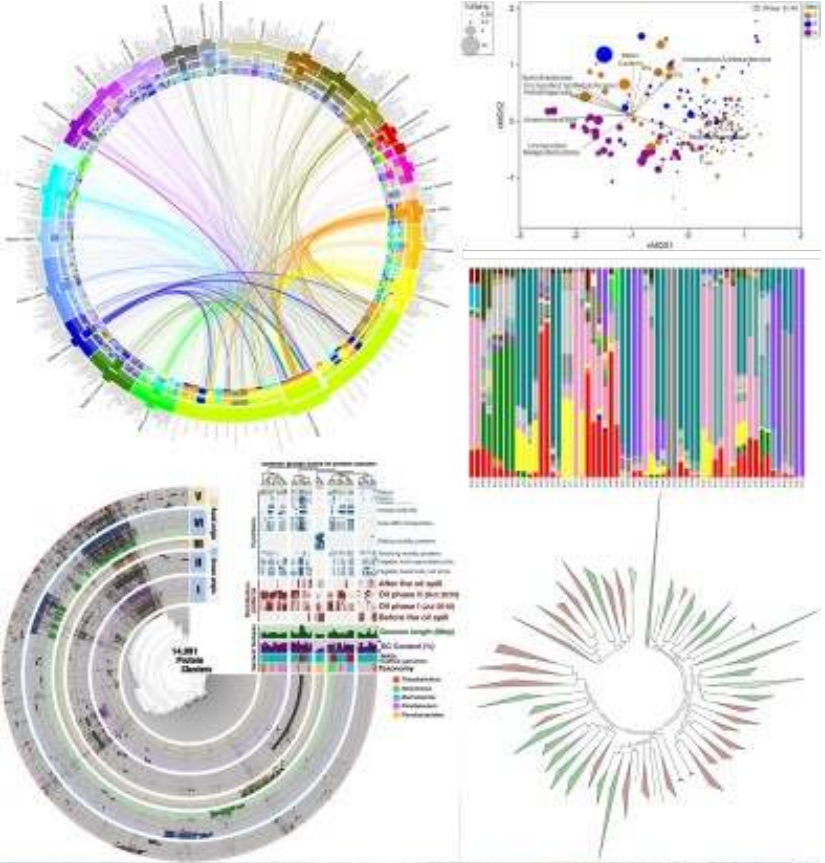


GIOVANNELLI LAB

DIPARTIMENTO DI BIOLOGIA

University of Naples Federico II

Environmental Metagenomic

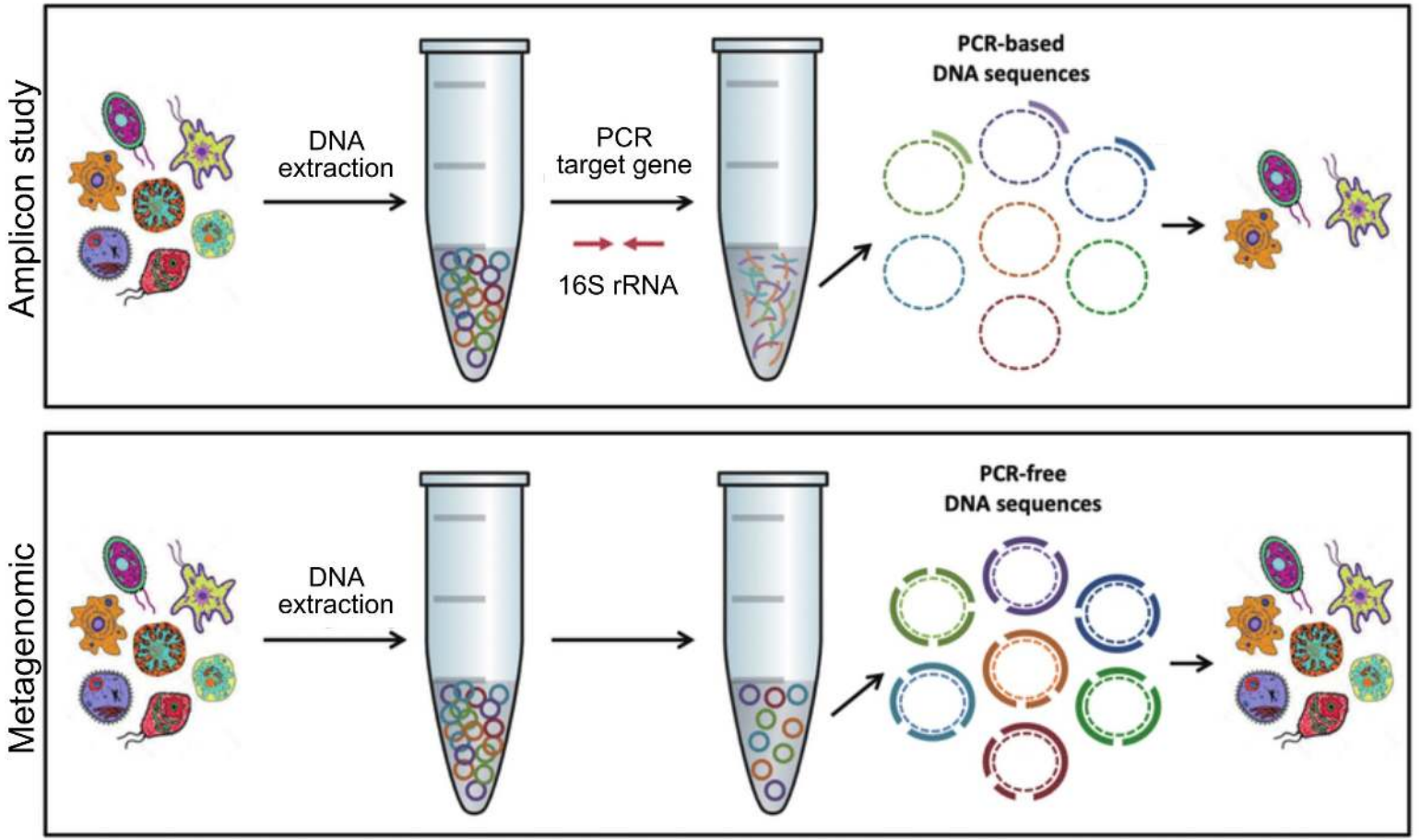
aa 2020-2021

INTRODUCTION TO SHOTGUN METAGENOMIC

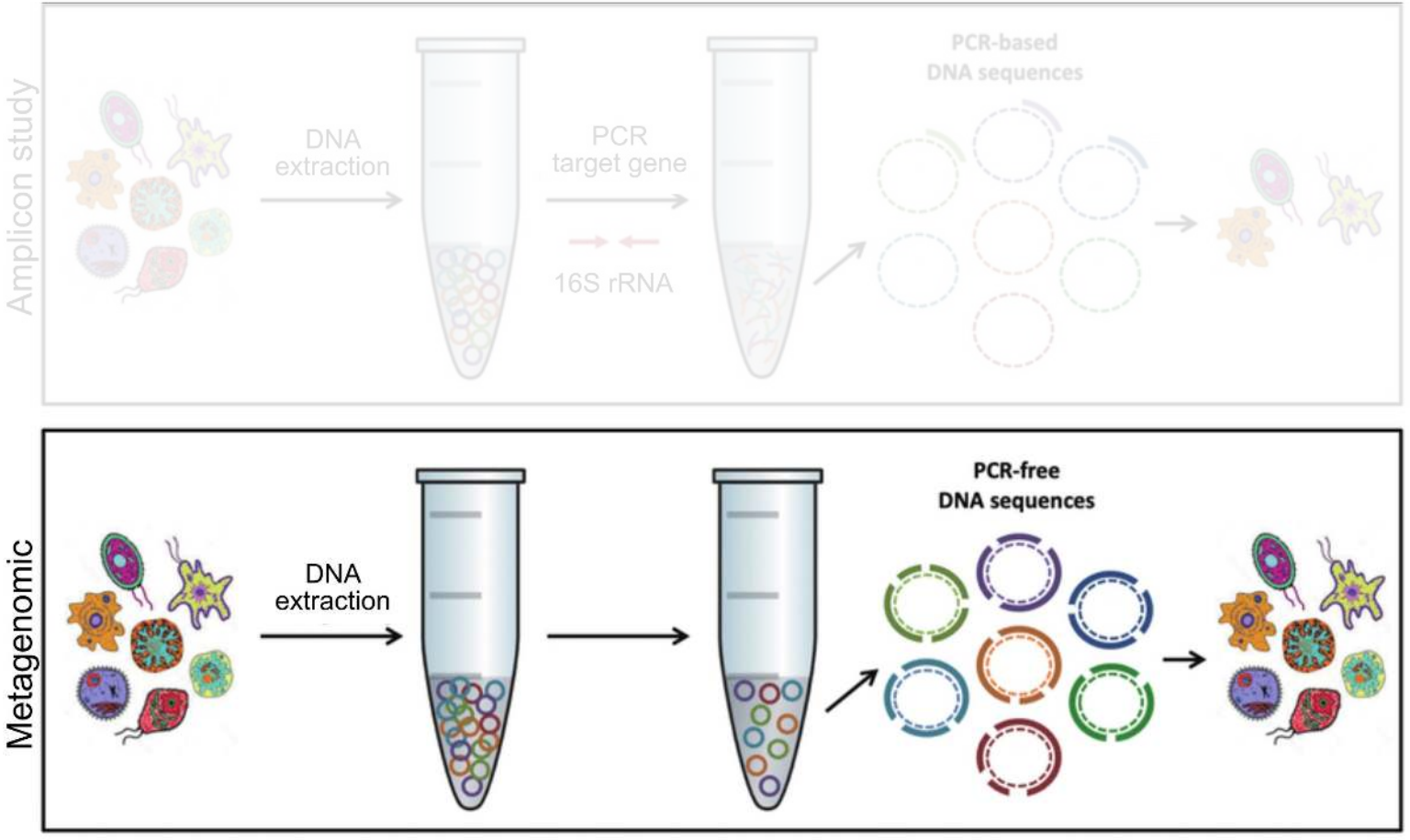
Donato Giovannelli

donato.giovannelli@unina.it • @d_giovannelli • http://dgiovannelli.github.io

Amplicon vs Metagenomic

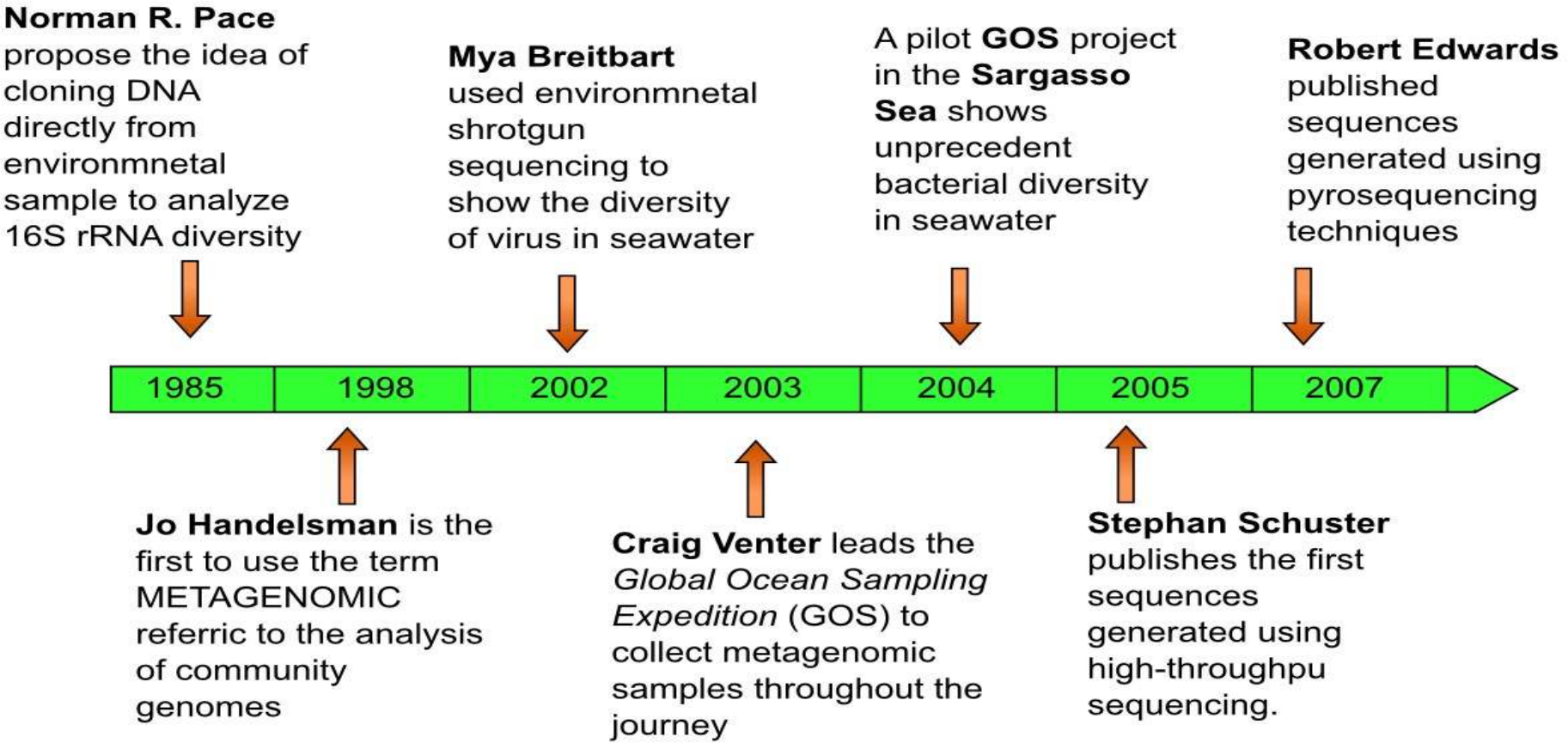


Amplicon vs Metagenomic



Microbial Diversity: metagenomic data

Brief history of Metagenomic

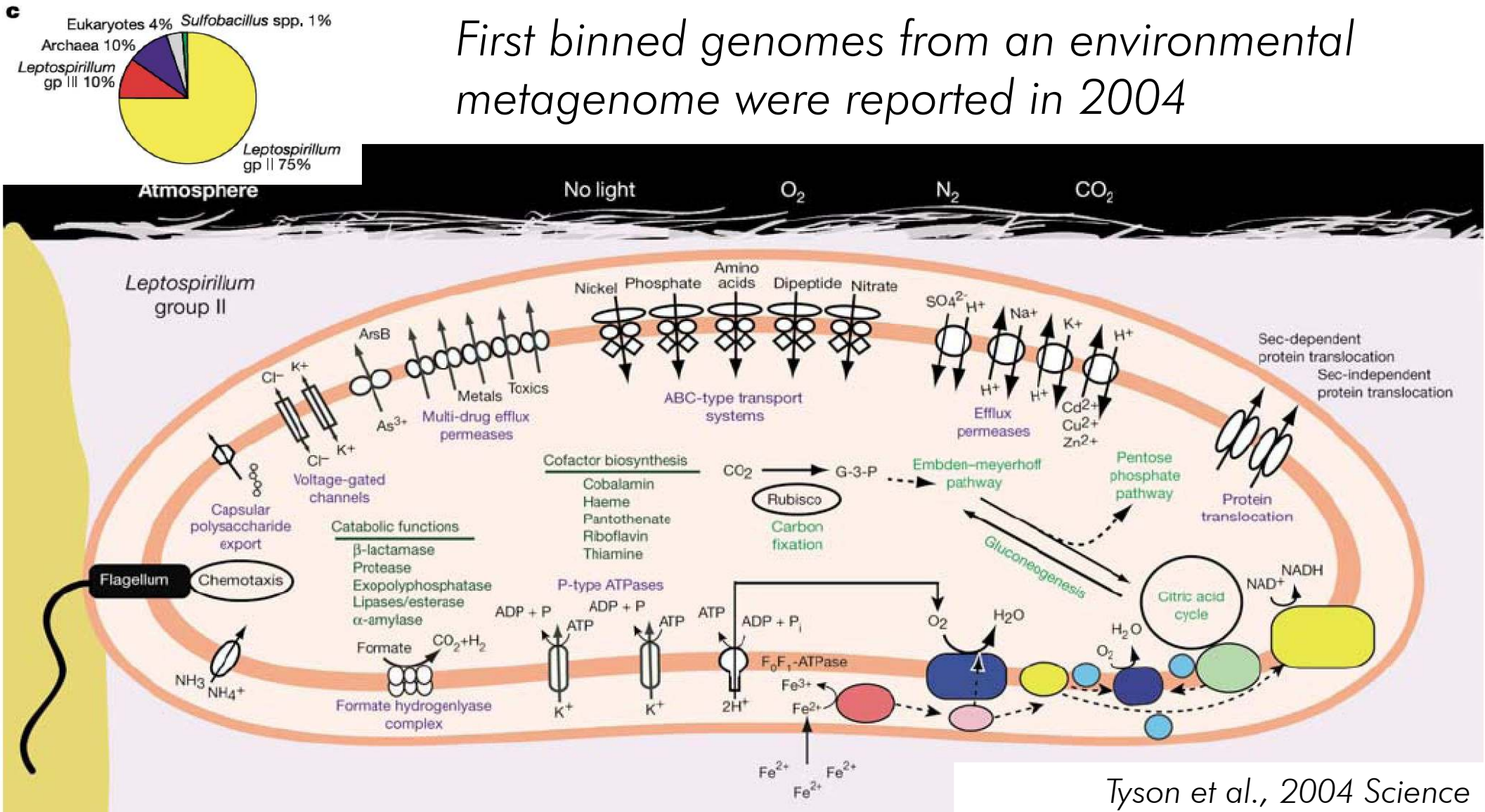


The history of metagenomics: An incomplete summary

a post by Matthew Schechter

https://merenlab.org/2020/07/27/history-of-metagenomics/

First binned genomes from an environmental metagenome were reported in 2004



Tyson et al., 2004 Science

The list of -omics is longer than you may expect

	Google pages
Genomic	*2,780,000*
Transcriptomic	*221,000*
Proteomic	*857,000*
Secretomic	*18,700*
Metabolomic	*29,000*
Metallomeic	*630*
Regulomic	*696*
Lipidomic	*4,620*

...

Metagenomic

Metagenomic reveals the functional (as well as taxonomic) diversity of a community

Being based on the sequencing of DNA, it reveals the metabolic potential, and does not provide any information on the active metabolisms

Metagenomic approaches, intended as the sequencing of the community DNA, can be applied to a sample as it is (directly after extraction of the bulk community DNA) or combined with a number of other techniques (like SIP, enrichments, etc..)

The ***depth of sequencing*** *(how many reads are obtained for each sample) and the* ***read length*** *(how many base pairs is each read), both depending on the sequencing technology used, very important parameters that can severely affect the final result*

Metagenomic basic approaches

Metagenomic analysis can happen at three different levels:

Metagenomic basic approaches

Metagenomic analysis can happen at three different levels:

***Read-based analysis**: it applies to the analysis of unassembled reads. Potentially provides a large amount (ecologically speaking) of coarse level information about the metabolic potential at the community level. Put another way, we can broadly assign a lot of function without much information about their coupling in different organisms*

Metagenomic basic approaches

Metagenomic analysis can happen at three different levels:

Read-based analysis: *it applies to the analysis of unassembled reads. Potentially provides a large amount (ecologically speaking) of coarse level information about the metabolic potential at the community level. Put another way, we can broadly assign a lot of function without much information about their coupling in different organisms*

Contig-based analysis: *it analyses the assembled reads to provide genetic context to the identified functions and might provide also taxonomic information. The amount of information obtained is more detailed but represent a smaller subset of the community (i.e. only the reads that assemble)*

Metagenomic basic approaches

Metagenomic analysis can happen at three different levels:

Read-based analysis: *it applies to the analysis of unassembled reads. Potentially provides a large amount (ecologically speaking) of coarse level information about the metabolic potential at the community level. Put another way, we can broadly assign a lot of function without much information about their coupling in different organisms*

Contig-based analysis: *it analyses the assembled reads to provide genetic context to the identified functions and might provide also taxonomic information. The amount of information obtained is more detailed but represent a smaller subset of the community (i.e. only the reads that assemble)*

Genome resolved metagenomic: *it analyses near-complete genomes assembled from metagenomes (MAGs). It provides highly detailed informations about a small subset of the community (only the contigs that bin into high quality genomes)*

Metagenomic jargon

Assembly: *linking together reads that are partially overlapping*

Contigs: *sequences of variable length obtained from assembling the reads*

Binning: *Clustering together contigs of variable length (bins) based on their identification as coming from a close group of highly related strains*

MAGs: *Metagenomes assembled genomes*

Single-copy marker gene: *ubiquitous conserved genes that are present with a single copy in known genomes*

Annotation: *Assigning taxonomy or function to a sequence*

Read-mapping: *counting the number of reads that recruit (align) to a specific sequence (MAG, contig or gene)*

Technical difficulties and challenges when using (metagen)omic

Study question / hypothesis

Initial sample	→	*Sample size and replicates*
Isolation and purification of starting material	→	*Quality and quantity. Extraction biases*
Technology/ Reads length and quantity / number of spots and resolution of gel	→	*Depth of sequencing and coverage, Sequencing errors, Accuracy and read quality*
Annotation and assembly	→	*Quality of reference database, Annotation accuracy, Assembly quality, Chimeric sequences/contigs*
Taxonomic binning	→	*Lateral gene transfer events*
Genomic reconstruction from metagenome	→	*Multi strain genomes*

Next Gen sequences are on average short (75-450 bp). This is less than the average gene (~1500 bp). What are the implication for our purpose?

Illumina	*50 to 350 bp Paired End*
Pacbio	*up to 1000 bp*
Nanopore	*up to 10 kbp*

Does size matters?

Next Gen sequences are on average short (75-450 bp). This is less than the average gene (~1500 bp). What are the implication for our purpose?

Illumina	*50 to 350 bp Paired End*
Pacbio	*up to 1000 bp*
Nanopore	*up to 10 kbp*

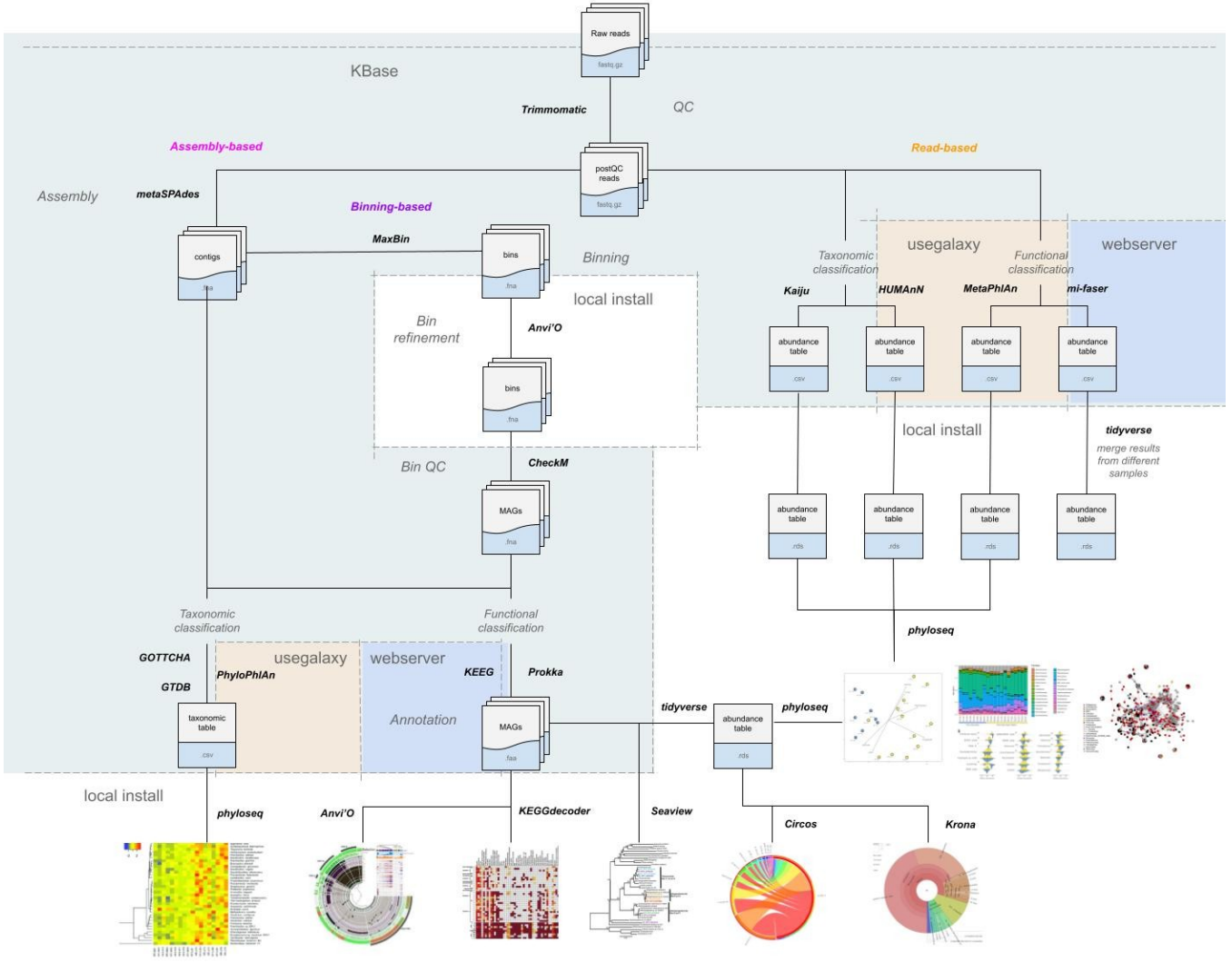
Does size matters?

YES! *In sequencing it makes a big difference. Generally speaking, the longer the reads, higher the amount of information it contains. You can deal with short reads increasing coverage (i.e. number of total transcripts per sample), but gene calling, taxonomic and functional annotation and contigs assembly is a computational challenge.*

NCBI BLAST nr +euk

Shotgun Metagenomic

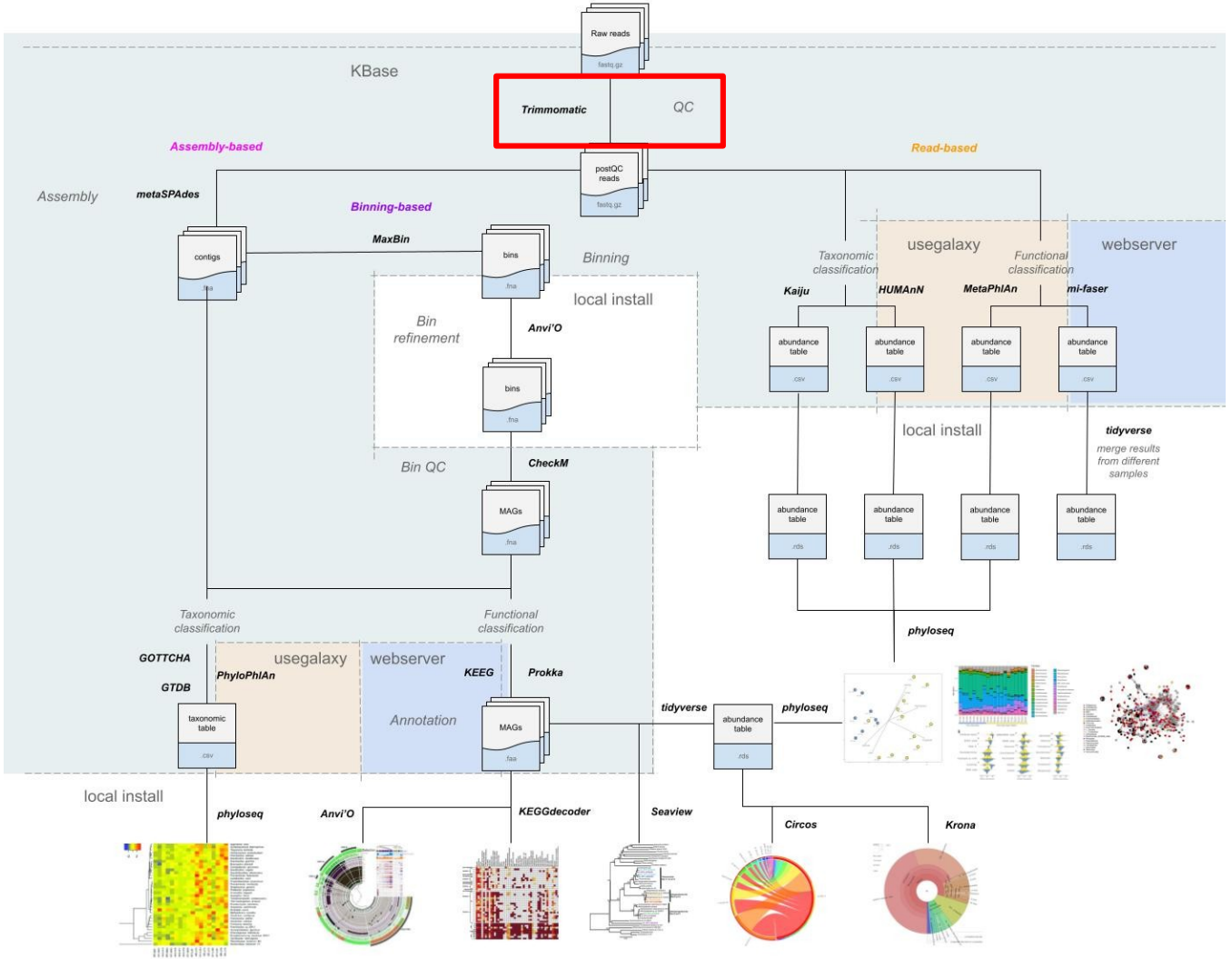
generic workflow - Giovannelli Lab 2021



NCBI BLAST nr +euk

Shotgun Metagenomic

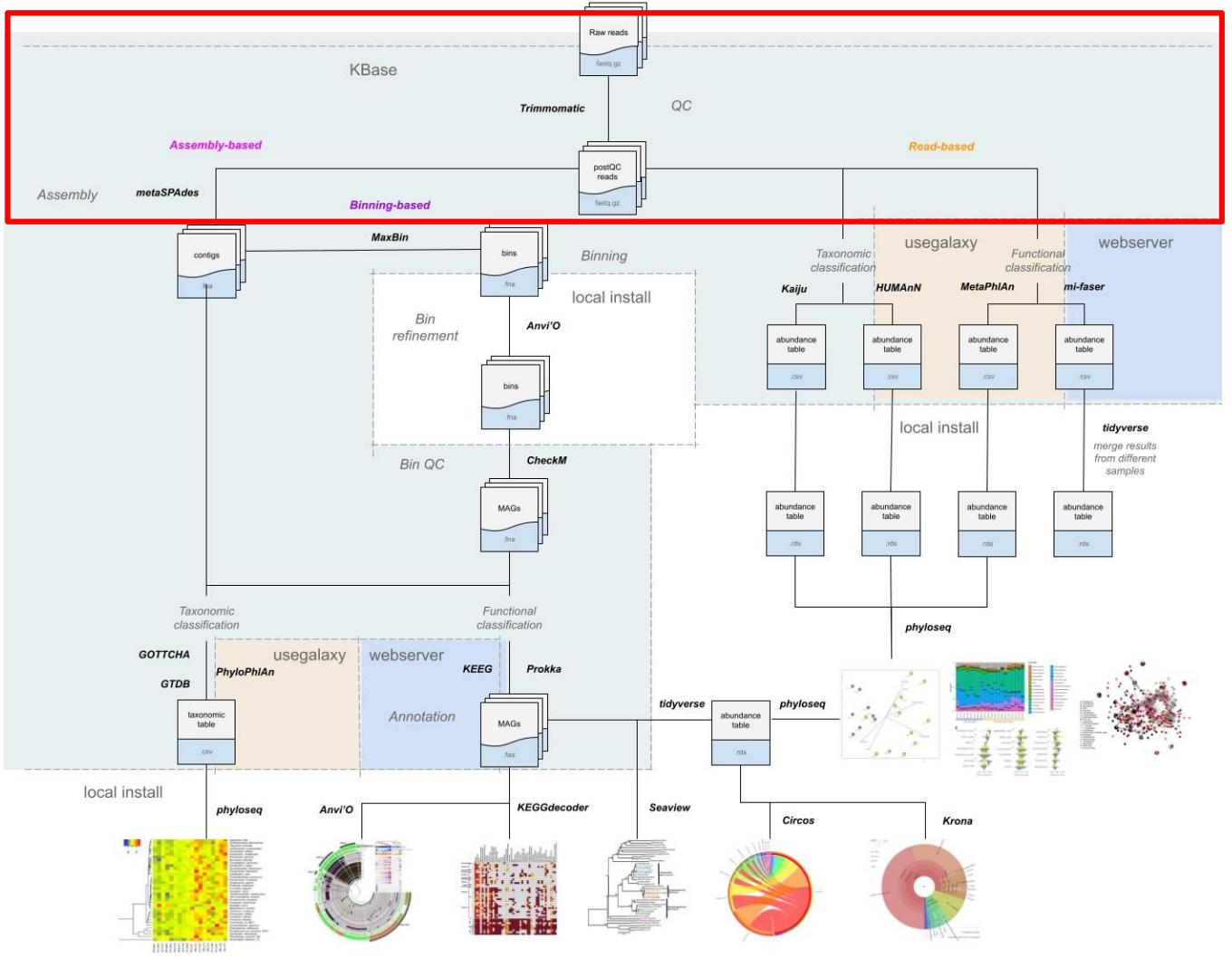
generic workflow - Giovannelli Lab 2021



NCBI BLAST nr +euk

Shotgun Metagenomic

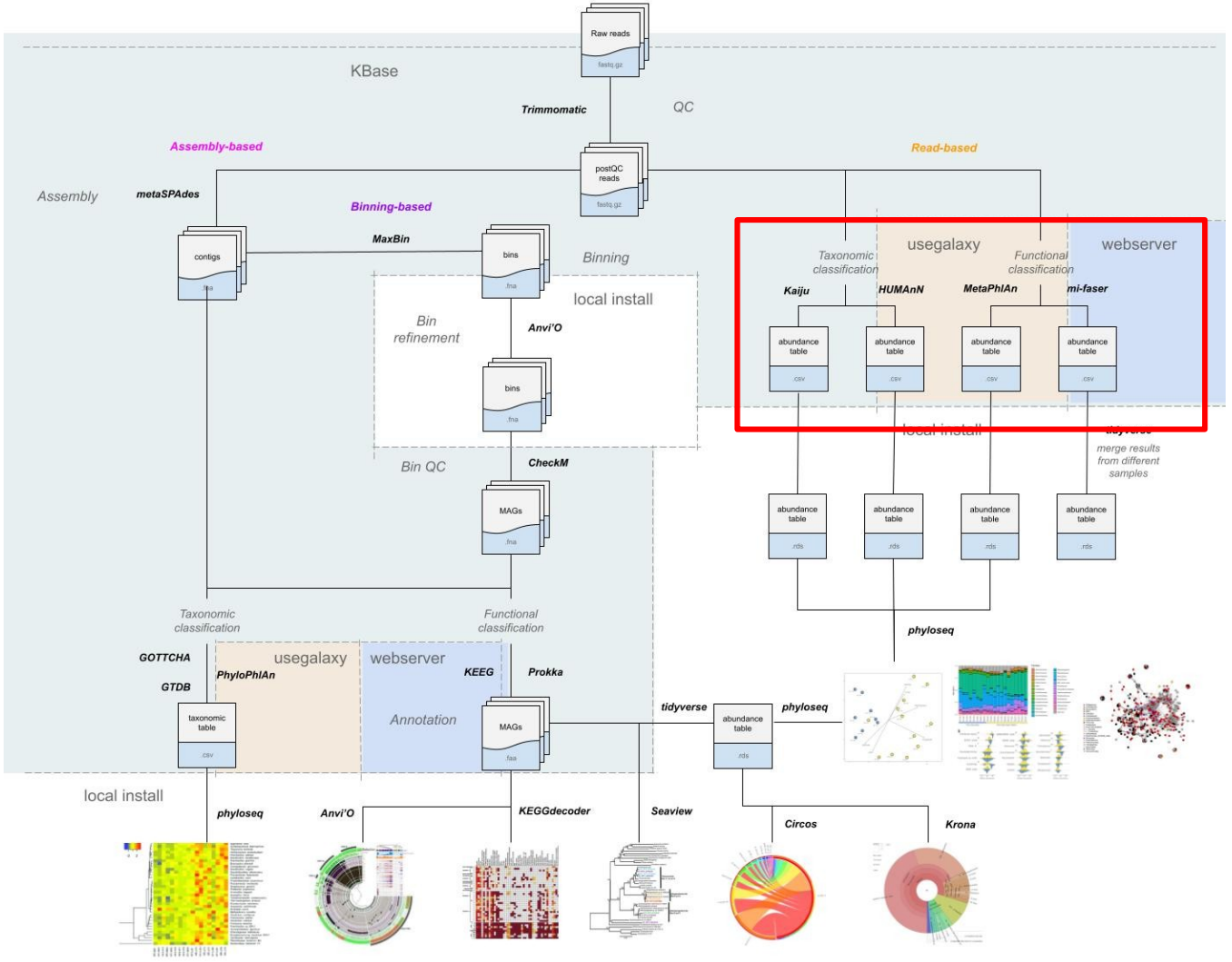
generic workflow - Giovannelli Lab 2021



NCBI BLAST nr +euk

Shotgun Metagenomic

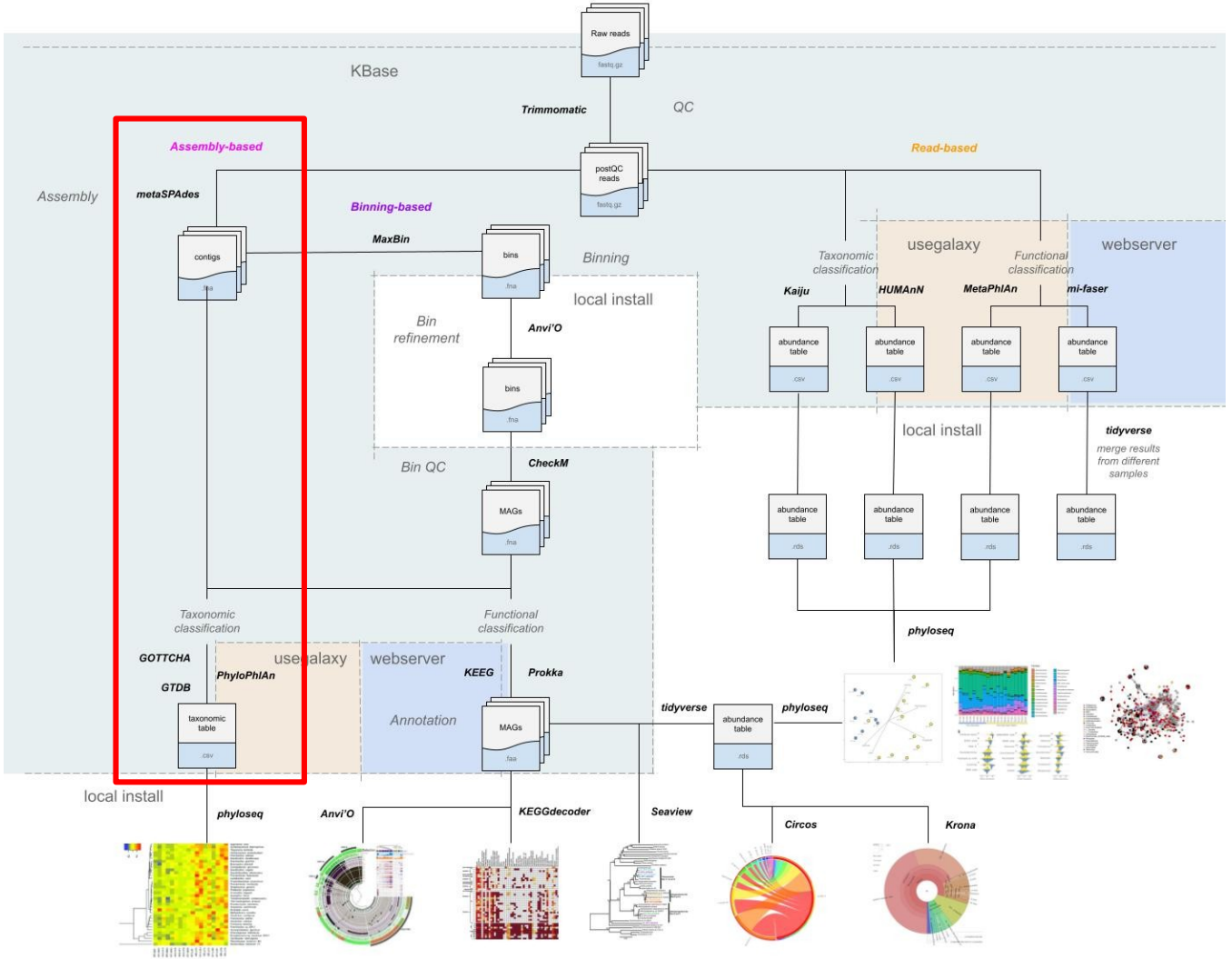
generic workflow - Giovannelli Lab 2021



NCBI BLAST nr +euk

Shotgun Metagenomic

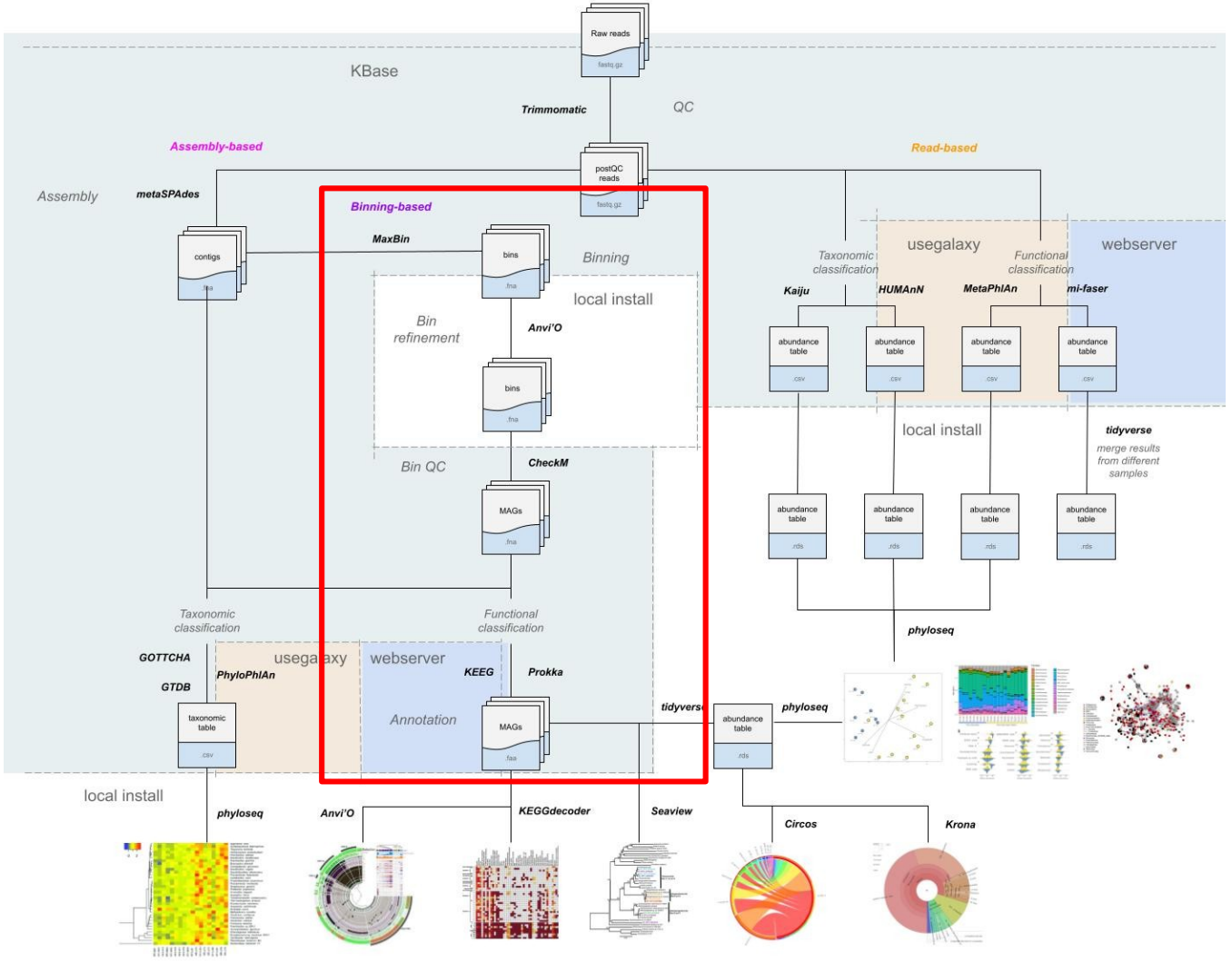
generic workflow - Giovannelli Lab 2021



NCBI BLAST nr +euk

Shotgun Metagenomic

generic workflow - Giovannelli Lab 2021



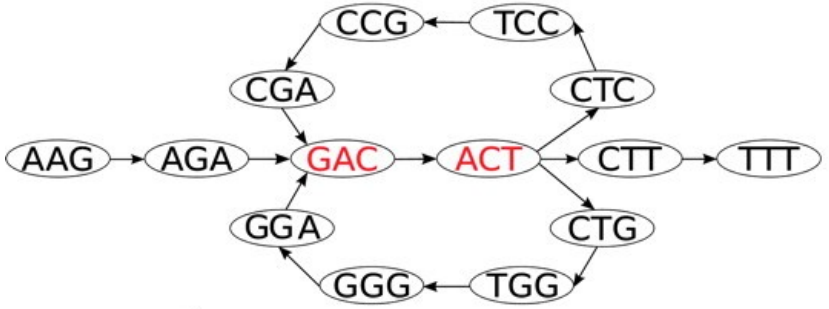
Read Assembly


Reference
TCCTAGAGATCCGCCTCTTAGCGGATATAATACAGCCGAATCTTAGCGGAATTGCCAGCACAG
Reads
CCTAGAGATCCG
GAGATCCGCCTC
ATCCGCCTCTTA
GCCTCTTAGCGG
CTTAGCGGATAT
TAGCGGATATAA
TATAATACAGCC
ACAGCCGAATCT
CCGAATCTTAGC
GAATCTTAGCGG
TCTTAGCGGAAT
AGCGGAATTGCC
GGAATTGCCAGC
AATTGCCAGCAC
TTGCCAGCACAG
Contigs
CCTAGAGATCCGCCTCTTAGCGG
CTTAGCGGATATAATACAGCCGAATCTTAGCGG
TCTTAGCGGAATTGCCAGCACAG

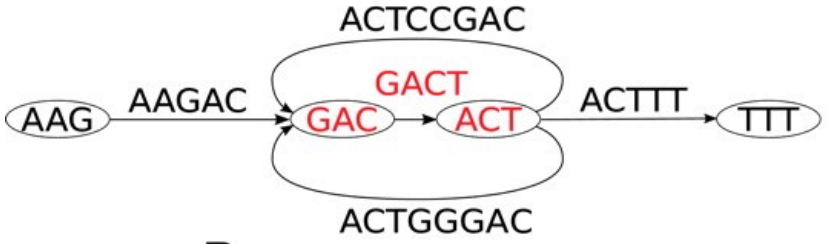
Velvet
Spades
IDBA-UD
Unicycler
AbySS

MEGAHIT
Trinity
MaSuRCA

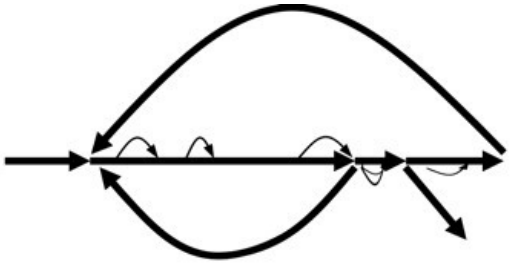
AAGACTCCGACTGGGACTTT



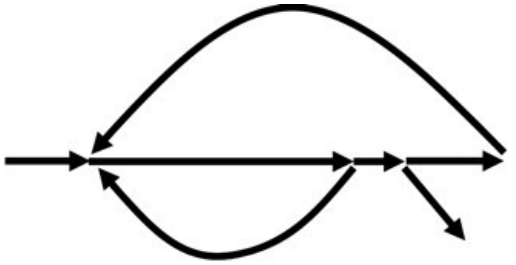
A de Bruijn graph of a sequence



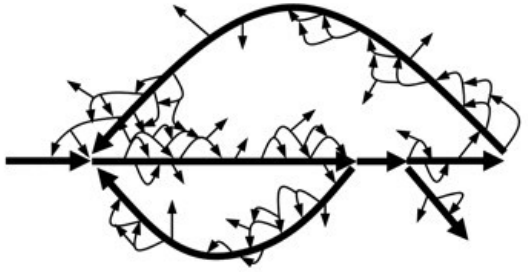
B condensed de Bruijn graph



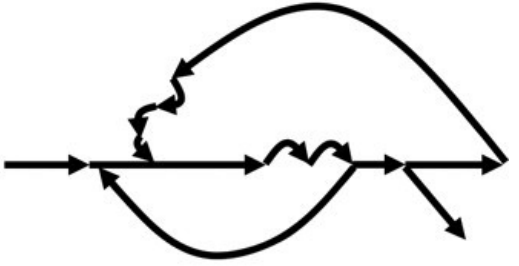
C de Bruijn graph of a genome



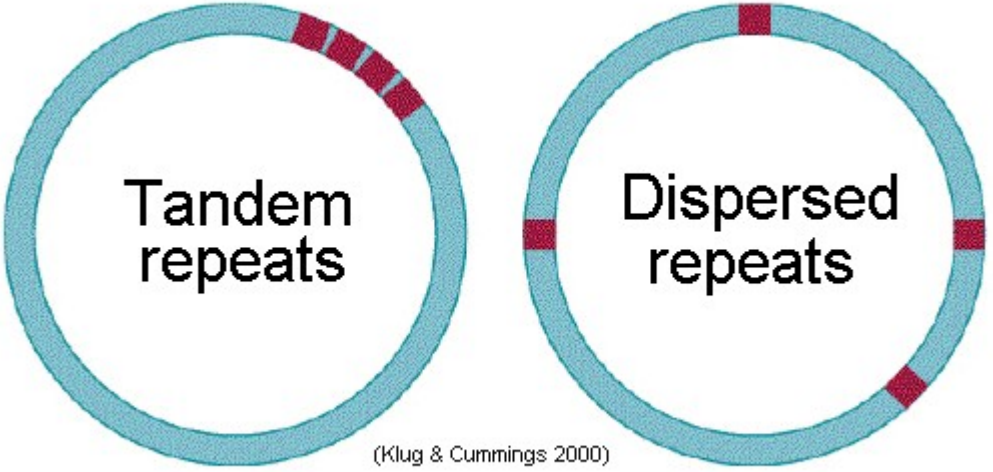
E repeat graph of a genome



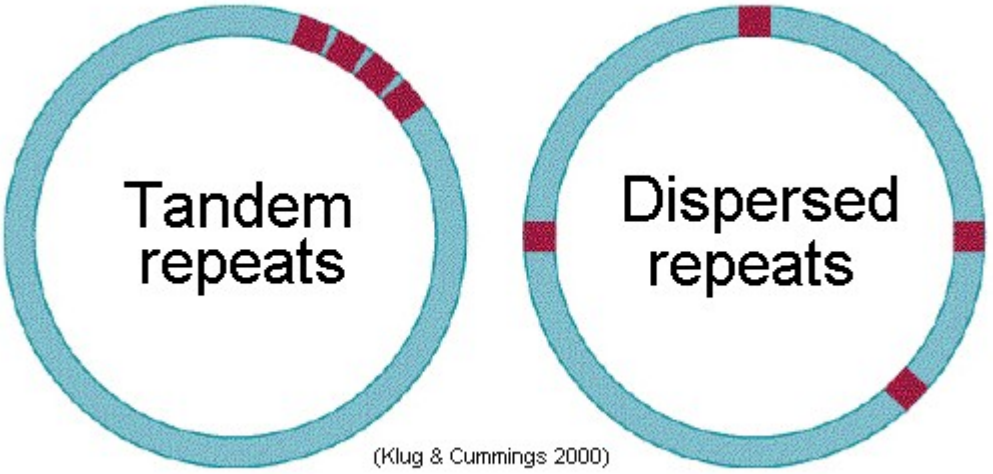
D de Bruijn graph of a set of reads



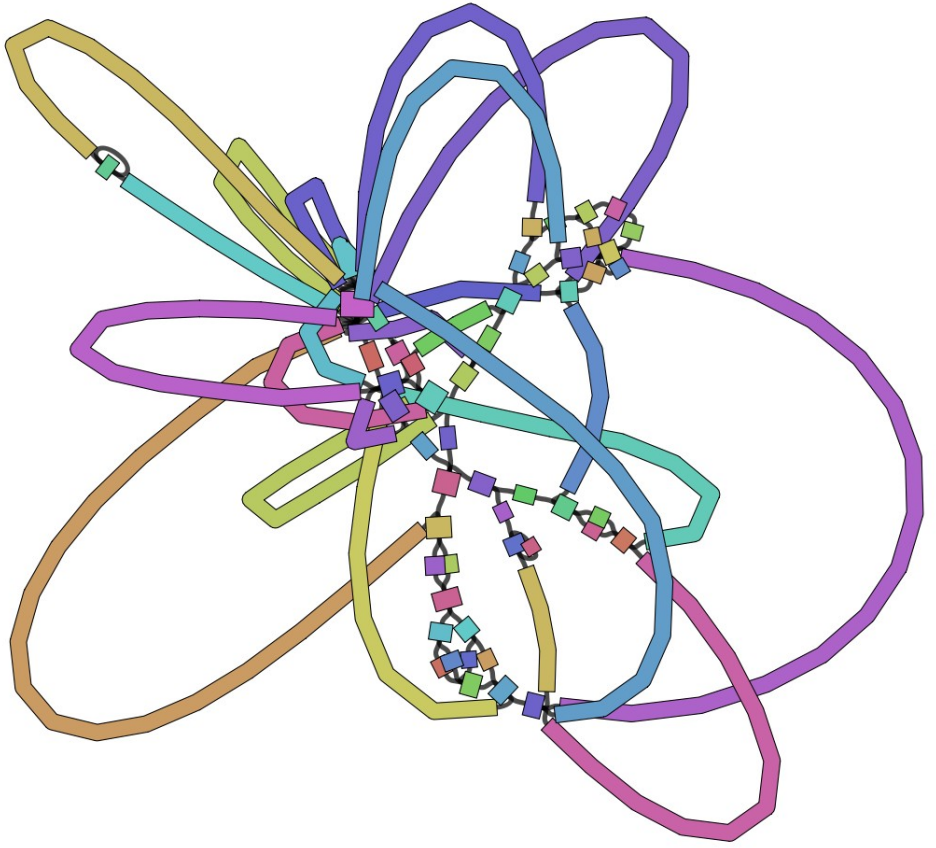
F repeat graph on a set of reads



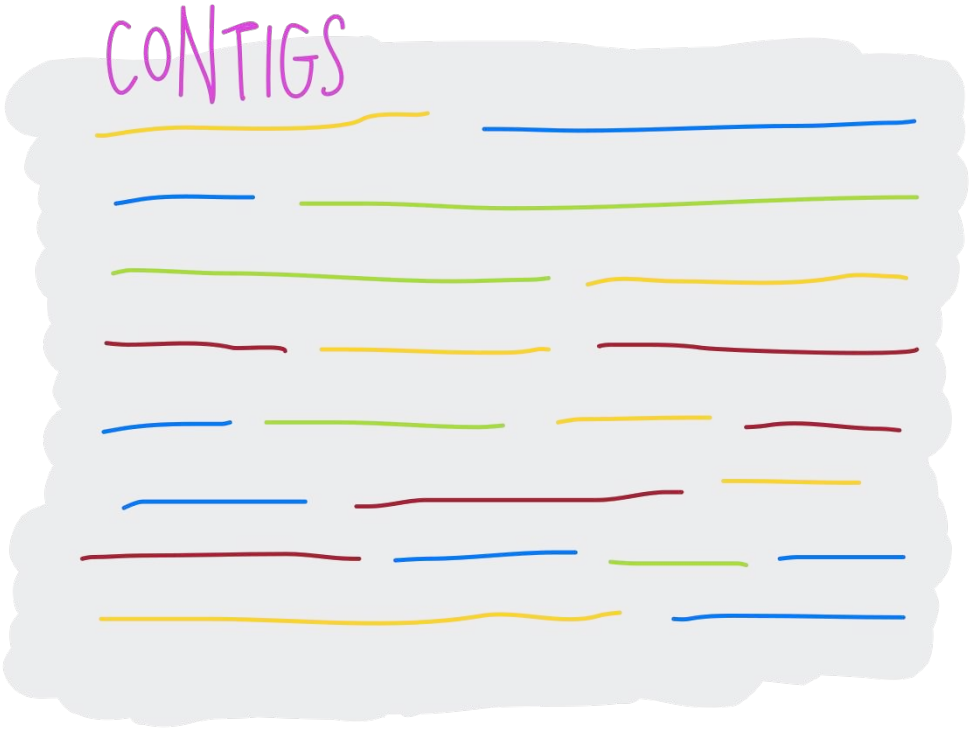
(Klug & Cummings 2000)

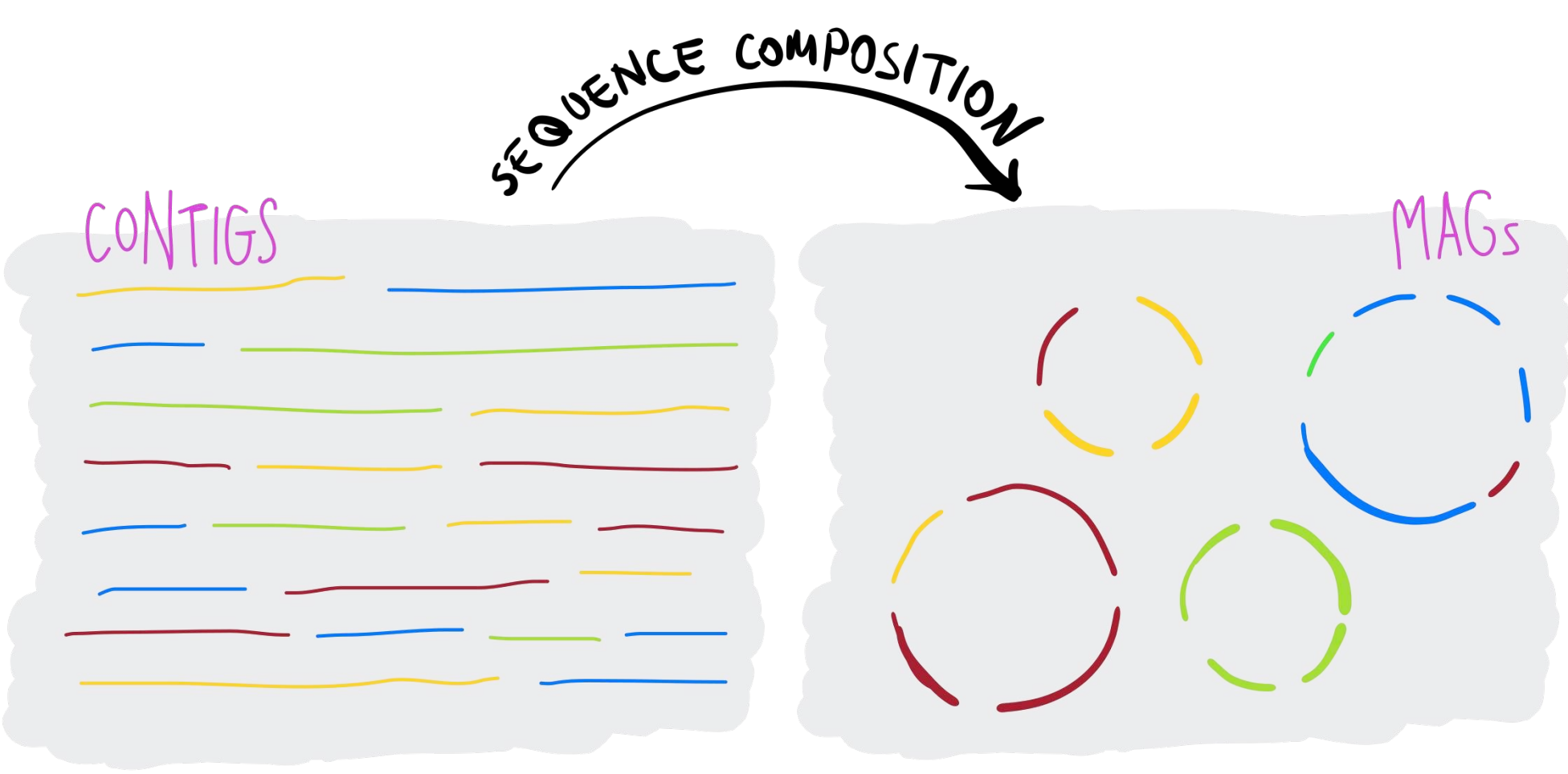


(Klug & Cummings 2000)



Contig binning


CONTIGS


SEQUENCE COMPOSITION
CONTIGS
MAGs

Peter A. Noble[1]
Robert W. Citek[2]
Oladele A. Ogunseitan[3]

[1]**Belle W. Baruch Institute for Marine Biology and Coastal Research, University of South Carolina, Columbia, SC, USA**
[2]**Department of Soil and Environmental Science, University of California at Riverside, Riverside, CA, USA**
[3]**Department of Environmental Analysis and Design, University of California at Irvine, Irvine, CA, USA**

Tetranucleotide frequencies in microbial genomes

A computational strategy for determining the variability of long DNA sequences in microbial genomes is described. Composite portraits of bacterial genomes were obtained by computing tetranucleotide frequencies of sections of genomic DNA, converting the frequencies to color images and arranging the images according to their genetic position. The resulting images revealed that the tetranucleotide frequencies of genomic DNA sequences are highly conserved. Sections that were visibly different from those of the rest of the genome contained ribosomal RNA, bacteriophage, or undefined coding regions and had corresponding differences in the variances of tetranucleotide frequencies and GC content. Comparison of nine completely sequenced bacterial genomes showed that there was a nonlinear relationship between variances of the tetranucleotide frequencies and GC content, with the highest variances occurring in DNA sequences with low GC contents (less than 0.30 mol). High variances were also observed in DNA sequences having high GC contents (greater than 0.60 mol), but to a much lesser extent than DNA sequences having low GC contents. Differences in the tetranucleotide frequencies may be due to the mechanisms of intercellular genetic exchange and/or processes involved in maintaining intracellular genetic stability. Identification of sections that were different from those of the rest of the genome may provide information on the evolution and plasticity of bacterial genomes.

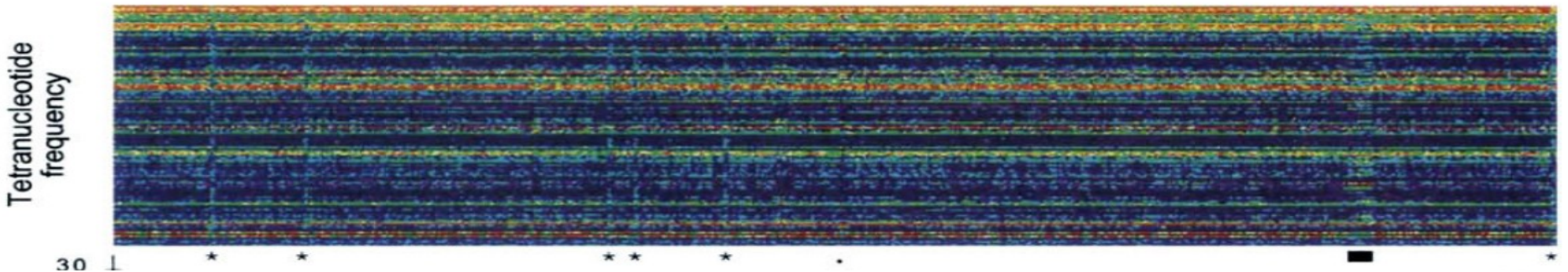


Figure 1. Fingerprints, variances of tetranucleotide frequencies, and GC values of sections of the *Haemophilus influenzae* Rd genome are consecutively ordered from the *Not*I restriction site [9]. Each column of the color image represents the fingerprint obtained from the analysis of one DNA sequence (*i.e.*, a 3000 bp section). Each row represents the frequency of a specific tetranucleotide and its complement. Tetranucleotides are arranged alphabetically on the *y*-axis. Each tetranucleotide is represented by a box, whose color is determined by its frequency, ranging from purple (low) to red (high). A star (*) identifies sections containing ribosomal RNA. The black bar identifies the location of the cryptic Mu-like bacteriophage. The variance and GC values were computed from the analysis of one section.

https://merenlab.org/momics/

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1

k=2

GTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3

k=2

GTTTTGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	3

k=2

GTTTTGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	1	0	0	0	0	1	1	1	0	0	1	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	1	1	0	0	0	0	1	1	1	0	0	1	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	1	1	0	0	0	0	1	1	1	0	0	2	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	1	1	0	0	0	1	1	1	1	0	0	2	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	2	1	0	0	0	1	1	1	1	0	0	2	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	2	1	0	0	0	1	1	1	1	0	0	2	4

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	6	0	1	2	1	1

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	0	0	1	2	1	1

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC
GAAGCAGAAAAGAAACTCCTTAATCATGCCAAAAC

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	0	0	1	2	1	1

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC
GAAGCAGAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	GA	CA	CC	CG	GC	AT	TA

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	6	0	1	2	1	1

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC
GAAGCAGAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
11	3	4	4	5	2	0	2	2	1

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	6	0	1	2	1	1

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC
GAAGCAGAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
11	3	4	4	5	2	0	2	2	1

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	6	0	1	2	1	1

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC
GAAGCAGAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
11	3	4	4	5	2	0	2	2	1

→ PALINDROMES :)

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y										
Z										
L										
K										
M										

k=2

ACTTCCGCAGTCGGGCATTACGCGTTGTGGAATGA

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z										
L										
K										
M										

k=2

ACTTGCGCAGTCGCGCATTACGCGTAGTGGAATAA

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L										
K										
M										

k=2

GGAGCGTTTTATTTAGTACCGGTTTTGAAGTTAAC

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K										
M										

k=2

GCCGCGAGCGGCCCGCGCCGGCTTCGGCGCCGCAC

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K	1	1	2	2	1	8	9	10	0	0
M										

k=2

GGGCCTGCGCCGGTCCAGTCACCCGGCTGCGACCT

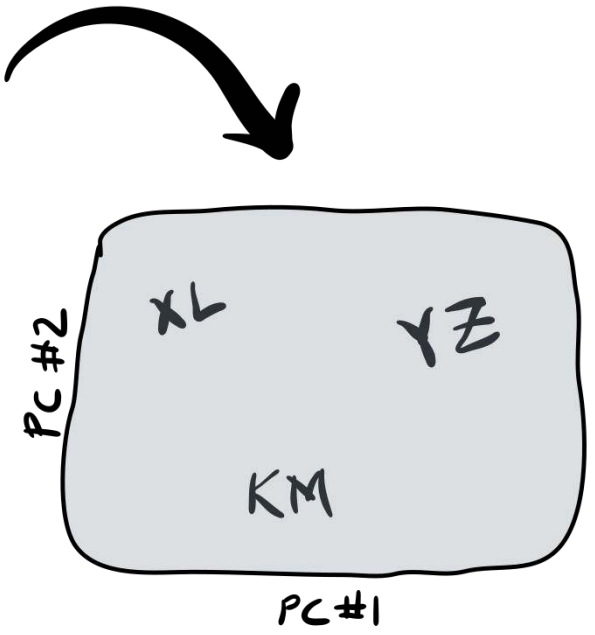
	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K	1	1	2	2	1	8	9	10	0	0
M	0	4	4	3	4	10	4	5	0	0

k=2

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K	1	1	2	2	1	8	9	10	0	0
M	0	4	4	3	4	10	4	5	0	0

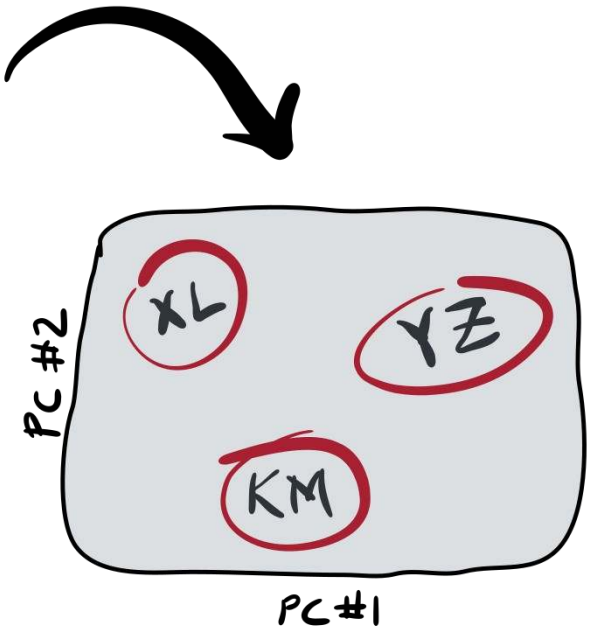
k=2

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K	1	1	2	2	1	8	9	10	0	0
M	0	4	4	3	4	10	4	5	0	0

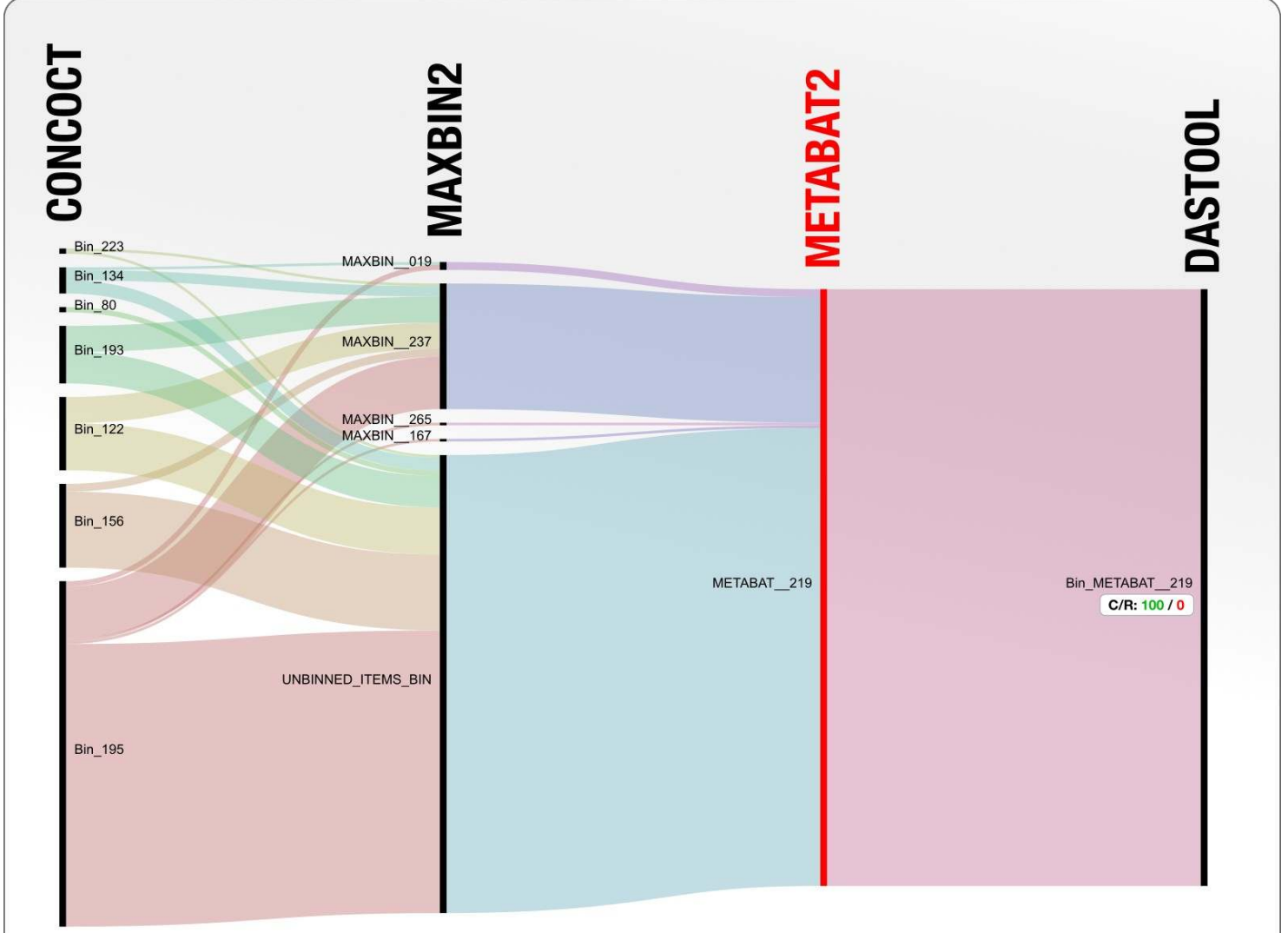


k=2

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K	1	1	2	2	1	8	9	10	0	0
M	0	4	4	3	4	10	4	5	0	0



k=2




```
./anvi-script-gen-alluvial --algorithm DASTOOL --bin Bin_METABAT__219
```

Checking Bin quality: CheckM

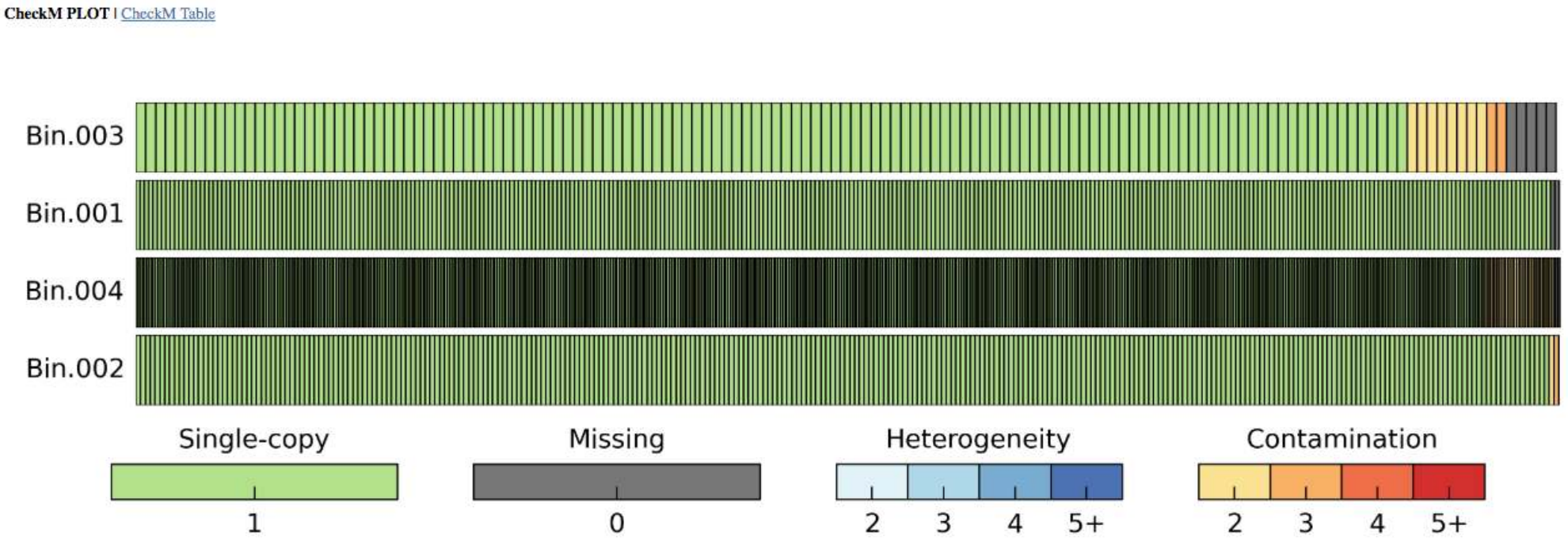
CheckM uses single-copy marker genes to estimate a bin completeness and contamination

Evaluating the quality of obtained MAGs is very important in order to minimize the likelihood of extracting chimeric information

A manual inspection and refinement of the MAGs usually can improve the quality of the obtained bins

Checking Bin quality: CheckM

CheckM uses single-copy marker genes to estimate a bin completeness and contamination



CONCOCT
MAXBIN2
METABAT2
DASTOOL
Surveys CONCOCT, MAXBIN, and METABAT results.
MANUAL
Interactive anvi'o refinement of the original CONCOCT Bin_131
Bin_131
C/R: 100 / 88
MAXBIN_212
C/R: 70 / 16
MAXBIN_214
C/R: 63 / 2
METABAT_99
C/R: 71 / 8
METABAT_140
C/R: 65 / 5
UNBINNED_ITEMS_BIN
Bin_Bin_131_sub
C/R: 98 / 17
Bin_METABAT_99
C/R: 72 / 8
Bin_131_1
C/R: 80 / 9
Bin_131_2
C/R: 94 / 4
Bin 131_1
Bin 131_2

Annotation: assigning functions

Annotation: assigning functions

Annotation of sequences (both for functional and taxonomic assignment) is usually done through sequence similarity

Basically, if a sequence is similar (within a certain pre-selected threshold) to a "known" sequence it gets annotated with the same function (or taxonomic assignment)

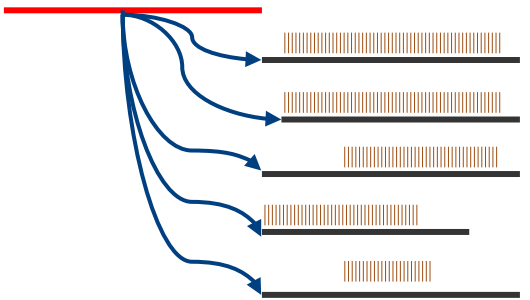
In practice, annotation is complex and can be approached in many different ways, and even similarity-based annotation can have many approaches

Several problem are derived from the functional and taxonomic annotation of sequences

- *Find related or similar sequences by mapping letters of two sequences, with some spacers (indels),*

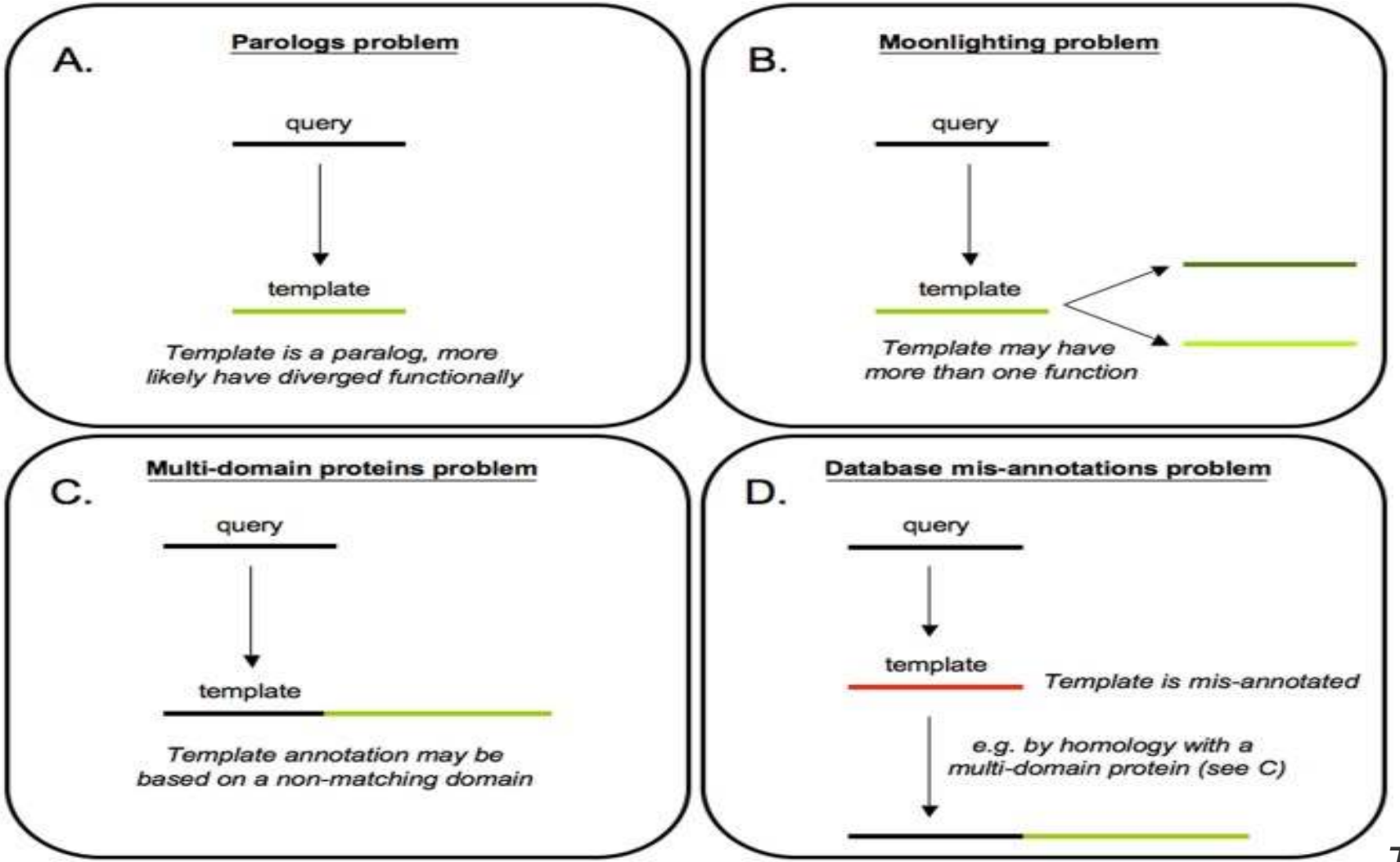
```
76 GGMLKPIEGGTYEVNEAMVEDLKIGVQGPHASNLGGILSNEIAKEIGKRAFIVDPVVVDE 135
   ||||||||||||||||||||||||||  |||| ||||||||||||  ||||||||||||
61 GGMLKPIEGGTYEVNEAMVEDLKIGFEGPHAXNLGGILSNEIAKKLGKRAFIVDPVVVDX 120
```

- *Parse similarities, determine "best hit"*



- *Examples of pairwise search tools - Basic Local Alignment Search Tool or BLAST, LAST, etc.*

Tully, USC



Tully, USC

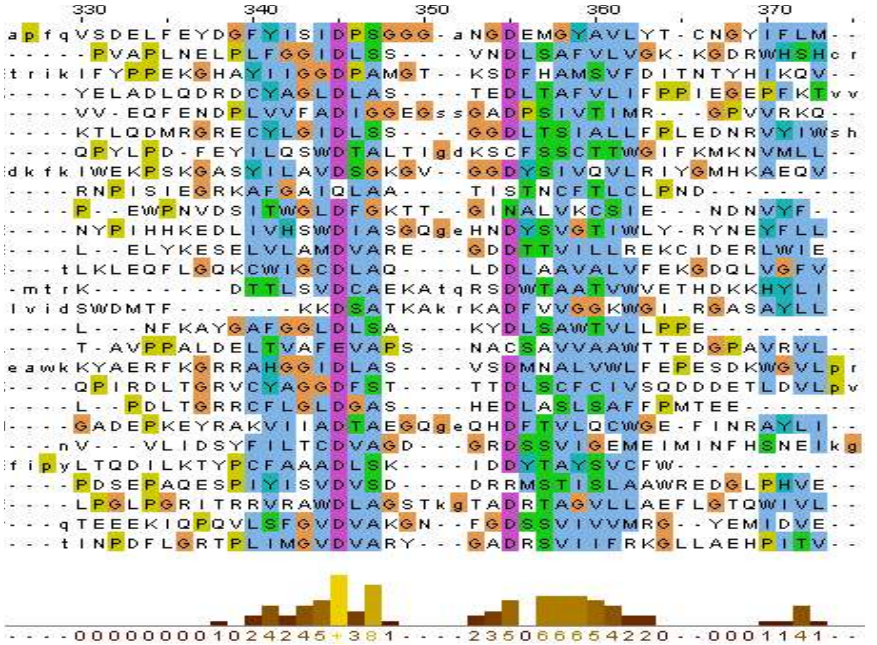
Punta & Ofran. PLOS Comp Biol. 2008

- *"Feature prediction" - location of genes and other features, such as noncoding (e.g., rRNA, tRNA), protein coding sequences (CDS), and more (e.g., regulatory sites, repeats, frameshifts)*
- *Next, an attempt to "name" and interpret function/role by comparing against functional "databases"*
  - *NCBI Guidelines on "names" (~20 rules) e.g.,*
  - *"concise name, not a description or phrase"*
  - *"The protein name should not contain specific characteristics of the protein, (e.g., subcellular location, domain structure…")*
- *Archives of accumulated biological data and knowledge*
  - *Genome, gene sequences, mutations*
  - *Gene regulation, expression, splice variants*
  - *Protein sequence, post-translational modifications*
  - *Protein tertiary structure, localization, networks*
  - *Enzyme kinetics, metabolites, metabolic networks*
  - *E.g., nr, swissprot, pdb, Pfam*

Tully, USC

Other approach to annotation

- Alternative to Blast
- Multiple sequence alignment (MSA) of known sequences – detect “regions of similarity” – build consensus
- Use structural and mechanistic information (catalytic sites)
- Generate profiles using Hidden Markov Models (HMMs) or Position-Specific Scoring Models (PSSMs)
- More sensitive than pairwise – detect distant relationships
- Example of profile/HMM search tools: RPS-BLAST, Hmmer

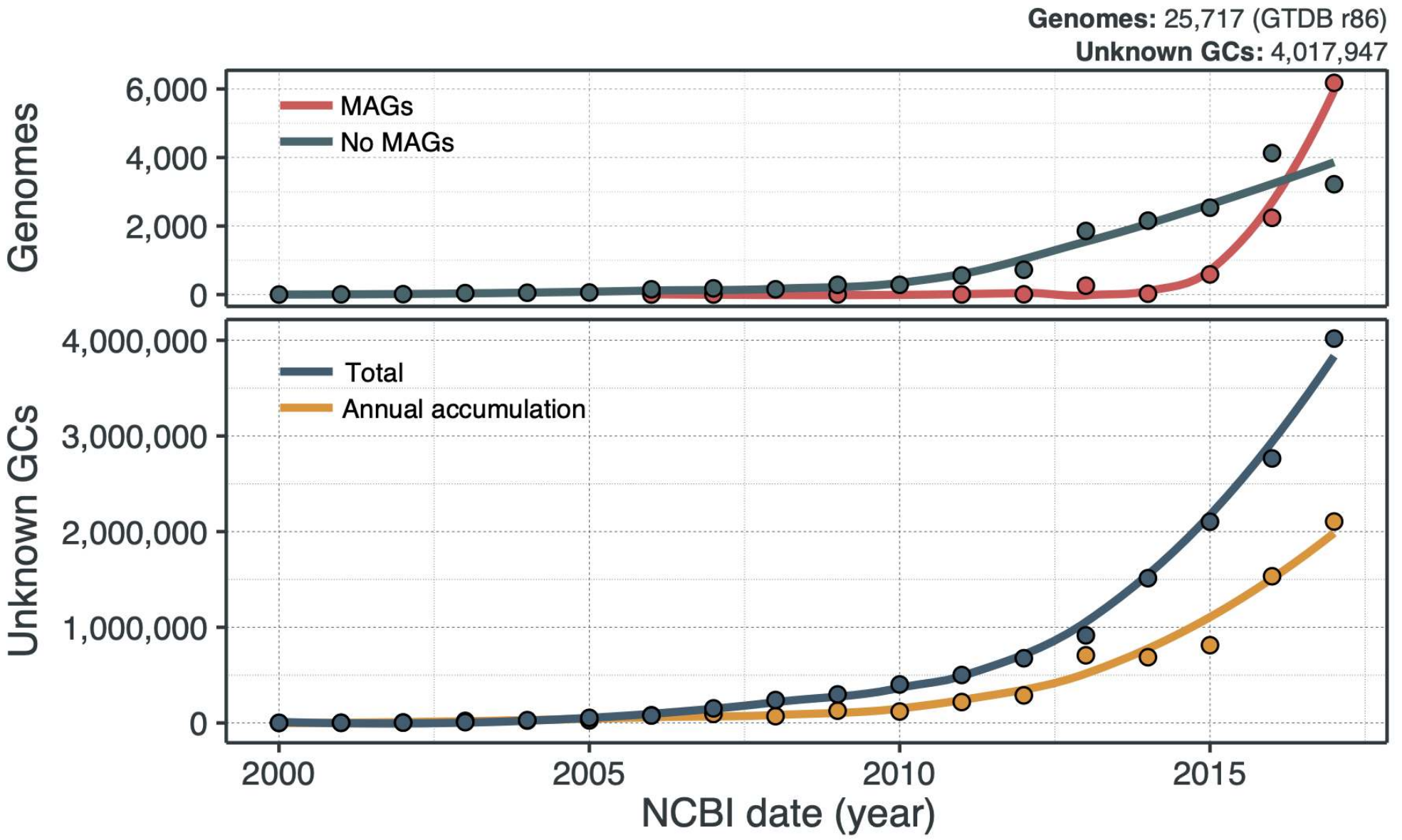


Tully, USC

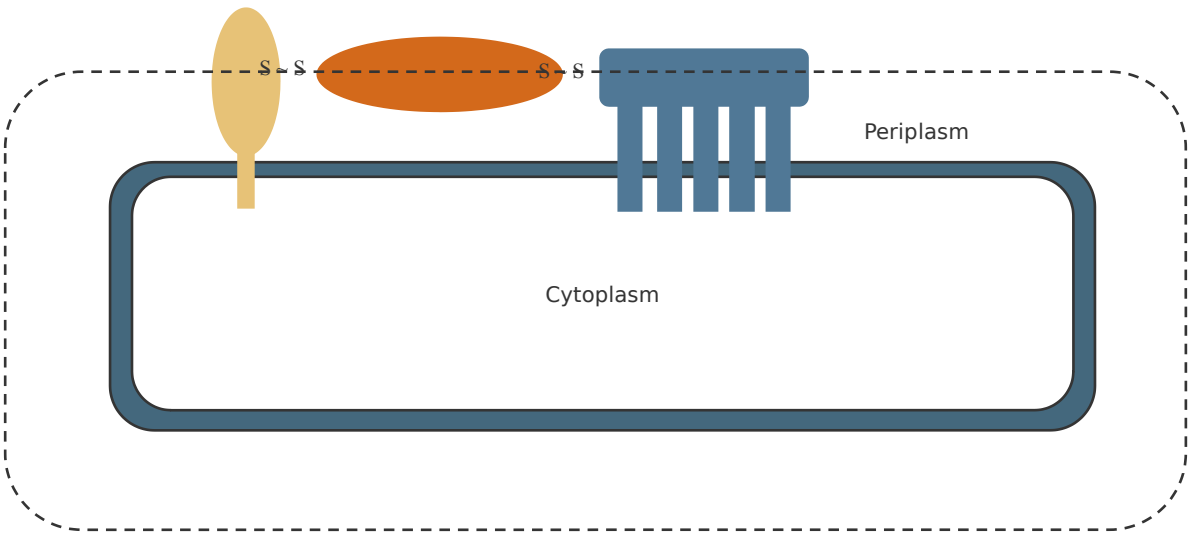
The ONLY way to confirm the function of a coding gene is through biochemical analyses and functional assays. No matter of similar two sequences are there might still be a functional difference.

A large portion of the identified genes codes for hypothetical proteins (proteins with no known function). In the E.coli genome these account for ~30% of the total coding genes. In metagenome this number can be higher.

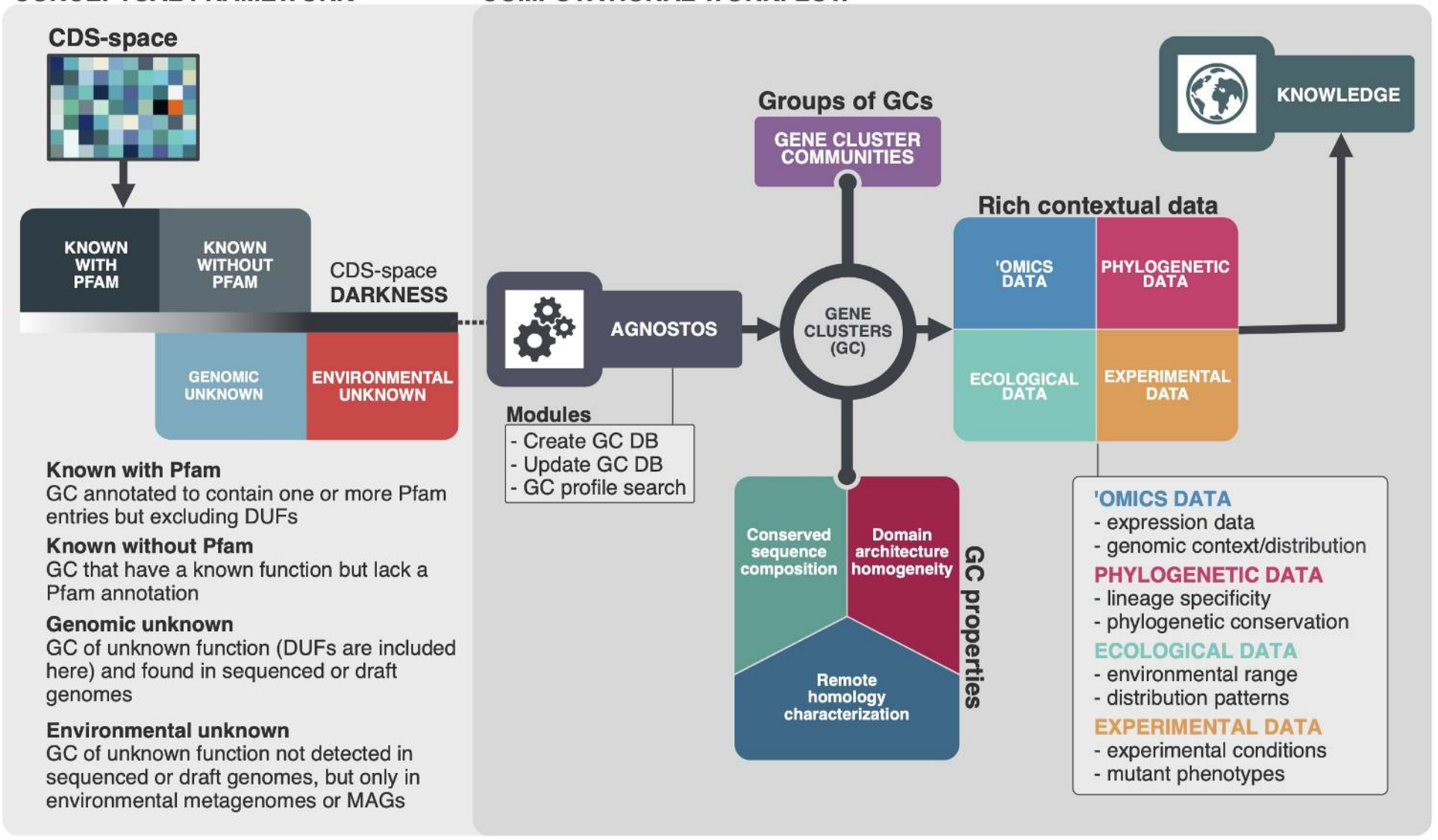
If the gene has never been seen before is usually called and ORFan

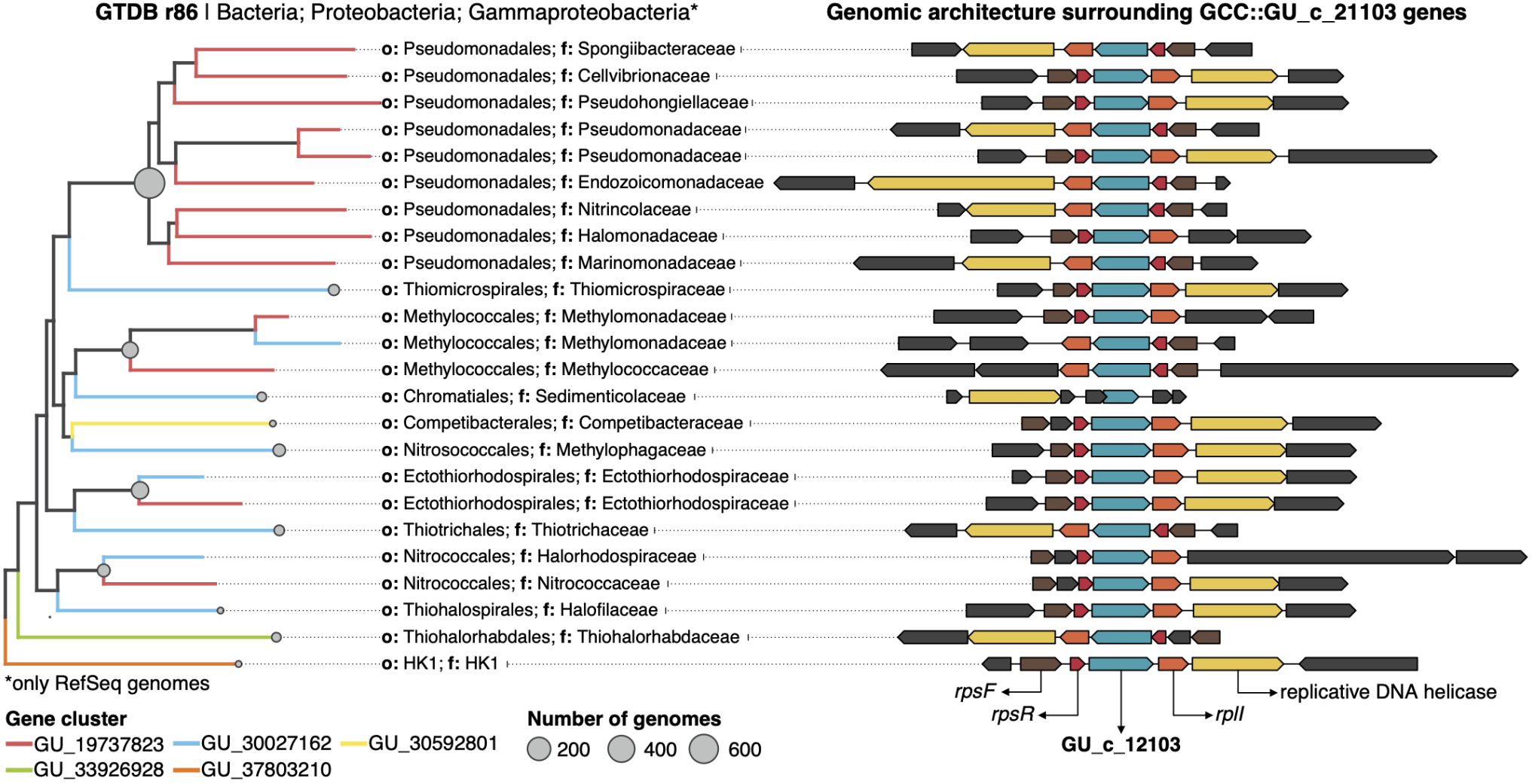

Genomes: 25,717 (GTDB r86)
Unknown GCs: 4,017,947
Genomes
6,000
4,000
2,000
0
MAGs
No MAGs
Unknown GCs
4,000,000
3,000,000
2,000,000
1,000,000
0
Total
Annual accumulation
2000
2005
2010
2015
NCBI date (year)

- *Gene Context*
- *Sub-cellular localization – PSORTb*
- *Topological features*
- *Prediction of binding residues – ex. CAZy: Carbohydrate Binding Motifs*



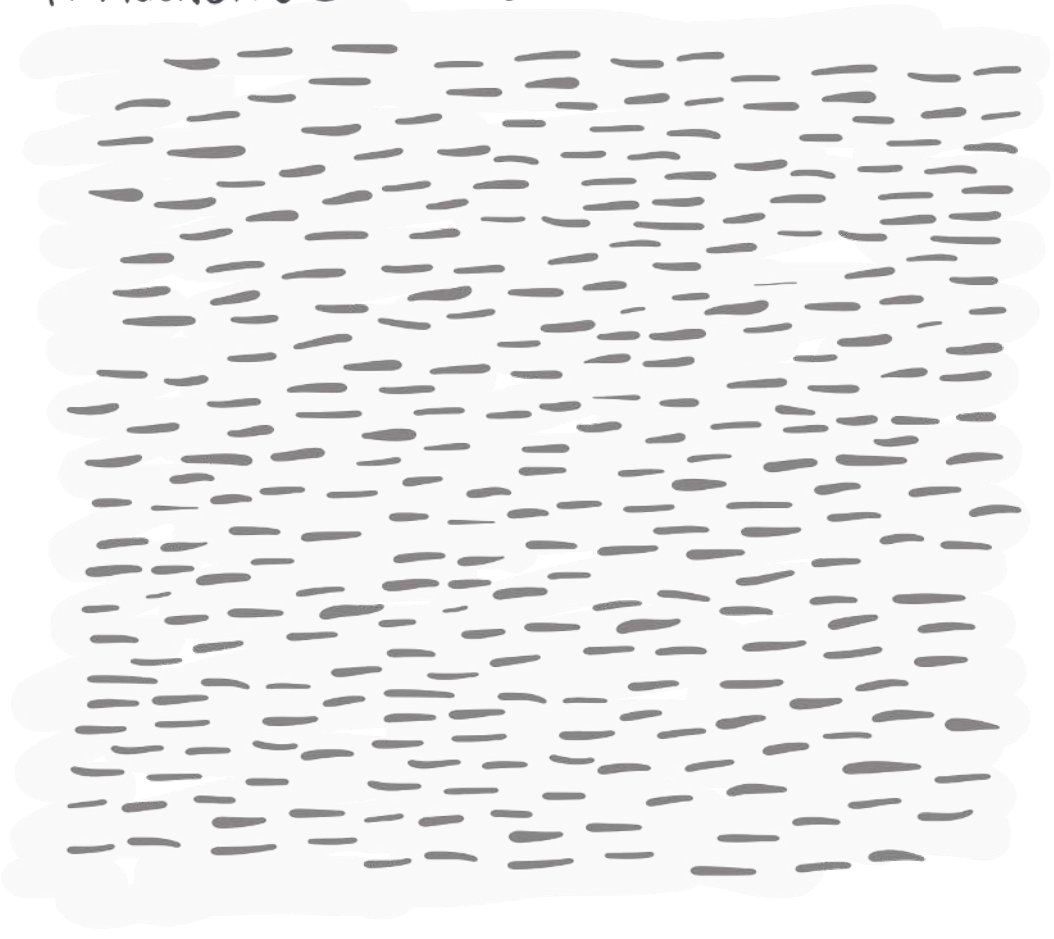
Tully, USC


CONCEPTUAL FRAMEWORK
COMPUTATIONAL WORKFLOW
CDS-space
KNOWN WITH PFAM
KNOWN WITHOUT PFAM
CDS-space DARKNESS
GENOMIC UNKNOWN
ENVIRONMENTAL UNKNOWN
Known with Pfam
GC annotated to contain one or more Pfam entries but excluding DUFs
Known without Pfam
GC that have a known function but lack a Pfam annotation
Genomic unknown
GC of unknown function (DUFs are included here) and found in sequenced or draft genomes
Environmental unknown
GC of unknown function not detected in sequenced or draft genomes, but only in environmental metagenomes or MAGs
AGNOSTOS
Modules
- Create GC DB
- Update GC DB
- GC profile search
Groups of GCs
GENE CLUSTER COMMUNITIES
GENE CLUSTERS (GC)
Conserved sequence composition
Domain architecture homogeneity
Remote homology characterization
GC properties
Rich contextual data
'OMICS DATA
PHYLOGENETIC DATA
ECOLOGICAL DATA
EXPERIMENTAL DATA
KNOWLEDGE
'OMICS DATA
- expression data
- genomic context/distribution
PHYLOGENETIC DATA
- lineage specificity
- phylogenetic conservation
ECOLOGICAL DATA
- environmental range
- distribution patterns
EXPERIMENTAL DATA
- experimental conditions
- mutant phenotypes

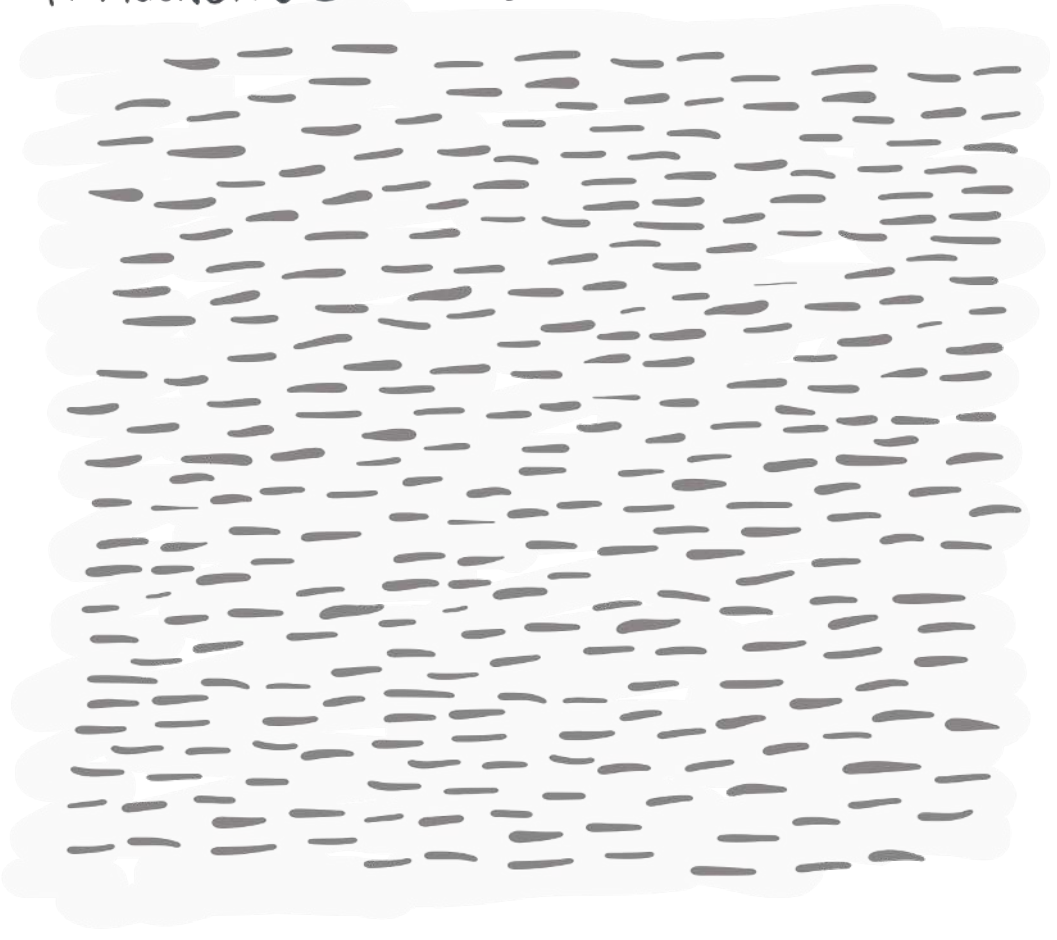

GTDB r86 | Bacteria; Proteobacteria; Gammaproteobacteria*
Genomic architecture surrounding GCC::GU_c_21103 genes
o: Pseudomonadales; f: Spongiibacteraceae
o: Pseudomonadales; f: Cellvibrionaceae
o: Pseudomonadales; f: Pseudohongiellaceae
o: Pseudomonadales; f: Pseudomonadaceae
o: Pseudomonadales; f: Pseudomonadaceae
o: Pseudomonadales; f: Endozoicomonadaceae
o: Pseudomonadales; f: Nitrincolaceae
o: Pseudomonadales; f: Halomonadaceae
o: Pseudomonadales; f: Marinomonadaceae
o: Thiomicrospirales; f: Thiomicrospiraceae
o: Methylococcales; f: Methylomonadaceae
o: Methylococcales; f: Methylomonadaceae
o: Methylococcales; f: Methylococcaceae
o: Chromatiales; f: Sedimenticolaceae
o: Competibacterales; f: Competibacteraceae
o: Nitrosococcales; f: Methylophagaceae
o: Ectothiorhodospirales; f: Ectothiorhodospiraceae
o: Ectothiorhodospirales; f: Ectothiorhodospiraceae
o: Thiotrichales; f: Thiotrichaceae
o: Nitrococcales; f: Halorhodospiraceae
o: Nitrococcales; f: Nitrococcaceae
o: Thiohalospirales; f: Halofilaceae
o: Thiohalorhabdales; f: Thiohalorhabdaceae
o: HK1; f: HK1
*only RefSeq genomes
Gene cluster
GU_19737823
GU_30027162
GU_30592801
GU_33926928
GU_37803210
Number of genomes
200
400
600
rpsF
rpsR
rplI
replicative DNA helicase
GU_c_12103

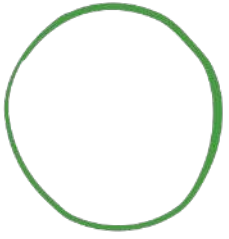
Read mapping: quantify functions

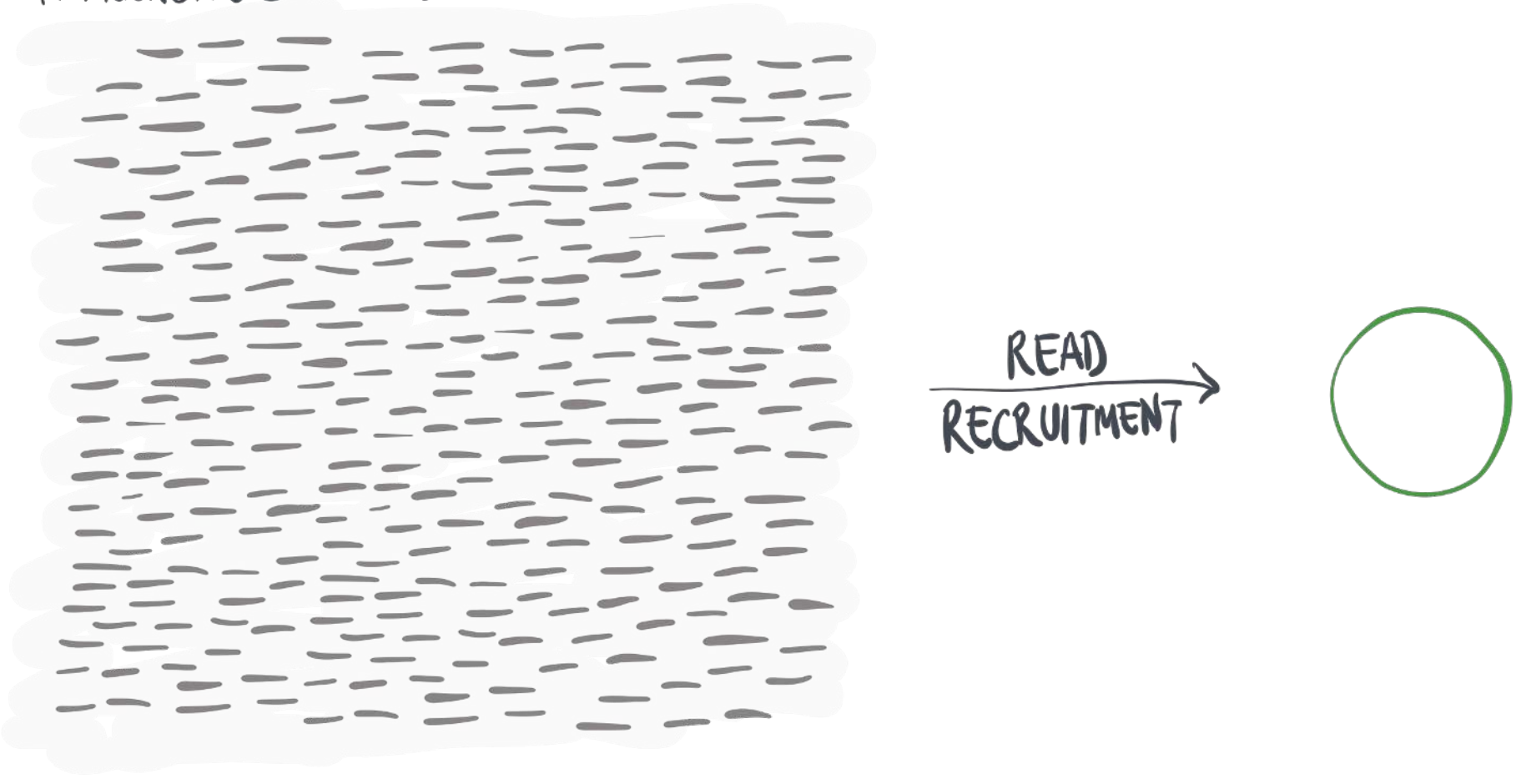
METAGENOMIC SHORT READS

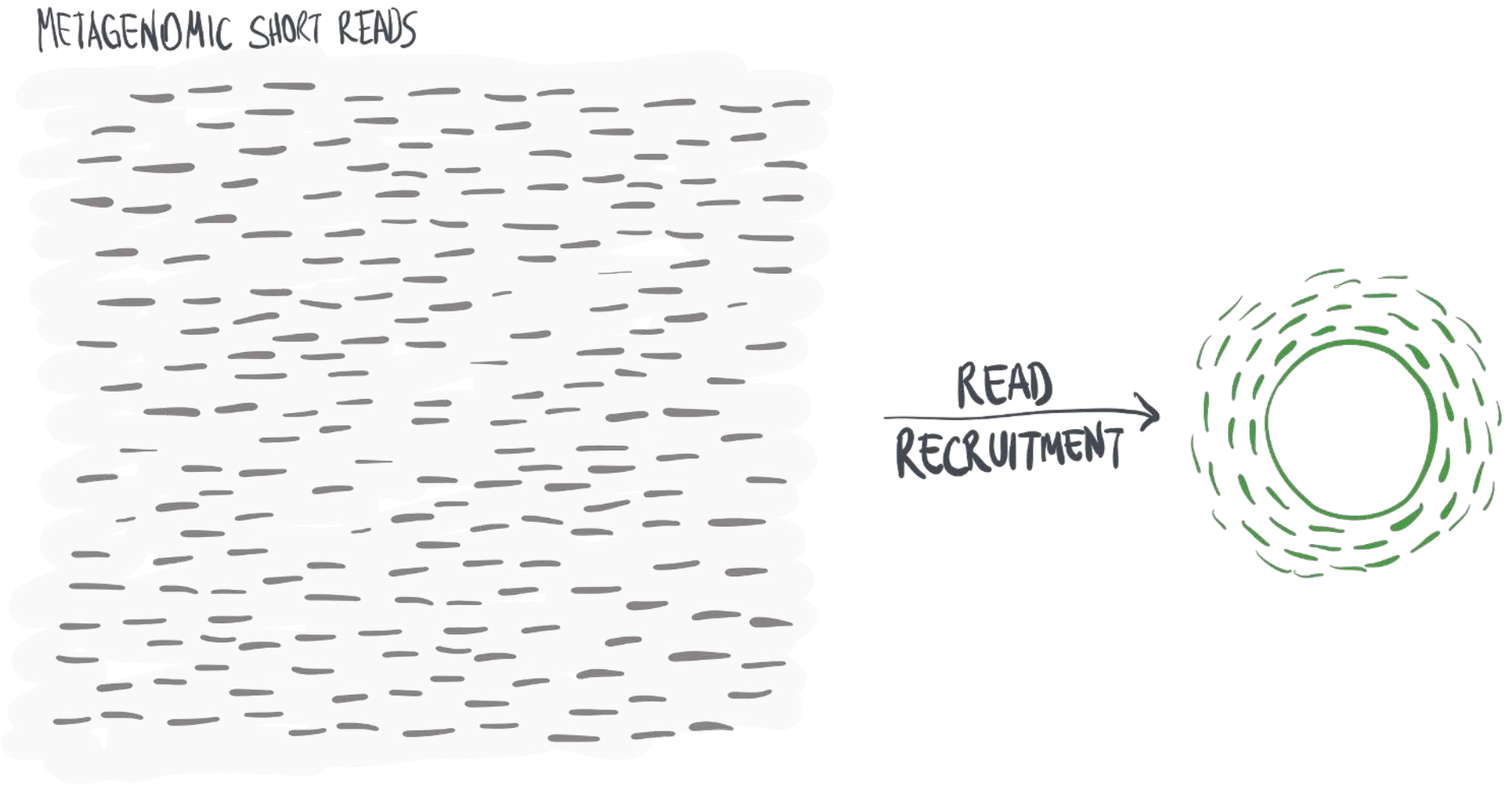


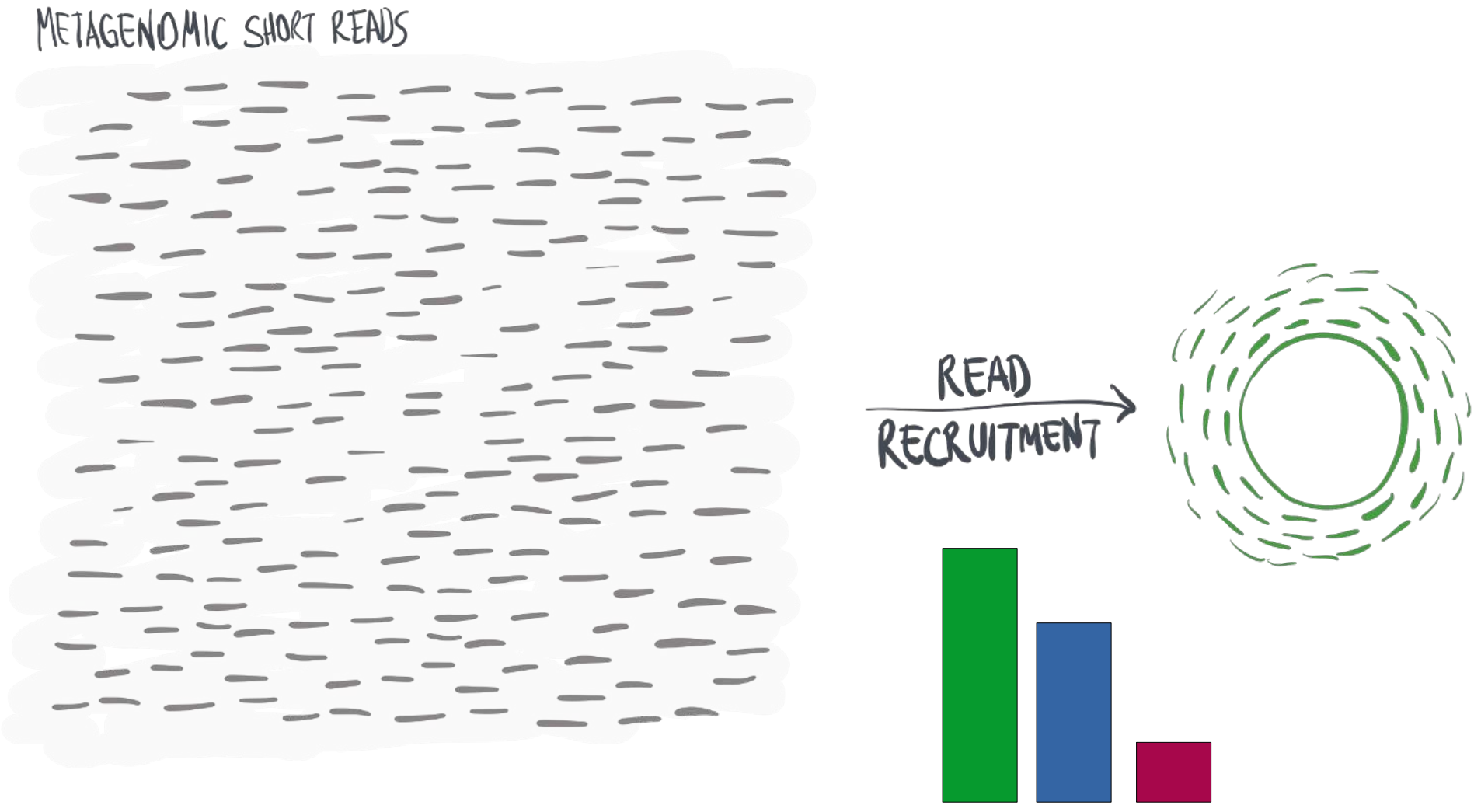
METAGENOMIC SHORT READS





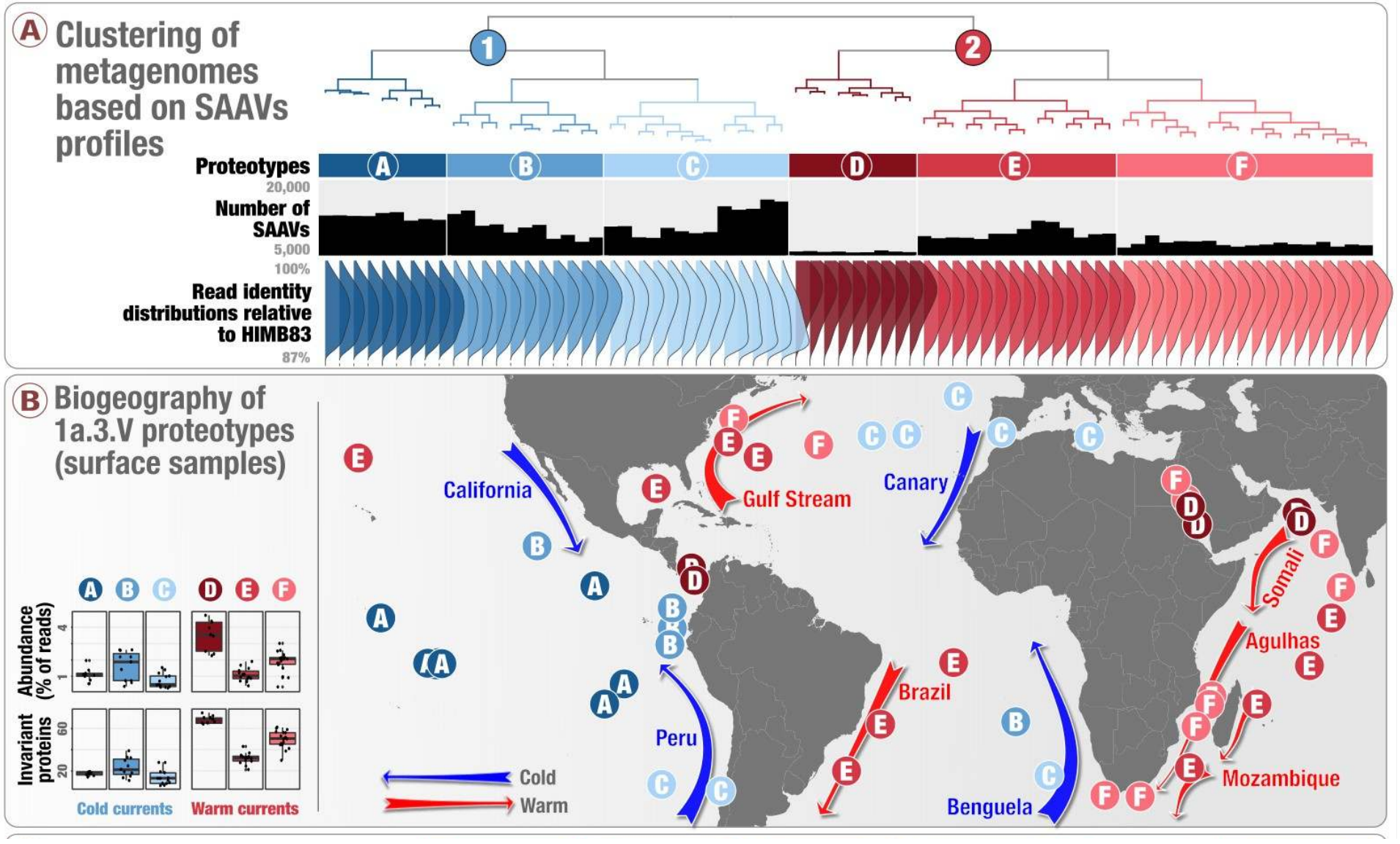


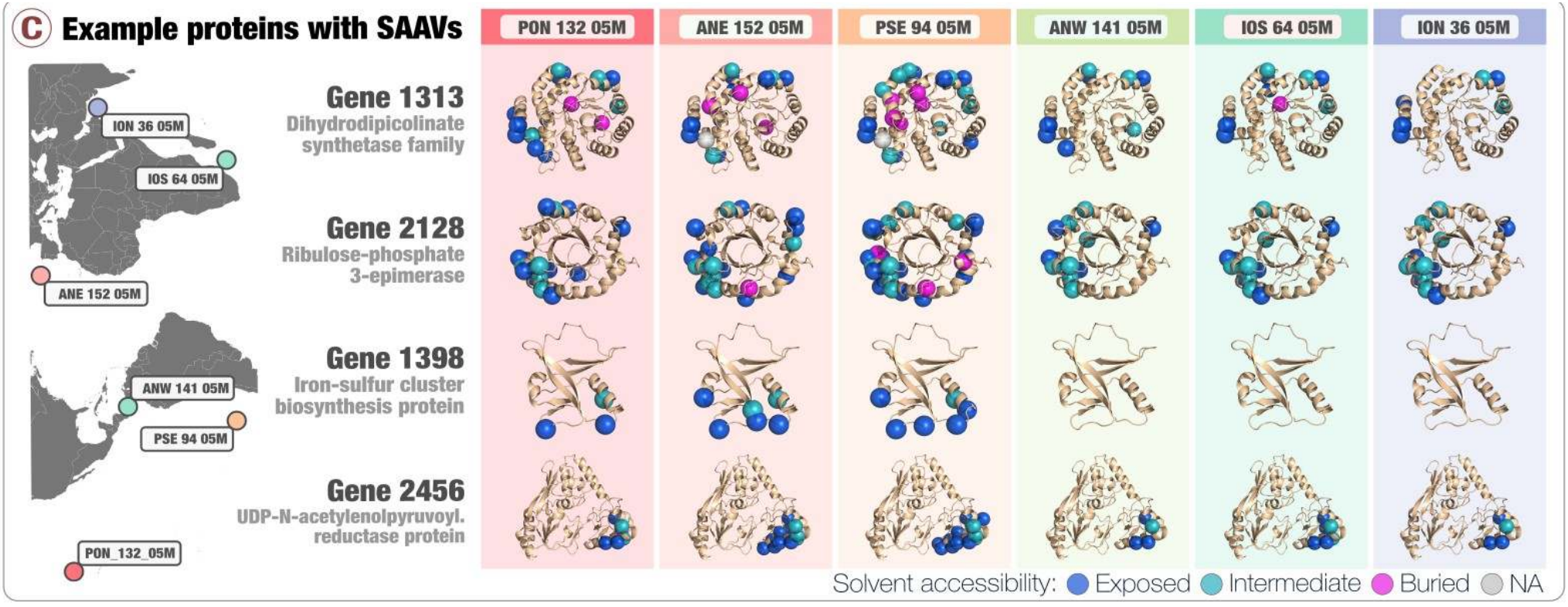




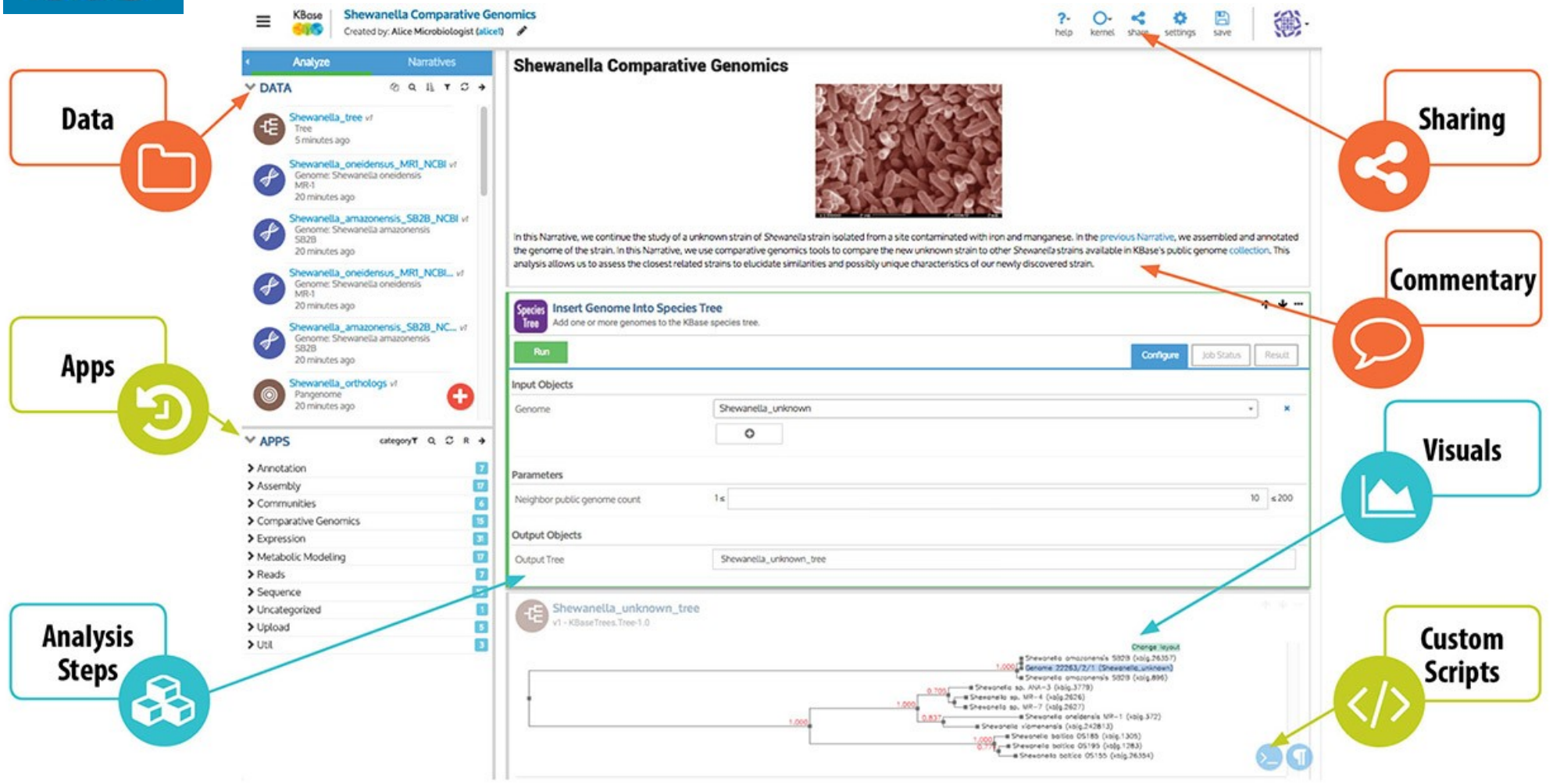
Results

- *A count table with the abundance of specific functions*
- *A count table with the abundance of specific MAGs*
- *The reconstruction of specific metabolic pathways*
- *Functional genes to be used for phylogenetic analyses*
- *Cluster of genes to be analyzed for their variability, adaptations or structural variations*


A Clustering of metagenomes based on SAAVs profiles
1
2
Proteotypes
A
B
C
D
E
F
20,000
Number of SAAVs
5,000
100%
Read identity distributions relative to HIMB83
87%
B Biogeography of 1a.3.V proteotypes (surface samples)
Abundance (% of reads)
4
1
Invariant proteins
60
20
Cold currents
Warm currents
California
Gulf Stream
Canary
Somali
Agulhas
Brazil
Peru
Benguela
Mozambique
Cold
Warm


C Example proteins with SAAVs
PON 132 05M
ANE 152 05M
PSE 94 05M
ANW 141 05M
IOS 64 05M
ION 36 05M
ION 36 05M
IOS 64 05M
ANE 152 05M
ANW 141 05M
PSE 94 05M
PON_132_05M
Gene 1313
Dihydrodipicolinate synthetase family
Gene 2128
Ribulose-phosphate 3-epimerase
Gene 1398
Iron-sulfur cluster biosynthesis protein
Gene 2456
UDP-N-acetylenolpyruvoyl. reductase protein
Solvent accessibility: Exposed Intermediate Buried NA

KBase - https://www.kbase.us/



***Practical**: read-based metagenomic analaysis*

Practical: *read-based metagenomic analysis*

- *Google Scholar (https://scholar.google.com/) – To find the relevant paper*
- *NCBI SRA (https://www.ncbi.nlm.nih.gov/sra) – To retrieve the accession number and sequences*
- *Kaiju (http://kaiju.binf.ku.dk/) – To taxonomically annotate the reads*
- *mi-faser (https://services.bromberglab.org/mifaser/) – To functionally annotate the reads*
- *KEGG Mapper (https://www.genome.jp/kegg/tool/map_pathway2.html) – To explore the metabolic pathways*
- *Excel or Google Sheet – To manipulate the results for plotting and statistical testing*
- *RawGraphs (https://app.rawgraphs.io) – To plot the results for visualization*
- *Powerpoint or Google Slides – To present the results*

Web server - Submit job

Please use the form to upload the sequencing file(s).
Once uploading is completed, press the Submit button at the bottom of the
Only submit one data set at a time.

Job Name

You can give a custom name to your submission.

e-mail

Receive a notification after your submission has been processed. [?]

File with sequencing reads *

Nucleotide sequences must be in compressed FASTA or FASTQ format [?]

Select file File name:
Start upload Progress:

☐ Upload a second file for paired-end sequencing

Options

Reference Database

- ○ RefSeq Genomes - proteins from completely assembled RefSeq genomes: Bacteria, A
- ○ proGenomes - proteins from the representative genomes in proGenomes: Bacteria, Arc
- ○ NCBI BLAST *nr* - non-redundant protein database: Bacteria, Archaea, Viruses
- ◉ NCBI BLAST *nr* +euk - as above, but also including fungi and microbial eukaryotes.

SEG filter

☑ Filter low complexity protein query sequences

Run mode

- ○ MEM - for maximum exact matches.
- ◉ Greedy - allows mismatches.

Minimum match length	11

Only applicable for Greedy mode:

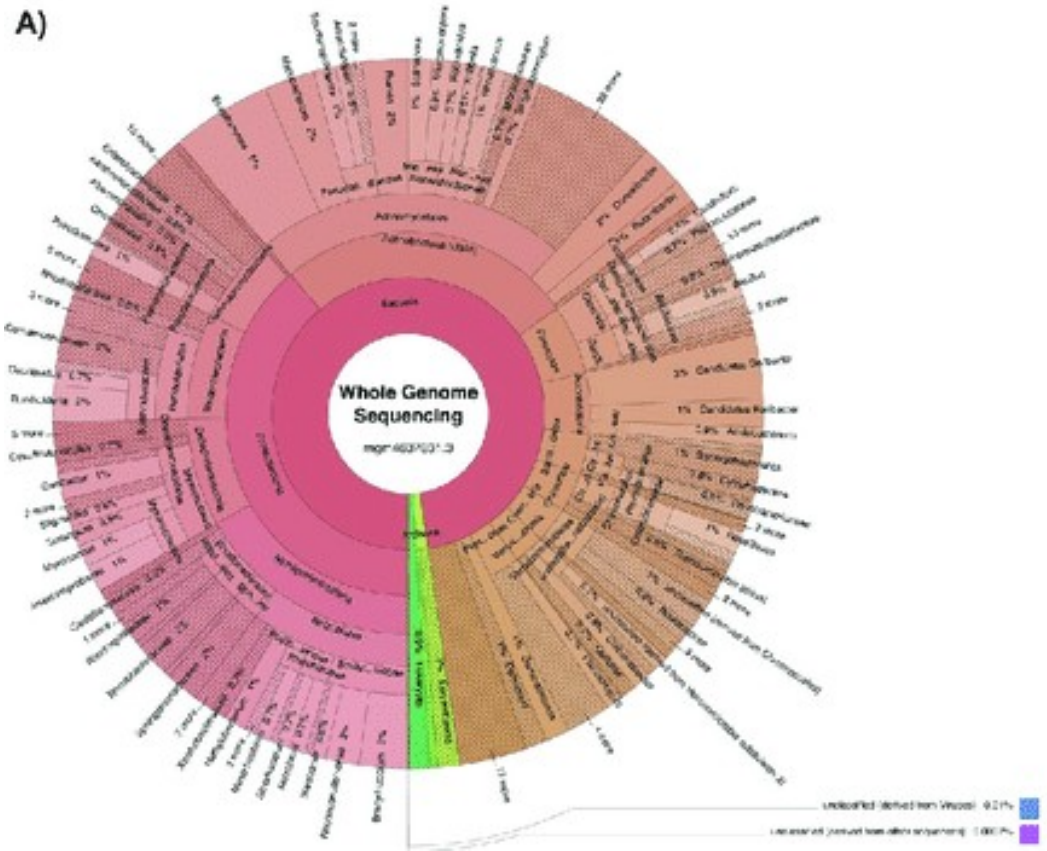
Minimum match score	75
Allowed mismatches	5

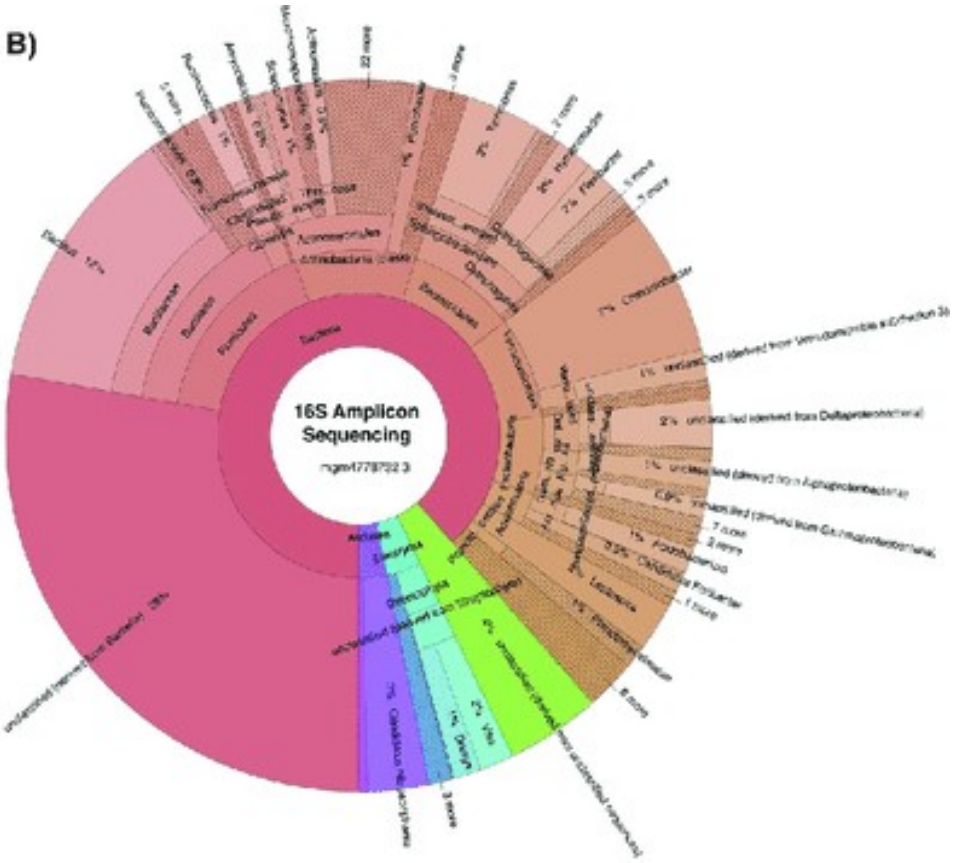
Submit

http://kaiju.binf.ku.dk/server

KAIJU

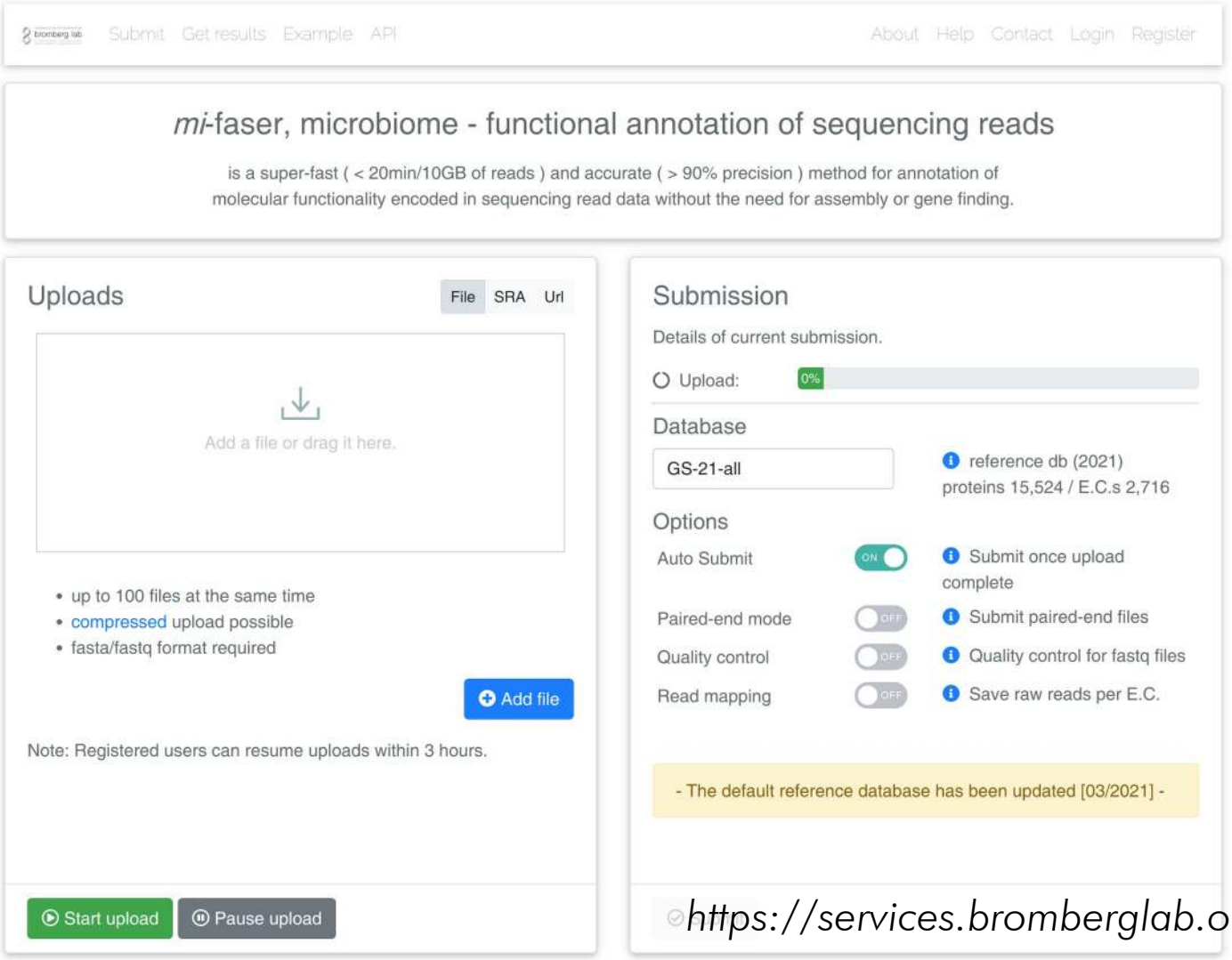
Krona





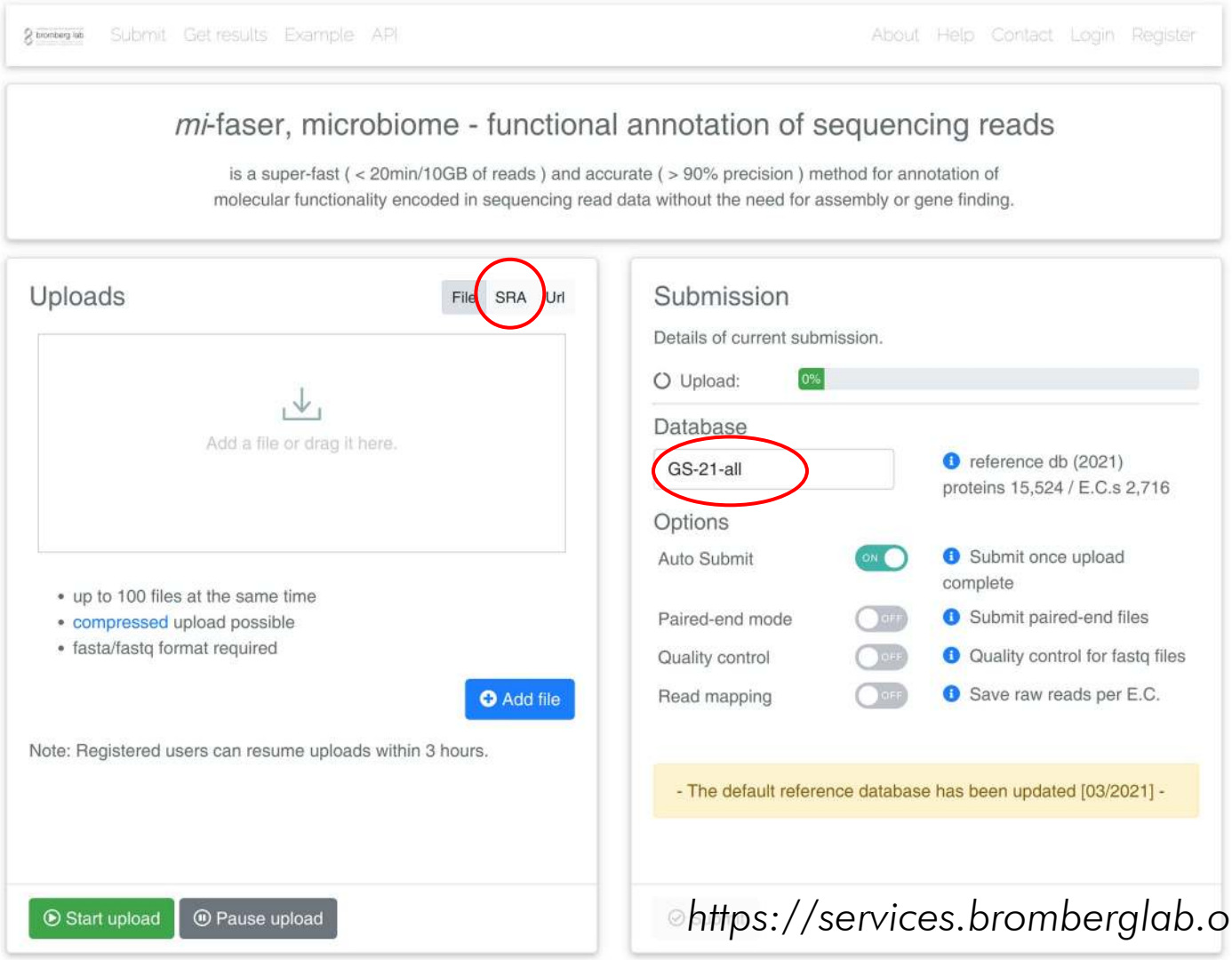
https://github.com/marbl/Krona/wiki

mi-faser



https://services.bromberglab.org/mifaser/submit

mi-faser



https://services.bromberglab.org/mifaser/submit

Submission *uYEHN9Ls2DI9KGNJ*

Summary statistics

Number of unique functions found:	746	
Reads that map unambiguously to a function:	310231	(96%)
Reads that map to multiple functions ( ):	11489	(4%)

Annotations | Multiple Mappings | Read Mappings | Quality Control

Submission details

label	CV88_metag
# reads	9537740
created	2020-04-17 09:39:42
runtime	00h : 10m : 50s
qa control	00h : 02m : 04s
file name	CV88_metag.fastq.gz
file size	1988MB
database	GS+
user	donato.giovannelli@gmail.com
your jobs	CV88_metag.fastq.gz

Annotated functions

regex Search

10 | Select all | Deselect all | Select Search | Show all | Filter selected | Download filtered (csv)

Enzyme E.C.#	Annotation	Read Count	Color
2.7.7.6	DNA-directed RNA polymerase	14011	
5.99.1.3	DNA topoisomerase (ATP-hydrolysing)	7867	
3.6.3.14	H^+-transporting two-sector ATPase	7675	
3.6.4.13	RNA helicase	6223	
4.2.1.11	phosphopyruvate hydratase	6018	
1.7.99.4	nitrate reductase	4935	
3.6.4.12	DNA helicase	4161	
2.7.2.3	phosphoglycerate kinase	4032	
6.3.4.4	adenylosuccinate synthase	3782	

Submission *uYEHN9Ls2DI9KGNJ*

Summary statistics

Number of unique functions found:	746	
Reads that map unambiguously to a function:	310231	(96%)
Reads that map to multiple functions ( ):	11489	(4%)

Annotations | Multiple Mappings | Read Mappings | Quality Control

Submission details

label	CV88_metag
# reads	9537740
created	2020-04-17 09:39:42
runtime	00h : 10m : 50s
qa control	00h : 02m : 04s
file name	CV88_metag.fastq.gz
file size	1988MB
database	GS+
user	donato.giovannelli@gmail.com
your jobs	CV88_metag.fastq.gz

Annotated functions

regex Search

10 | Select all | Deselect all | Select Search | Show all | Filter selected | Download filtered (csv)

Enzyme E.C.#	Annotation	Read Count	Color
2.7.7.6	DNA-directed RNA polymerase	14011	
5.99.1.3	DNA topoisomerase (ATP-hydrolysing)	7867	
3.6.3.14	H⁺-transporting two-sector ATPase	7675	
3.6.4.13	RNA helicase	6223	
4.2.1.11	phosphopyruvate hydratase	6018	
1.7.99.4	nitrate reductase	4935	
3.6.4.12	DNA helicase	4161	
2.7.2.3	phosphoglycerate kinase	4032	
6.3.4.4	adenylosuccinate synthase	3782	

KEGG Mapper – Search&Color Pathway

- About KEGG Mapper
- Reconstruct Pathway (and Brite, Module)
- Search Pathway (and Brite, Module, Network, Disease)
- Search&Color Pathway (and Brite, Module)
- Color Pathway Join Brite
- Convert ID Annotate Sequence
- BlastKOALA Map Taxonomy KEGG

Target databases: Pathway, Brite hierarchy, Module

Search mode: ○ Reference
● Organism-specific hsa Enter: org , ko, ec, rn, hsadd
Optional use of outside ID: Select

Enter objects one per line followed by bgcolor, fgcolor:

Examples: Select

Or upload file:

Choose File No file chosen

If necessary, change default bgcolor: pink

☑ Include aliases

☐ Use uncolored diagrams

☐ Search pathways containing all the objects (AND search)

Exec Clear

https://www.genome.jp/kegg/

KEGG Mapper Search Result

Pathway (130) | Brite (0) | Module (212)

Sort by the pathway list

Show matched objects

map01100 Metabolic pathways (473)

map01110 Biosynthesis of secondary metabolites (188)

map01120 Microbial metabolism in diverse environments (141)

map00520 Amino sugar and nucleotide sugar metabolism (49)

map00230 Purine metabolism (35)

map00541 O-Antigen nucleotide sugar biosynthesis (33)

map00270 Cysteine and methionine metabolism (27)

map00010 Glycolysis / Gluconeogenesis (27)

map00620 Pyruvate metabolism (26)

map00720 Carbon fixation pathways in prokaryotes (25)

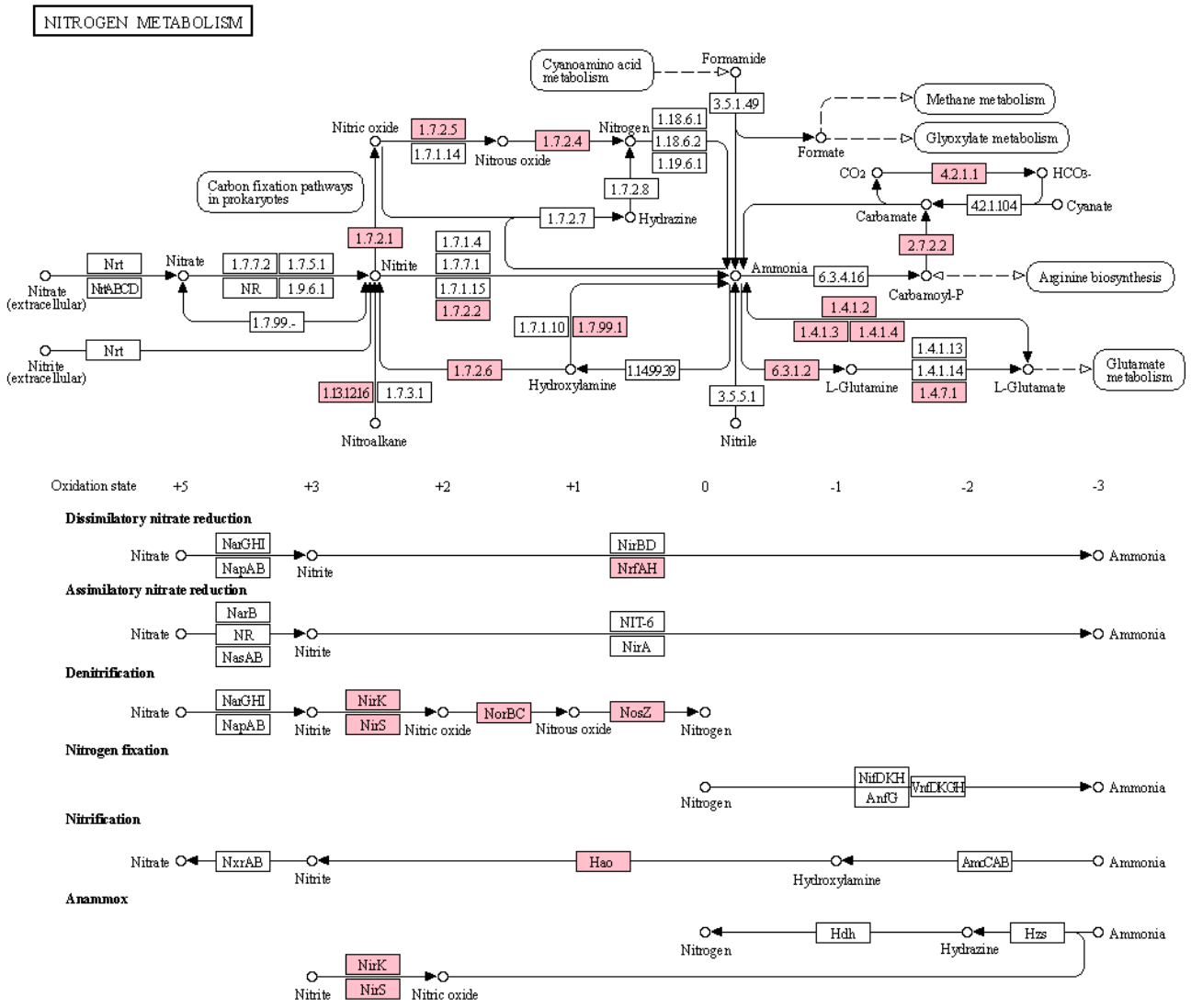
map00240 Pyrimidine metabolism (24)

map00250 Alanine, aspartate and glutamate metabolism (23)

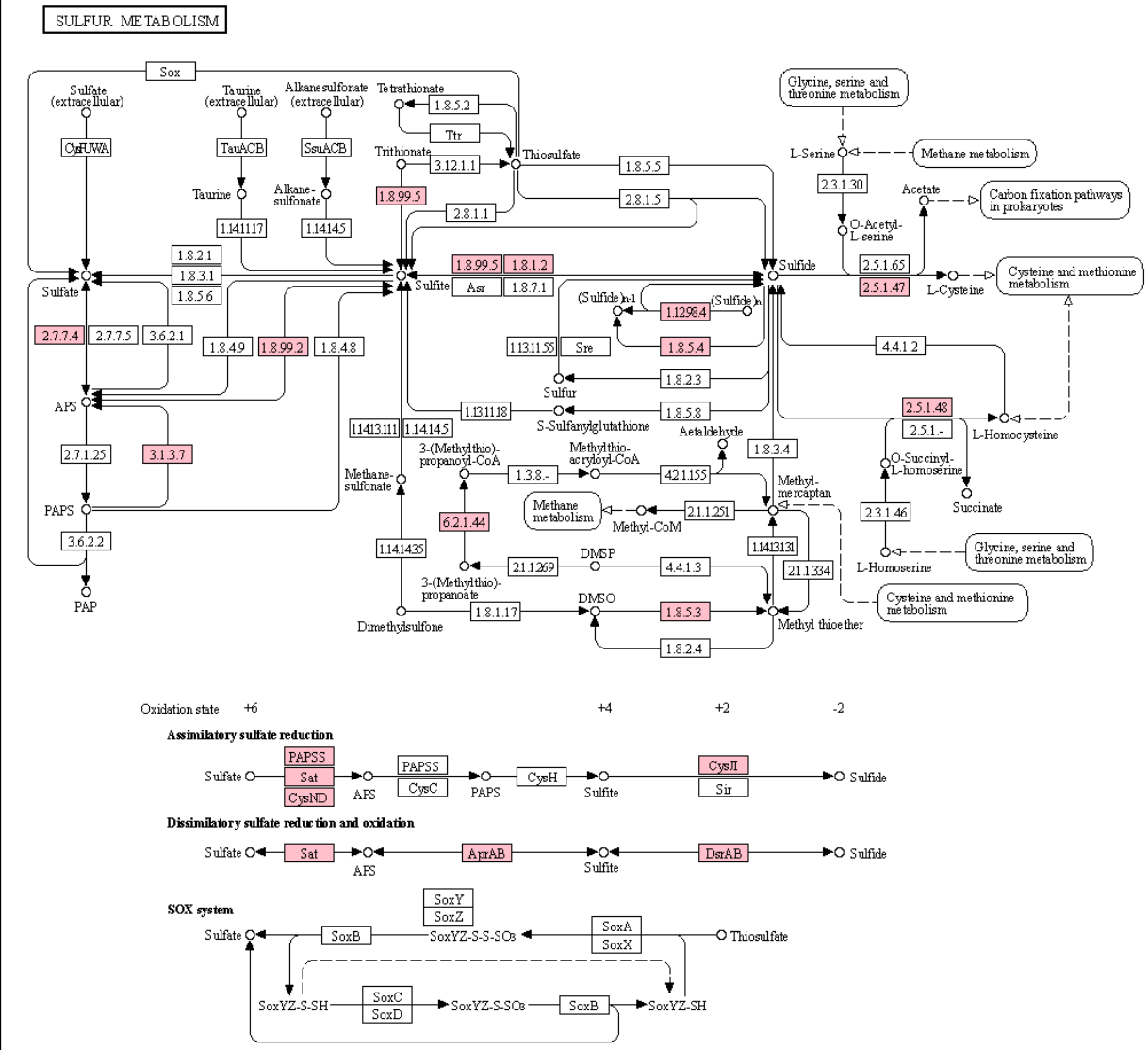
map00630 Glyoxylate and dicarboxylate metabolism (23)

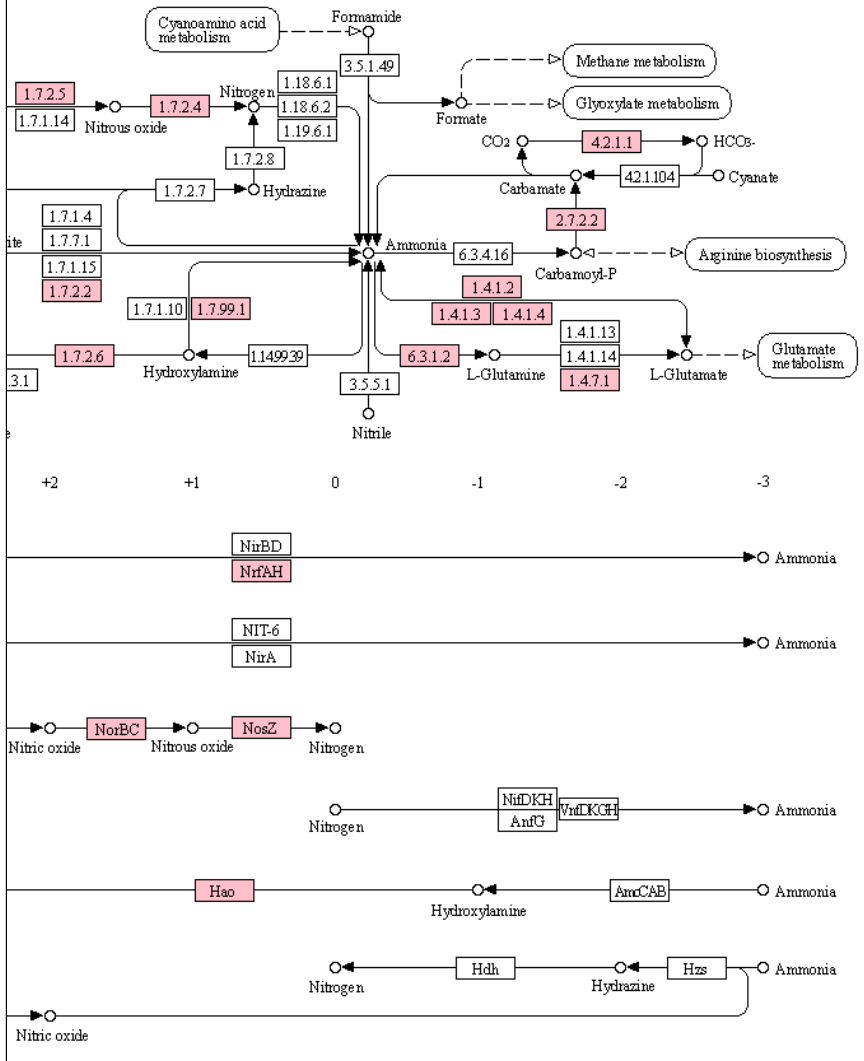
map00680 Methane metabolism (21)

map00260 Glycine, serine and threonine metabolism (20)

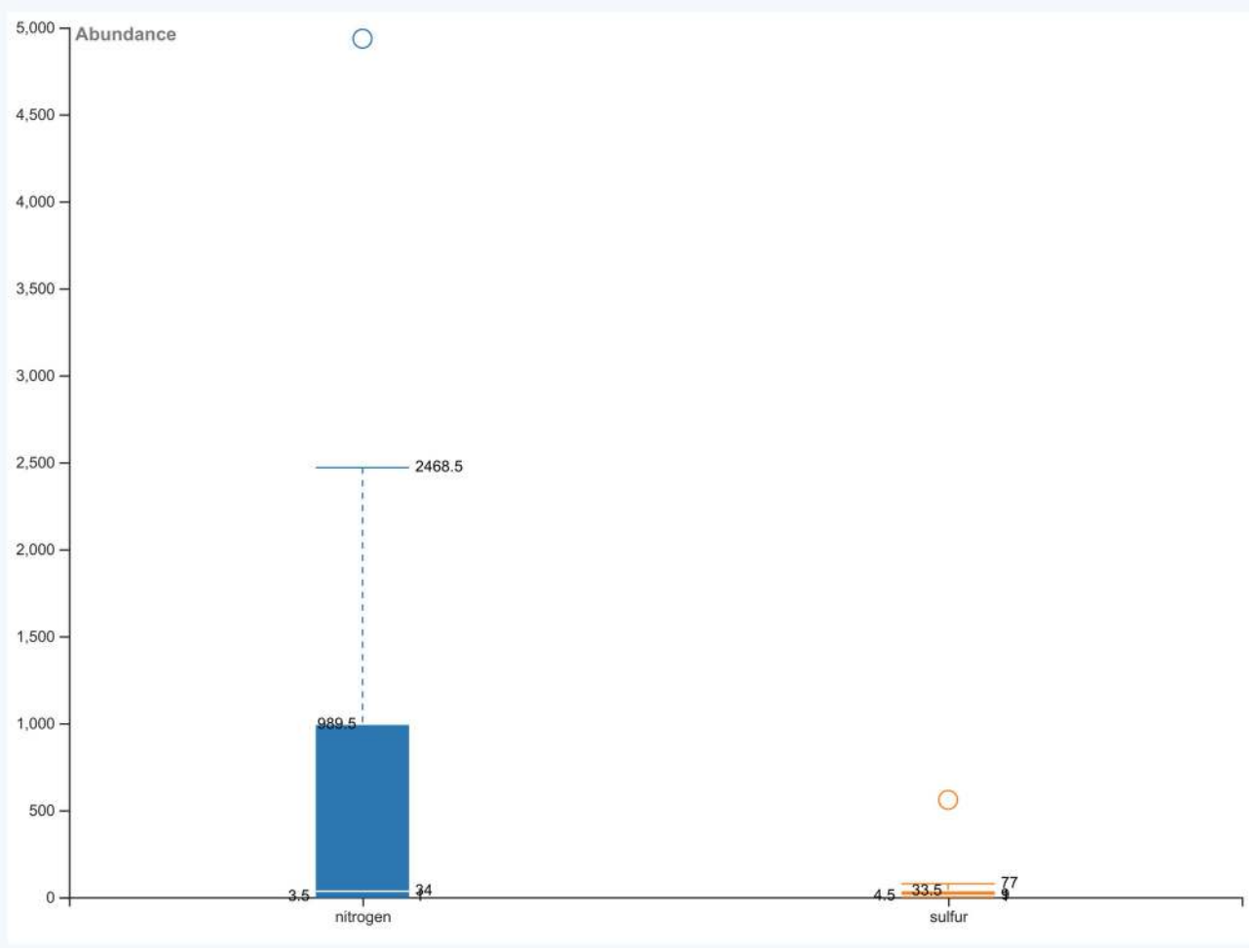


00910 1/9/20

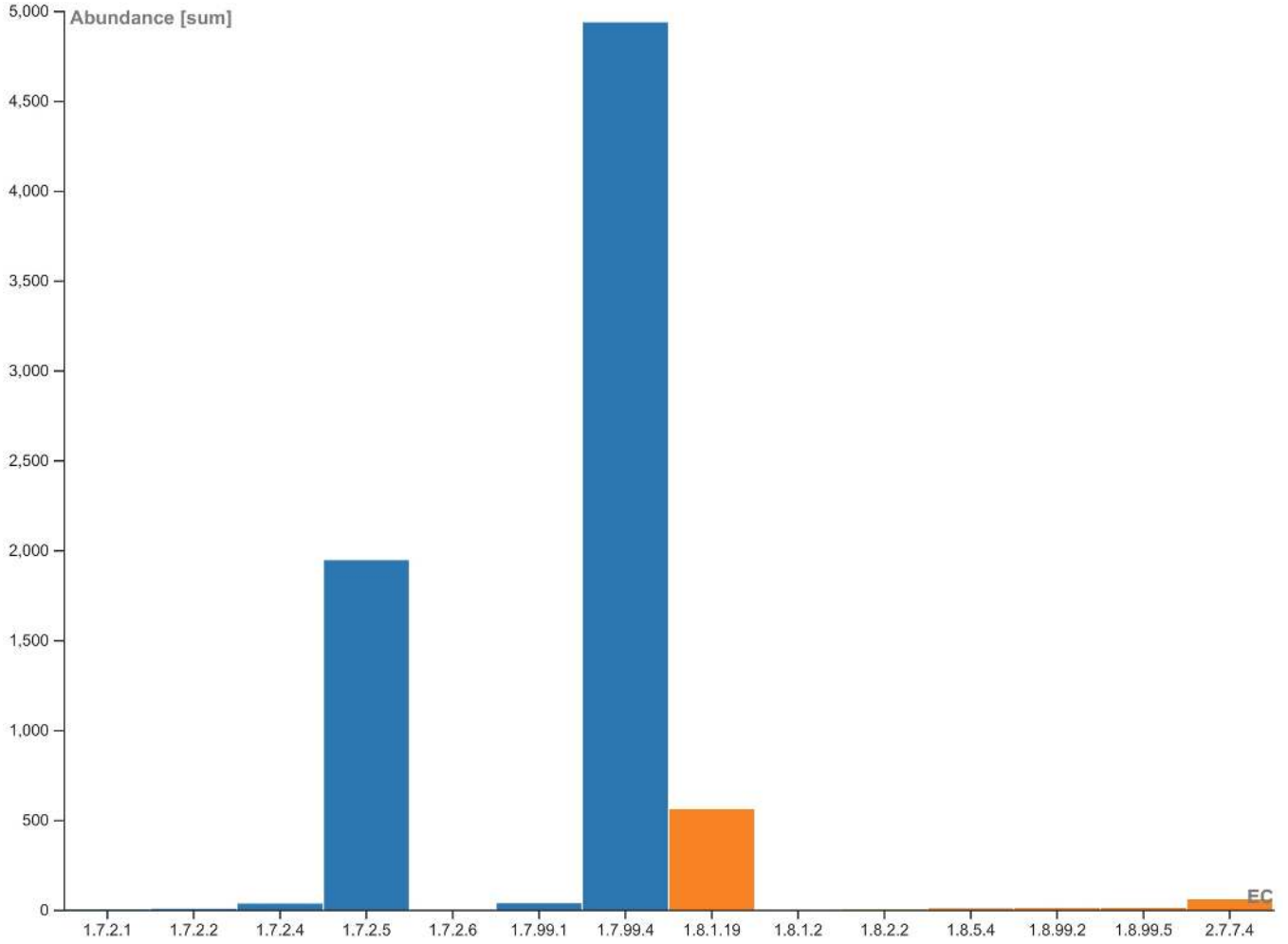

SULFUR METABOLISM
Sox
Sulfate (extracellular)
Taurine (extracellular)
Alkanesulfonate (extracellular)
Tetrathionate
CysTUWA
TauABC
SsuABC
Taurine
Alkane-sulfonate
Trithionate
Thiosulfate
Glycine, serine and threonine metabolism
L-Serine
Methane metabolism
Acetate
Carbon fixation pathways in prokaryotes
O-Acetyl-L-serine
Sulfide
Sulfite
Sulfate
Asr
Sre
Cysteine and methionine metabolism
L-Cysteine
APS
PAPS
PAP
Sulfur
S-Sulfanylglutathione
Aetaldehyde
L-Homocysteine
O-Succinyl-L-homoserine
Succinate
Methane-sulfonate
3-(Methylthio)-propanoyl-CoA
Methylthio-acryloyl-CoA
Methyl-mercaptan
Methane metabolism
Methyl-CoM
L-Homoserine
Glycine, serine and threonine metabolism
DMSP
3-(Methylthio)-propanoate
DMSO
Dimethylsulfone
Methyl thioether
Cysteine and methionine metabolism
Oxidation state
+6
+4
+2
-2
Assimilatory sulfate reduction
PAPSS
Sat
CysND
CysC
CysH
CysJI
Sir
Dissimilatory sulfate reduction and oxidation
AprAB
DsrAB
SOX system
SoxY
SoxZ
SoxB
SoxYZ-S-S-SO3
SoxA
SoxX
Thiosulfate
SoxYZ-S-SH
SoxC
SoxD
00920 2/4/20
(c) Kanehisa Laboratories

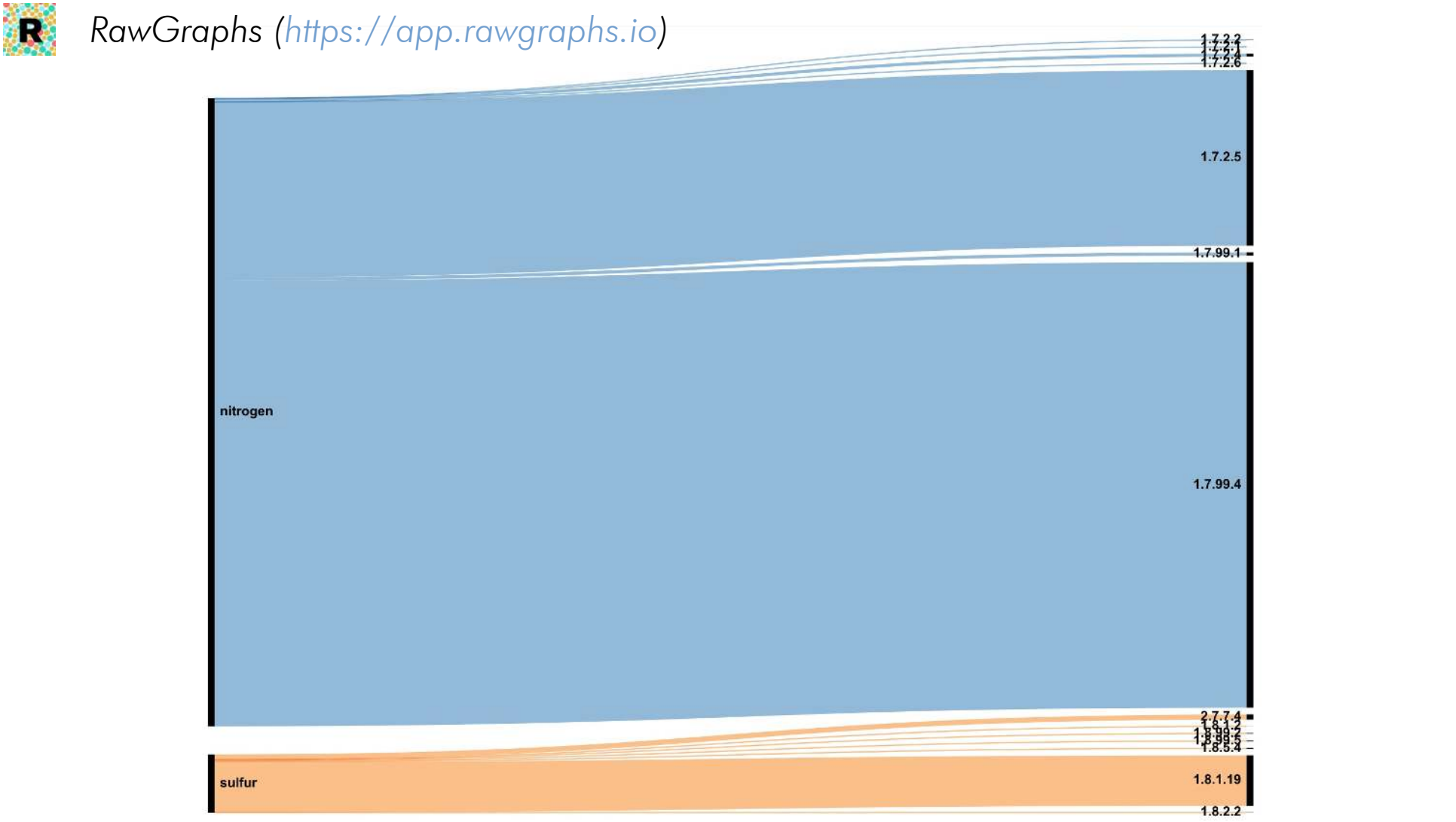

Cyanoamino acid metabolism
Formamide
Nitrogen
Nitrous oxide
Formate
Methane metabolism
Glyoxylate metabolism
CO2
HCO3-
Cyanate
Carbamate
Hydrazine
Ammonia
Carbamoyl-P
Arginine biosynthesis
Hydroxylamine
L-Glutamine
L-Glutamate
Glutamate metabolism
Nitrile
+2
+1
0
-1
-2
-3
NirBD
NrfAH
NIT-6
NirA
Ammonia
Nitric oxide
NorBC
NosZ
Nitrogen
NifDKH
AnfG
VnfDKGH
Hao
AmoCAB
Hdh
Hzs
Nitric oxide

RawGraphs (https://app.rawgraphs.io)



RawGraphs (https://app.rawgraphs.io)




RawGraphs (https://app.rawgraphs.io)
nitrogen
sulfur
1.7.2.5
1.7.99.1
1.7.99.4
1.8.1.19
1.8.2.2