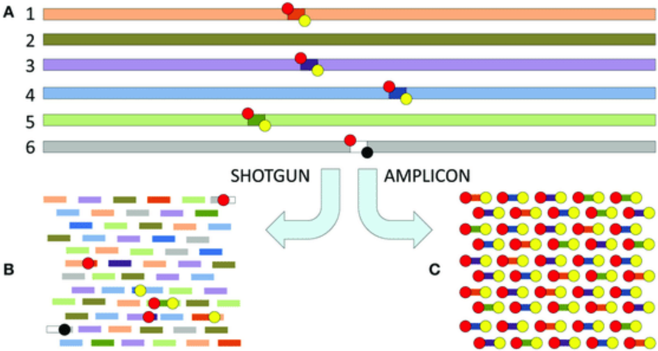
(SHOTGUN) METAGENOMICS

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

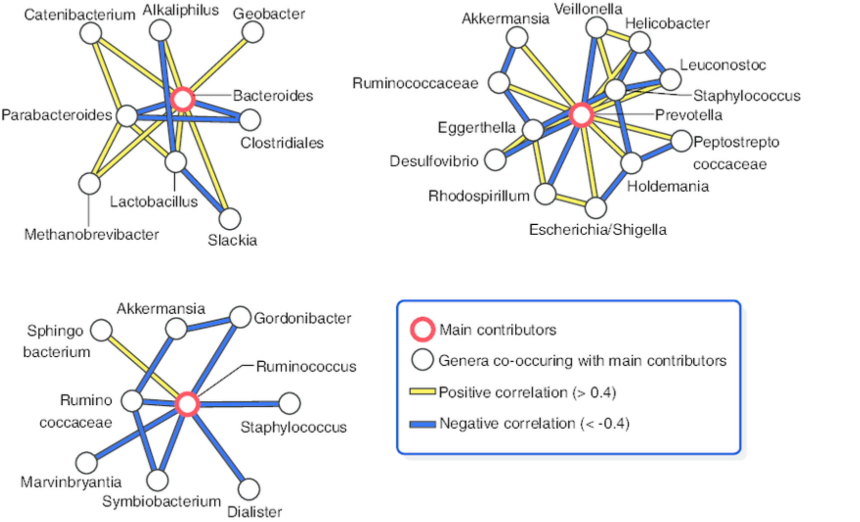
Project MetaHIT (2008-2012)

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 → 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

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

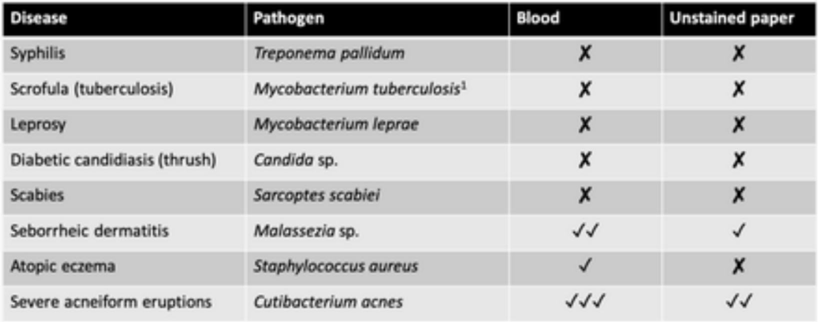
  74,244,610 reads mapped to the human reference genome
  ancestry analysis

  494,378,566 other reads

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

Bioinformatics analysis

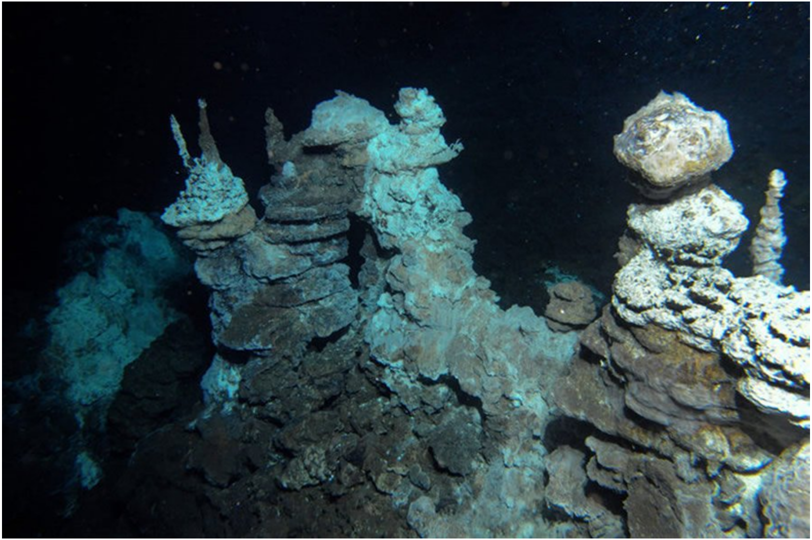
Alignment of reads against database of bacterial genomes



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

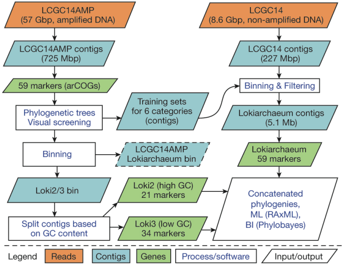


Lokiarchaeota

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


LCGC14AMP (57 Gbp, amplified DNA)
LCGC14 (8.6 Gbp, non-amplified DNA)
LCGC14AMP contigs (725 Mbp)
LCGC14 contigs (227 Mbp)
59 markers (arCOGs)
Binning & Filtering
Phylogenetic trees Visual screening
Training sets for 6 categories (contigs)
Lokiarchaeum contigs (5.1 Mb)
Binning
LCGC14AMP Lokiarchaeum bin
Lokiarchaeum 59 markers
Loki2/3 bin
Loki2 (high GC) 21 markers
Concatenated phylogenies, ML (RAxML), BI (Phylobayes)
Split contigs based on GC content
Loki3 (low GC) 34 markers
Legend
Reads
Contigs
Genes
Process/software
Input/output

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

Shotgun sequencing for community samples

- Metagenomics

  potentially sequences all fragmented DNA in a community

  → includes all microorganisms and viruses

  → gives access to all genes across the entire genomes

- Metatranscriptomics

  potentially sequences all fragmented RNA in a community

  → 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

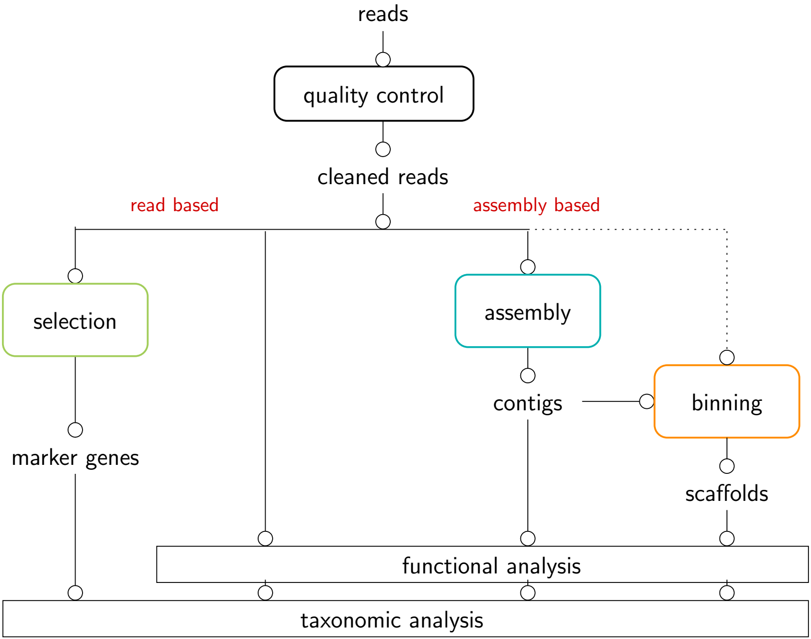
- To assemble, or not

  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

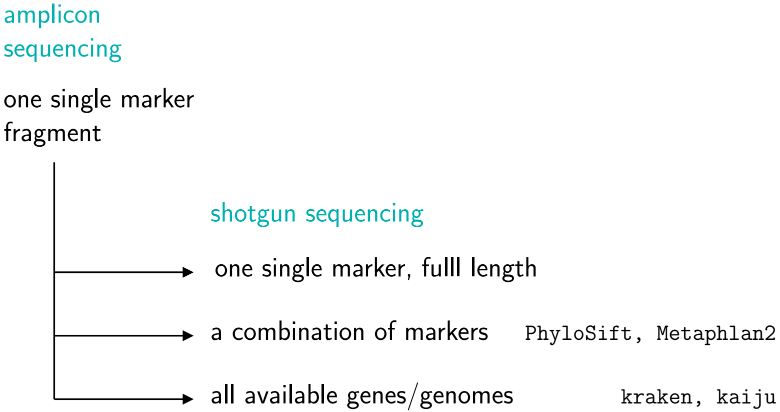
Elements of choice

	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

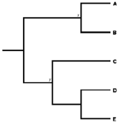
a combination of markers `PhyloSift, Metaphlan2`

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

  `rRNAselector 2011, SortMeRNA 2012`

- processing of the extracted reads

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

  reconstruction of the full sequence of the marker gene before classification : `Emirge 2011, MATAM 2017`

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)

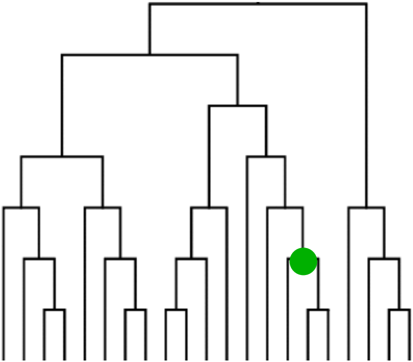
- markers and quasi-markers

  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 database (Bowtie2)
- calculation of the relative abundance of each taxonomic unit

  priority to (strict) markers

  quasi-markers are added only if the number of (strict) markers is < 200

  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_Acidoba
```

```
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 strucural annotation, no phylogenetic markers

- comparison against the database : should be very efficient

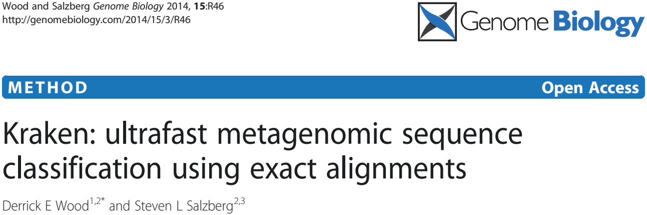
  alignment-free approaches

Kraken

- database : complete bacterial, archaeal, and viral genomes in RefSeq NCBI

- comparison : k-mer composition approach

- classification : discriminative k-mers



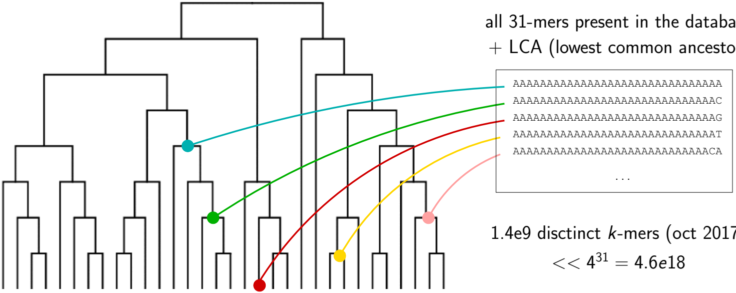
Wood and Salzberg *Genome Biology* 2014, **15**:R46
http://genomebiology.com/2014/15/3/R46

Genome **Biology**

METHOD **Open Access**

Kraken: ultrafast metagenomic sequence classification using exact alignments

Derrick E Wood[1,2*] and Steven L Salzberg[2,3]



all 31-mers present in the database

+ LCA (lowest common ancestor)

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

1.4e9 disctinct k-mers (oct 2017)

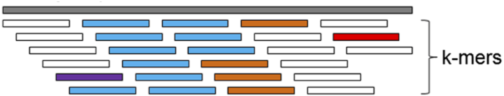
$<< 4^{31} = 4.6e18$

all completed microbial genomes of the RefSeq database

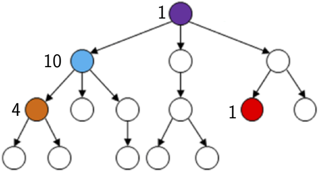
47,768 bacteria + 1,034 archaea + 7,530 viruses

Precomputed database

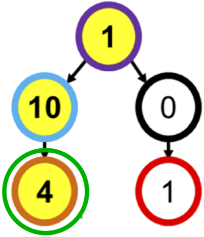
1. short read → overlapping k-mers



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



3. assignation of the read



Read assignation

Performances of Kraken

- very fast
- excellent results with known/poor results with unknow species
- high memory demanding

  500 GB of disk space to build the database 200 GB to store it
- Minikraken : reduced databases

  DB 4GB : 2.7% of k-mers from the original database DB 8GB : 5% of k-mers from the original database
- Centrifuge : space-efficient evolution of Kraken

  Burrows-Wheeler Transform

Similar tools following the same paradigm

- LMAT, 2013

  Scalable metagenomic taxonomy classification using a reference genome database. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. Bioinformatics

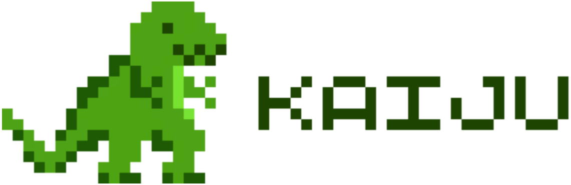
- Clark, 2015

  CLARK : fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers R. Ounit, S. Wanamaker, T.J. Close, S. Lonardi BMC Genomics. 2015 ; 16(1) : 236

- One codex (commercial, free demo version)

  web server based on kraken algorithm, registration required

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

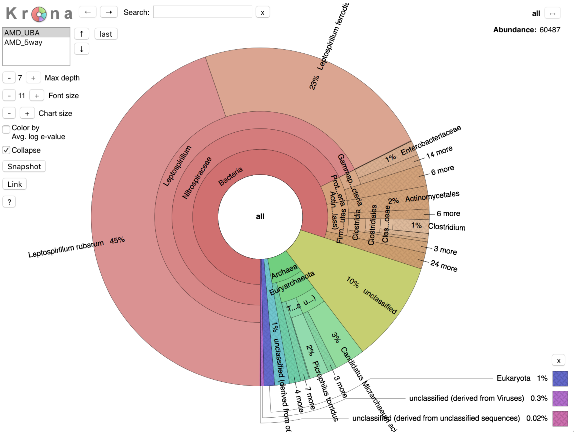
  NCBI RefSeq, proGenomes, non-redundant BLAST protein database (optionally also including fungi and microbial eukaryotes)

- comparison between the reads and the database

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

- classification

Visualisation – krona chart



Search:

AMD_UBA
AMD_5way

7 Max depth

11 Font size

☑ Collapse

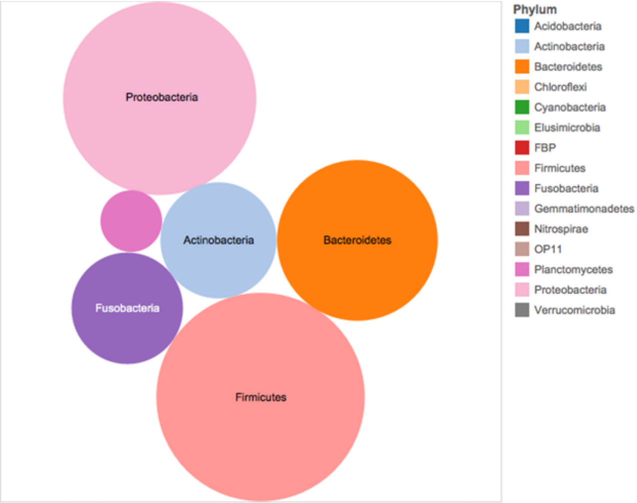
□ Color by Avg. log e-value

all

Abundance: 60487

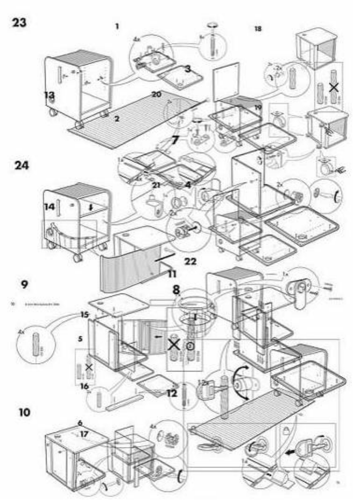
Ondov BD, Bergman NH, and Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12(1) :385, 2011

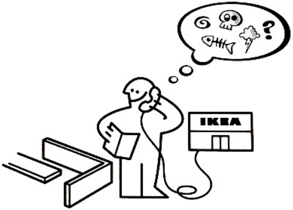
Vizualisation – bubble plot

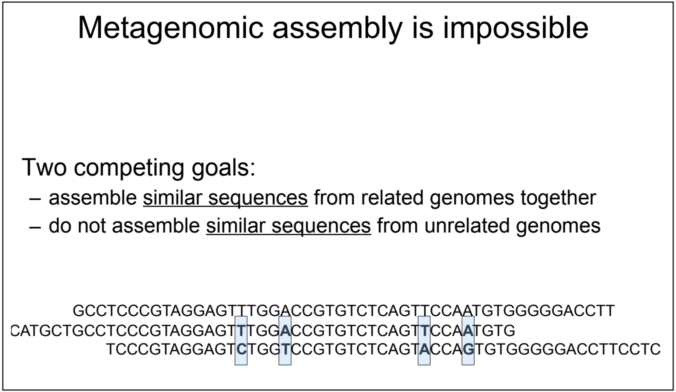


Getting Started
What do you know?
Use a Pedigree Chart and Family Group Sheets to record known data:
"To-Do" Lists
Make your own forms

Assembly







Metagenomic assembly is impossible

Two competing goals:

- assemble <u>similar sequences</u> from related genomes together
- do not assemble <u>similar sequences</u> from unrelated genomes

```
      GCCTCCCGTAGGAGTTTGGACCGTGTCTCAGTTCCAATGTGGGGGACCTT
CATGCTGCCTCCCGTAGGAGTTTGGACCGTGTCTCAGTTCCAATGTG
         TCCCGTAGGAGTCTGGTCCGTGTCTCAGTACCAGTGTGGGGGACCTTCCTC
```

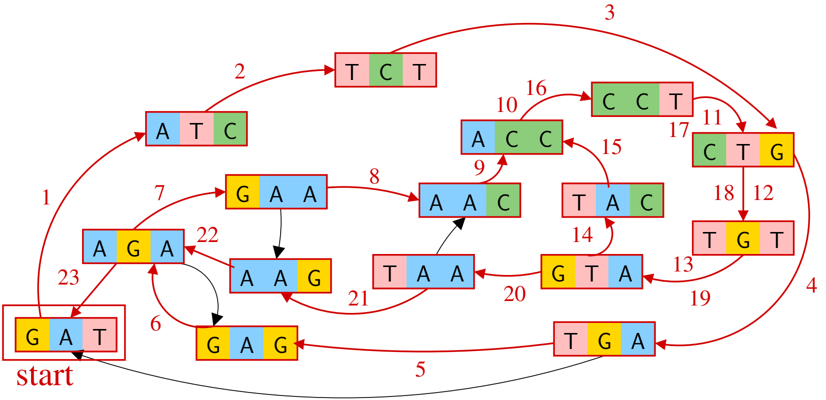
Mihai Pop, Sergey Koren, Dan Sommer

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



R_1 CTGAGAACCTGT CCTGTAAGAT R_2

R_4 GATCTGA R_3 CTGTACCT

GATCTGAGAACCTGTACCTGTAAGAT

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

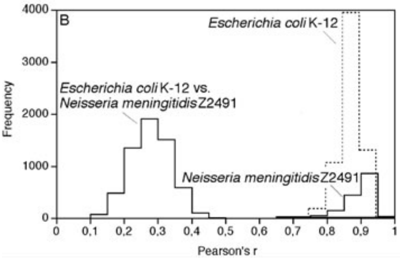
- taxonomy classification
  analogous to read-based approaches
- functional annotation
  this afternoon
- binning

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

Contig coverage

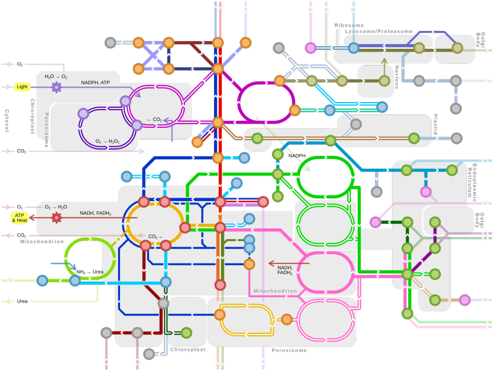
- reads are mapped on the contigs
- similar coverage = similar abundance
- two contigs with similar coverage potentially come from same underlying source population in the community

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

Functional analysis



Functional analysis

- how to annotate genes in genomes ?

Functional analysis

- how to annotate genes in genomes ?
- how to adapt these approaches to metagenomic/metranscriptomic reads/contigs ?

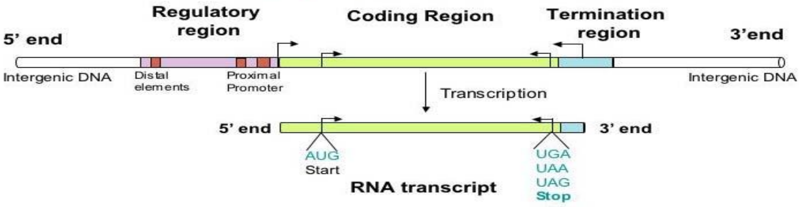
Functional analysis

Three main approaches

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

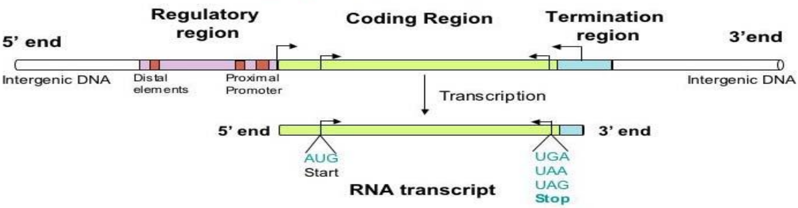
Prediction of coding regions

how can we find genes in prokariotic genomes ?



Prediction of coding regions

how can we find genes in prokariotic 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

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

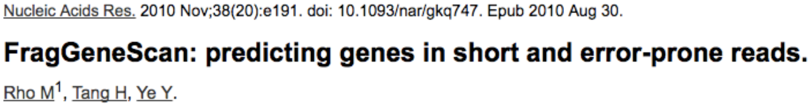
Codon Usage Frequence Table – *E. coli*

1st column : observed frequency

2nd column : theoretical frequency

- short reads : codon usage bias
- contigs : ORF + codon usage bias
- Hidden Markov Models + incomplete ORFs +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

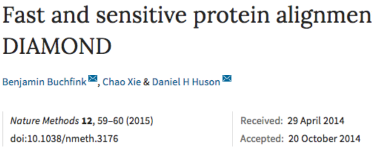
http://exon.gatech.edu/meta_gmhmmp.cgi

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

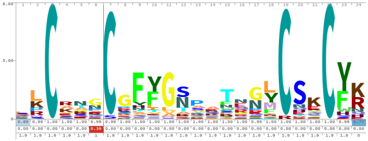
Benjamin Buchfink, Chao Xie & Daniel H Huson

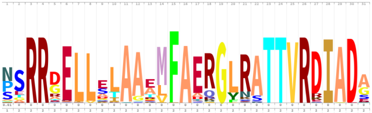
Nature Methods **12**, 59–60 (2015)
doi:10.1038/nmeth.3176

Received: 29 April 2014
Accepted: 20 October 2014

Motif based annotation

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





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

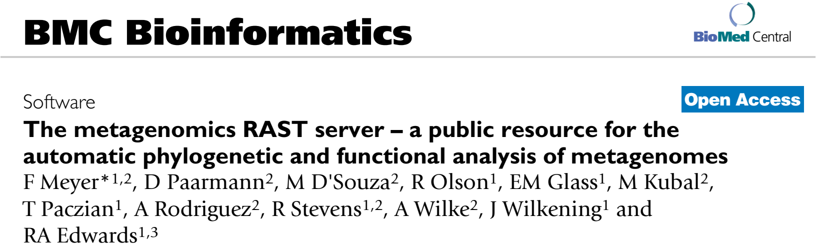
Taxonomic+functional analyses

MG-RAST

Metagenomics Rapid Annotation using Subsystem Technology

MG-RAST

- developed since 2007 (University of Chicago)
- supports amplicons (16S, 18S, and ITS), metagenomics and metatranscriptomics



BMC Bioinformatics

BioMed Central

Software

Open Access

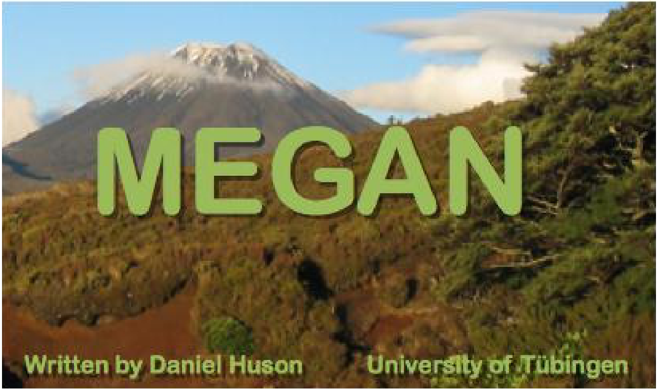
The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes

F Meyer*[1,2], D Paarmann[2], M D'Souza[2], R Olson[1], EM Glass[1], M Kubal[2], T Paczian[1], A Rodriguez[2], R Stevens[1,2], A Wilke[2], J Wilkening[1] and RA Edwards[1,3]

- Cleaning of the sequencing reads
- Taxonomic classification
  - rRNA selection SortmeRNA/Silva
  - RDP classifier
- Functional annotation
  - protein coding gene calling : FragGeneScan (prokaryotes)
  - comparison to GenBank, SEED, Uniprot, KEGG, IMG and eggNOGs with BLAT
- Usage : web interface http://metagenomics.anl.gov
- 315,470 metagenomes containing 1,147 billion sequences and 153.91 Tbp processed for 24,415 registered users.

MEGAN

MEtaGenome ANalyzer



MEGAN

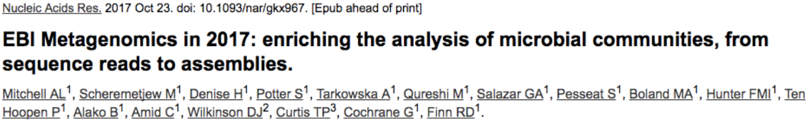
- developed since 2007 (U. Tübingen)
- last release : MEGAN CE, 2017
- Databases : NCBI nr + NCBI taxonomy
- Alignment of the reads on the database : Diamond
- Taxonomic classification : LCA, lowest common ancestor against NCBI nr
- Functional analysis : mapping to KEGG, SEED, EggNOG and InterPro2GO
- local installation

MGnify

EBI metagenomics

Submit, analyse, discover and compare microbiome data

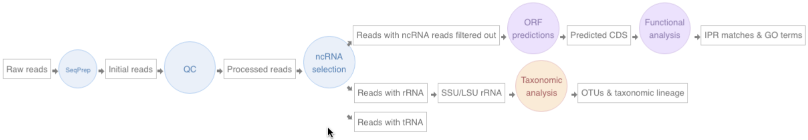
- first public release in 2013
- close integration with the ENA (European Nucleotide Archive)



Nucleic Acids Res. 2017 Oct 23. doi: 10.1093/nar/gkx967. [Epub ahead of print]

EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies.

Mitchell AL[1], Scheremetjew M[1], Denise H[1], Potter S[1], Tarkowska A[1], Qureshi M[1], Salazar GA[1], Pesseat S[1], Boland MA[1], Hunter FMI[1], Ten Hoopen P[1], Alako B[1], Amid C[1], Wilkinson DJ[2], Curtis TP[3], Cochrane G[1], Finn RD[1].



Version 4.1 (jan 2018)

- cleaning and trimming of the short reads : `Trimmomatic`
- identification of ncRNAs : `infernal`
- taxonomic analysis
  Mapseq on 16S and 18S rRNA reads (SILVA database)
- functional analysis
  gene finding : `FragGeneScan` + `Prodigal`
  annotation : `InterPro` + `InterProScan` + `InterPro2GO`

Data submission : ENA Webin

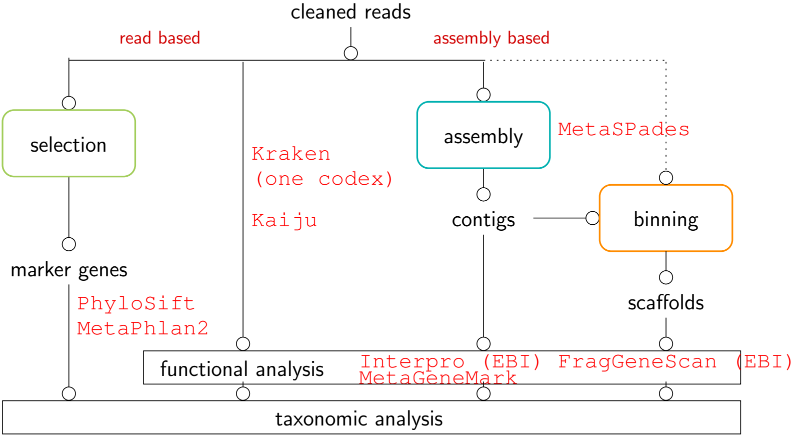
- data is stably archived
- accession numbers (prerequisite for many publications)
- active submission helpdesk
- training materials

"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."

Usage

- webserver : `https://www.ebi.ac.uk/metagenomics`
  upload of the data, analyses on the cloud
- programmatic access : REST API
- krona visualisation
- maybe very slow (several days)

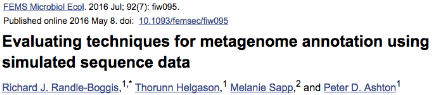
Conclusion


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

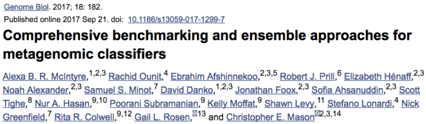


FEMS Microbiol Ecol. 2016 Jul; 92(7): fiw095.
Published online 2016 May 8. doi: 10.1093/femsec/fiw095

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

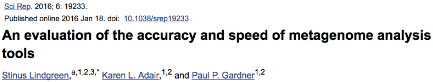


Genome Biol. 2017; 18: 182.
Published online 2017 Sep 21. doi: 10.1186/s13059-017-1299-7

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)