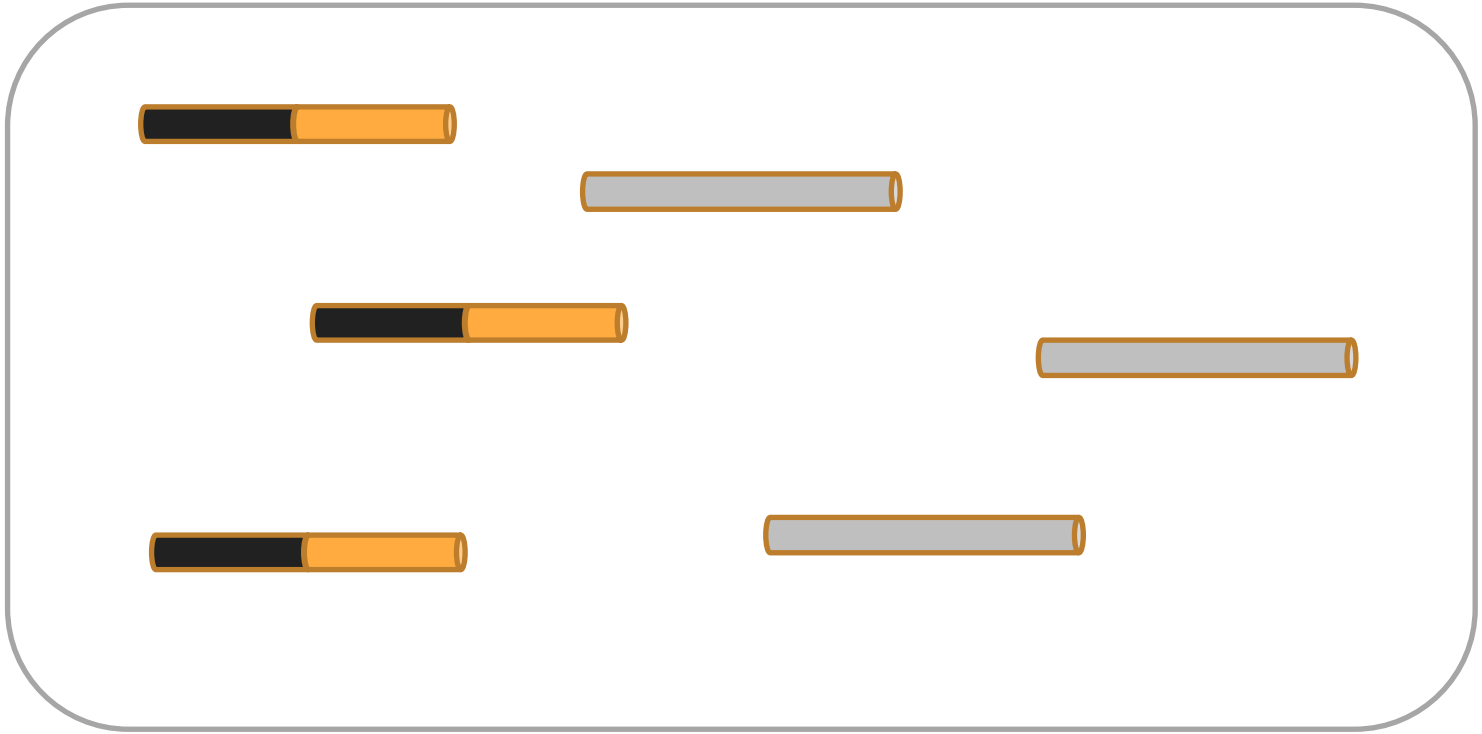
Polyproteins and their peptides

```
                 SAKNVHVPYTQASSGFEMWKNNSGRPLQETAPFGCKIAVNPLRAVDCSYGNIPISIDI
                 PNAAFIRTSDAPLVSTVKCEVSECTYSADFGGMATLQYVSDREGQCPVHSHSSTATLQ
                 ESTVHVLEKGAVTVHFSTASPQANFIVSLCGKKTTCNAECKPPADHIVSTPHKNDQEF
                 QAAISKTSWSWLFALFGGASSLLIIGLMIFACSMMLTSTRR"
mat_peptide      7647..8438
                 /locus_tag="SINVgp3"
                 /product="capsid (c) protein"
                 /protein_id="NP_740673.1"
mat_peptide      8439..8630
                 /locus_tag="SINVgp3"
                 /product="e-3 structural protein"
                 /protein_id="NP_740674.1"
mat_peptide      8631..9899
                 /locus_tag="SINVgp3"
                 /product="e-2 structural protein"
                 /protein_id="NP_740675.1"
mat_peptide      9900..10064
                 /locus_tag="SINVgp3"
                 /product="6k structural protein"
                 /protein_id="NP_740676.1"
mat_peptide      10065..11381
                 /locus_tag="SINVgp3"
                 /product="e-1 structural protein"
                 /protein_id="NP_740677.1"
gene             7647..10111
                 /locus_tag="SINVgp4"
                 /db_xref="GeneID:13165406"
CDS              join(7647..10028,10028..10111)
                 /locus_tag="SINVgp4"
                 /note="Truncated verstion of structural polyprotein that
                 will be produced when frameshifting occurs at nt 10028;
```

A typical flavivirus genome

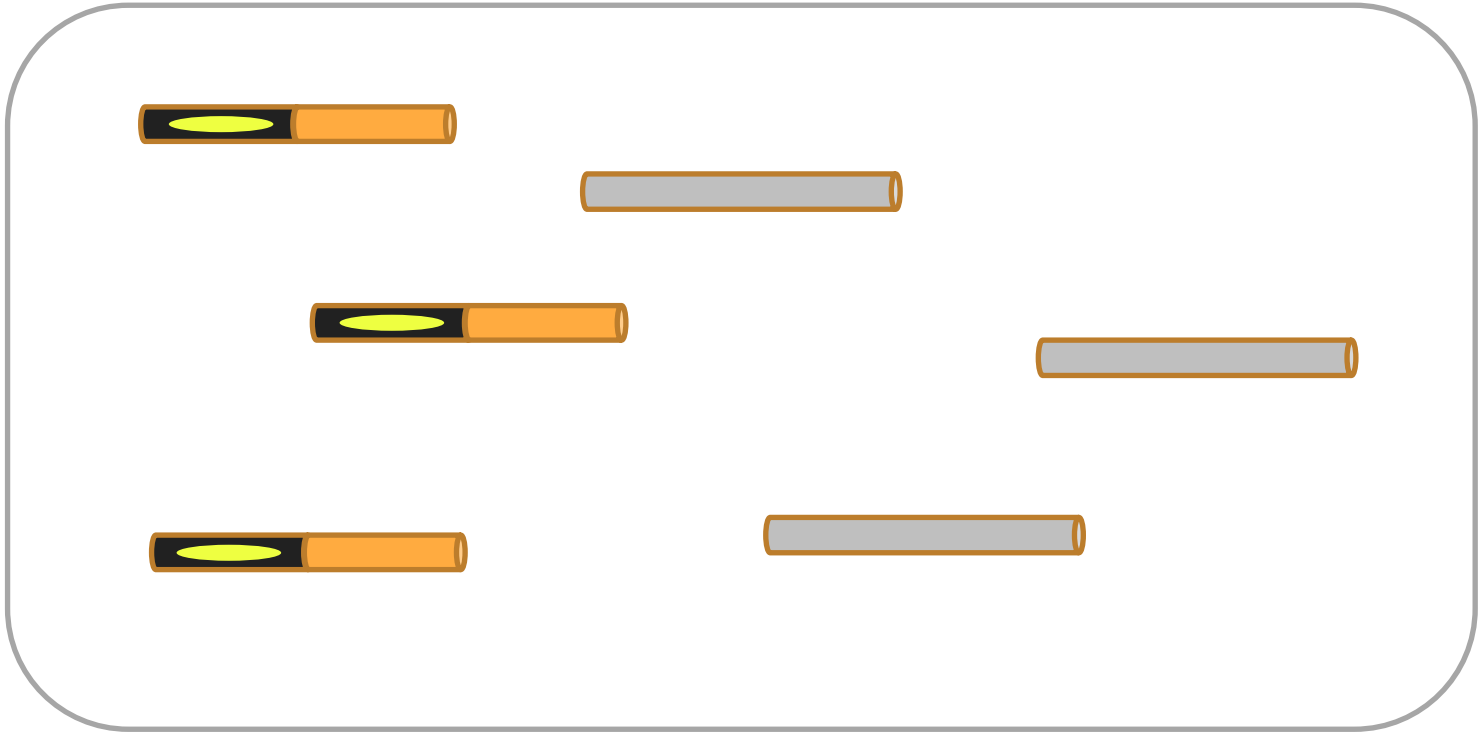


Automatic polyprotein annotation: concept



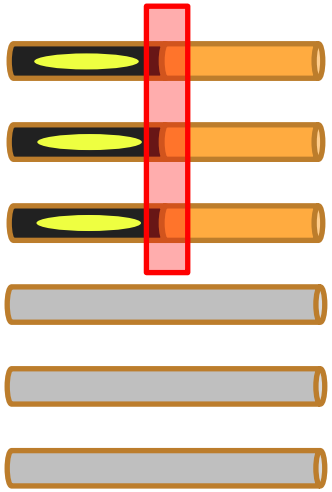
1. Extraction of homologous polyproteins from a genus or a homologous cluster

Automatic polyprotein annotation: concept



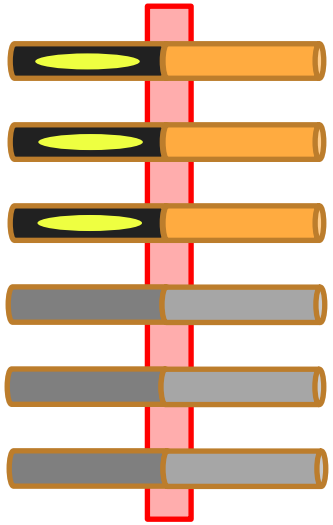
2. Validation of peptide annotation with domains

Automatic polyprotein annotation: concept



3. Multiple alignment, validation of conservation of sequence and cleavage sites

Automatic polyprotein annotation: concept



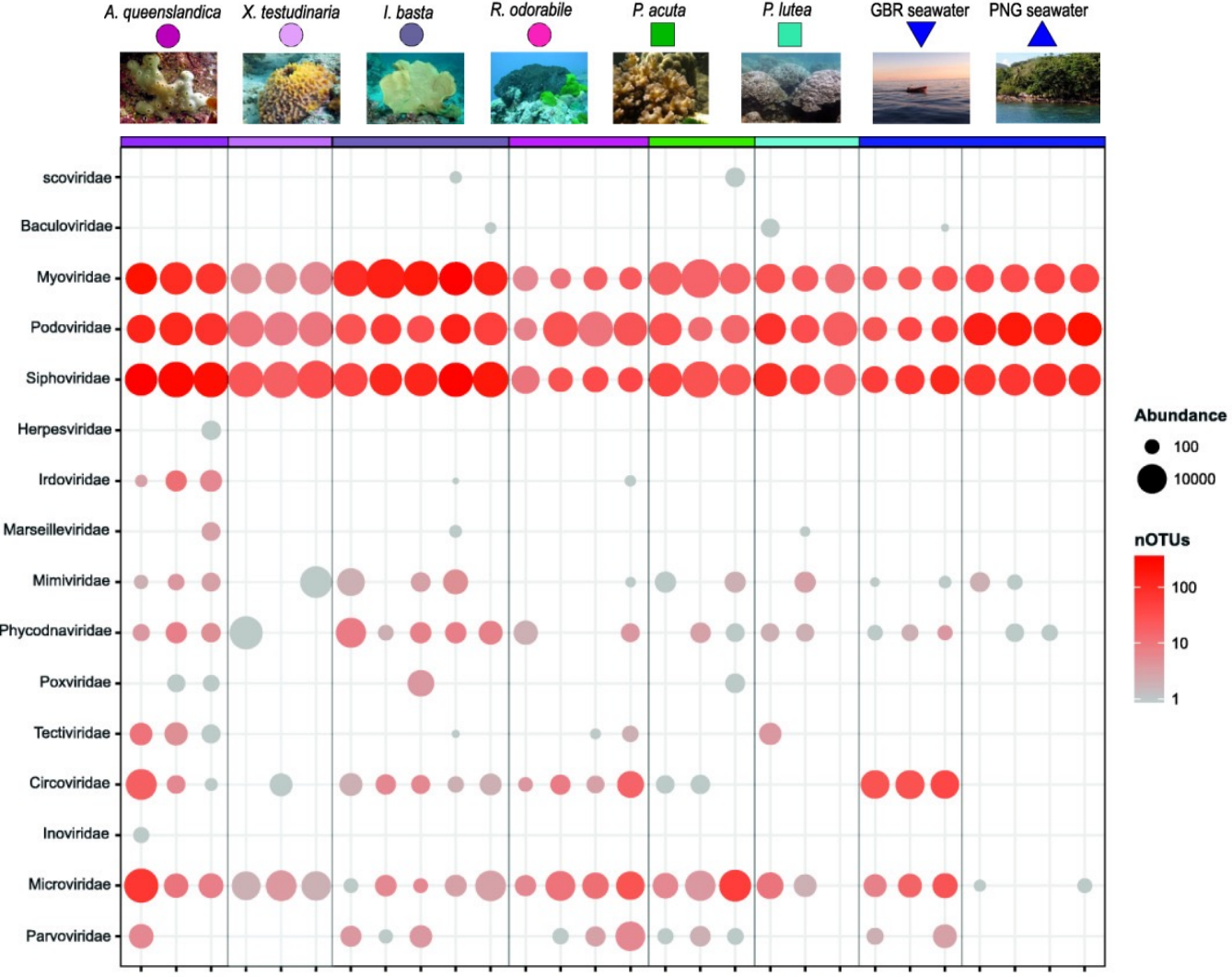
4. Propagation of cleavage sites to unannotated proteins

VOGDB

Virus Orthologous Groups

http://vogdb.org

Hellinger et al., in prep.



Composition of viruses across holobiont species and reef environments

Laffy et al., Env Microbiol 2018

Take home

- Grouping of all viral proteins into orthologous groups and families.
- Virus genomes require annotation check and re-annotation
- Many families are present in prokaryotic and eukaryotic viruses
- Applications e.g. in viral ecology, virus evolution, study of host interactions…
- Lots of features in database and web interface to come