Bibliography: https://paperpile.com/shared/7b9DPA

Phylometagenomics

Christophe Dessimoz

Reviews in Quantitative Biology
UNIL Nov 2018

Outline

- Introduction
- Mapping methods
- Assembly methods
- Examples
- Challenges

Related reviews

- *Mallick et al. Genome Biology (2017) 18:228*
  very general, also discusses metatranscriptomics and integration

- *Mande et al. Briefings Bioinf (2012) 13:6*
  Methods-oriented. Binning, LCA. Somewhat outdated.

- Hernandes Coutinho et al. Trends Microbiology 2018

- Breitwieser et al. Briefings in Bioinformatics 2017
  focus on bioinformatic methods for assembly

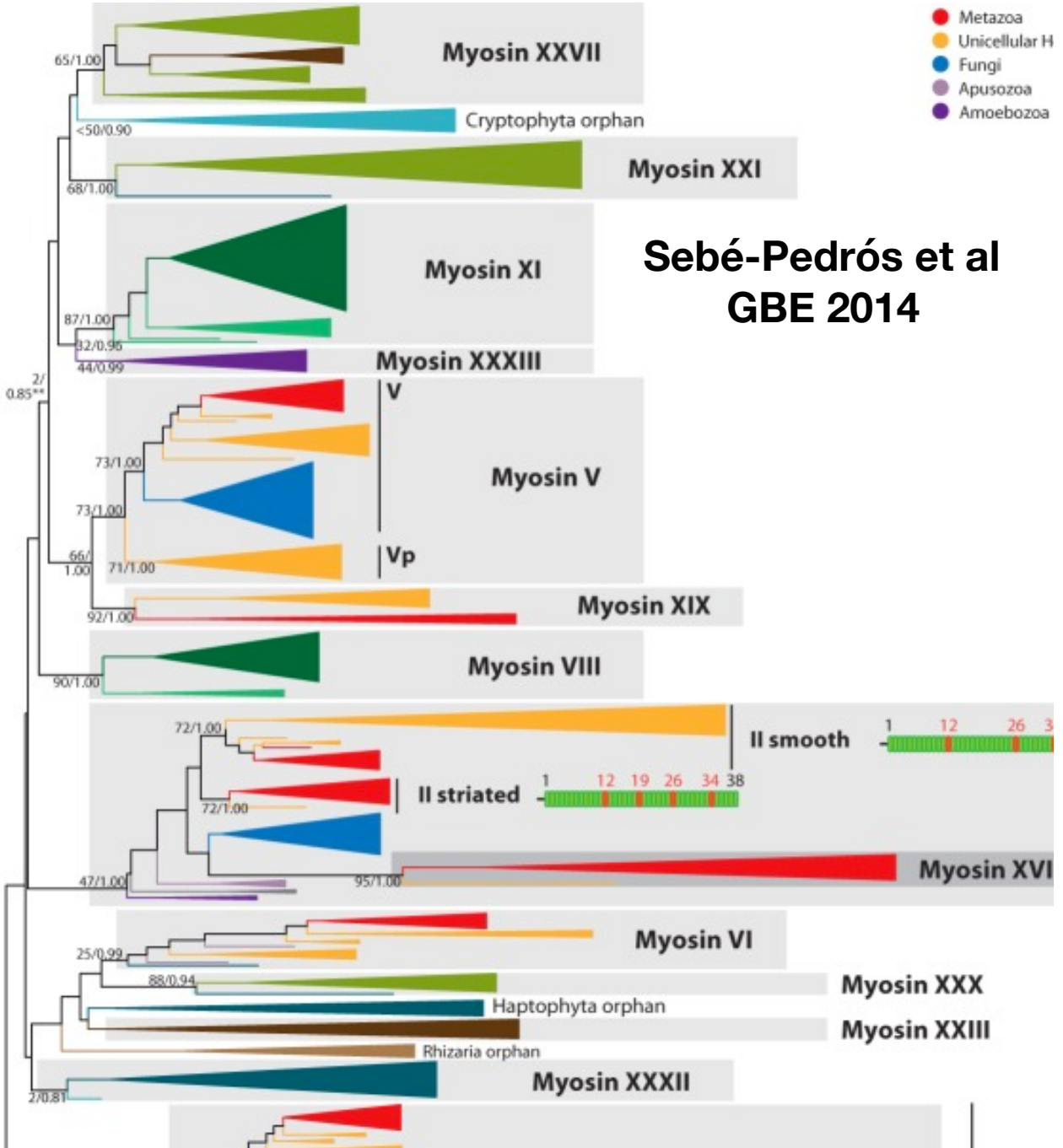
Phylogenomics: def. 1

J. Eisen 1998:

To understand the function of genes, computing similarities is not enough.

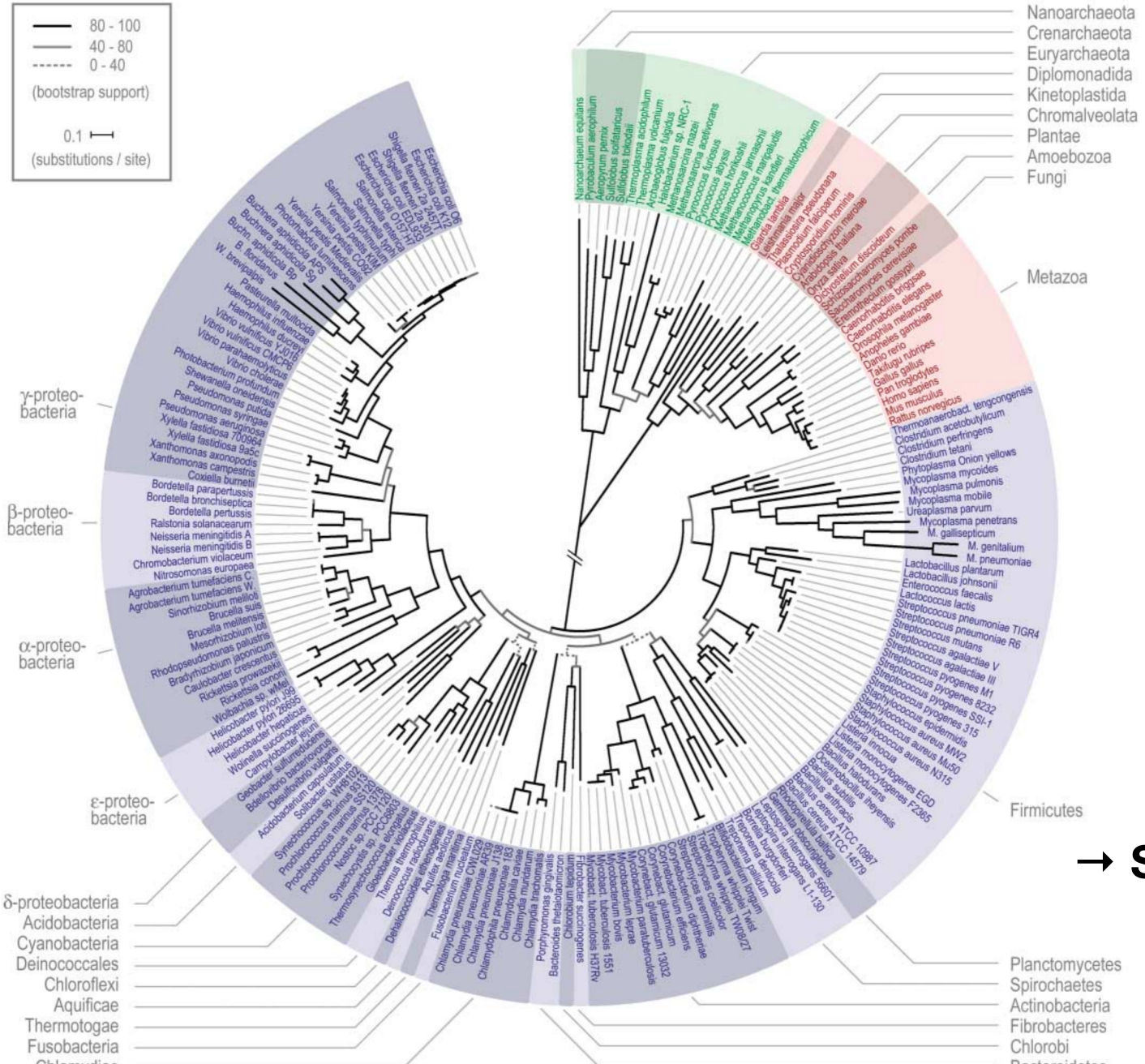
Mapping function evolution onto gene trees is key.

→ Gene-centric view



Sebé-Pedrós et al GBE 2014

Phylogenomics: def. 2



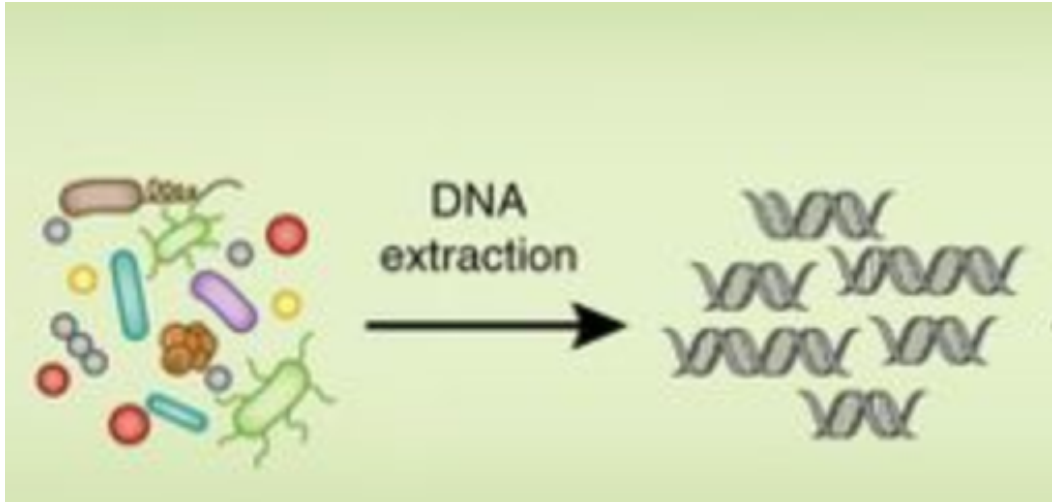
Use “genome-wide” information (i.e. many orthologous markers) to infer species phylogenies

As opposed to e.g. just 16S RNA or a handful of amplified markers

→ Species-centric view

Ciccarelli et al.
Science 2006

Metagenomics



Quince et al, Nat Biotechnology 2017

Advantages:

- **more data at once**
- **not limited to cultivable genomes (0.1-1%; Garza & Dulith 2015)**

Disadvantage:

- **DNA read mixture can be tricky to untangle!**

Questions

Compare to known stuff!

- Taxonomic classification: what species is in my sample?
- Pathway analysis: what metabolic pathways are in my sample?

Identify new stuff!

- Can we identify new species and even entire phyla?
- Can we broaden gene families & identify new families?
  - for 3D structure reconstruction, using contact map predictions
  - for protein design (e.g. enzymes in biotechnology)
  - to discover new antibiotics
  - Steinegger et al. Biorxiv 2018

Outside the scope!

- What are the genes that are expressed? Metatranscriptomics (Carredec et al. Nature Comm 2018)
- Integration with other omics data (e.g. Metabolome McHardy et al 2013; Proteomics Grassl et al 2016)

Comparing to known sequences

→ mapping reads to

taxonomic classification

- 16S/18S RNA
- nuclear marker genes

community-level gene content analysis

- markers of metabolic pathways
- virulence factors
- antibiotic resistance genes

in general

- Reference database with all known genes!

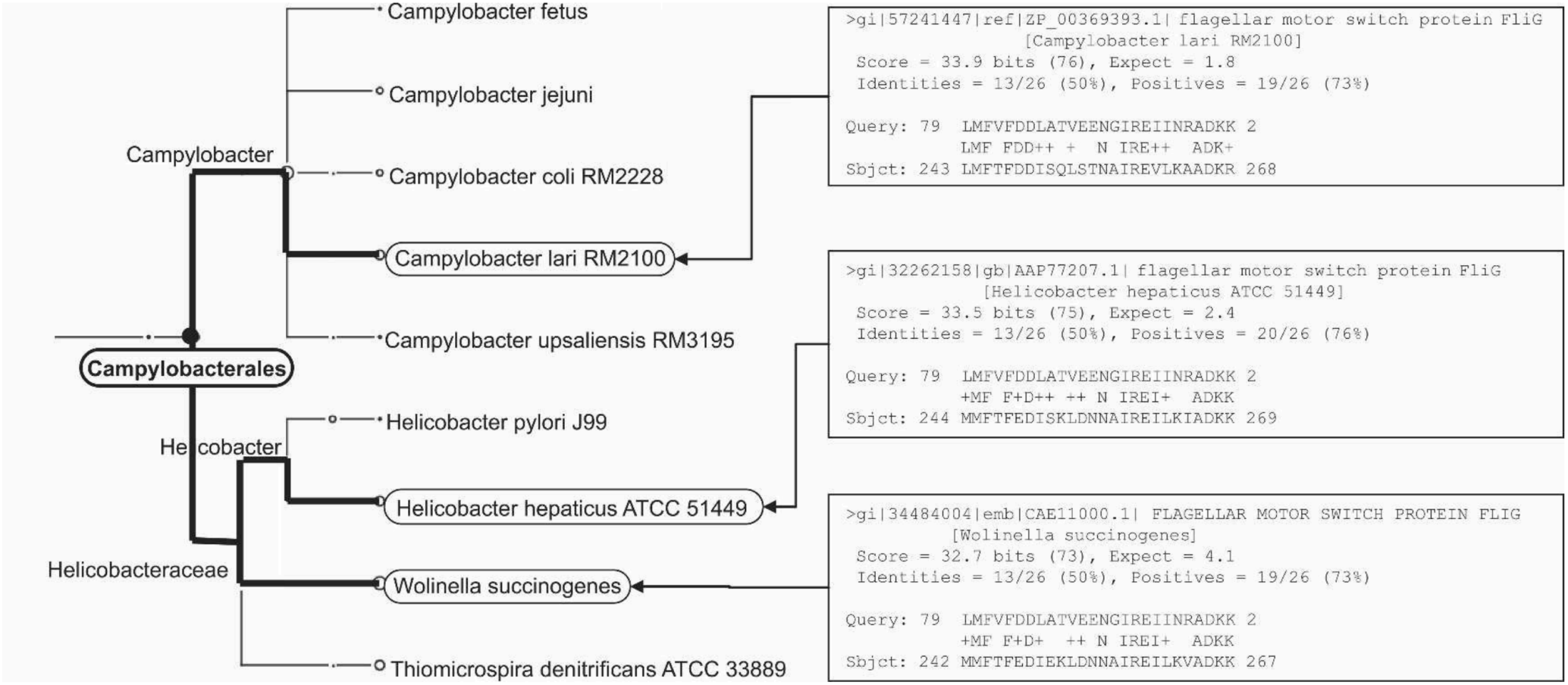
Table 1 Tools for metagenomic strain analysis

Tool	Data type	Requires reference genome?	Method to resolve ambiguous reads	New strain detection	Recommended minimal coverage	Reference
Oligotyping	16S	–	–	–	–	[42]
Long-read 16S	16S	–	–	–	–	[119]
Minimum entropy decomposition	16S	–	–	–	–	[39]
OTU subpopulations	16S	–	–	–	–	[40]
LEA-seq	16S	–	–	–	–	[41]
DADA2	16S	No	Poisson modeling of sequence errors (the Divisive Amplicon Denoising Algorithm)	Yes	–	[43]
UNOISE2	16S	No	Abundance-based identification of sequencing errors	Yes	–	[44]
Deblur	16S	No	Abundance-based identification of sequencing errors	Yes	–	[45]
Megan	WGS	Yes	Lowest common ancestor algorithm	No	Not reported	[120]
GSMer	WGS	Yes	Unique strain-level markers (k-mers)	No	0.25×	[121]
WG-FAST	WGS	Yes	SNVs	No	3×	[122]
StrainPhlAn	WGS	Yes	SNVs within species-level marker genes	Yes	10×	[6]
PanPhlAn	WGS	Yes	Unique combinations of species-level marker genes	Yes	1×	[6]
MIDAS	WGS	Yes	Unique strain-level marker genes	Yes	Not reported	[37]
Sigma	WGS	Yes	Likelihood-based	No	0.027×	[123]
PathoScope	WGS	Yes	Likelihood-based	No	Less than 1×	[124]
ConStrains	WGS	Yes	Inferred haplotype-like SNP profiles	Yes	10×	[4]
LSA	WGS	De novo assembly	SVD-based K-mer clustering	Yes (not validated)	25 ~ 50×	[125]
CNV-based methods	WGS	Yes	Target gene or region copy number variation			[126]

CNV **copy number variation,** ***LEA-Seq*** **low-error amplicon sequencing,** ***OTU*** **operational taxonomic unit,** ***SNP*** **single nucleotide polymorphism,** ***SNV*** **single-nucleotide variant,** ***SVD*** **singular-value decomposition,** ***WGS*** **whole-genome sequencing**

- Mallick et al. Genome Biology (2017)

LCA algorithm



Huson et al. 2007

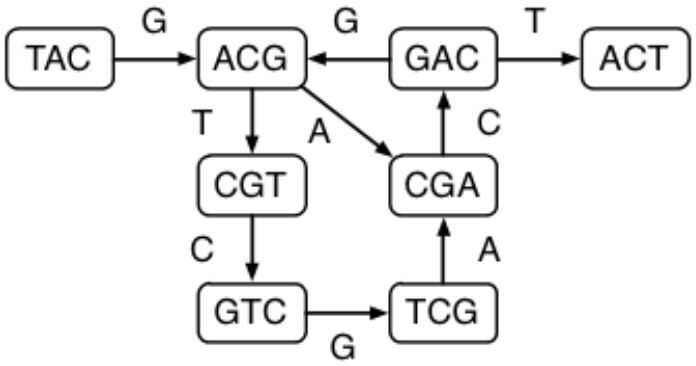
Assembling genes

(sometimes called “co-assembly”)

DNA-level

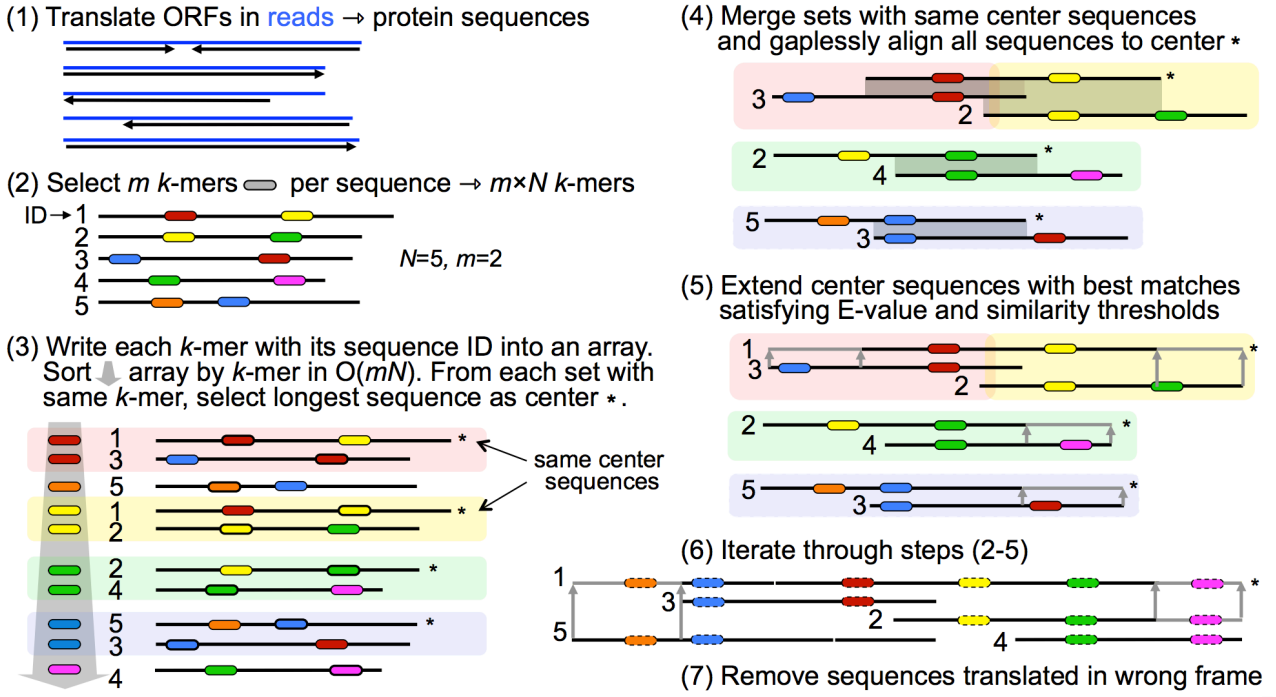
e.g. Megahit (Li et al. 2015)

Assembly based on succinct de bruijn graphs (Bowe et al. WABI 2012)



Protein-level

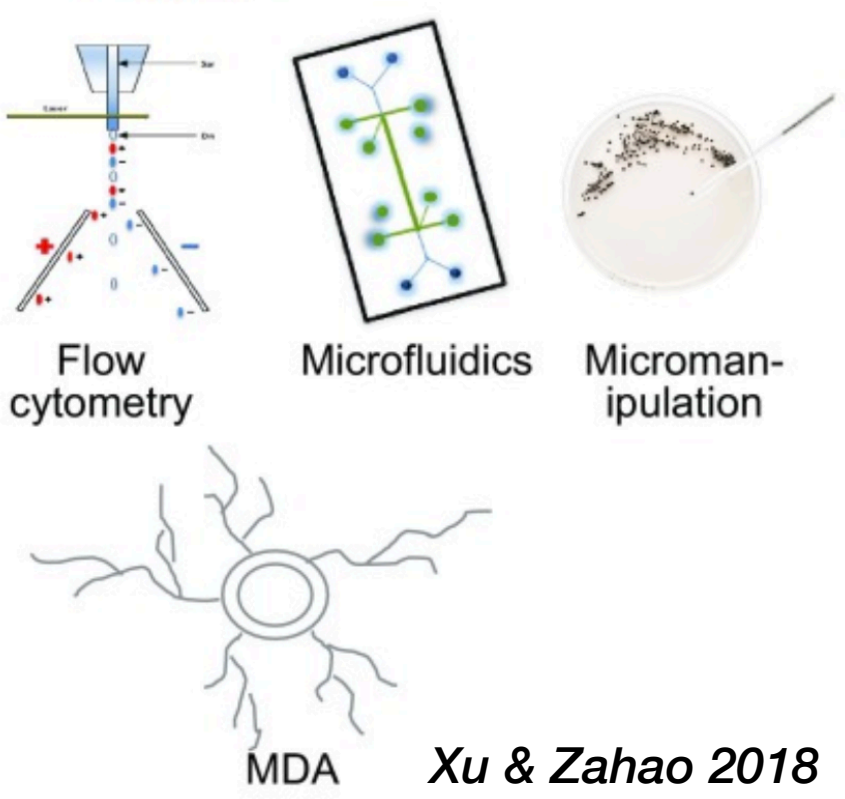
PLASS (Steingger et al. 2018)



https://alexbowe.com/succinct-debruijn-graphs/

Assembling genomes

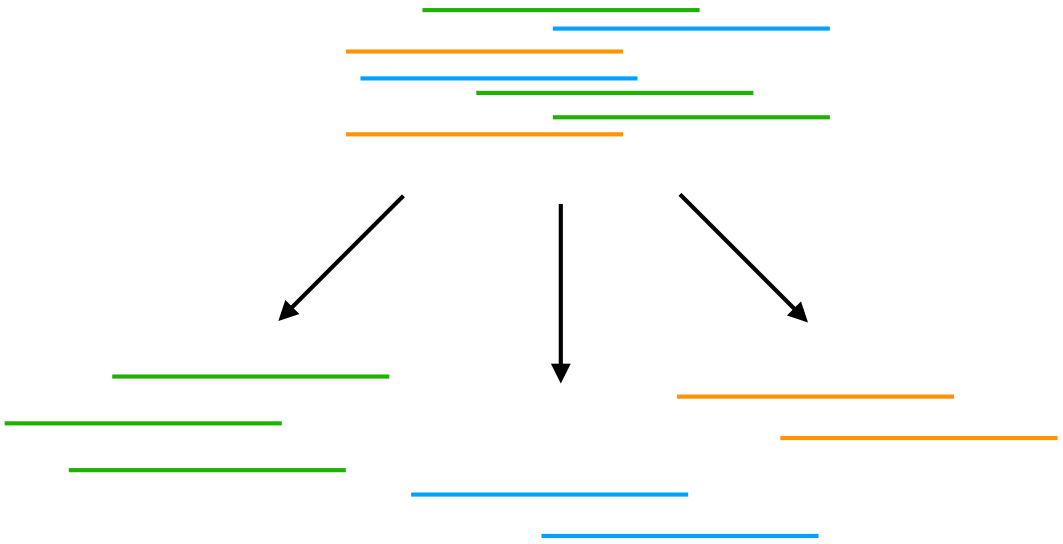
- Metagenomic Assembled Genomes (MAGs)
  - Co-Assembly, followed by binning and scaffolding
  - Binning, followed by assembly
- Single Amplified Genomes
  - reviewed in Xu & Zahao 2018



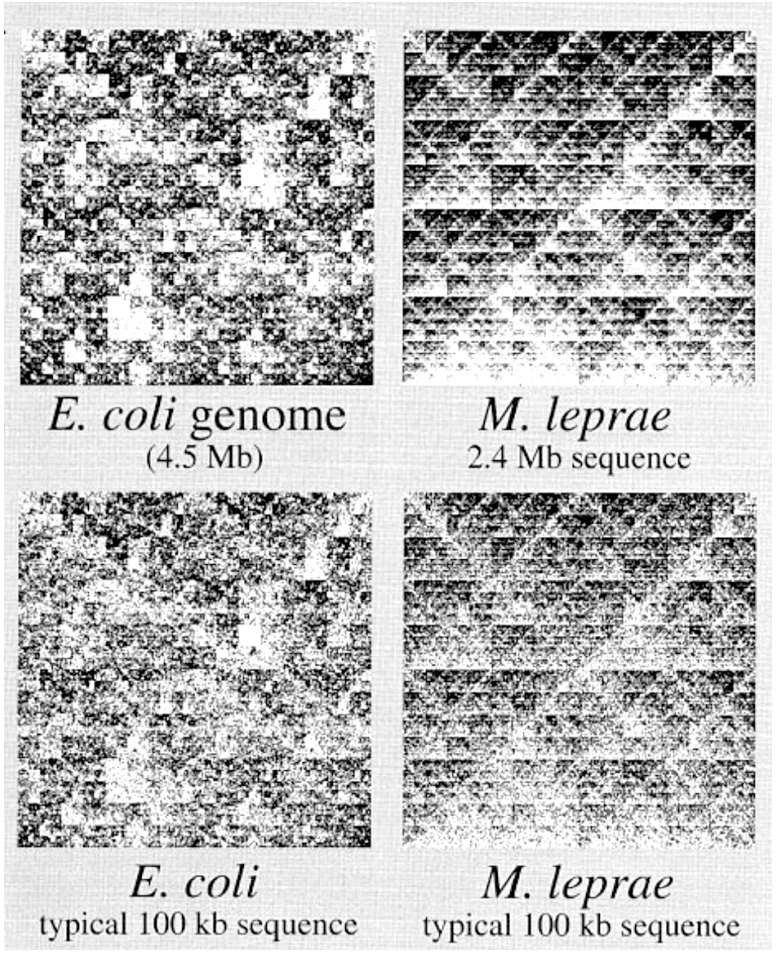
Xu & Zahao 2018

Binning

partly reviewed in Sedlar et al. 2017



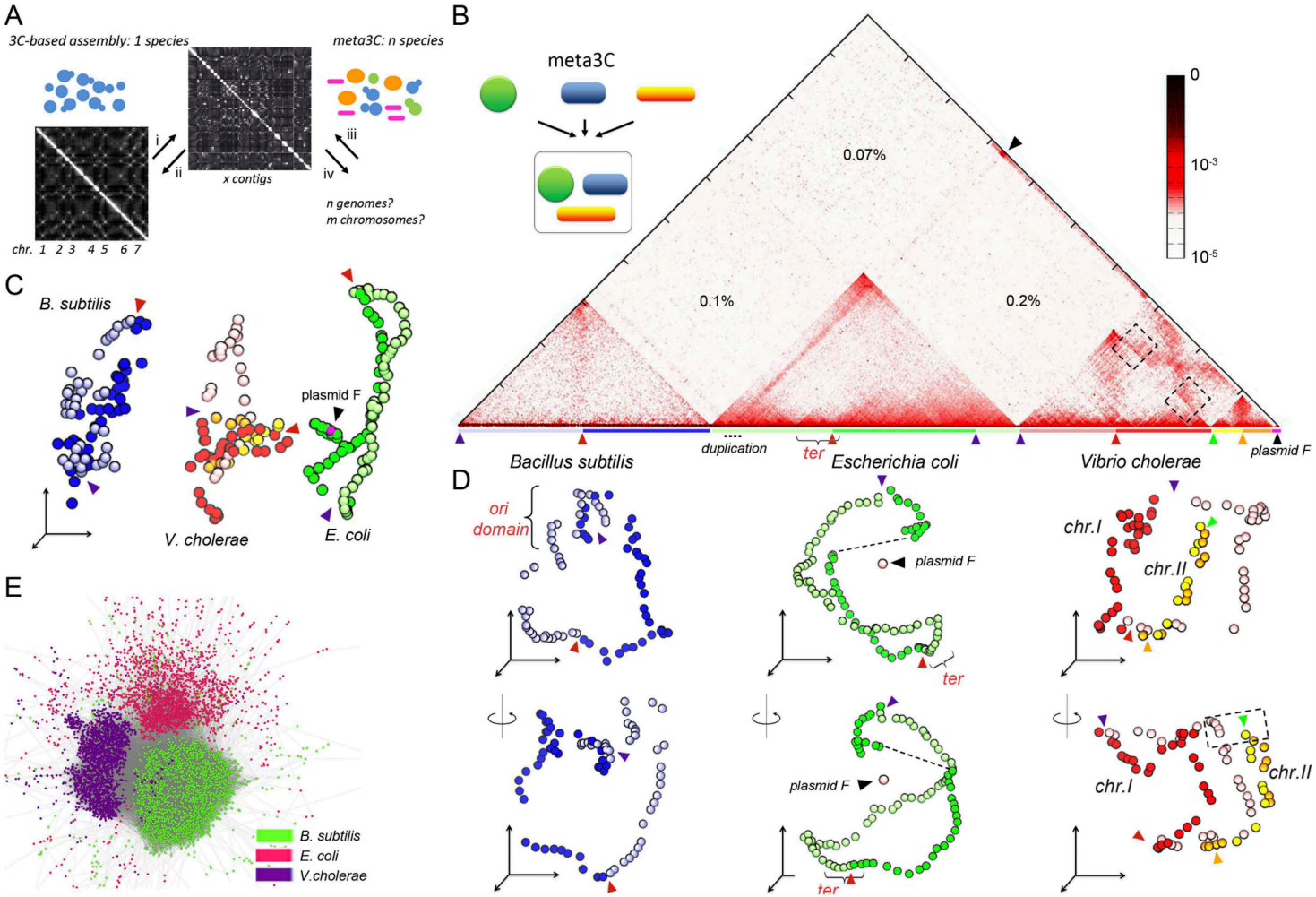
1. Use nucleotide distribution (GC content, k-mer spectrum)



Deschavannes et al. 1999
Karlin & Burge 1995

Binning (con't)

Experimental techniques



Long reads

- Frank et al. Sci Reports 2016

3C / Hi-C

- Koszul (Marbouty et al. life 2014, Marbouty et al. Science Advances 2017)
- Steward et al. Nat Comm 2018 (also using binning).
- Press et al. biorxiv 2017
- Burton et al. G3 2014 (Shendure lab)
- Darling: Beitel et al PeerJ 2014; Liu and Darling F1000 2015 (review); DeMaere & Darling Biorxiv 2018

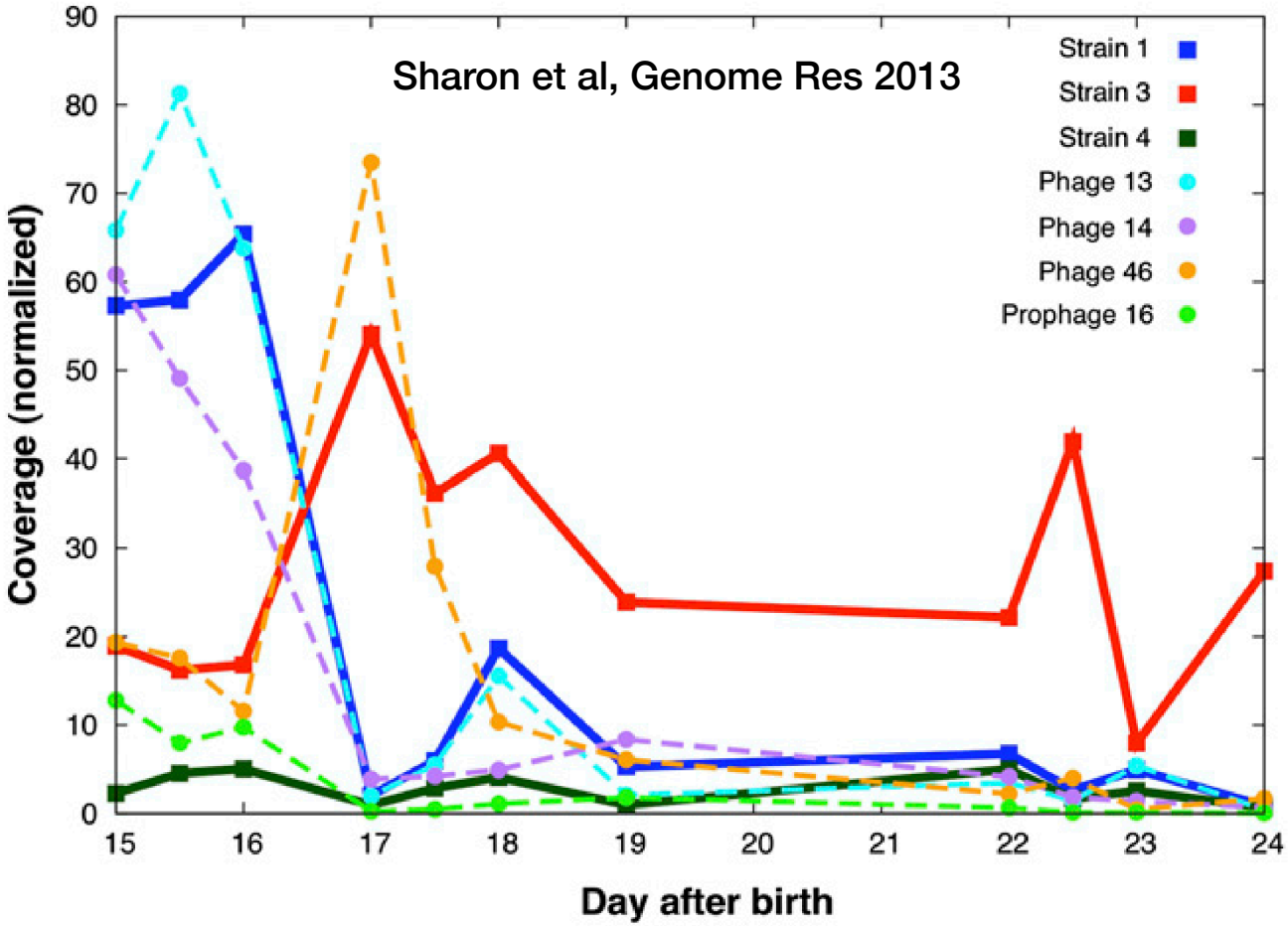
DNA methylation

- Beaulaurier et al. 2017

Marbouty et al. life 2014

Binning (con't)

Differential coverage



- Sharon et al, Genome Res 2013
    - metagenomics from birth to young age!
- Albertsen et al. Nat Biotech. 2013
- Alneberg J, Bjarnason B, de Bruijn I, Schirmer M, Quick J, Ijaz U, et al. Nature Methods 2014
    - CONCOCT? "Gaussian mixture models to predict the cluster membership of each contig while automatically determining the optimal number of clusters in the data through a variational Bayesian approach"
- Imelford et al. 2014, Wu et al 2016, Kang et al. 2015
- Lu et al. 2017 COCACOLA
- BinSanity Graham et al. PeerJ 2017

Binning (con't)

BinSanity Graham et al. PeerJ 2017

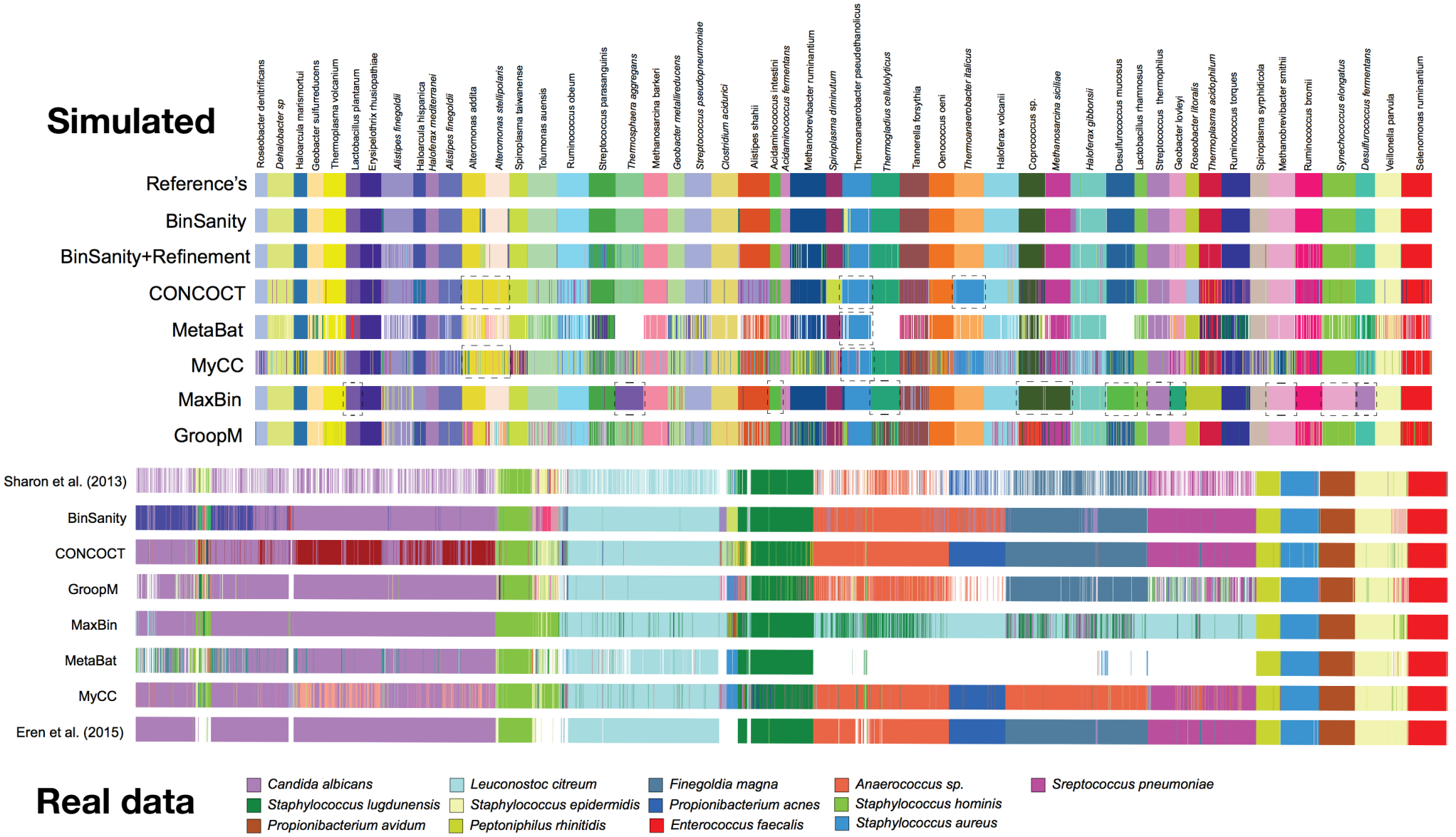
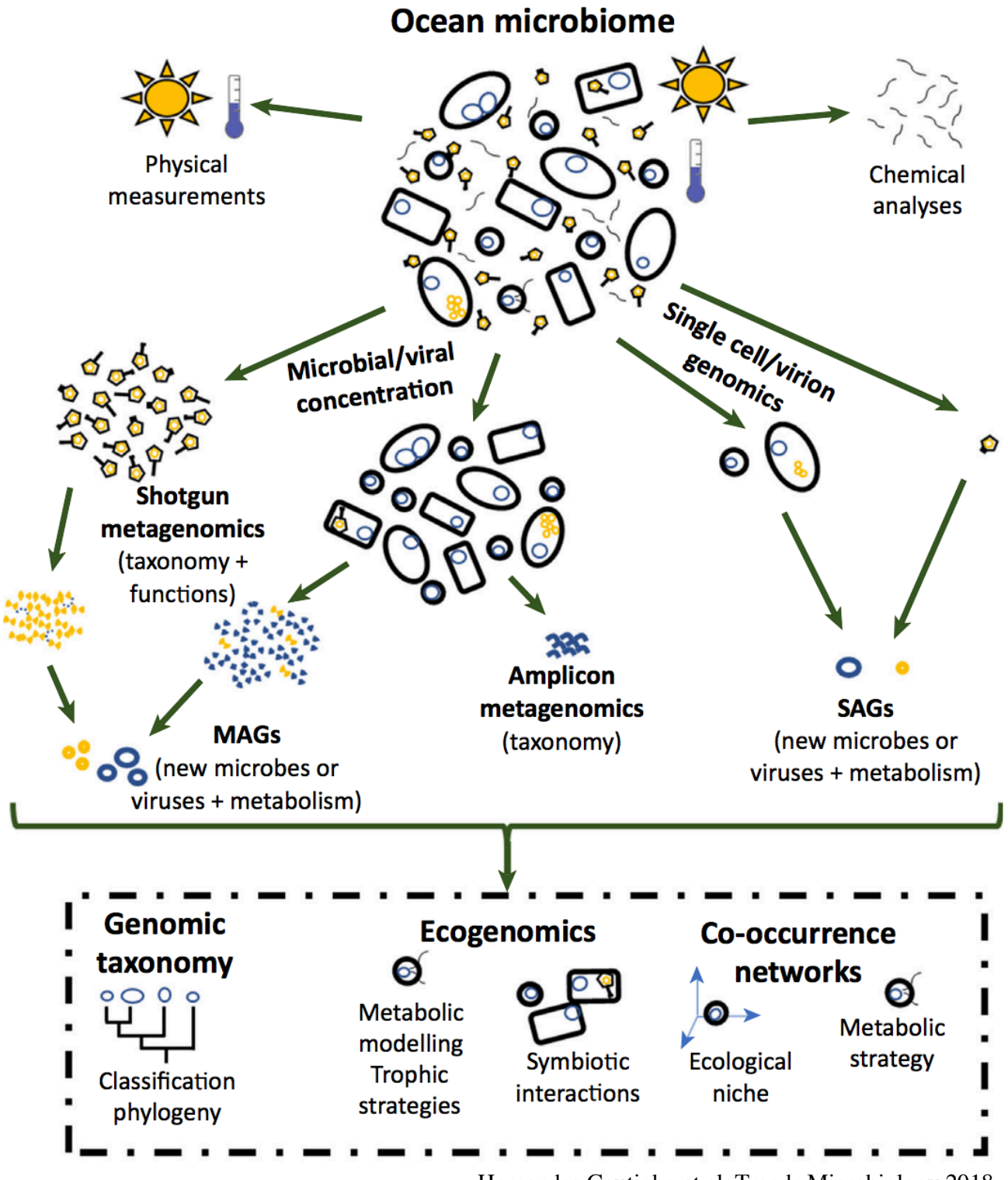
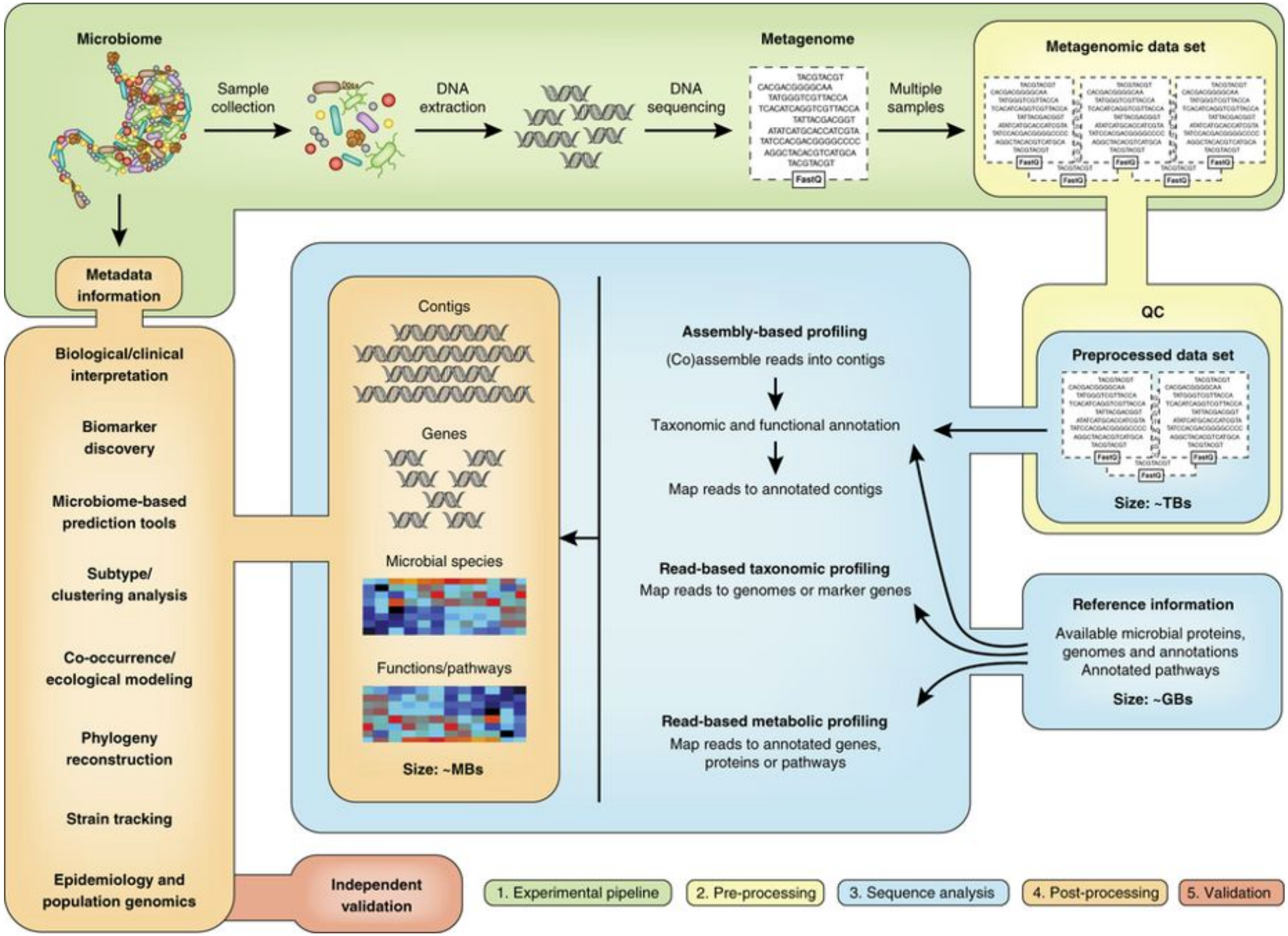


Figure 6 Clustering of the infant gut metagenome by BinSanity, CONCOCT, GroopM, MaxBin, MetaBat, MyCC, *Eren et al. (2015)* and *Sharon et al. (2013)*. The image was generated through Anvi'o.

Summarising



Hernandes Coutinho et al. Trends Microbiology 2018

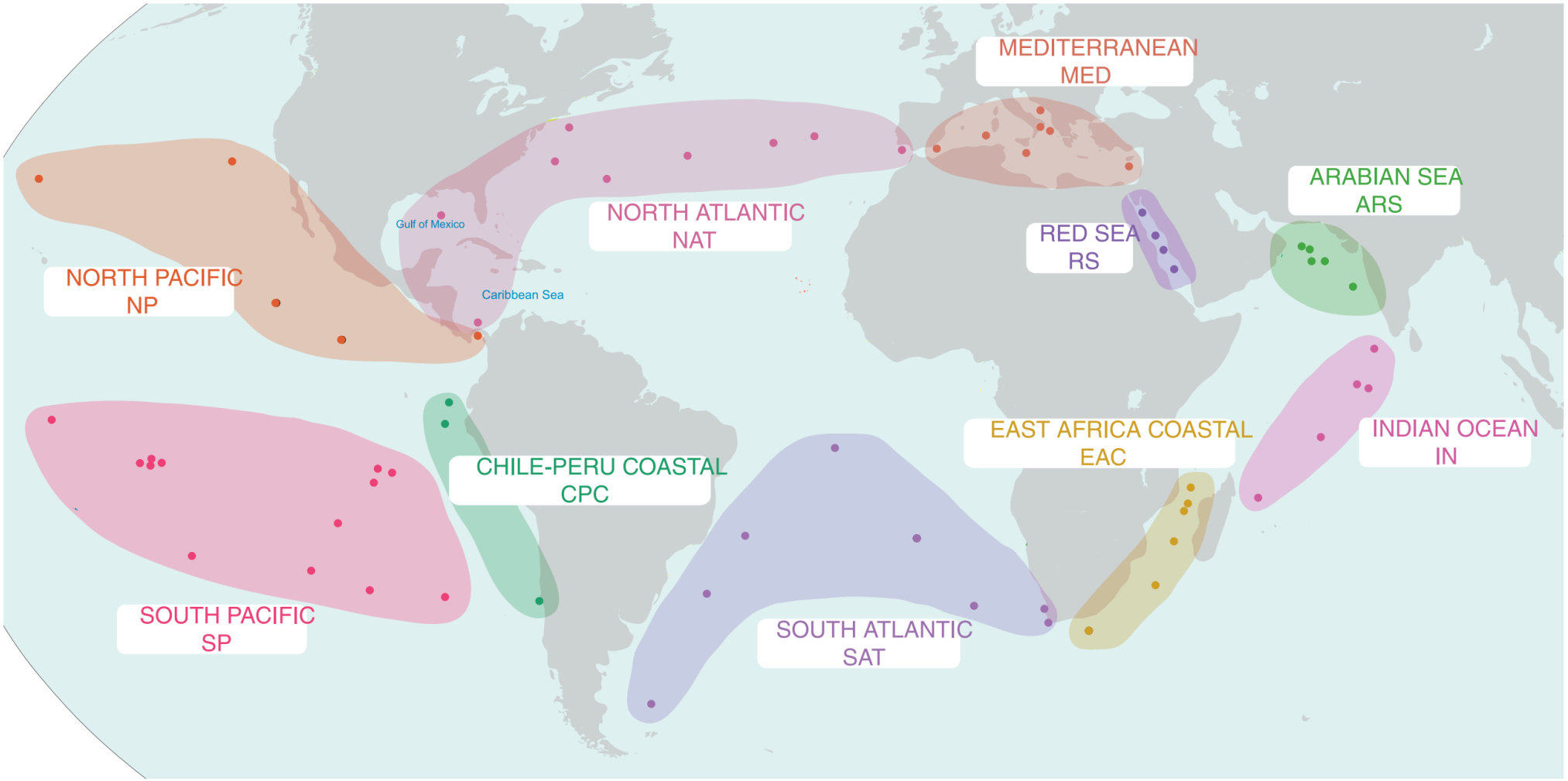


Quince et al, Nat Biotechnology 2017

Example

The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans

Tara Oceans Expedition



Tully et al. Scientific Data 2018

Samples were collected from multiple size fractions, commonly 'viral' (<0.22 µm), 'girus' (0.22–0.8 µm), 'bacterial' (0.22–1.6 µm), and 'protistan' (0.8–5.0 µm)

Co-assembly with Megahit

In total, over 102 billion paired-end reads were assembled into >562 million contigs

"However, this assembly procedure does not resolve issues with abundant organisms with high degrees of strain heterogeneity within a single sample"

[some extra filtering and assembly done at this stage…]

Binning with BinSanity

"Then due to computational limitations imposed during the BinSanity binning method, the secondary contigs from each province were size selected (≥4–14 kb cutoffs) to choose approximately 100,000 contigs for binning (Table 2). Approximately 6 million secondary contigs remain un-binned and are available for analysis.

Province	No. of Secondary Contigs	Size Cutoff (kb)	No. of Binned Contigs	No. of Draft Genomes
Mediterranean	660,937	7.5	95,506	360
Red Sea	328,325	5.0	84,936	180
Arabian Sea	525,636	6.0	99,649	194
Indian Monsoon	285,238	4.0	93,760	72
East Africa Coastal Current	613,778	7.0	91,053	208
South Atlantic	1,373,173	11.5	96,972	360
Chile Peru Coastal	857,548	5.5	95,557	146
South Pacific	807,193	14.0	104,598	536
North Pacific	943,809	7.0	96,396	254
North Atlantic	804,316	8.5	104,848	321
SUM	7,199,953	-	963,275	2,631

Tree inference

"Two sets of single-copy markers recalcitrant to horizontal gene transfer were identified and used to construct phylogenetic trees; a set of 16 generally syntenic markers identified in Hug, et al. 29 and an alternative set of 25 markers"

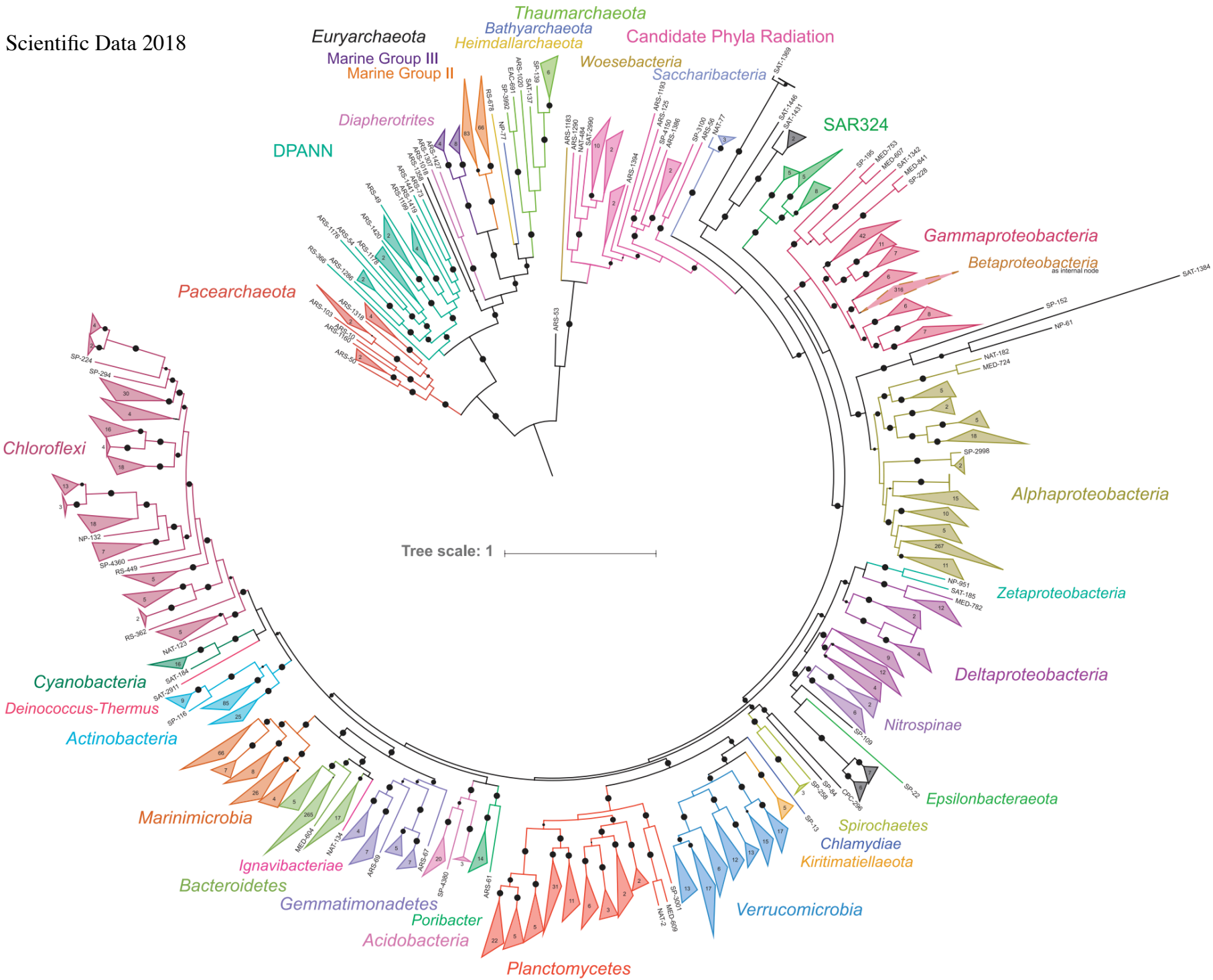
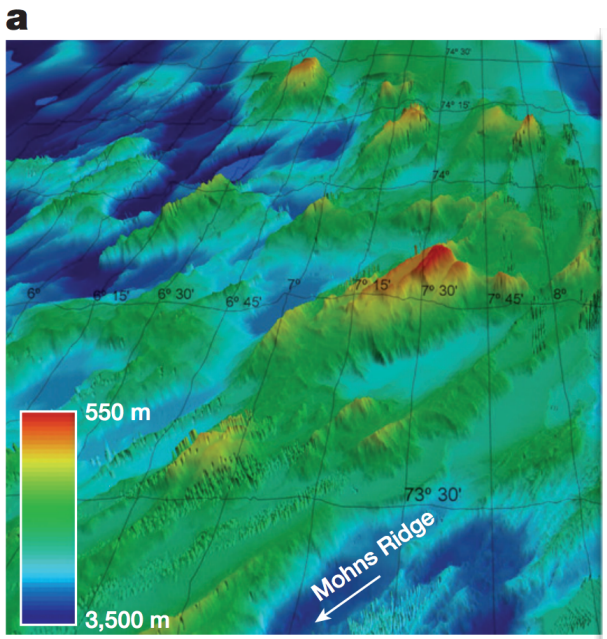


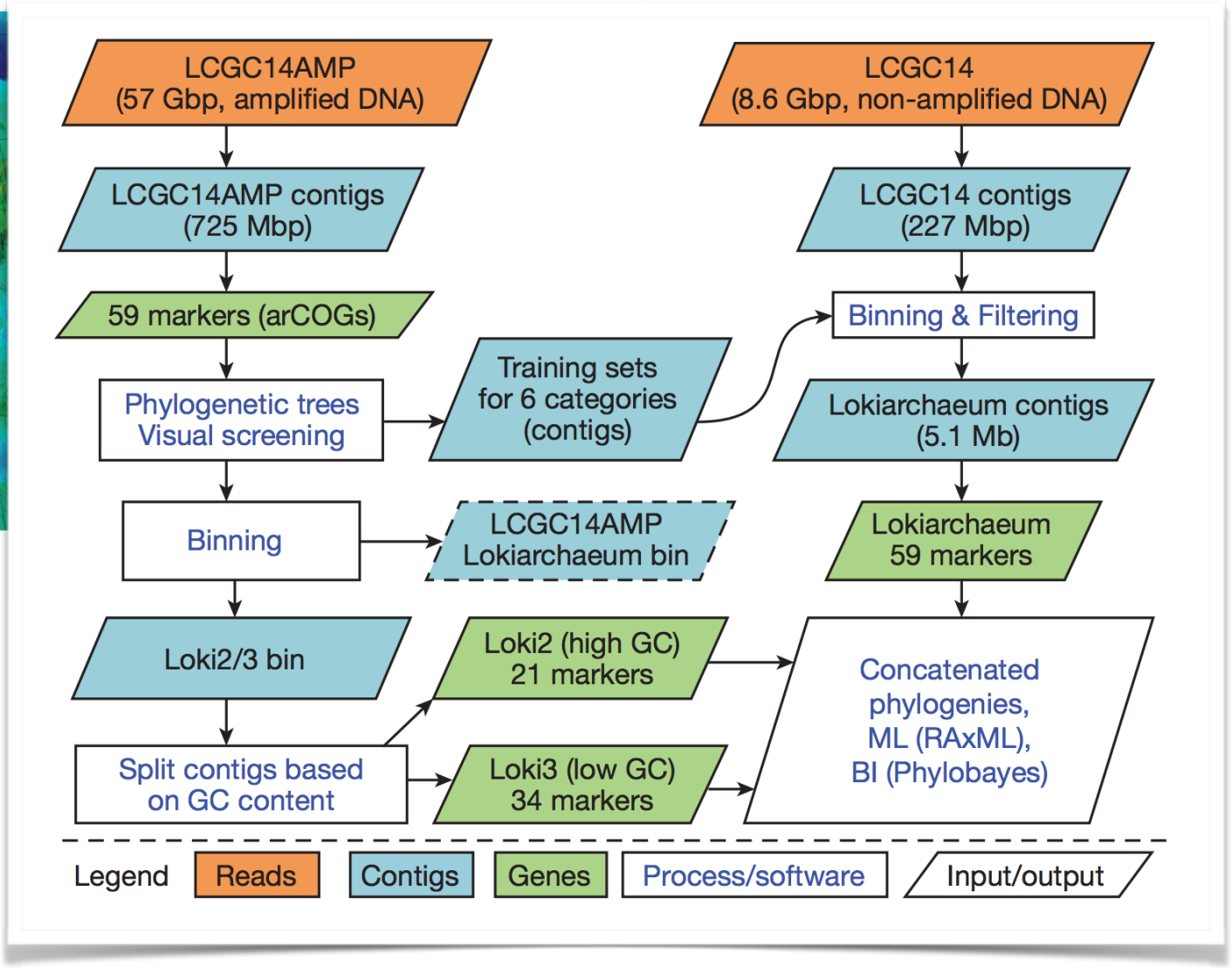
Figure 2. A maximum likelihood tree of the TOBG draft genomes based on 16 concatenated single-copy phylogenetic markers. Bootstrap values >0.75 are shown. Circle size representing the bootstrap value is scaled

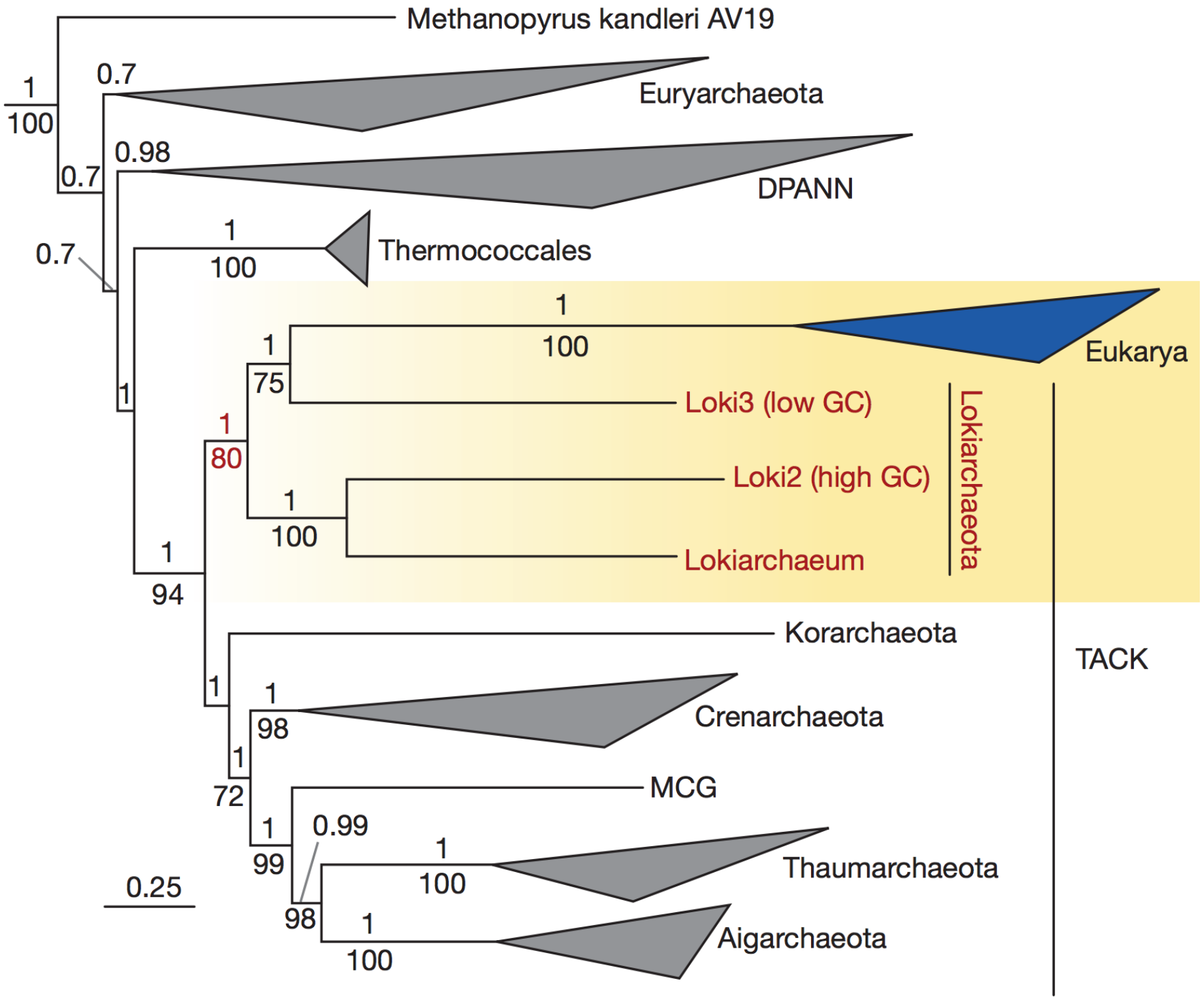
New phyla!

Lokiarchaea

- Spang et al. Nature 2015
- Da Cunha et al. PLOS Genetics 2017: reanalysis and dispute on phylogenetic conclusions drawn
- Zaremba-Niedzwiedzka et al., Nature 2017: Asgard clade

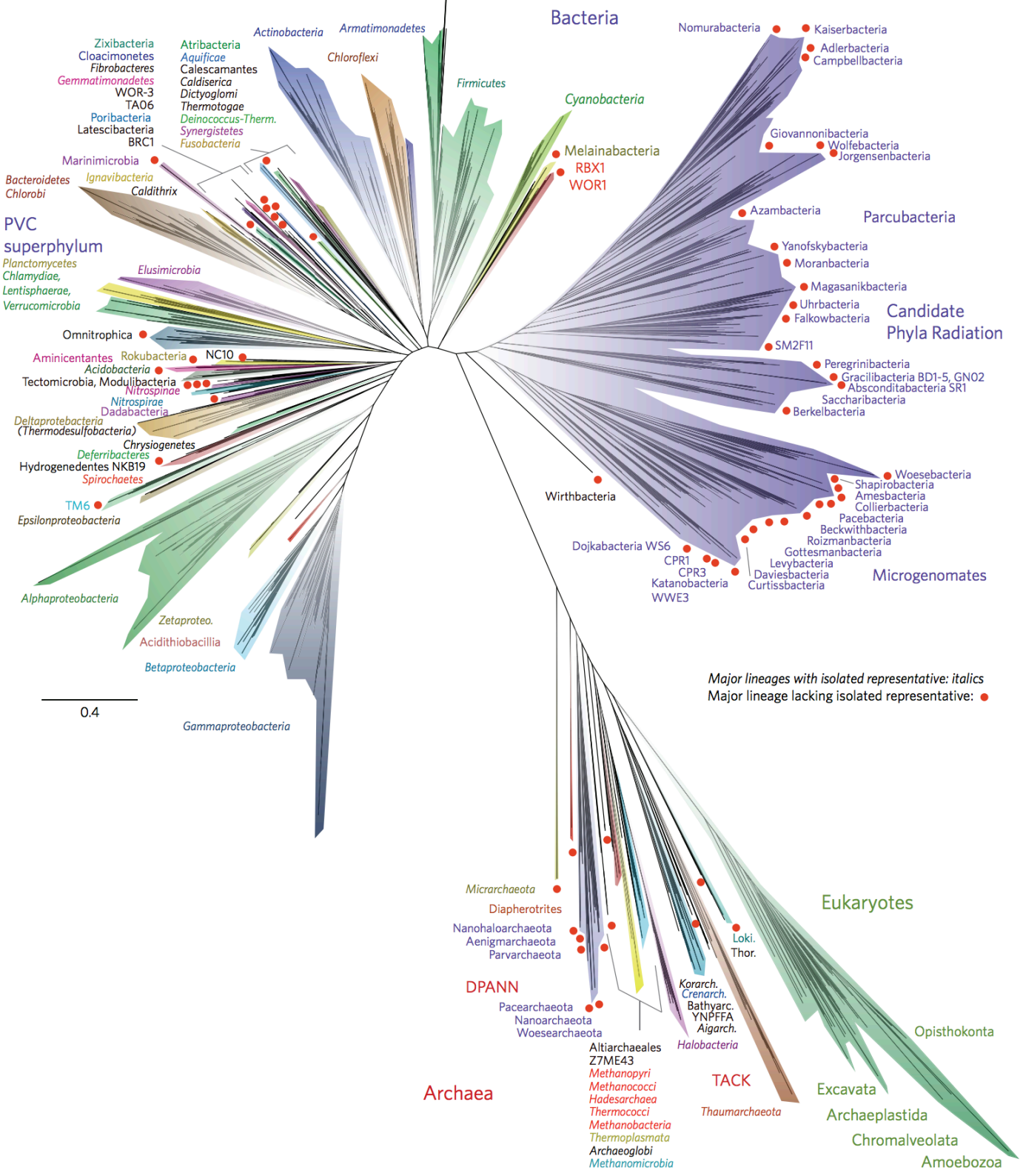





Methanopyrus kandleri AV19
Euryarchaeota
DPANN
Thermococcales
Eukarya
Loki3 (low GC)
Loki2 (high GC)
Lokiarchaeum
Lokiarchaeota
Korarchaeota
Crenarchaeota
MCG
Thaumarchaeota
Aigarchaeota
TACK
0.25

Bacteria

- Hug et al. Nature Microbiology 2016
- Parks et al. Nature Microbiology 2017
- Bernard et al. GBE 2018



Virus Diversity

- **Review: Simmonds et al Rev Nat Micro 2017**
- Review: Koonin, 2018, Koonin & Dolja 2018
- Review: Hernandes Coutinho et al. 2018

Virus diversity

The discrepancy between the number of potential taxa into which viruses in environmental samples could be classified and the number currently recognized by the ICTV is striking. A recent analysis of dsDNA virus sequences that were characterized as part of the *Tara* Oceans expedition from 43 surface ocean sites worldwide identified 5,476 distinct dsDNA virus populations[21], but only 39 of these corresponded to virus groups that have been classified by the ICTV. Most of these populations were both abundant and widely dispersed geographically, but

Some open challenges

- Mapping methods based on few markers:
  limited resolution (too little or too much variation)
- Assembly -> fragmented or chimeric assemblies
  - Once genomes have been assembled, still the same issues of genome annotation and analysis
- Marker-based analyses typically use single copy genes -> ignores paralogs
- Eukaryotic genomes heavily underrepresented