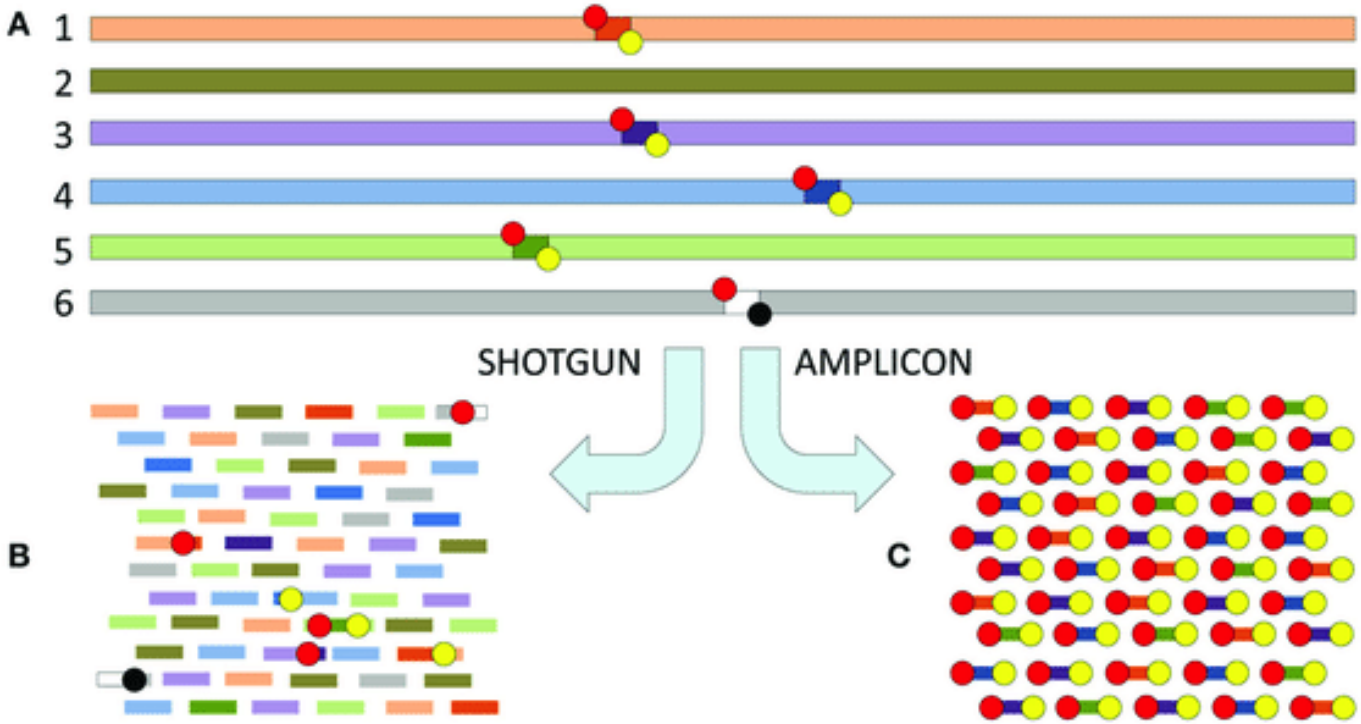
(SHOTGUN) METAGENOMICS

Hélène Touzet

helene.touzet@univ-lille.fr

CNRS, Bonsai, CRIStAL

obtained directly from the samples without culturing microbes in the laboratory



total genomic DNA of a sample
high sequencing depth

amplicon/targeted/16S rRNA

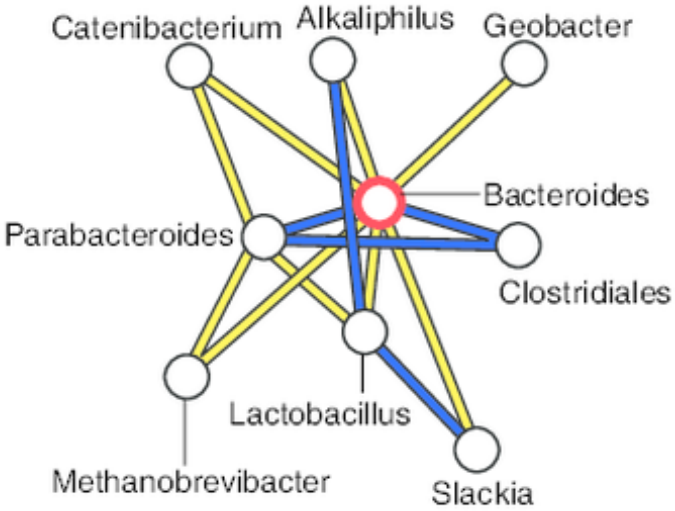
Project MetaHIT (2008-2012)

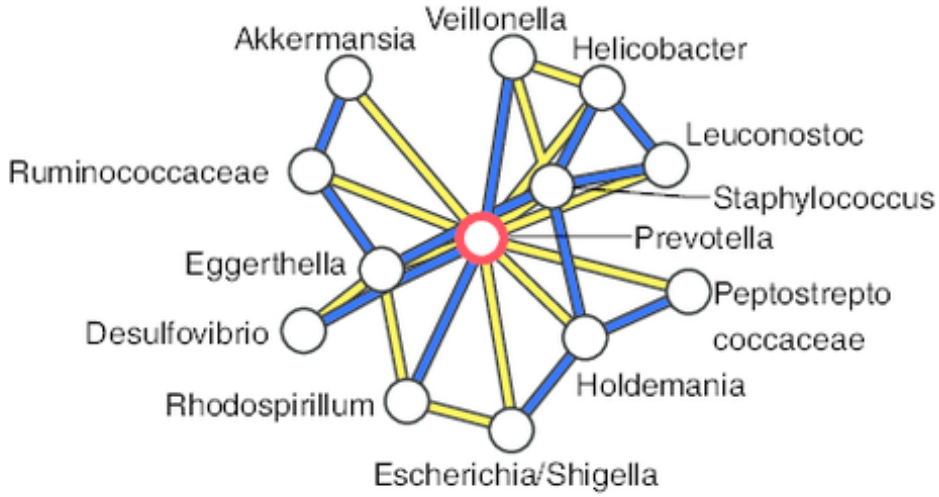
METAgenomics of the Human Intestinal Tract

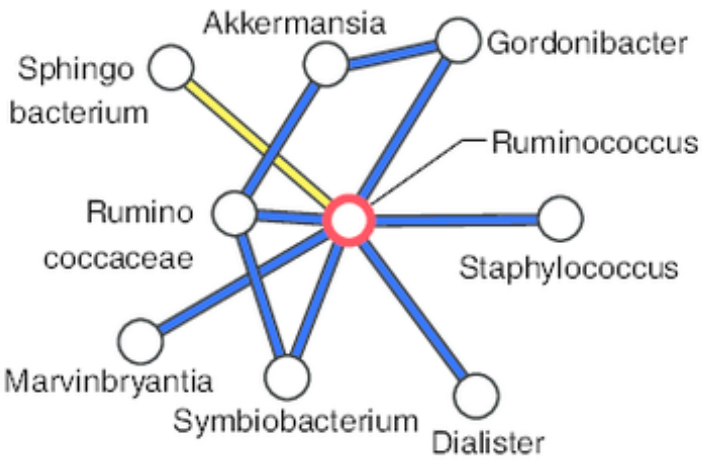
- 124 individuals

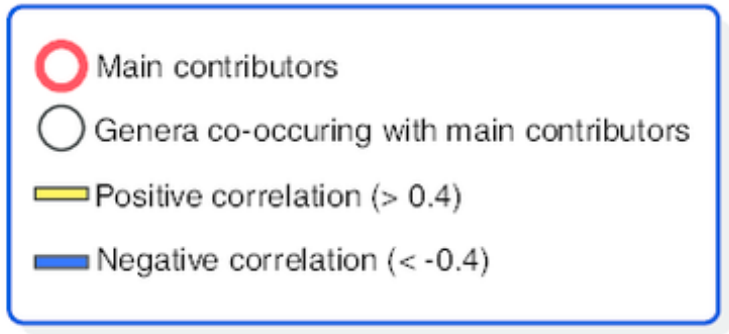
  healthy, overweight and obese individual human adults, as well as inflammatory bowel disease (IBD)

- sequencing of stool samples $\rightarrow$ 540 Gb of DNA
- 3 million different genes
- a person carries, on average, 540000 genes, a value that corresponds to some 160 species



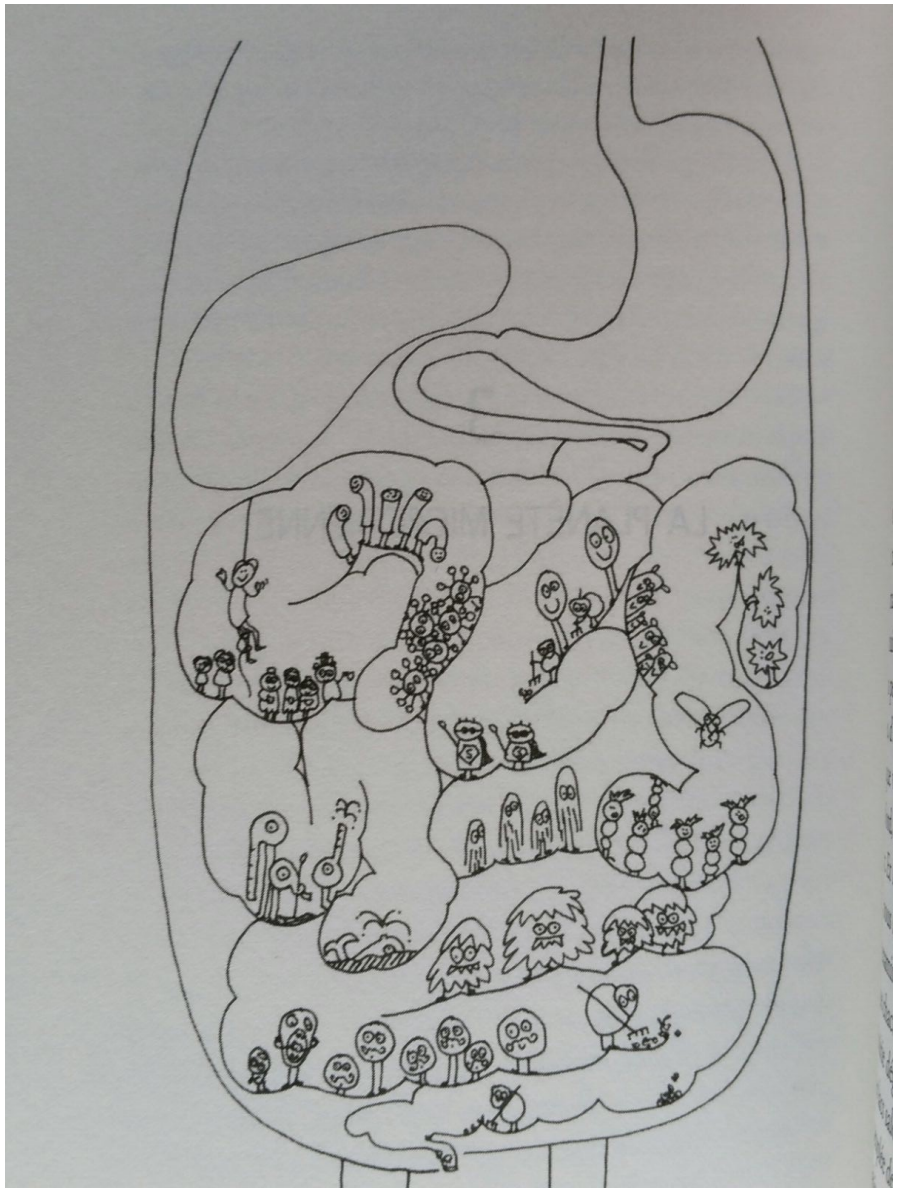






- type 1 : high levels of Bacteroides
- type 2 : few Bacteroides but Prevotella are common
- type 3 : high levels of Ruminococcus


Giulia
Enders
LE CHARME
DISCRET
DE
L'INTESTIN
TOUT
SUR UN ORGANE
MAL AIMÉ...
ACTES SUD




N'AYANT PU ME CORROMPRE
ILS M'ONT ASSASSINE.

Historical sample

- Sample : Jean-Paul Marat, blood stain from the newspaper *L'Ami du peuple*

- DNA sequencing : HiSeq 4000, paired-end

  568,623,176 reads in total

  74,244,610 reads mapped to the human reference genome
  ancestry analysis

  494,378,566 other reads

  among them 9,788,947 quality controlled and cleaned reads
  metagenomic analysis

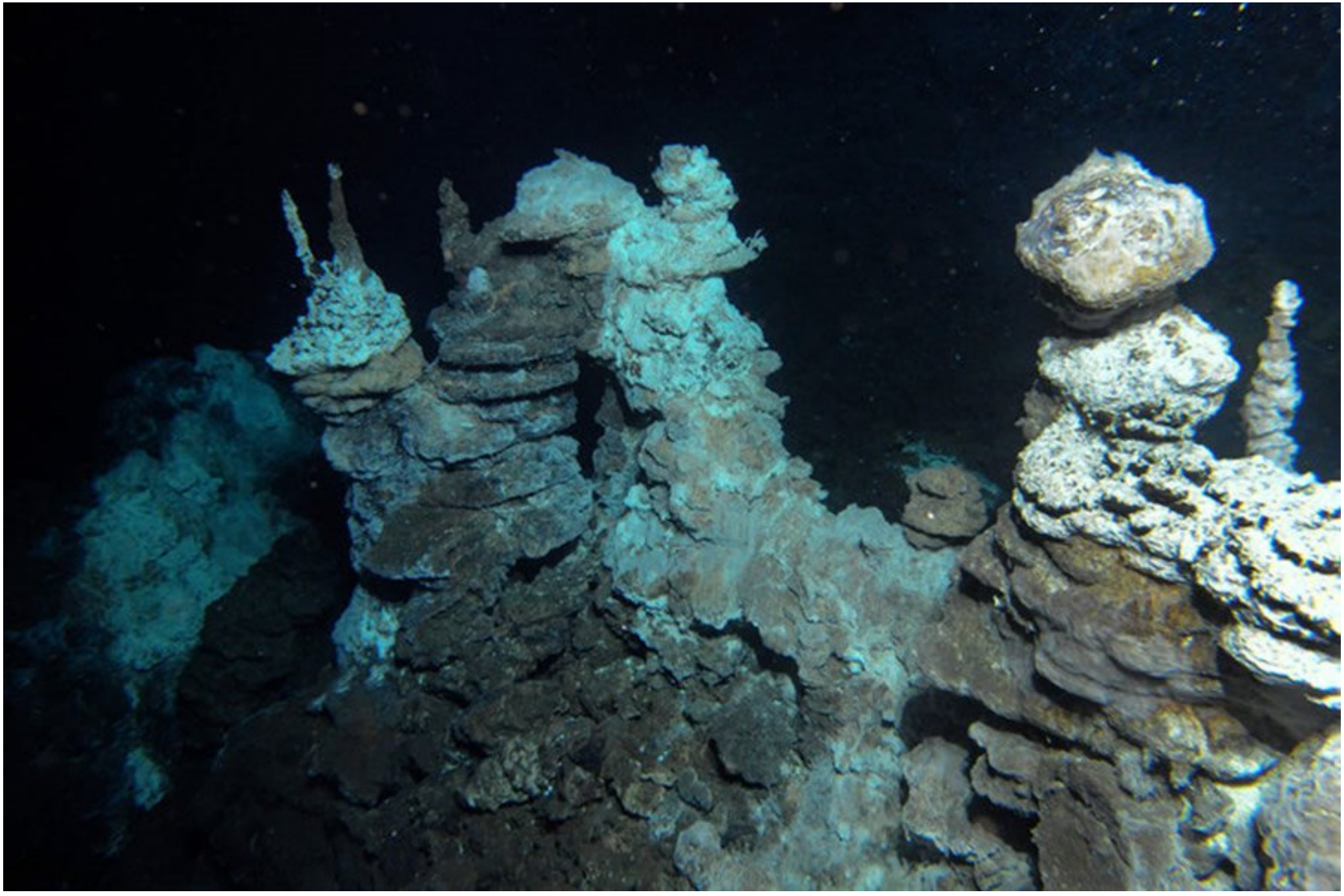
Bioinformatics analysis

Alignment of reads against database of bacterial genomes

Disease	Pathogen	Blood	Unstained paper
Syphilis	*Treponema pallidum*	✗	✗
Scrofula (tuberculosis)	*Mycobacterium tuberculosis*[1]	✗	✗
Leprosy	*Mycobacterium leprae*	✗	✗
Diabetic candidiasis (thrush)	*Candida* sp.	✗	✗
Scabies	*Sarcoptes scabiei*	✗	✗
Seborrheic dermatitis	*Malassezia* sp.	✓✓	✓
Atopic eczema	*Staphylococcus aureus*	✓	✗
Severe acneiform eruptions	*Cutibacterium acnes*	✓✓✓	✓✓

Marat may have suffered from a primary fungal infection (seborrheic dermatitis), superinfected with bacterial opportunistic pathogens

Metagenomic analysis of a blood stain from the French revolutionary Jean-Paul Marat (1743-1793)
`https://www.biorxiv.org/content/10.1101/825034v1.full`
See also (in French) `https://www.lemonde.fr/blog/realitesbiomedicales/2019/11/08/des-biologistes-moleculaires-font-parler-le-sang-du-revolutionnaire-marat`

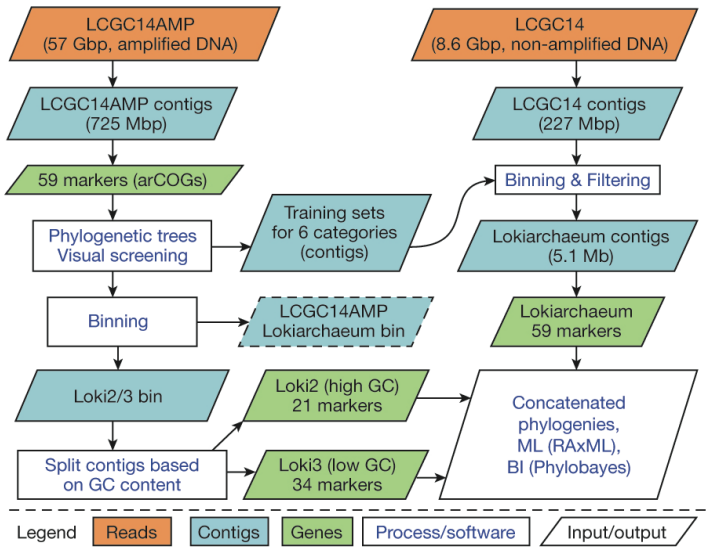


Lokiarchaeota

a novel candidate archaeal phylum

- sample : deep marine sediments near Loki's castle (Norvege)
- amplicon sequencing (16S) : new archea
- shotgun sequencing : Illumina HiSeq 2500, SRP045692

  assembly : 5,381 protein coding genes, 32% new, 26% archea, 29% bacteria, 3.3% eukaryotes



Complex archaea that bridge the gap between prokaryotes and eukaryotes
Nature volume 521, pages173–179(2015)

How was obtained the first SARS-CoV-2 genome ?

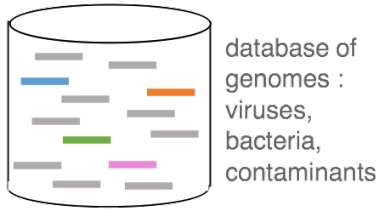
nasal sample

sequencing

sequencing reads (paired-end Illumina)

↓

filtering



↓

genome, 30.000 nt (january 2020) ← assembly

Shotgun sequencing for community samples

- Metagenomics

  potentially sequences all fragmented DNA in a community

  $\rightarrow$ includes all microorganisms and viruses

  $\rightarrow$ gives access to all genes across the entire genomes

- Metatranscriptomics

  potentially sequences all fragmented RNA in a community

  $\rightarrow$ activity of the genes

Amplicon sequencing


fast and cost-effective


captures a large diversity of microorganisms


benefits from well-designed computational tools


requires PCR (primers, amplification)


restrained to taxonomic classification and profiling


low taxonomic resolution

Shotgun sequencing versus amplicon sequencing

who is there ?

more complete taxonomic information
no bias due to PCR amplification
access to the full genomes and genes
captures genomes which lack amplicon targets (viruses, . . .)

what are they doing ?

functional potential of the community
analysis of gene functions, metabolic pathways, etc.

more expensive

new challenges in terms of data processing, storage and analysis : size of the data, uneven coverage

Content of this lecture

- Taxonomic analysis

  Some general ideas, principles and tools

- Functional analysis

  Some general ideas, principles and tools

- Not presented today : Richness, comparative analysis

Key concepts

- To select, or not

  focusing on some marker genes
  one single marker or a combination of markers

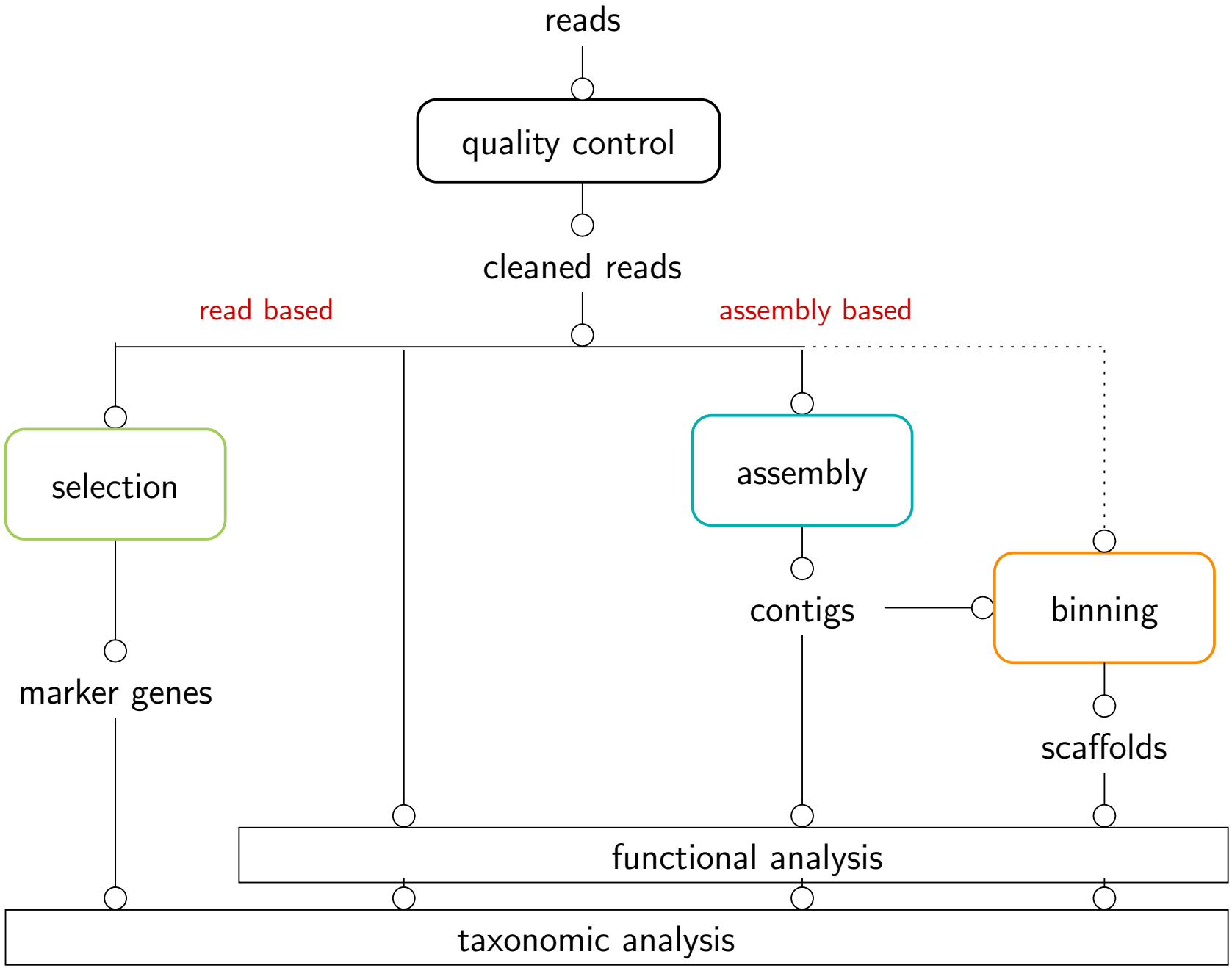
- To assemble, or not

  reconstructing the original sequences from short reads

- To bin, or not

  gathering sequences that are intended to belong to the same species, or the same strain

Many routes, many strategies, many tools


reads
quality control
cleaned reads
read based
assembly based
selection
assembly
contigs
binning
marker genes
scaffolds
functional analysis
taxonomic analysis

Elements of choice

	selection	all reads	assembly
Biological question			
presence/absence of known species	⋆⋆⋆	⋆⋆⋆	⋆
discovery of novel species	⋆		⋆⋆⋆
functional analysis		⋆	⋆⋆
Complexity of the community	H/M/L	M/L	L
Requirements			
computational time	++	+	+++
sequencing depth	+	+	+++
bioinformatics skills	+	+	+++

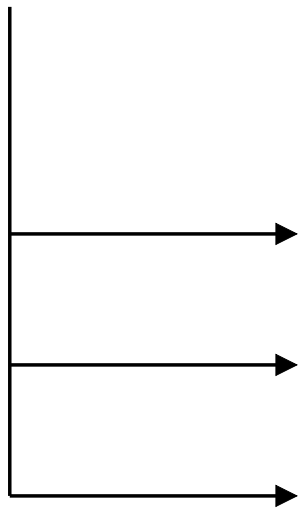
H : high, M : medium, L : low
Computational time : from a few minutes to a few days/weeks
Read-based approaches : web servers or pipelines

Taxonomic classification

- input : short reads from a single shotgun metagenomic sequencing experiment (FASTA or FASTQ files)
- output : list of detected microbes and their abundances

amplicon sequencing

one single marker fragment

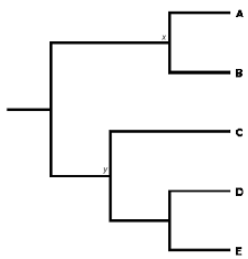


shotgun sequencing

- one single marker, fulll length
- a combination of markers `PhyloSift, Metaphlan2`
- all available genes/genomes `kraken, kaiju`

which data to use for the marker(s) ?
reference database with a taxonomy

how to compare the reads to the database ?
comparison engine



how to classify a read ?
supervised binning

Approach 1 : One single marker

- choice of the phylogenetic marker

  ubiquitous in the environment/showing some differences between species

  16S rRNA (prokaryote), 18S rRNA (eukaryote), ITS (fungi)

- database : `Silva`, `Greengenes`, ...

- comparison to the database

  identification of the reads corresponding to the marker

- processing of the extracted reads

  direct classification of the raw reads : `Qiime2`, `MAPseq`

Approach 2 : Multiple markers

- how to choose the markers ?
- selection of a few universal phylogenetics markers

  `PhyloSift`
- selection of clade-specific markers

  `Metaphlan2`

PhyloSift

- 37 families of "elite" marker genes
  congruent phylogenetic histories
  represent about 1% of an average bacterial genome
- 16S and 18S ribosomal RNA genes
- mitochondrial gene families
- eukaryote-specific gene families
- viral gene families

PeerJ. 2014; 2: e243.
Published online 2014 Jan 9. doi: 10.7717/peerj.243

PhyloSift: phylogenetic analysis of genomes and metagenomes

Aaron E. Darling,[1,2] Guillaume Jospin,[2] Eric Lowe,[2] Frederick A. Matsen, IV,[5] Holly M. Bik,[2] and Jonathan A. Eisen[3,4]

Metaphlan2

Metagenomic Phylogenetic Analysis

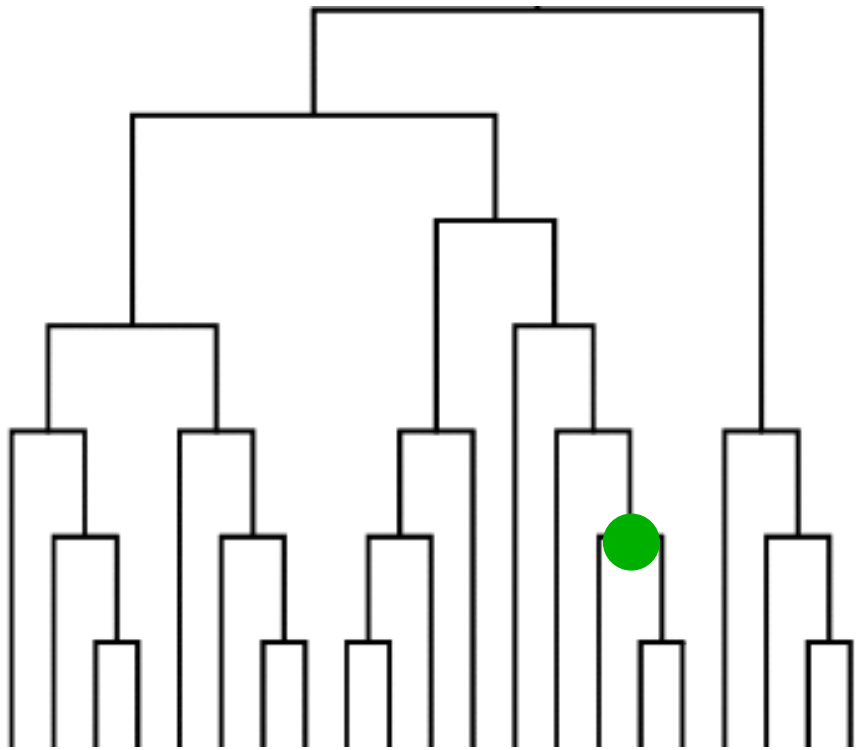
- successor of Metaphlan (2012, Human Microbiome Project)
- markers and quasi-markers

  coding sequences that unequivocally identify specific microbial clades at the species level or higher taxonomic levels

  markers : specific of the clade

  quasi-markers : show a minimal number of sequence hits in genomes outside the clade

pre-computed database of markers
and pseudo markers
+ clades (LCA in the taxonomy)



short
read

bacteria : 770,000 markers + 130,000 pseudomarkers from 13,000 genomes

archaea : 460,000 markers + 4,600 pseudomarkers from 300 genomes

eukaryotes : 22,400 markers + 2,550 pseudomarkes from 110 genomes

virus : 38,800 markers + 23,000 pseudomarkers from 3500 genomes

Metaphlan2 — pipeline

- mapping of short reads on the marker + pseudomarker database (Bowtie2)
- computation of the relative abundance of each taxonomic unit from presence of markers and pseudo-markers

  normalization of the total number of reads in each clade by the nucleotide length of its markers
- unclassified subclades : reads belonging to clades with no available sequenced genomes are reported as an unclassified subclade of the closest ancestor for which there is available sequence data

```
SampleID Metaphlan2_Analysis k_Bacteria 100.0
k_Bacteria|p_Acidobacteria 55.60886 k_Bacteria|p_Verrucomicrobia
36.2624 k_Bacteria|p_Proteobacteria 7.09312
k_Bacteria|p_Actinobacteria 1.03562
k_Bacteria|p_Acidobacteria|c_Acidobacteriia 55.60886
k_Bacteria|p_Verrucomicrobia|c_Opitutae 36.2624
k_Bacteria|p_Proteobacteria|c_Gammaproteobacteria  3.60559
k_Bacteria|p_Proteobacteria|c_Alphaproteobacteria  3.48753
k_Bacteria|p_Actinobacteria|c_Actinobacteria  1.03562
k_Bacteria|p_Acidobacteria|c_Acidobacteriia|o_Acidobacteriales  55.60886
k_Bacteria|p_Verrucomicrobia|c_Opitutae|o_Puniceicoccales  36.2624
k_Bacteria|p_Proteobacteria|c_Gammaproteobacteria|o_Pseudomonadales  3.6
k_Bacteria|p_Proteobacteria|c_Alphaproteobacteria|o_Rhodobacterales  3.4
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales  1.03562
k_Bacteria|p_Acidobacteria|c_Acidobacteriia|o_Acidobacteriales|f_Acidobac
```

```
SampleID Metaphlan2_Analysis k_Bacteria 100.0
k_Bacteria|p_Acidobacteria 55.60886 k_Bacteria|p_Verrucomicrobia
36.2624 k_Bacteria|p_Proteobacteria 7.09312
k_Bacteria|p_Actinobacteria 1.03562
k_Bacteria|p_Acidobacteria|c_Acidobacteriia 55.60886
k_Bacteria|p_Verrucomicrobia|c_Opitutae 36.2624
k_Bacteria|p_Proteobacteria|c_Gammaproteobacteria  3.60559
k_Bacteria|p_Proteobacteria|c_Alphaproteobacteria  3.48753
k_Bacteria|p_Actinobacteria|c_Actinobacteria  1.03562
k_Bacteria|p_Acidobacteria|c_Acidobacteriia|o_Acidobacteriales  55.60886
k_Bacteria|p_Verrucomicrobia|c_Opitutae|o_Puniceicoccales  36.2624
k_Bacteria|p_Proteobacteria|c_Gammaproteobacteria|o_Pseudomonadales  3.6
k_Bacteria|p_Proteobacteria|c_Alphaproteobacteria|o_Rhodobacterales  3.4
k_Bacteria|p_Actinobacteria|c_Actinobacteria|o_Actinomycetales  1.03562
k_Bacteria|p_Acidobacteria|c_Acidobacteriia|o_Acidobacteriales|f_Acidobac
```

Kingdom|Phylum|Class|Order|Family|Genus|Species|Strain

Approach 3 : all possible genes/genomes

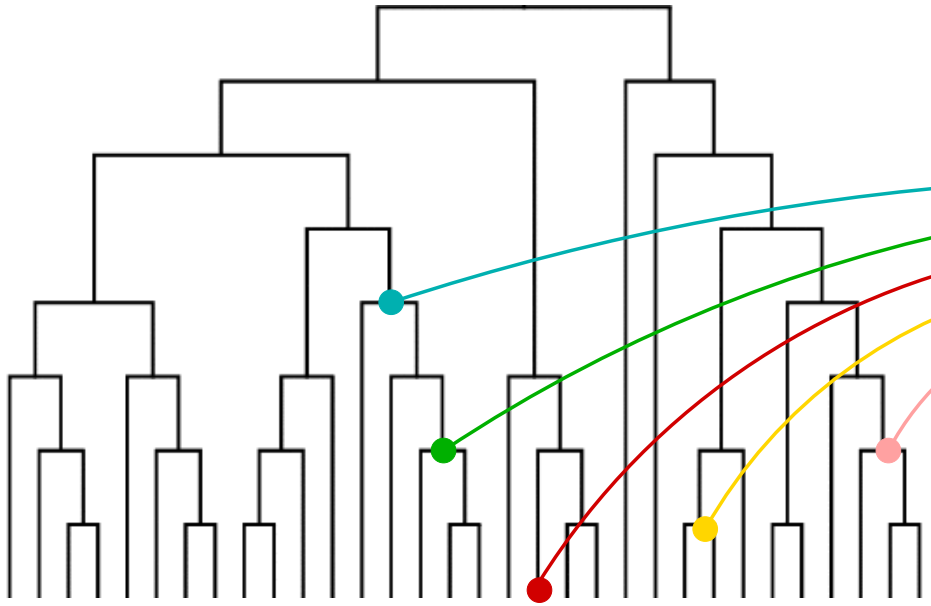
- database : reference genomes + taxonomy

  no structural annotation, no phylogenetic markers

- comparison of reads against the database : should be very efficient

- main principle : split the data into k-mers (words of length k)
  - Data : genomes sequences, reads sequences
  - No prior knowledge on the genomes
- Examples : Kraken, Centrifuge, One codex, LMAT...

Example : Kraken



all completed microbial genomes of the RefSeq database

47,768 bacteria + 1,034 archaea + 7,530 viruses

all 31-mers present in the database

+ LCA (lowest common ancestor)

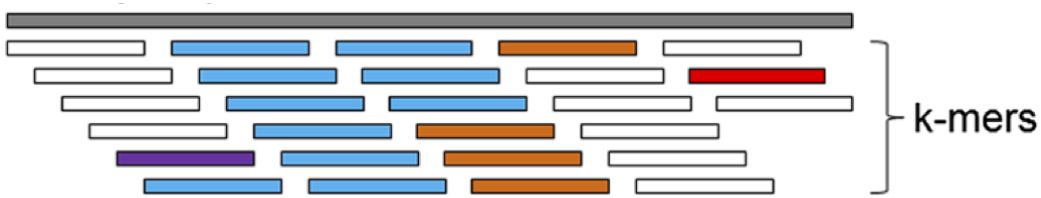
```
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAC
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAG
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAT
AAAAAAAAAAAAAAAAAAAAAAAAAAAAACA
...
```

1.4e9 disctinct k-mers (oct 2017)

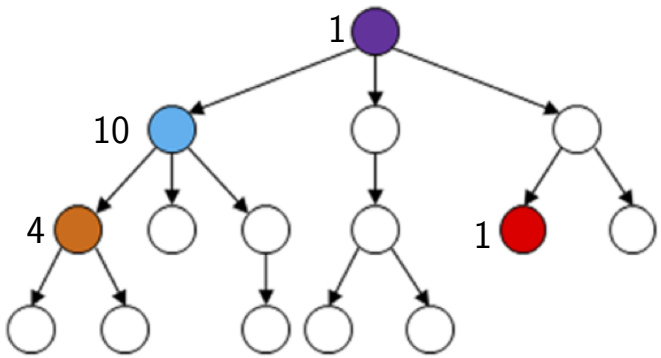
$<< 4^{31} = 4.6e18$

Precomputed database

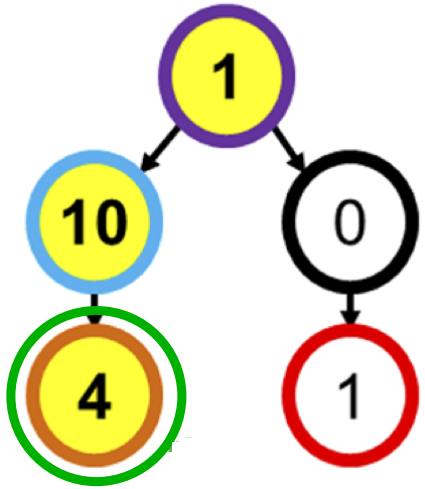
1. short read → overlapping k-mers



2. identification of the LCA in the taxonomy for each k-mer

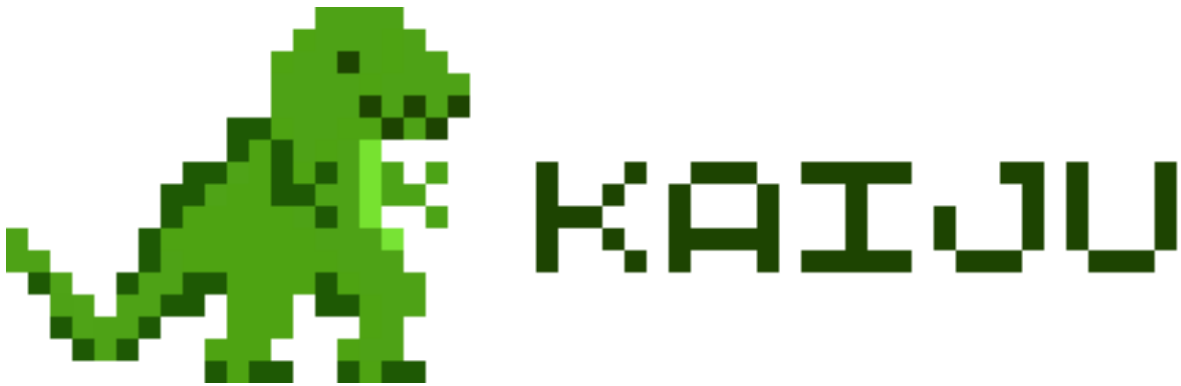


3. assignation of the read



Read assignation

Kaiju



Menzel, P. et al. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun. 7 :11257

- protein-level classification : reads are translated into amino acid sequences

- database

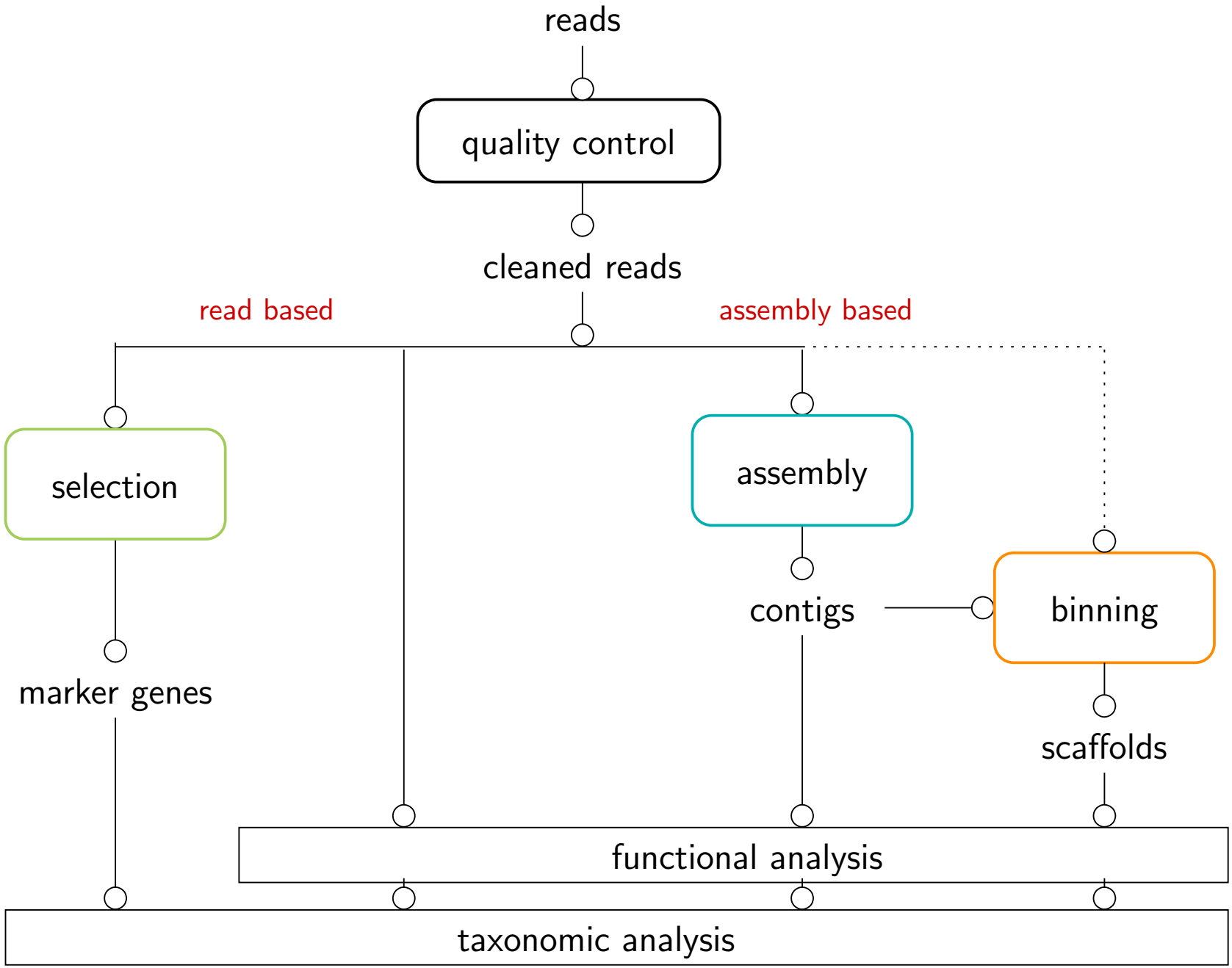
  NCBI RefSeq, proGenomes, non-redundant BLAST protein database (optionally also including fungi and microbial eukaryotes)

- comparison between the reads and the database

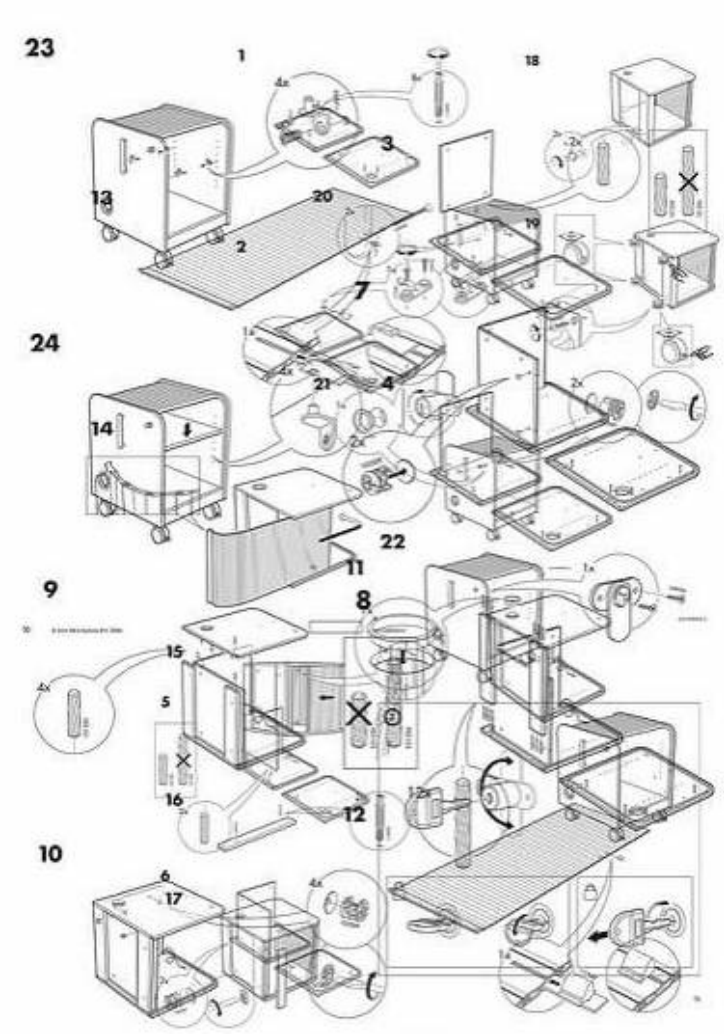
  maximum exact matches (MEMs), optionally allowing mismatches
  Burrows-Wheeler Transform

- classification


Getting Started


reads
quality control
cleaned reads
read based
assembly based
selection
assembly
contigs
binning
marker genes
scaffolds
functional analysis
taxonomic analysis

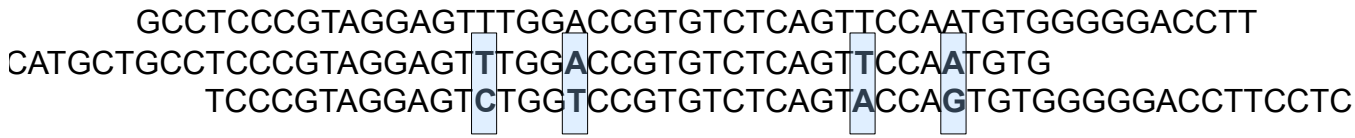
Assembly



Metagenomic assembly is impossible

Two competing goals:

- assemble similar sequences from related genomes together
- do not assemble similar sequences from unrelated genomes



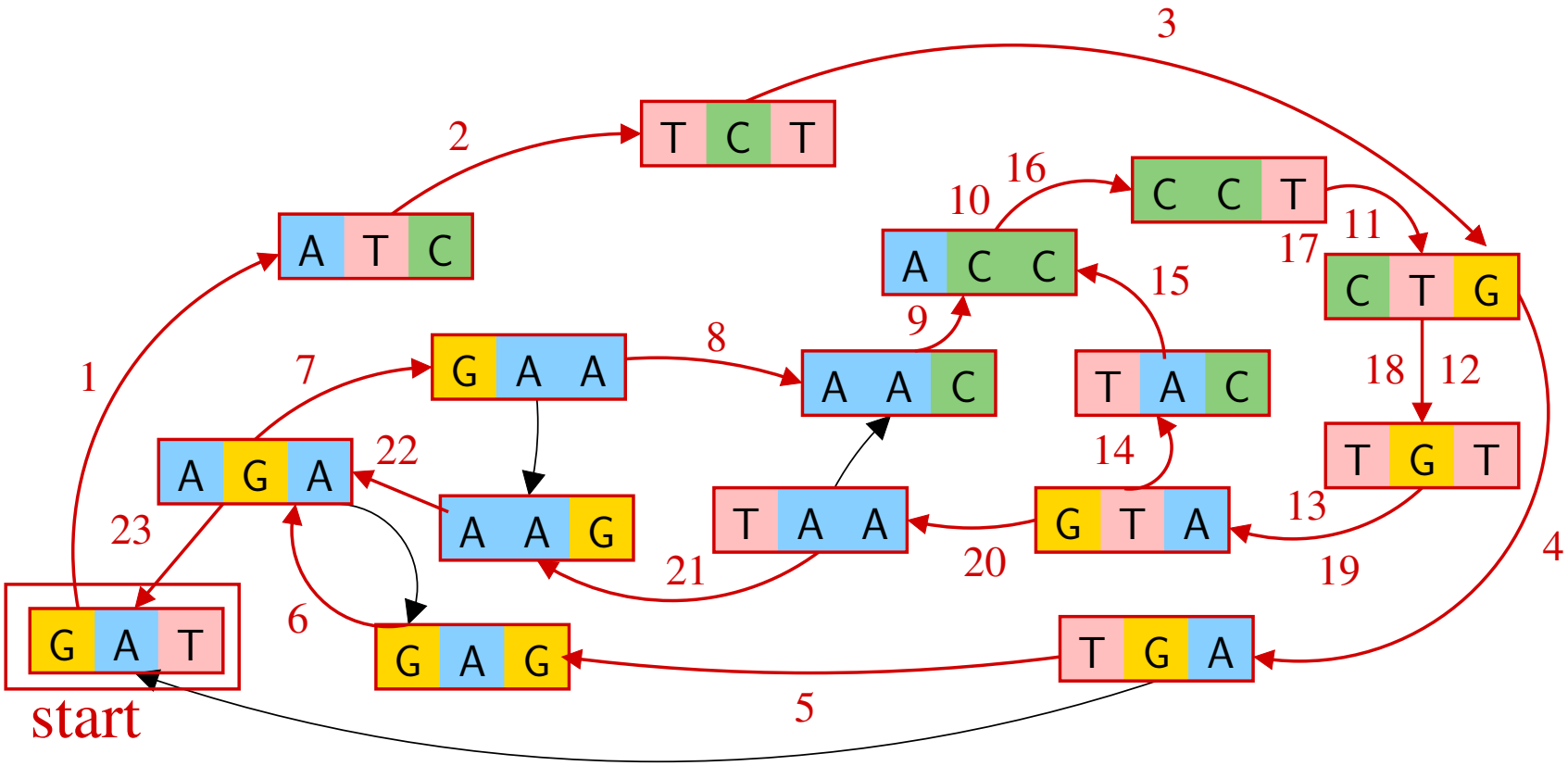
Mihai Pop, Sergey Koren, Dan Sommer

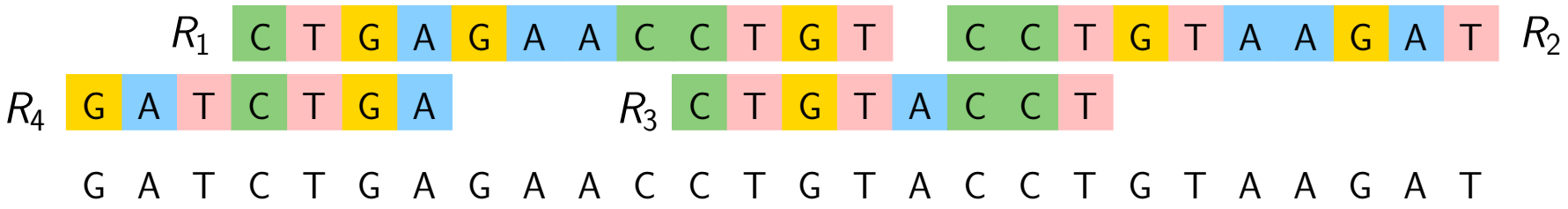
Why it is so difficult

- presence of multiple closely related strains or species : hard to distinguish sequencing errors and poymorphisms
- uneven abundance of organisms present in the sample : this causes uneven sequencing depth of organisms present in the sample
- presence of intragenomic repeats + intergenomics repeats (horizontal transfer) : risk of chimera creation
- size of the data : Gb $\rightarrow$ Tb

De Bruijn Graph (reminder)

- rationale
  - the genome can be reconstructed from the k-mers it contains
  - reads are decomposed into k-mers
- graph
  - nodes : k-mers present in the reads
  - arcs : overlaps of length $k-1$ between k-mers
- contig : simple path in the graph


start
G A T
A T C
T C T
C C T
A C C
C T G
G A A
A A C
T A C
A G A
A A G
T A A
G T A
T G T
G A G
T G A
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

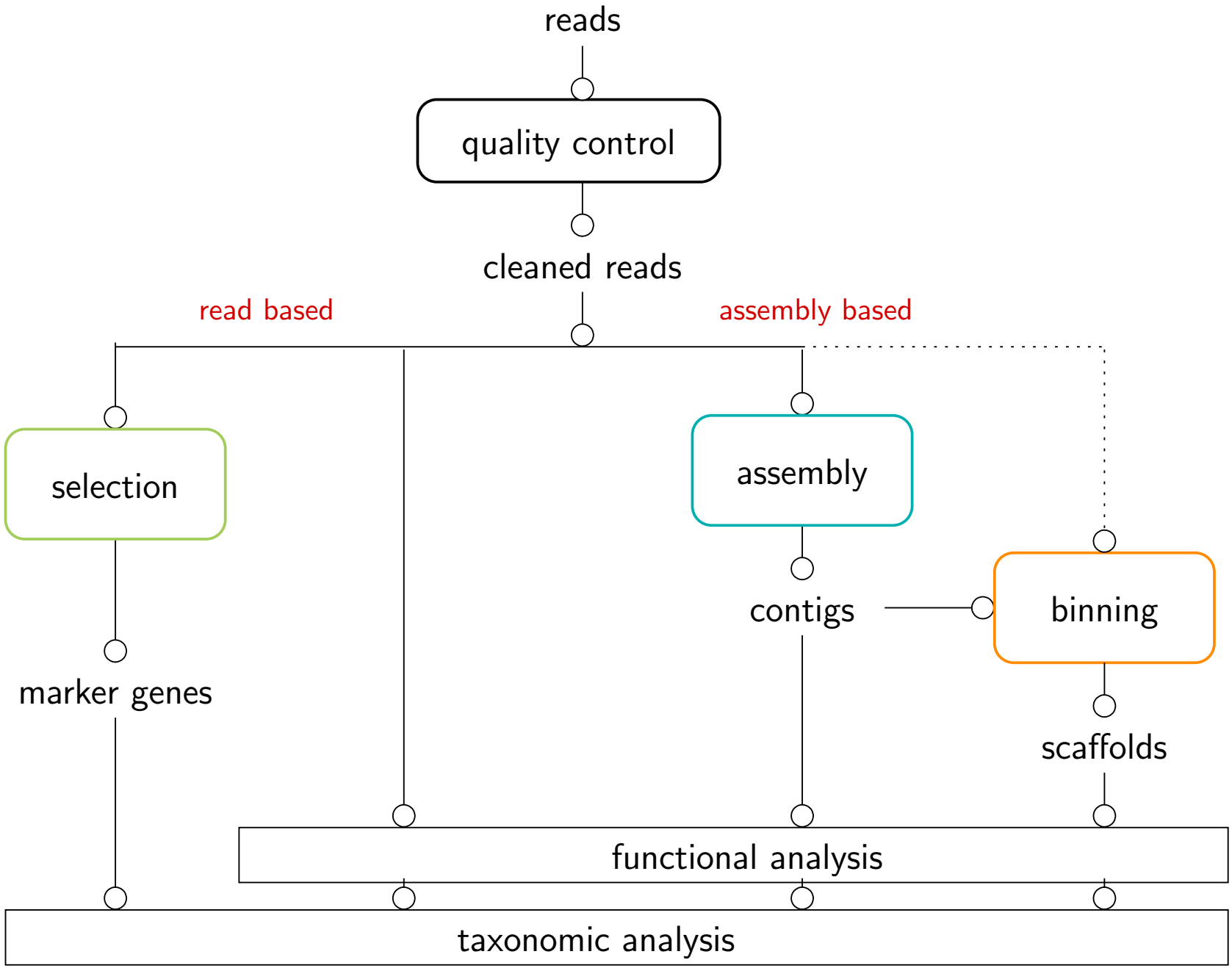

R_1 C T G A G A A C C T G T
C C T G T A A G A T R_2
R_4 G A T C T G A
R_3 C T G T A C C T
G A T C T G A G A A C C T G T A C C T G T A A G A T

Application to community samples

- de Bruijn graph + multi-k principle
  $k = 21 \rightarrow k = 55 \rightarrow k = 77$
- efficient construction and storage of the De Bruijn Graphs
- careful handling of mismatches
- careful extension of paths in the De Bruiijn Graphs
- intergenomic repeats solving with abundance
- metagenomics : MEGAHIT (2015), MetaSPAdes (2016)
- metatranscriptomics : MEGAHIT (2015)

What to do with contigs

- taxonomy classification : analogous to read-based approaches
- functional annotation : later in this lecture
- binning : just now

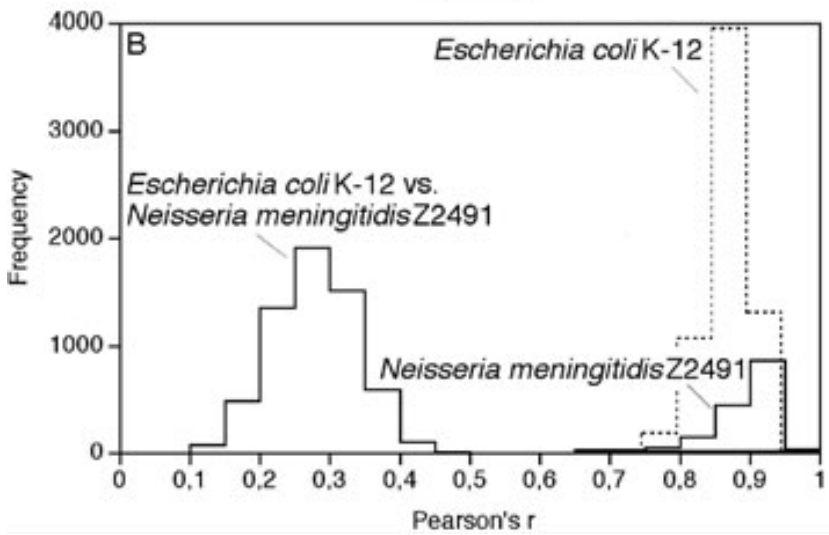

reads
quality control
cleaned reads
read based
assembly based
selection
assembly
contigs
binning
marker genes
scaffolds
functional analysis
taxonomic analysis

Binning

- gathering sequences that are intended to belong to the same species, or the same strain
- taxonomy dependent (supervised binning, taxonomic binning)
  - database search, sequence comparison
  - known species
  - `Phylosift`, `Metaphlan2`, `MG-Rast`, `MEGAN`, `MGnify`...
- taxonomy independent (inherent statistics)
  - sequence composition : nucleotide composition, codon usage
  - contig coverage
  - hybrid : machine learning

Nucleotide composition

Tetranucleotide usage patterns

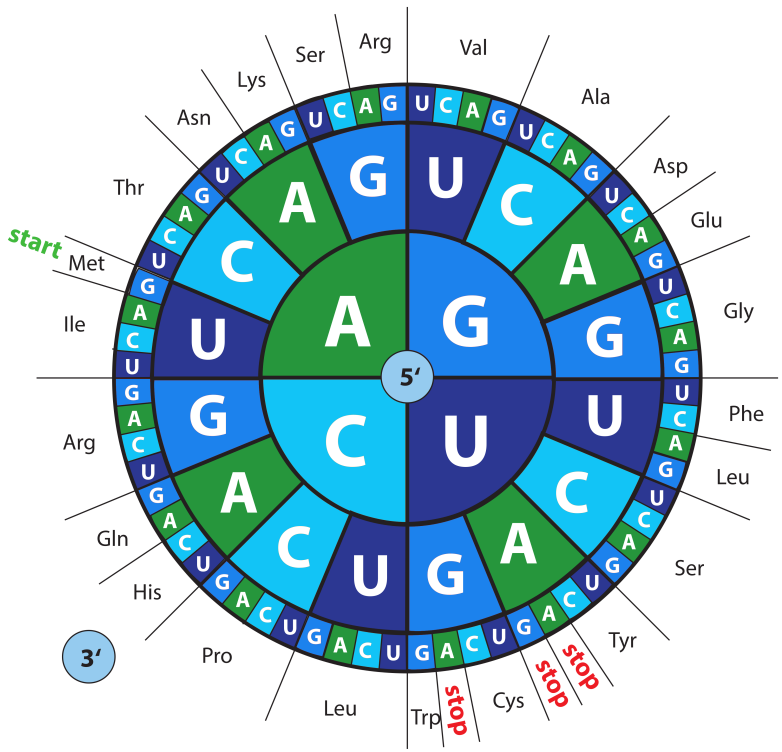


- *Escherichia coli* and *Neisseria meningitidis*
- overlapping fragments of 40kb
- for each fragment, for each tetranucleotide : Z-score observed frequency/theoretical frequency
- histograms of Pearson's correlation coefficients : pairwise comparisons of the fragment's tetranucleotide-derived z-scores

Application of tetranucleotide frequencies for the assignment of genomic fragments. Environmental Microbiology (2004) 6(9), 938–947

Codon usage

- the genetic code is redundant : several codons can code for the same amino acid
- each species tends to show a preference for particular synonymous codons
- clustering of sequences according to their codon bias



AAA	3.5	1.3	CAA	1.3	1.4	GAA	4.3	1.6	TAA	*	*
AAG	1.1	1.6	CAG	3.0	1.7	GAG	1.8	1.8	TAG	*	*
AAC	2.4	1.4	CAC	1.1	1.5	GAC	2.2	1.7	TAC	1.4	1.4
AAT	1.4	1.3	CAT	1.2	1.4	GAT	3.2	1.5	TAT	1.5	1.3
AGA	0.1	1.6	CGA	0.3	1.7	GGA	0.6	1.8	TGA	*	*
AGG	0.1	1.8	CGG	0.4	2.0	GGG	1.0	2.2	TGG	1.4	1.8
AGC	1.6	1.7	CGC	2.4	1.8	GGC	3.2	2.0	TGC	0.7	1.6
AGT	0.7	1.5	CGT	2.5	1.6	GGT	2.8	1.8	TGT	0.5	1.5
ACA	0.5	1.4	CCA	0.8	1.5	GCA	2.0	1.7	TCA	0.6	1.4
ACG	1.4	1.7	CCG	2.6	1.8	GCG	3.6	2.0	TCG	0.8	1.6
ACC	2.5	1.5	CCC	0.4	1.6	GCC	2.5	1.8	TCC	0.9	1.5
ACT	0.9	1.4	CCT	0.6	1.5	GCT	1.6	1.6	TCT	0.9	1.4
ATA	0.3	1.3	CTA	0.3	1.4	GTA	1.1	1.5	TTA	1.1	1.3
ATG	2.5	1.5	CTG	5.7	1.6	GTG	2.7	1.8	TTG	1.2	1.5
ATC	2.7	1.4	CTC	1.0	1.5	GTC	1.5	1.6	TTC	1.8	1.4
ATT	2.8	1.3	CTT	0.9	1.4	GTT	1.9	1.5	TTT	1.9	1.2

Codon Usage Frequence Table for *E. coli*

1st column : observed frequency

2nd column : theoretical frequency

Examples of the usage of Serine codons in different organisms

Codon	*E.coli*	*D.melanogaster*	*H.sapiens*	*S.cerevisiae*
AGT	3	1	10	5
AGC	20	23	34	4
TCG	4	17	9	1
TCA	2	2	5	6
TCT	34	9	13	52
TCC	37	48	28	33

(rounded percentages– source : D. Gautheret)

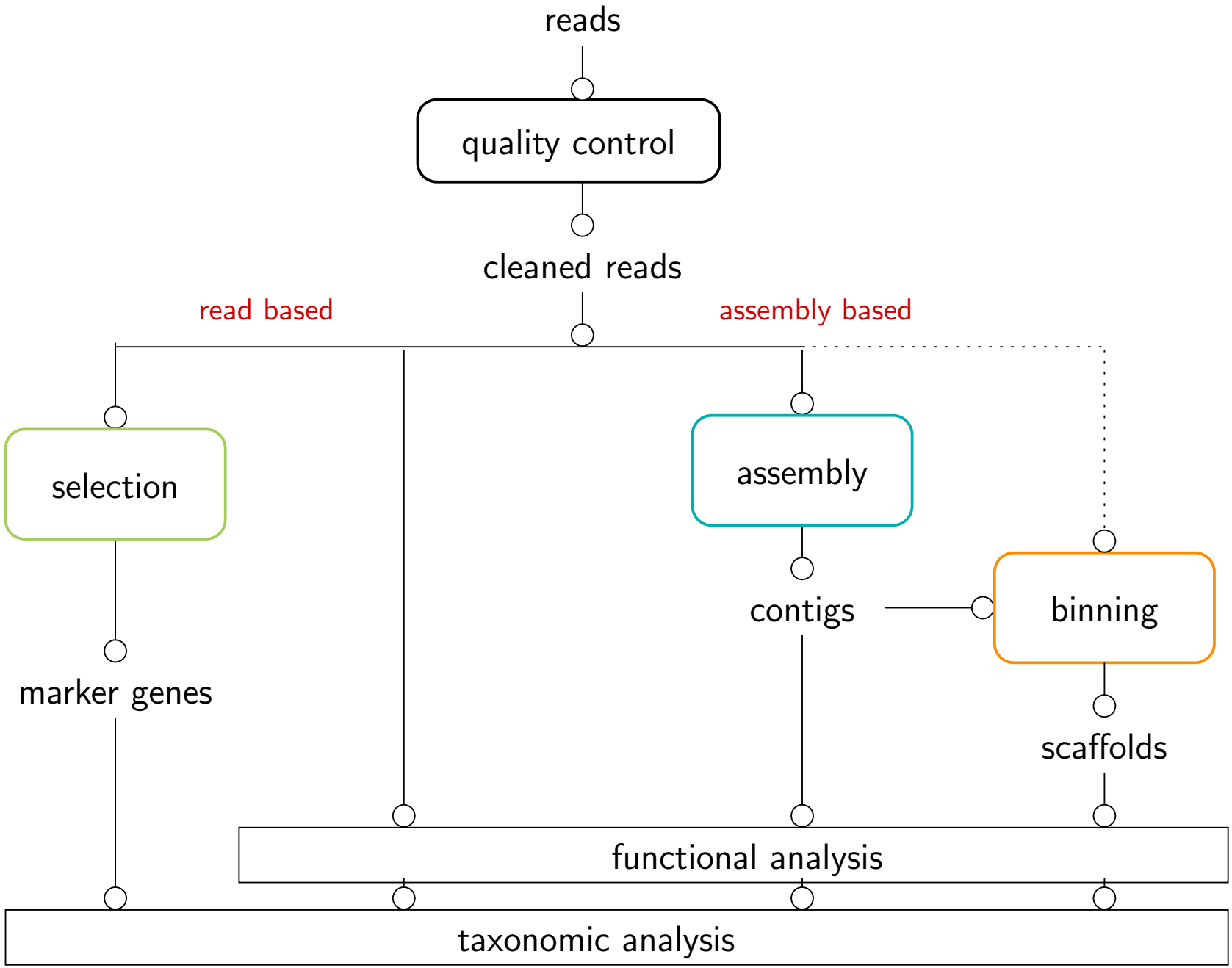
Contig coverage

- reads are mapped on the contigs
- similar coverage $=$ similar abundance
- two contigs with similar coverage potentially come from same underlying source population in the community

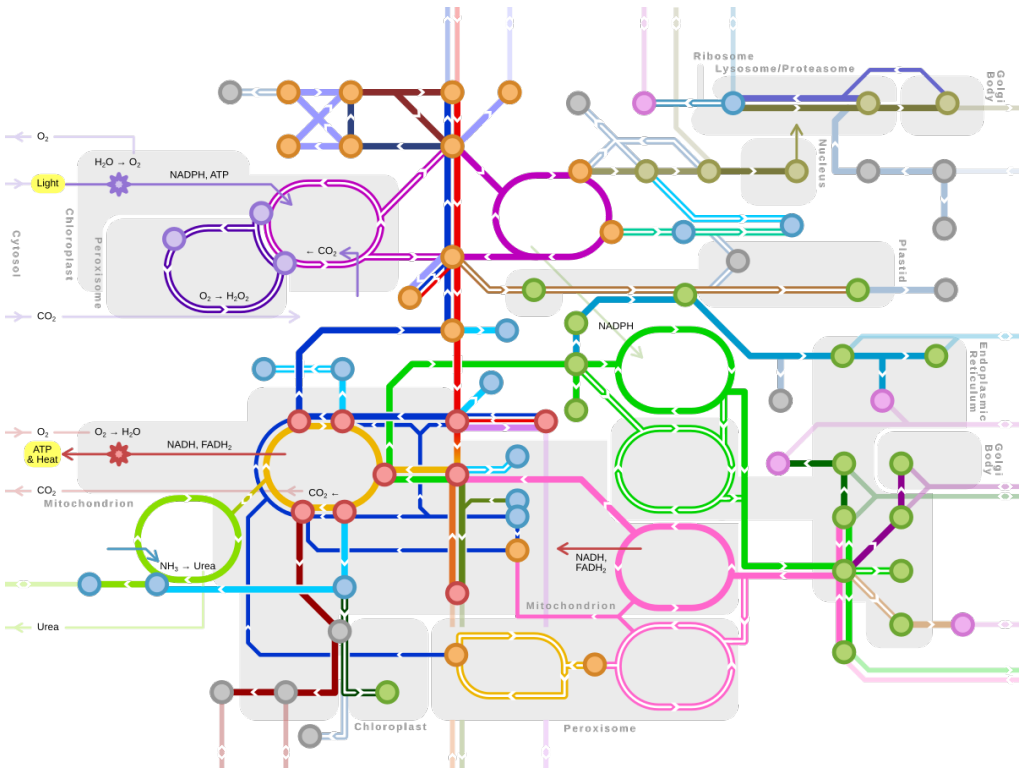
Hybrid approaches

- cocacola (2017)
  COCACOLA : binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge
  Bioinformatics, Volume 33, Issue 6, 15, pages 791–798
- concoct (2014)
  Binning metagenomic contigs by coverage and composition Nature Methods volume 11, pages 1144–1146
- MyCC (2016)
  Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. Sci Rep. 2016 ; 6 : 24175.
- MetaBat (2015)
  MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 2015 ; 3 : e1165.


BESTWALLSITE.COM


reads
quality control
cleaned reads
read based
assembly based
selection
assembly
contigs
binning
marker genes
scaffolds
functional analysis
taxonomic analysis

Functional analysis



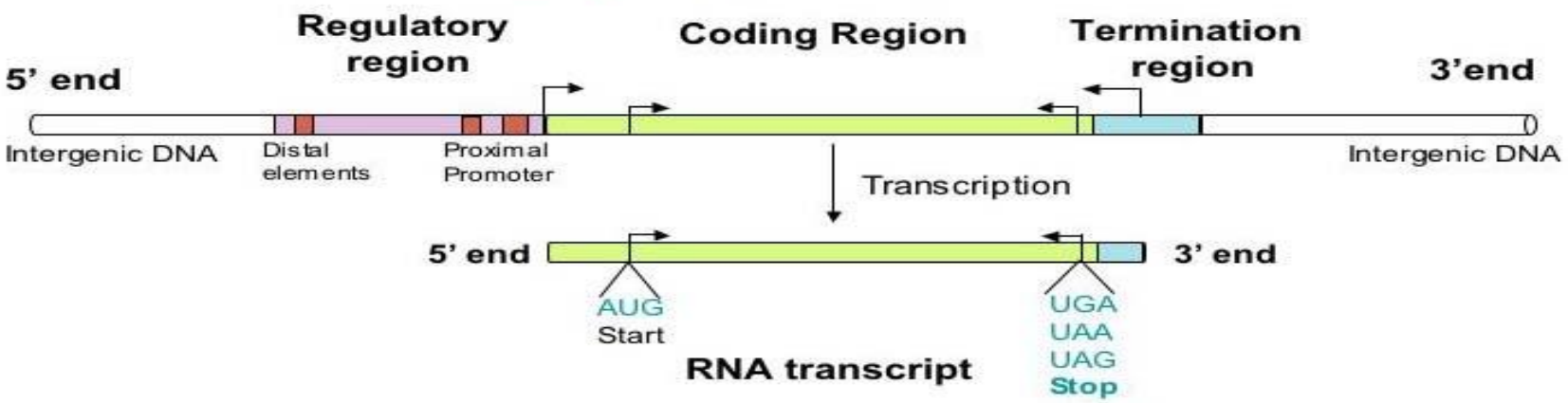
Functional analysis : how to annotate genes in genomes ?

Three main approaches

- *de novo* prediction of coding regions
- homology based annotation
- motif based annotation

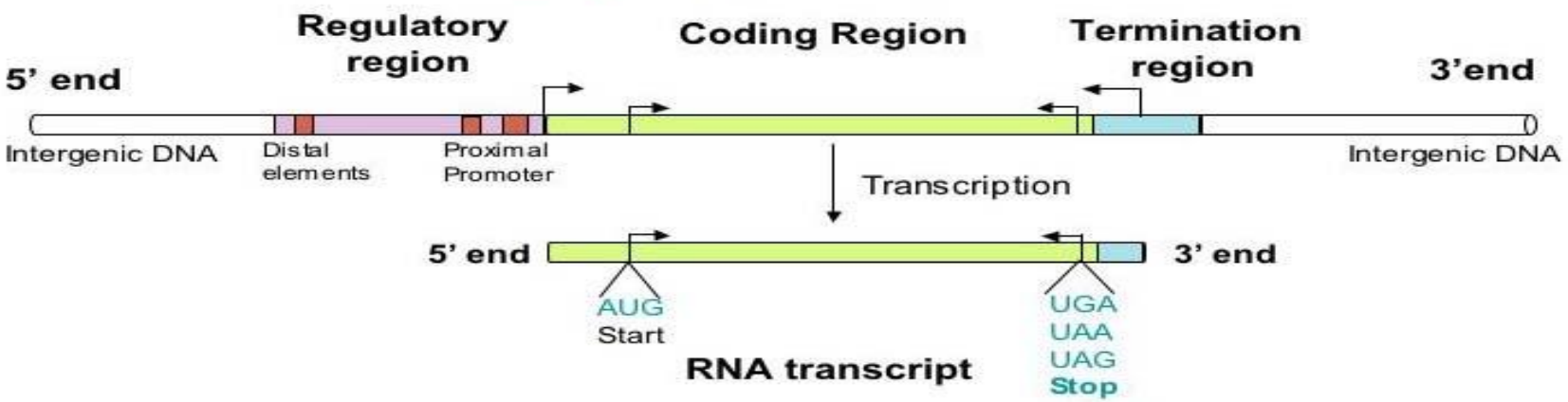
Prediction of coding regions

how can we find genes in prokaryotic genomes ?



Prediction of coding regions

how can we find genes in prokaryotic genomes ?



- identification of ORFs (start + stop codon)
- codon usage bias
  differences in the frequency of occurrence of synonymous codons in coding DNA compared to non-coding DNA, and between species

How to adapt these approaches to metagenomic/metranscriptomic reads/contigs ?

- short reads : codon usage bias
- contigs : ORF (codons start, stop) + codon usage bias
- +resistant to sequencing errors

`FragGeneScan`

Nucleic Acids Res. 2010 Nov;38(20):e191. doi: 10.1093/nar/gkq747. Epub 2010 Aug 30.

FragGeneScan: predicting genes in short and error-prone reads.

Rho M[1], Tang H, Ye Y.

`MetaGeneMark`

`http://exon.gatech.edu/meta_gmhmmp.cgi`

Homology based annotation

- alignment of short reads/contigs to a large database of annotated protein sequences
- databases : `Eggnog`, `SEEDS`, `KEGG`, `Interpro`, `swissprot`, ...
- choice of the alignment tool, DNA/protein

  pre-NGS tools : `BlastX`, `BLAT` especially designed for gene or genome comparison

  `Diamond` : optimized to deal with short reads
  order of magnitude faster than `BlastX` for this kind of data (x 1000)

Fast and sensitive protein alignment using DIAMOND

Benjamin Buchfink ✉, Chao Xie & Daniel H Huson ✉

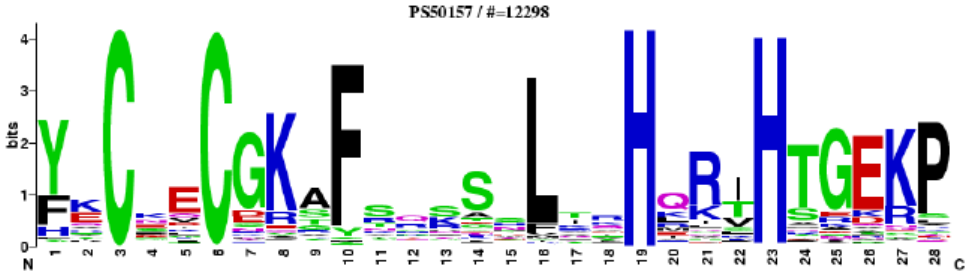
Nature Methods **12**, 59–60 (2015)
doi:10.1038/nmeth.3176

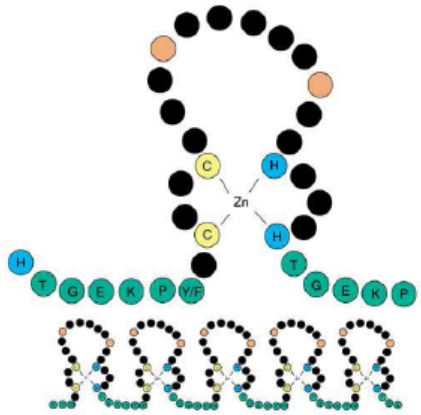
Received: 29 April 2014
Accepted: 20 October 2014

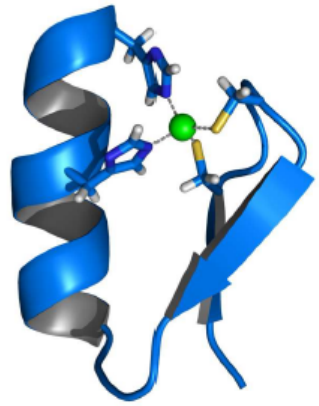
Motif based annotation

- motif : signature for a given protein family
- models : prosite expression, matrix, profile Hidden Markov Model

```
TYY1_HUMAN YVCPFDGCNKKFAQSTNLKSHILT--H
YKQ8_CAEEL YKCT--VCRKDISSSESLRTHMFKQHH
BASO_HUMAN FQCD--ICKKTFKNACSVKIHHKN-MH
ZG2-9_XENL FVCT--VCGKTYKYKHGLNTHLHS--H
P43_XENBO  LKCSVPGCKRSFRKKRALRIHVSE--H
IKAR_MOUSE FECN--MCGYHSQDRYEFSSHITRGEH
TRA1_CAEEL YKCEFADCEKAFSNASDRAKHQNR-TH
ZN10_HUMAN YKCN--QCGIIFSQNSPFIVHQIA--H
XFIN_XENLA FRCS--ECSRSFTHNSDLTAHMRK--H
TF3A_BUFAM CKCETENCNLAFTTASNMRLHFKR-AH
ZG58_XENLA FVCT--ECNLSFAGLANLRSHQHL--H
P43_XENBO  YRCSYEDCQTVSPTWTALQTHLKK--H
TSH_DROME  FRCV--WCKQSFPTLEALTTHMKDSKH
ZN76_HUMAN FRCGYKGCGRLYTTAHHLKVHERA--H
TF3A_BUFAM YRCPRENCDRTYTTKFNLKSHILT-FH
SUHW_DROAN YACK--ICGKDFTRSYHLKRHQKYSSC
ZN76_HUMAN YTCPEPHCGRGFTSATNYKNHVRI--H
SRYC_DROME FKCN--YCPRDFTNFPNWLKHTRR-RH
EVI1_HUMAN YRCK--YCDRSFSISSNLQRHVRN-IH
```







modélisation : motif Prosite

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

Interpro

Protein sequence analysis & classification

- `http:// www.ebi.ac.uk/interpro`
- developed at EBI since 1999 (version 70)
- signatures for protein families, domains and functional sites collected from 14 databases

  35 020 entries based on 48 938 signatures
- mappings of InterPro entries to Gene Ontology (GO) terms (InterPro2GO)

Pipelines for read-based strategies

Taxonomic+functional analyses

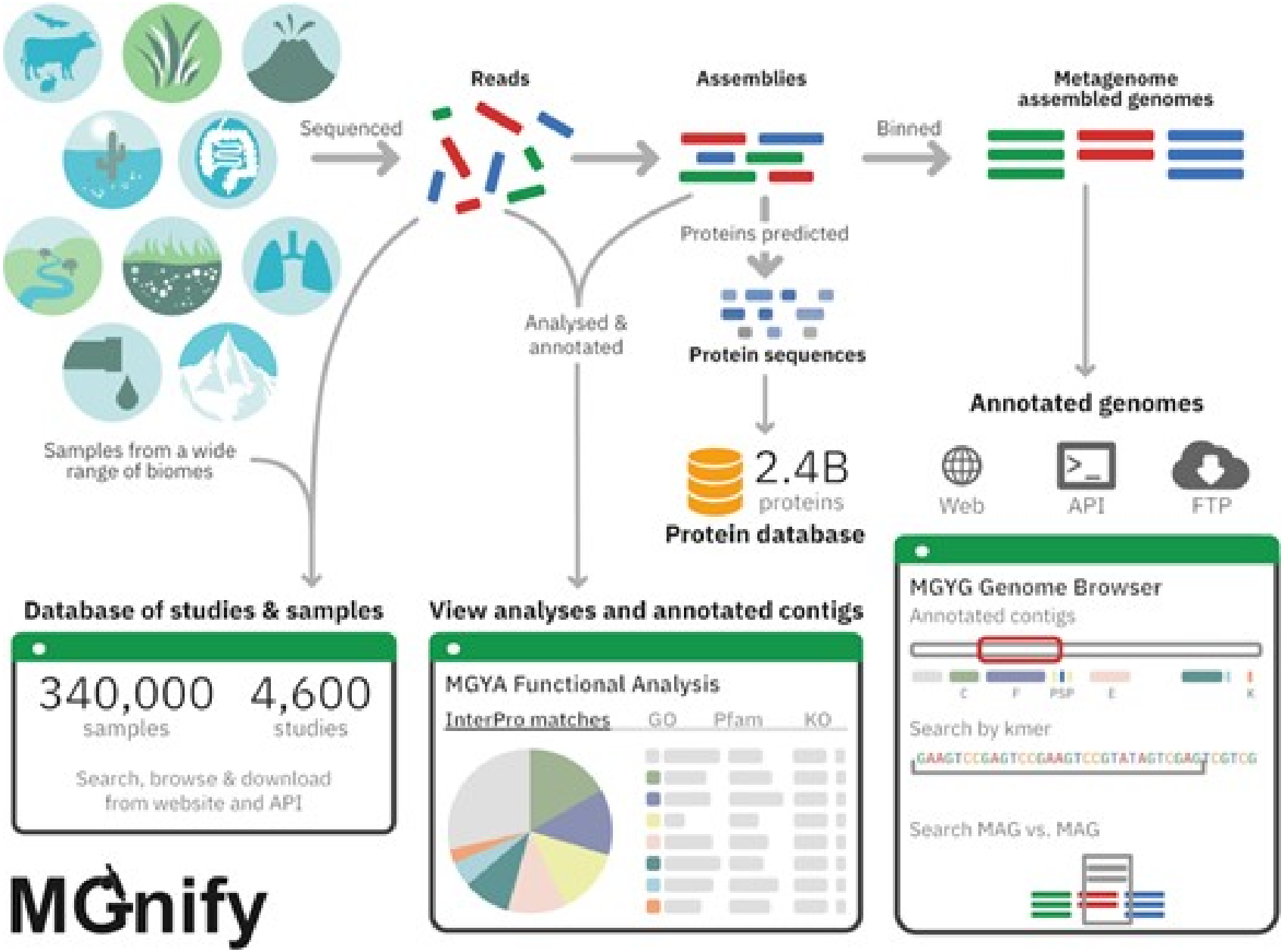
- MG-RAST
  - developed since 2007 (University of Chicago)
  - supports amplicons (16S, 18S, and ITS), metagenomics and metatranscriptomics
  - `http://metagenomics.anl.gov`
- MEGAN
  - developed since 2007 (U. Tübingen)
  - Alignment of the reads on the database : Diamond
  - Taxonomic classification : LCA, lowest common ancestor against NCBI nr
  - Functional analysis : mapping to KEGG, SEED, EggNOG and InterPro2GO
- MGnify

MGnify

EBI metagenomics

Submit, analyse, discover and compare microbiome data since 2013

`https://www.ebi.ac.uk/metagenomics`


Reads
Assemblies
Metagenome assembled genomes
Sequenced
Binned
Proteins predicted
Analysed & annotated
Protein sequences
Samples from a wide range of biomes
2.4B proteins
Protein database
Annotated genomes
Web
API
FTP
Database of studies & samples
340,000 samples
4,600 studies
Search, browse & download from website and API
View analyses and annotated contigs
MGYA Functional Analysis
InterPro matches
GO
Pfam
KO
MGYG Genome Browser
Annotated contigs
Search by kmer
Search MAG vs. MAG
MGnify

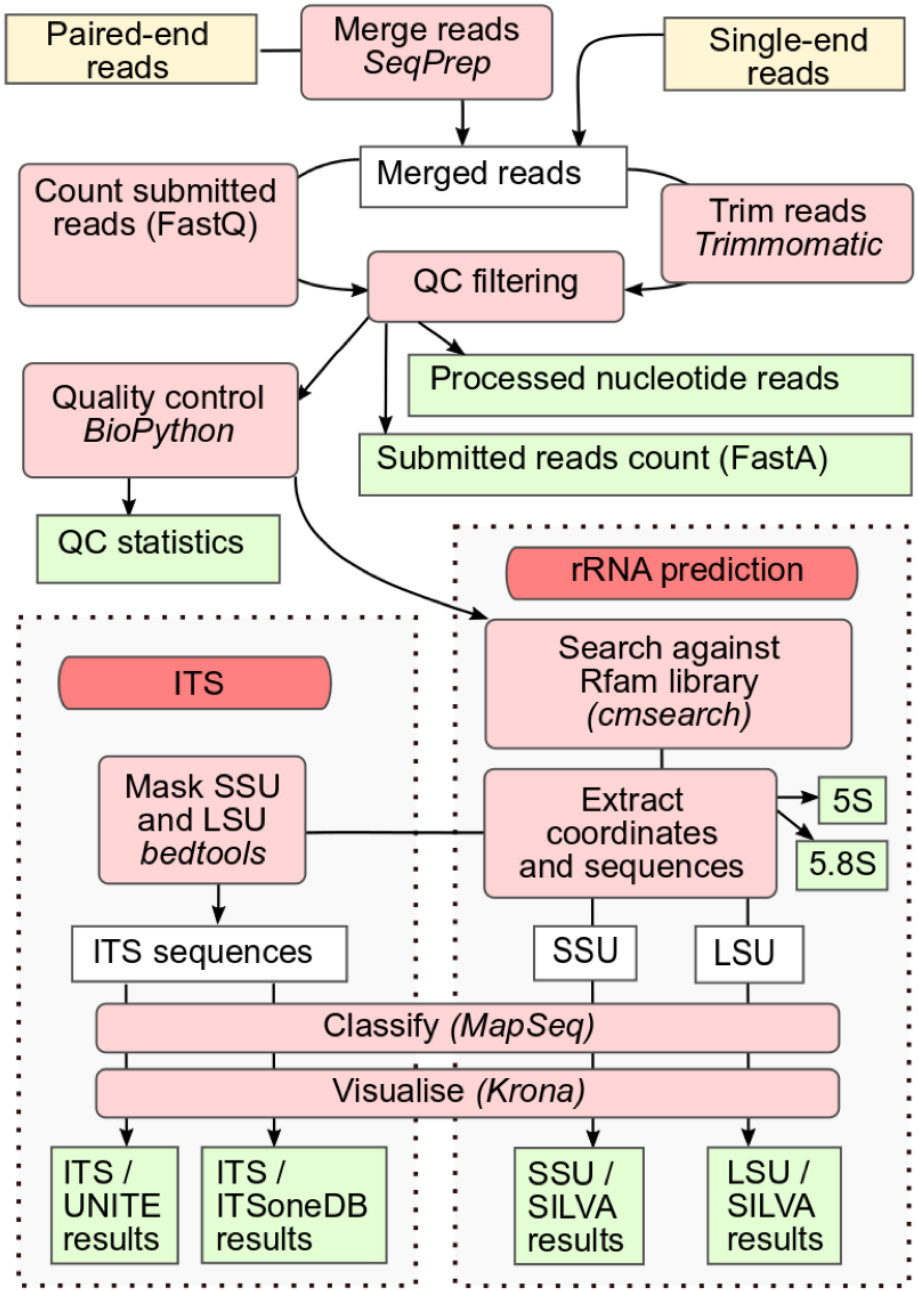
What can you do with MGnify ?

- Submit microbiome studies for analysis : amplicon, metagenomic, metatranscriptomic or assembled data for analysis
- Request analysis of any publicly available data
- Explore a diverse range of analysed microbiome studies
- Visualise and download analysis results
- Access the raw data from the European Nucleotide Archive (ENA).

Data structuration in MGnify

- MGYS `xxxxxx` – study / project
  Each project contains one or more (biological) samples
- sample (ENA identifier)
  Each sample can have one or more experiments associated with it (such as metagenomic, amplicon or metatranscriptomic).
- run (ENA identifier)
  Set of reads for one experiment
- MGYA `xxxxxx` – analysis
  Results obtained from processing a run file

MGnify : amplicon analysis pipeline



Analysis MGYA00383254

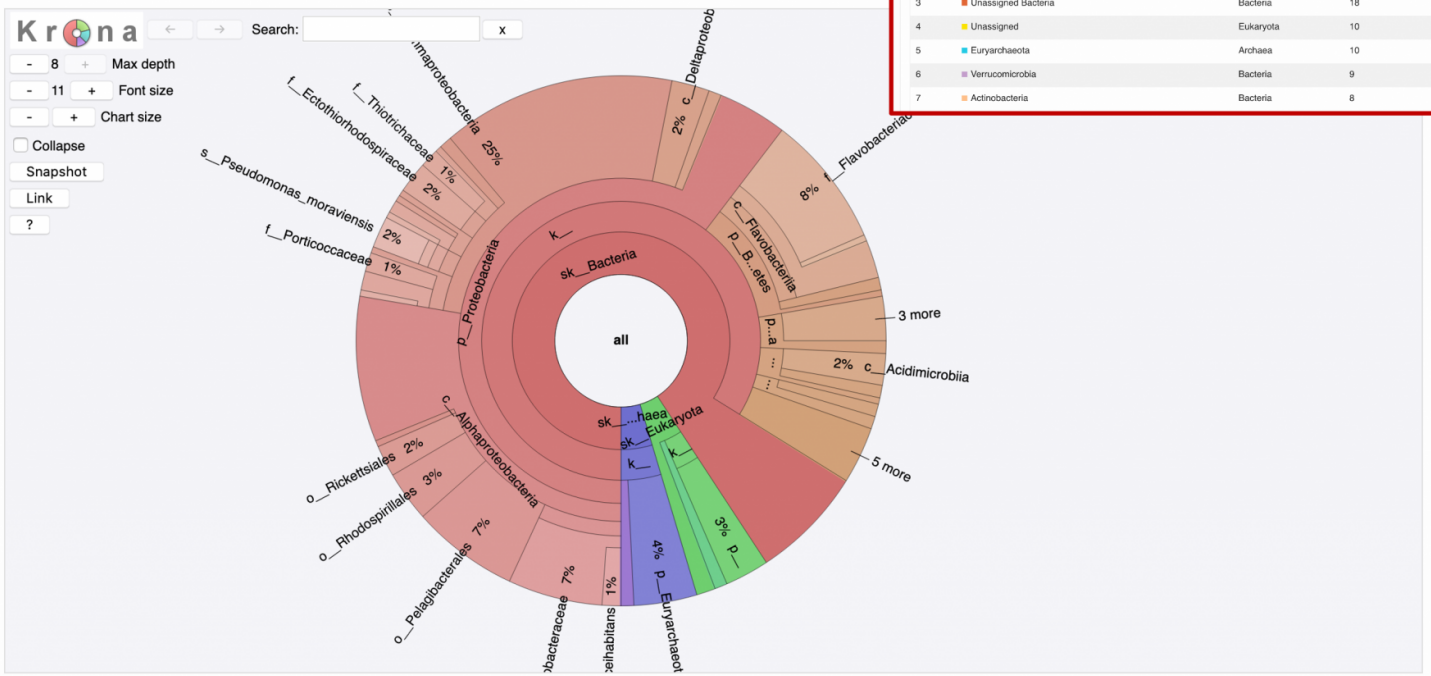
Other analyses

Overview | Quality control | Taxonomic analysis | Functional analysis | Pathways/Systems

rRNA

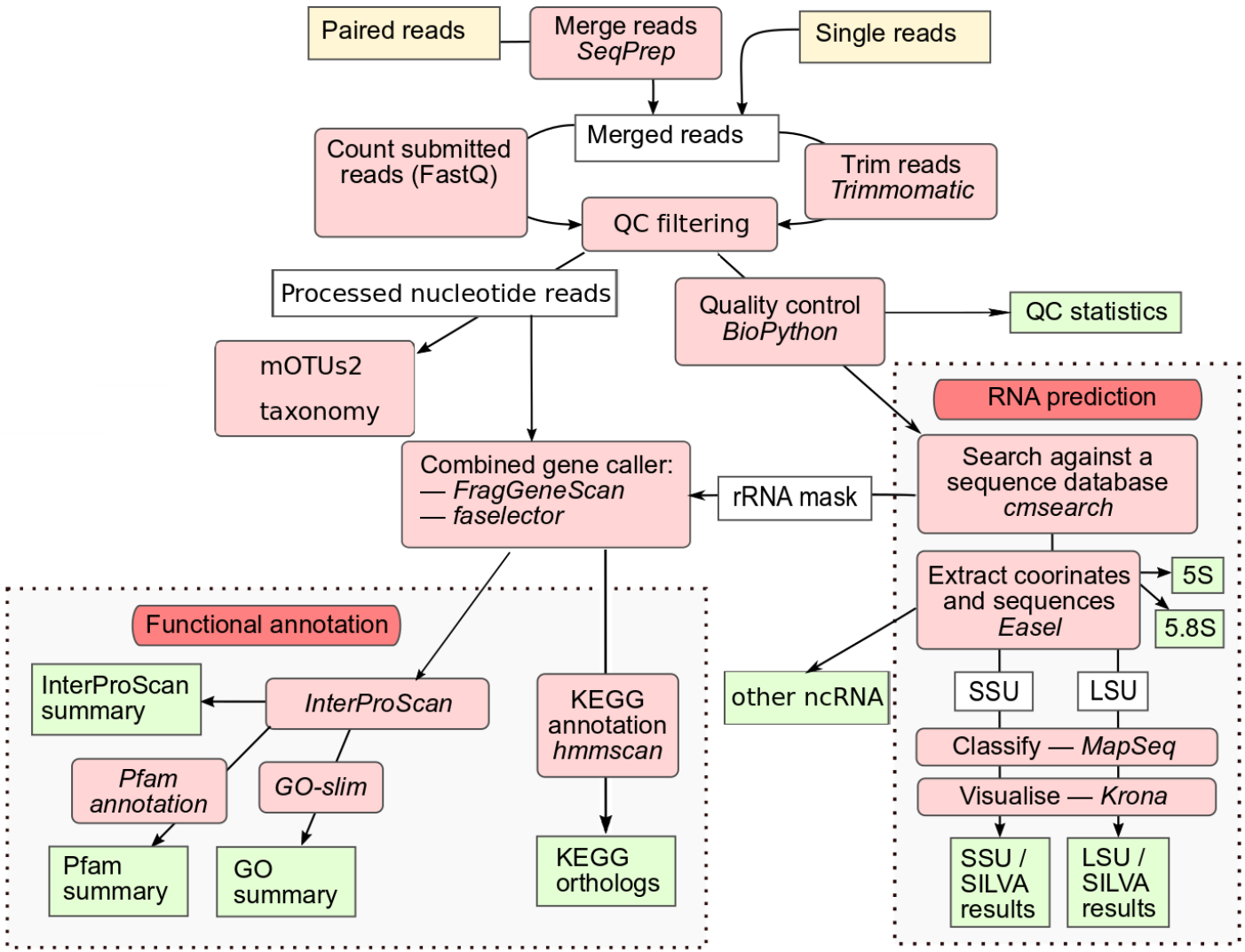
- (•) small subunit rRNA
- ( ) large subunit rRNA

These are the results from the taxonomic analysis steps of our pipeline. You can switch between differer
stacked and interactive krona charts). The data used to build these charts can be found under the "Down



	Phylum	Domain	Unique OTUs	%
1	Proteobacteria	Bacteria	157	60.38
2	Bacteroidetes	Bacteria	31	11.92
3	Unassigned Bacteria	Bacteria	18	6.92
4	Unassigned	Eukaryota	10	3.85
5	Euryarchaeota	Archaea	10	3.85
6	Verrucomicrobia	Bacteria	9	3.46
7	Actinobacteria	Bacteria	8	3.08

MGnify : raw reads analysis pipeline



Analysis MGYA00383254

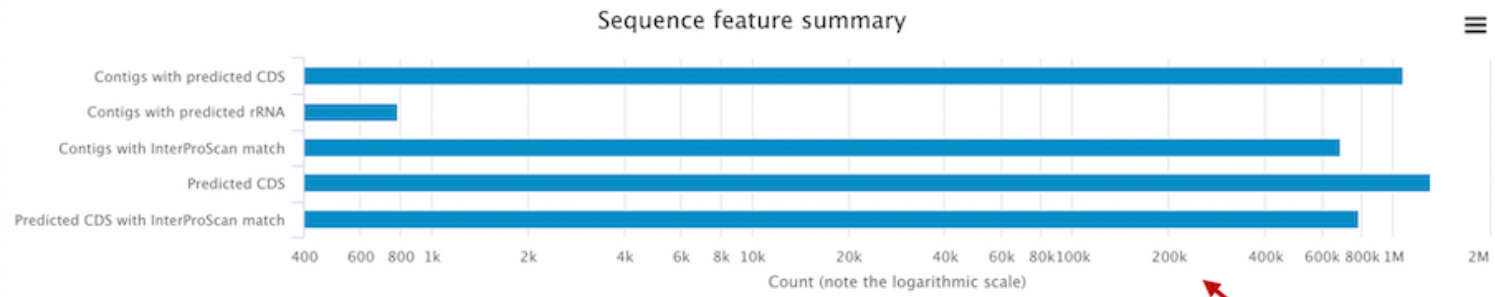
Other analyses

Overview | Quality control | Taxonomic analysis | **Functional analysis** | Pathways/Systems | Contig Viewer | Download

These charts present the functional analysis outputs of our pipeline, which focus on InterPro, Pfam, KEGG orthologue and GO term annotations. These summarise the functional content of the sequences in the sample. The full set of results files can be found under the "Download" tab.

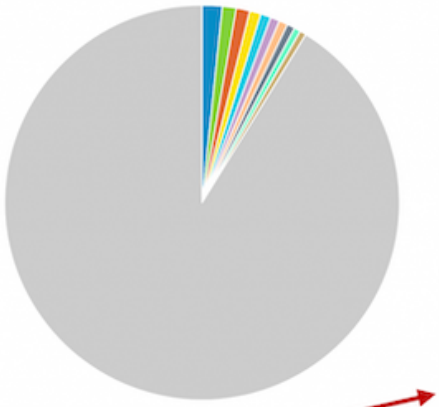
InterPro | GO Terms | Pfam | KO

Tabs for the various functional analyses



Sequence stats

InterPro match summary



InterPro results

Show: 25 results from a total of 9887 (page 1 of 396).

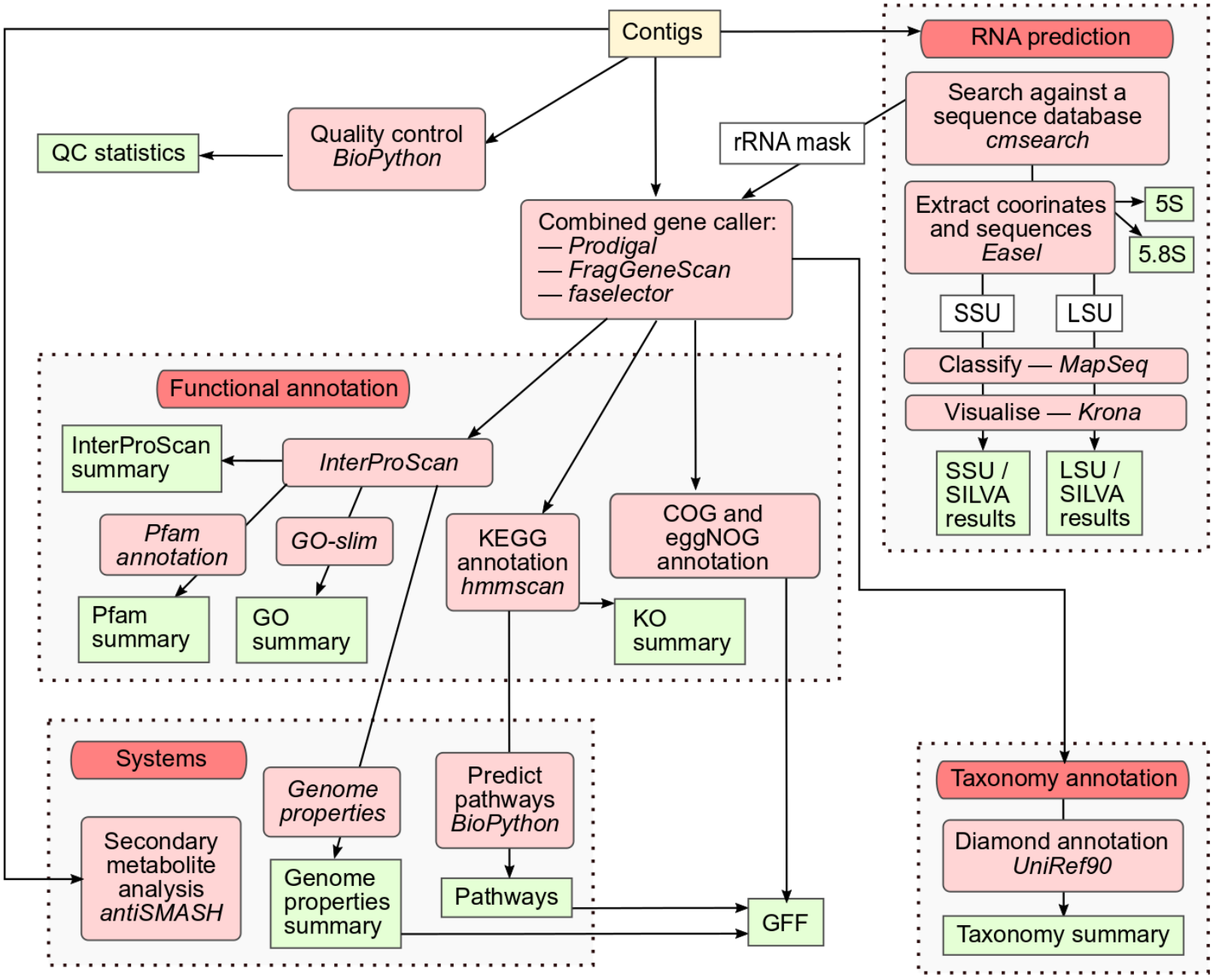
Download results

	Entry name	ID	pCDS matched	%
1	Aldolase-type TIM barrel	IPR013785	18415	1.65
2	FAD/NAD(P)-binding domain superfamily	IPR036188	12737	1.14
3	Rossmann-like alpha/beta/alpha sandwich fold	IPR014729	12327	1.10
4	Pyridoxal phosphate-dependent transferase, major region, subdomain 1	IPR015421	10430	0.93
5	ABC transporter-like	IPR003439	9415	0.84
6	Nucleotide-diphospho-sugar transferases	IPR029044	8346	0.75
7	Alpha/Beta hydrolase fold	IPR029058	8022	0.72

Mgnify : assembly analysis pipeline

- Assembly
  - submission of raw reads (with host sequences removed) to ENA
  - quality control + additional host contamination removal process
  - assembly with metaSPAdes (paired reads) or SPAdes (single reads)

- Contig analysis : assembly pipeline


Contigs
RNA prediction
Search against a sequence database *cmsearch*
rRNA mask
QC statistics
Quality control *BioPython*
Combined gene caller:
— *Prodigal*
— *FragGeneScan*
— *faselector*
Extract coorinates and sequences *Easel*
5S
5.8S
SSU
LSU
Classify — *MapSeq*
Visualise — *Krona*
SSU / SILVA results
LSU / SILVA results
Functional annotation
InterProScan summary
InterProScan
Pfam annotation
GO-slim
KEGG annotation *hmmscan*
COG and eggNOG annotation
Pfam summary
GO summary
KO summary
Systems
Genome properties
Predict pathways *BioPython*
Secondary metabolite analysis *antiSMASH*
Genome properties summary
Pathways
GFF
Taxonomy annotation
Diamond annotation *UniRef90*
Taxonomy summary


Contig Browser
IGV
ERZ782894.1-NODE-1-length
92 kb
Cursor Guide
Center Line
Track Labels
Save SVG
Legend
40 kb
50 kb
60 kb
70 kb
80 kb
90 kb
100 kb
110 kb
120 kb
Functional annotation
antiSMASH
Display options
Assembly contigs
Contig length (bp)
500
188460
COG Category
C
KEGG ortholog
K00161
GO
GO:1901575
Pfam
PF02086
InterPro
IPR015200
antiSMASH
terpene
Show contigs with:
COG
KEGG ortholog
GO
Pfam
InterPro
antiSMASH
Contigs
Filter rows:
Clear
Show: 25 results from a total of 19625 (page 1 of 785).
C COG K KEGG ortholog G GO P Pfam I InterPro A antiSMASH
Download results
Name
Length (bp)
Coverage
Features
ERZ782894.1-NODE-1-length-188460-cov-14.862063 188460 14.862063
ERZ782894.2-NODE-2-length-126717-cov-14.758404 126717 14.758404
ERZ782894.3-NODE-3-length-109593-cov-14.193832 109593 14.193832
ERZ782894.4-NODE-4-length-108695-cov-14.205145 108695 14.205145
ERZ782894.5-NODE-5-length-108166-cov-14.842088 108166 14.842088
ERZ782894.12-NODE-12-length-91999-cov-11.047083 91999 11.047083
ERZ782894.13-NODE-13-length-90972-cov-14.970688 90972 14.970688
86637 10.637477
79538 25.07818
ERZ782894.16-NODE-16-length-77689-cov-15.268890 77689 15.26889
ERZ782894.17-NODE-17-length-77058-cov-11.042258 77058 11.042258
Filter options

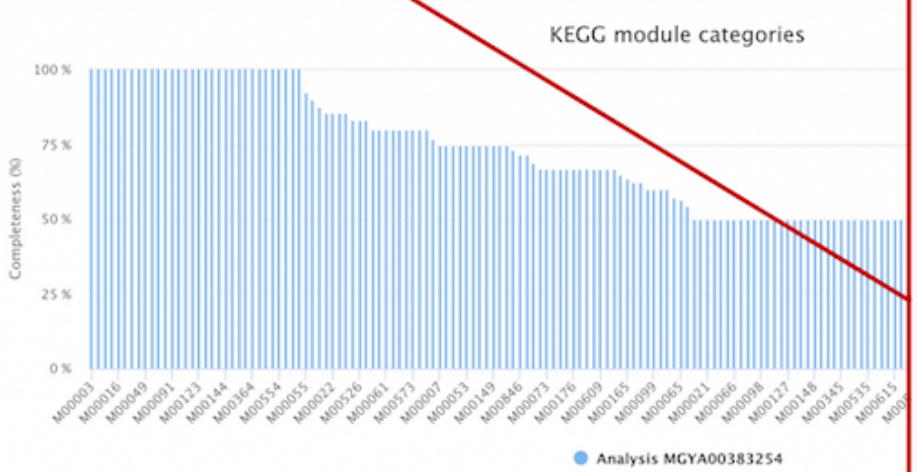
Analysis MGYA00383254

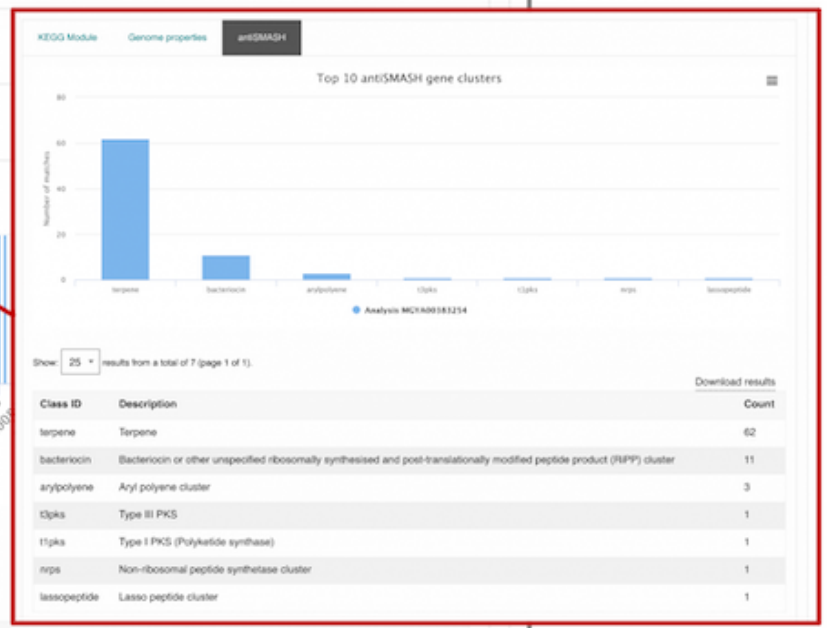
Other analyses

Overview | Quality control | Taxonomic analysis | Functional analysis | **Pathways/Systems** | Contig Viewer | Download

These are the results from the biochemical pathways and systems predictions steps of our pipeline. These summarise the KEGG Module, Genome Properties and antiSMASH annotations in this assembly. The full set of results files may be found under the "Download" tab.

KEGG Module | Genome properties | antiSMASH

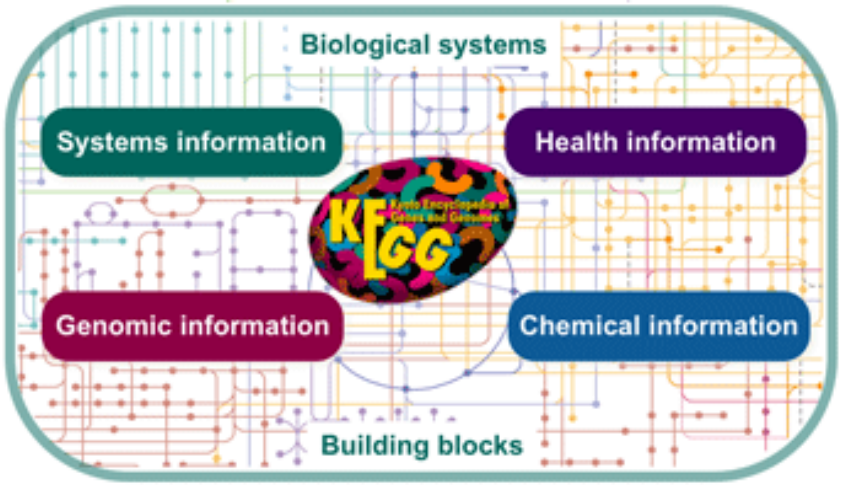




Show: 25 results from a total of 189 (page 1 of 8).

Class ID	Name	Description	Completeness	Matching KO	Missing KO
M00003	Gluconeogenesis, oxaloacetate => fructose-6P	Pathway modules; Carbohydrate metabolism; Central carbohydrate metabolism	100	10	0
M00005	PRPP biosynthesis, ribose 5P => PRPP	Pathway modules; Carbohydrate metabolism; Central carbohydrate metabolism	100	1	0
M00010	Citrate cycle, first carbon oxidation, oxaloacetate => 2-oxoglutarate	Pathway modules; Carbohydrate metabolism; Central carbohydrate metabolism	100	4	0
M00015	Proline biosynthesis, glutamate => proline	Pathway modules; Amino acid metabolism; Arginine and proline metabolism	100	3	0
M00016	Lysine biosynthesis, succinyl-DAP pathway, aspartate => lysine	Pathway modules; Amino acid metabolism; Lysine metabolism	100	9	0

Pathways : KEGG



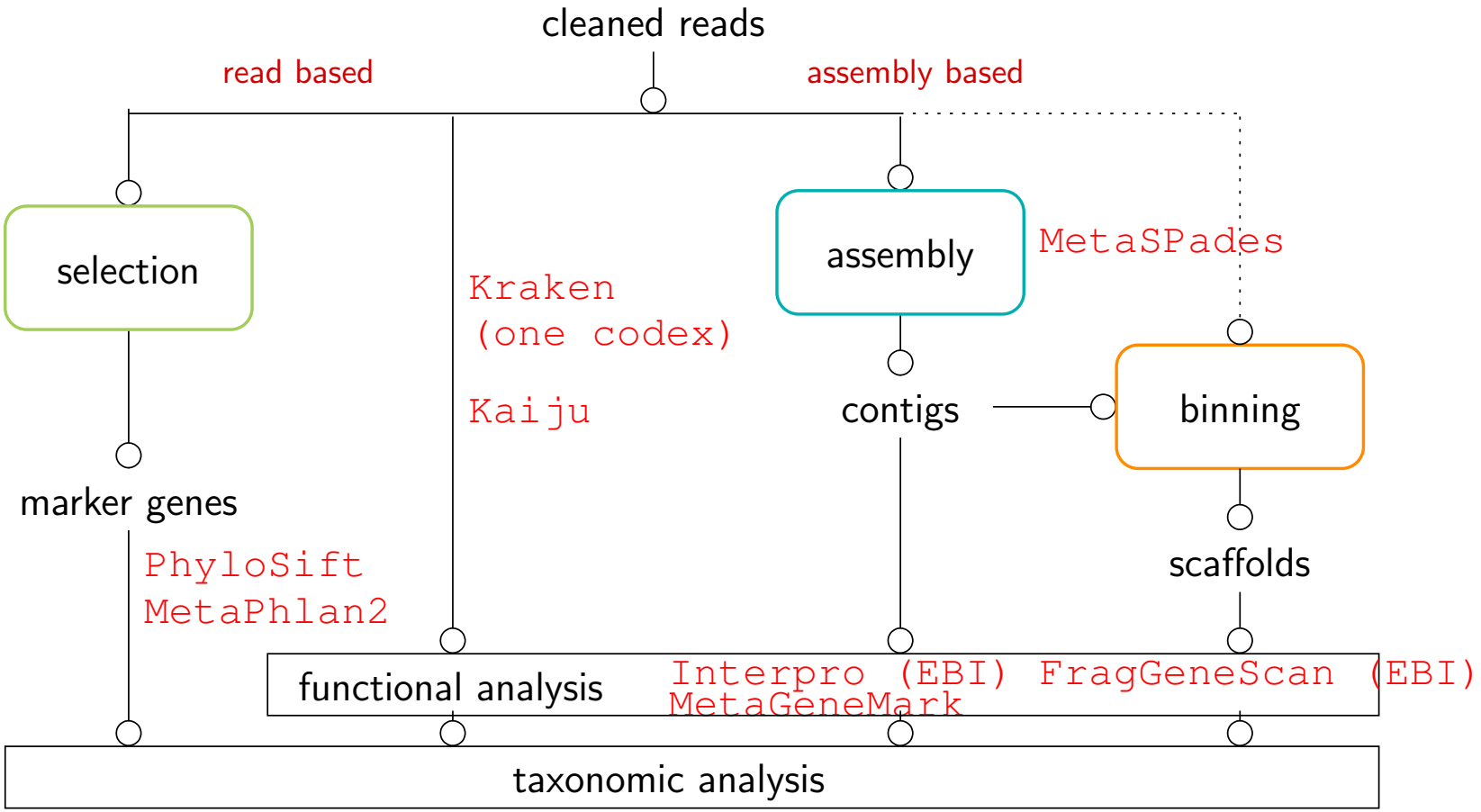
- `http://www.genome.jp/kegg`
- collection of databases : metabolic pathways, genomes, genes, diseases, ...
- KO entries : group of genes representing functional orthologs in the molecular networks
- available in MGnify for assemblies

Data submission : ENA Webin

- data is stably archived
- accession numbers (prerequisite for many publications)
- active submission helpdesk
- training materials

"A major aim in the development of this resource has been to encourage metagenomics researchers to openly share their data as widely as possible, and to also describe their data in sufficient detail such that other scientists are able to extract maximum value from it."

Conclusion


cleaned reads
read based
assembly based
selection
assembly
MetaSPades
Kraken
(one codex)
Kaiju
contigs
binning
marker genes
scaffolds
PhyloSift
MetaPhlan2
functional analysis
Interpro (EBI)
FragGeneScan (EBI)
MetaGeneMark
taxonomic analysis

- fast evolving field
- influence of the nature of the data
  - sequencing technology and quality of the data
  - complexity of the community
  - coverage
- balance between performances and usability

Taxonomic classification

which tool is the best ? with which parameters ?

FEMS Microbiol Ecol. 2016 Jul; 92(7): fiw095.
Published online 2016 May 8. doi: 10.1093/femsec/fiw095

Evaluating techniques for metagenome annotation using simulated sequence data

Richard J. Randle-Boggis,[1,*] Thorunn Helgason,[1] Melanie Sapp,[2] and Peter D. Ashton[1]

2016, MEGAN (older version), MG-RAST, One Codex

Genome Biol. 2017; 18: 182.
Published online 2017 Sep 21. doi: 10.1186/s13059-017-1299-7

Comprehensive benchmarking and ensemble approaches for metagenomic classifiers

Alexa B. R. McIntyre,[1,2,3] Rachid Ounit,[4] Ebrahim Afshinnekoo,[2,3,5] Robert J. Prill,[6] Elizabeth Hénaff,[2,3] Noah Alexander,[2,3] Samuel S. Minot,[7] David Danko,[1,2,3] Jonathan Foox,[2,3] Sofia Ahsanuddin,[2,3] Scott Tighe,[8] Nur A. Hasan,[9,10] Poorani Subramanian,[9] Kelly Moffat,[9] Shawn Levy,[11] Stefano Lonardi,[4] Nick Greenfield,[7] Rita R. Colwell,[9,12] Gail L. Rosen,[✉13] and Christopher E. Mason[✉2,3,14]

2017, 11 tools (including CLARK, Kraken, LMAT, Metaphlan2, PhyloSift, MGAN+Diamond)

Sci Rep. 2016; 6: 19233.
Published online 2016 Jan 18. doi: 10.1038/srep19233

An evaluation of the accuracy and speed of metagenome analysis tools

Stinus Lindgreen,[a,1,2,3,*] Karen L. Adair,[1,2] and Paul P. Gardner[1,2]

2016, 14 tools (including CLARK, MetaPhan2, One codex, EBI, MG-Rast, kraken, LMAT, Megan)

Shotgun sequencing versus amplicon sequencing

Comparing 16S rRNA Marker Gene and Shotgun Metagenomics Datasets in the American Gut Project Using State of the Art Tools, E.R. Hyde, J. Sanders, A. Tripathi, Q. Zhu, R. Knight, 2017

"There is some consistency between the 16S and shotgun metagenomics approaches although some obvious differences are noted."

Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing, Michael Tessler et al., Scientific Reports 2017

"Overall the amplicon data were more robust across both biodiversity and community ecology analyses at different taxonomic scales."

Assembly and binning

CAMI Challenge

- community-driven initiative
- 700 newly sequenced microorganisms and 600 novel viruses and plasmids
- 3 artificial communities
  low, medium, high complexity
  presence of multiple, closely related strains, plasmid and viral sequences and realistic abundance profiles
- assemblers : MEGAHIT, Minia, Meraga, A*, Ray Meta, Velour
- binners : MyCC, MaxBin 2.0, MetaBAT, MetaWatt, CONCOCT2
- `https://data.cami-challenge.org`, `https://data.cami-challenge.org/cami2`

Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software Nature Methods volume 14, pages 1063–1071(2017)