(SHOTGUN)
METAGENOMICS

Hélène Touzet, CNRS, CRIStAL
helene.touzet@univ-lille.fr
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
Shotgun sequencing versus amplicon sequencing

😊 *Who is there?* more complete taxonomic information
no bias due to PCR amplification (primers)
access to the full genomes and genes
captures genomes which lack amplicon targets, such as viruses

😊 *What are they doing?* functional information
analysis of gene functions, metabolic pathways, etc. to
identify the functional potential of the community

😢 more expensive

😢 new challenges in terms of data processing, storage and analysis
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
short reads

individual reads

filtering

16S reads
- cf amplicons

selection of markers
- PhyloSift
- Metaphlan2

all reads
- MG-rast
- MEGAN
- EBI
- kraken
- centrifuge
- ...

assembly

one genome per organism

possible / not possible?

this afternoon

---

- yellow: taxonomic classification
- green: functional analysis
Taxonomic classification

Input

short reads from a single shotgun metagenomic sequencing experiment (FASTA or FASTQ)

Output

list of detected microbes and their relative abundances
Three main approaches

1. focus on a single phylogenetic marker
2. utilization of multiple markers
3. whole information contained in the reads
Approach 1 : One marker

• choice of the phylogenetic marker
  ○ ubiquitous in the environment
  ○ showing some differences between species

• identification of the reads corresponding to the marker

• processing of the extracted reads
  ○ direct classification of the raw reads
    Qiime, 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?

- all possible genes for all possible genomes
  - redundant information
  - high running-time due to the size of the database
    → not such a good idea

- selection of a few universal phylogenetics markers
  phylosift

- Selection of clade-specific marker sequences
  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


**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. Eisen3,4
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


MetaPhlAn2 for enhanced metagenomic taxonomic profiling.
Truong DT¹, Franzosa EA²,³, Tickle TL²,³, Scholz M¹, Weingart C², Pasolli E¹, Tett A¹, Huttenhower C²,³, Segata N¹.
A map of diversity in the human microbiome

Lactobacillus species (L. gasseri, L. jensenii, L. crispatus, L. iners) are predominant but mutually exclusive in the vagina.

Streptococcus epidermidis colonizes external body sites.

Propionibacterium acnes lives on the skin and nose of most people.

Streptococcus dominates the oral cavity with S. mitis > 75% in the cheek.

Many Corynebacterium species characterize different body sites: C. matruchoti the plaque C. accolens the nose C. croppenstedtii the skin.

Several Prevotella species are present in the gastrointestinal tract. P. copii is present in 19% of the subjects and dominates the intestinal flora when present.

Bacteroides is the most abundant genus in the gut of almost all healthy subjects.

Campylobacter includes opportunistic pathogens, but members live in the oral cavities of most healthy people in the cohort.

E. coli is present in the gut of the majority of healthy subjects but at very low abundance.

Commensal microbes

Potential pathogens

The four most abundant phyla
- Actinobacteria
- Bacteroidetes
- Firmicutes
- Proteobacteria

Low abundance phyla
- Chloroflexi
- Cyanobacteria
- Euryarchaeota
- Tenericutes
- Thermi
- Verrucomicrobia

National Institutes of Health
Human Microbiome Project

Intensity of external colors denotes species prevalence in each body site.

Bar lengths indicate microbial abundance (colored by body site of greatest prevalence).
pre-computed database of markers and pseudo markers + clades (LCA in the taxonomy)

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 pseudomarkers from 110 genomes
virus: 38,800 markers + 23,000 pseudomarkers from 3500 genomes
Metaphlan2 – pipeline

1. mapping of short reads on the marker database (Bowtie2)

2. 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

3. 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__Firmicutes 75.07942
k__Bacteria|p__Bacteroidetes 24.06956
k__Bacteria|p__Actinobacteria 0.85102
k__Bacteria|p__Firmicutes|c__Negativicutes 51.02993
k__Bacteria|p__Bacteroidetes|c__Bacteroidia 24.06956
k__Bacteria|p__Firmicutes|c__Bacilli 24.0495
k__Bacteria|p__Actinobacteria|c__Actinobacteria 0.85102
k__Bacteria|p__Firmicutes|c__Negativicutes|o__Selenomonadales 51.02993
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidiales 24.06956
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales 24.0495
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales 0.85102
k__Bacteria|p__Firmicutes|c__Negativicutes|o__Selenomonadales|f__Veillonellaceae
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Prevotellaceae
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales|f__Streptococcaceae
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae
SampleID Metaphlan2_Analysis
k__Bacteria 100.0
k__Bacteria|p__Firmicutes 75.07942
k__Bacteria|p__Bacteroidetes 24.06956
k__Bacteria|p__Actinobacteria 0.85102
k__Bacteria|p__Firmicutes|c__Negativicutes 51.02993
k__Bacteria|p__Bacteroidetes|c__Bacteroidia 24.06956
k__Bacteria|p__Firmicutes|c__Bacilli 24.0495
k__Bacteria|p__Actinobacteria|c__Actinobacteria 0.85102
k__Bacteria|p__Firmicutes|c__Negativicutes|o__Selenomonadales 51.02993
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales 24.06956
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales 24.0495
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales 0.85102
k__Bacteria|p__Firmicutes|c__Negativicutes|o__Selenomonadales|f__Veillonellaceae
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Prevotellaceae
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales|f__Streptococcaceae
k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae

Kingdom|Phylum|Class|Order|Family|Genus|Species|Strain
Installation and usage

- local installation only (python) 😞
  https://bitbucket.org/biobakery/metaphlan2

- tutorial
  https://bitbucket.org/biobakery/biobakery/wiki/metaphlan2

- hclust2: construction of heatmap

- graphlan: visualisation with cladograms
Approach 3 : Raw information

- reference genomes + taxonomy
- no annotation, no phylogenetic markers
- no alignment between the read and the database

```
GTGCATAGTGAAAAACGGAGTTGCTGACGACGAAAGCGACCESSGCTCTGTTGTCATTTGGG
CGAGGCCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
GGATTTAAGTGTAGTAAACTGTGATTAACGGAGACCGGTTTTAAGACAGGAGTTTCGG
CTGTCTGTTGTCATTTGGG
CGAGGCGGACACTGATTGACACCGGTTTTTGACAGAAGGTTAGGGGATAGGTTAAATTG
```
Kraken

- exact k-mer matching
- complete bacterial, archaeal, and viral genomes in RefSeq NCBI

**METHOD** Open Access

**Kraken: ultrafast metagenomic sequence classification using exact alignments**

Derrick E Wood¹,²* and Steven L Salzberg²,³

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

© 2014 Wood and Salzberg; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
all completed microbial genomes of the RefSeq database
47,768 bacteria + 1,034 archaea + 7,530 viruses

Precomputed database

1.4e9 distinct k-mers (oct 2017)
\( \ll 4^{31} \approx 4.6e18 \)

all 31-mers present in the database + LCA (lowest common ancestor)
1. short read $\rightarrow$ 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 unknown 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

Centrifuge: rapid and sensitive classification of metagenomic sequences

Daehwan Kim, Li Song, Florian P. Breitwieser, and Steven L. Salzberg
Installation and usage

- **kraken**: local installation only (C++) 😞
  http://www.ccb.jhu.edu/software/kraken

- **centrifuge**: local installation only (C++) 😞
  http://www.ccb.jhu.edu/software/centrifuge

- command line interface
Similar tools following the same paradigm

- **LMAT, 2013**

- **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, 2017**
  both web server and local installation (CLI)
Functional classification

Input

short reads from a single shotgun metagenomic sequencing experiment (FASTA or FASTQ)

Output

functions of the genes present in the community
How to annotate mixtures of short reads?

• ab initio approaches
  prediction of coding regions in short reads
  FragGeneScan

• homology-based approaches
  alignment of short reads to a large database of annotated sequences
  choice of the alignment tool, DNA /protein: BlastX, Diamond, ...
  choice of the database: EggnoG, SEEDS, KEGG, Interpro, ...

• built-in pipelines: MG-RAST, MEGAN, EBI, ...
FragGeneScan

- predict protein-coding regions from environmental sequences
- HMM combining codon usage bias + start/stop codon models (like Glimmer or Genemark)
  sequencing error models
- included in the MG-Rast and EBI pipelines


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

Rho M¹, Tang H, Ye Y.
Alignment tools

Query : large set of short DNA reads
Reference : large protein database

- Pre-NGS tools : BlastX, BLAT
- Diamond
  - optimized to deal with sort reads
  - order of magnitude faster than BlastX for this kind of data
    ($\times 1000$)

Fast and sensitive protein alignment using DIAMOND

Benjamin Buchfink, Chao Xie & Daniel H Huson

doi:10.1038/nmeth.3176
Received: 29 April 2014
Accepted: 20 October 2014
Interpro

- [http://www.ebi.ac.uk/interpro](http://www.ebi.ac.uk/interpro)
- developed at EBI since 1999
- signatures for protein families, domains and functional sites collected from 14 databases
- mappings of InterPro entries to Gene Ontology (GO) terms (InterPro2GO)

<table>
<thead>
<tr>
<th>Database</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATH-Gene3D</td>
<td>1366</td>
</tr>
<tr>
<td>CDD</td>
<td>2535</td>
</tr>
<tr>
<td>HAMAP</td>
<td>2216</td>
</tr>
<tr>
<td>PANTHER</td>
<td>7367</td>
</tr>
<tr>
<td>Pfam</td>
<td>16109</td>
</tr>
<tr>
<td>PIRSF</td>
<td>3223</td>
</tr>
<tr>
<td>PRINTER</td>
<td>1974</td>
</tr>
<tr>
<td>ProDom</td>
<td>1307</td>
</tr>
<tr>
<td>PROSITE patterns</td>
<td>1289</td>
</tr>
<tr>
<td>PROSITE profiles</td>
<td>1161</td>
</tr>
<tr>
<td>SFLD</td>
<td>143</td>
</tr>
<tr>
<td>SMART</td>
<td>1263</td>
</tr>
<tr>
<td>SUPERFAMILY</td>
<td>1595</td>
</tr>
<tr>
<td>TIGRFAMs</td>
<td>4445</td>
</tr>
</tbody>
</table>
KEGG

- collection of databases: genomes, genes, metabolic pathways, diseases, ...
- KO entries: group of genes representing functional orthologs in the molecular networks

![Diagram of KEGG](image)
EggNOG

- [http://eggnogdb.embl.de](http://eggnogdb.embl.de)
- database of orthologous groups of proteins at different taxonomic levels + functional annotation
- clustering of the 9.6 million proteins from 2031 genomes + 352 viruses → OG (orthologous groups)

eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences

Jaime Huerta-Cepas¹, Damian Szklarczyk²,³, Kristoffer Forslund¹, Helen Cook⁴, Davide Heller²,³, Mathias C. Walter⁵, Thomas Rattei⁶, Daniel R. Mende⁷, Shinichi Sunagawa¹, Michael Kuhn⁸, Lars Juhl Jensen⁴, Christian von Mering²,³,* and Peer Bork¹,⁹,¹⁰,*
SEEDS

- http://theseed.org
- initiated in 2003 for genome annotation
- hierarchical way to organize gene families
- > 1300 protein families
## Pipelines for functional analysis

<table>
<thead>
<tr>
<th></th>
<th>aligner</th>
<th>EggNOG</th>
<th>Interpro</th>
<th>KEGG</th>
<th>SEED</th>
</tr>
</thead>
<tbody>
<tr>
<td>MG-RAST</td>
<td>BLAT</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>MEGAN CE</td>
<td>BlastX</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>Diamond</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EBI</td>
<td>Interproscan</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MG-RAST
Metagenomics Rapid Annotation using Subsystem Technology
MG-RAST

- developed since 2007 (University of Chicago)
- extension of RAST (The Project to Annotate 1000 Genomes)
  Rapid Annotation using Subsystems Technology
- subsystems: SEED framework
- also supports amplicons (16S, 18S, and ITS) and metatranscriptomics

BMC Bioinformatics

Software

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

F Meyer*1,2, D Paarmann2, M D'Souza2, R Olson1, EM Glass1, M Kubal2, T Paczian1, A Rodriguez2, R Stevens1,2, A Wilke2, J Wilkening1 and RA Edwards1,3
• cleaning of the sequencing reads
• rRNA detection: SortmeRNA/ Silva + RDP classifier
• Protein coding gene calling: FragGeneScan (prokaryotes)
• Comparison to GenBank, SEED, IMG, UniProt, KEGG, and eggNOGs with BLAT
Usage

• easy accessibility via a web-based interface
  http://metagenomics.anl.gov

• account identification

• storage on the MG-rast server
  sequencing data + metadata
  status : private, public, ...

• priority queue depending on the status, may be very slow

315,470 metagenomes containing 1,147 billion sequences and
153.91 Tbp processed for 24,415 registered users.
- Green bars, indicating completed tasks, can be expanded via mouseclick
- Blue bars indicate tasks currently being computed on
- Orange bars represent the next tasks to be queued
- Gray tasks are waiting for completion of another task they depend on
- Red bars indicate an error

<table>
<thead>
<tr>
<th>task</th>
<th>status</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>qc_stats</td>
<td></td>
<td>07/12/2017 à 05:34:44</td>
</tr>
<tr>
<td>adapter trim</td>
<td></td>
<td>07/12/2017 à 05:41:10</td>
</tr>
<tr>
<td>preprocess</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>dereplication</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>screen</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>rna detection</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>rna clustering</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>rna sims blat</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>genecalling</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>Task</td>
<td>Date and Time</td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td>aa filtering</td>
<td>07/12/2017 à 06:28:07</td>
<td></td>
</tr>
<tr>
<td>aa clustering</td>
<td>07/12/2017 à 06:36:09</td>
<td></td>
</tr>
<tr>
<td>aa sims blat</td>
<td>07/12/2017 à 11:24:55</td>
<td></td>
</tr>
<tr>
<td>aa sims annotation</td>
<td>07/12/2017 à 11:24:57</td>
<td></td>
</tr>
<tr>
<td>ma sims annotation</td>
<td>07/12/2017 à 06:16:22</td>
<td></td>
</tr>
<tr>
<td>index sim seq</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>md5 abundance</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>lca abundance</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>source abundance</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>dark matter extraction</td>
<td>07/12/2017 à 11:31:20</td>
<td></td>
</tr>
<tr>
<td>abundance cassandra load</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>done stage</td>
<td>(not started)</td>
<td></td>
</tr>
<tr>
<td>notify job completion</td>
<td>(not started)</td>
<td></td>
</tr>
</tbody>
</table>
MEGAN

MEGAN - MEtaGenome ANalyzer
Written by Daniel H. Huson
Original design by D.H. Huson & Stephan C. Schuster
www-ab.informatik.uni-tuebingen.de/software/megan
• developed since 2007 (U. Tübingen)

• last release: MEGAN CE, 2017
This paper introduces MEGAN Community Edition (CE), which is a major update of our MEGAN software. This release contains a large number of new features and has been substantially rewritten so as to support the analysis of many samples (hundreds) and many reads (billions). This release includes a number of command line tools, in particular blast2rma, daa2rma and Meganizer, which can all be used to prepare input files for MEGAN CE. In addition, we recommend the use of DIAMOND for ultra-fast alignment of reads against NCBI-nr and MeganServer to allow web access to MEGAN files.

RMA files
MEGAN CE analysis requires that sequencing reads are first aligned against a suitable reference database, such as NCBI-nr in the case of a protein-based analysis, or Genbank for a DNA-based analysis, or the Silva database, say, when aligning 16S rRNA reads. MEGAN CE can import reads and alignments in a number of different file formats and computes a compressed and indexed binary file in so-called RMA format that contains all reads, alignments, taxonomic and functional classifications. The file is indexed to allow quick access to reads and alignments by taxonomic or functional assignment. MEGAN CE also provides a command line program called blast2rma for computing RMA files from BLAST-like alignments. With MEGAN CE, we introduce a new version of the RMA format that requires much less disk space than previous versions.

Meganizer
Our alignment program DIAMOND produces “diamond files” in a binary output format called DAA (“Diamond alignment archive”), from which both tabular and SAM format can be extracted. We provide a new program called Meganizer that analyses all reads present in a given diamond file, performs taxonomic and functional analysis of them, and then appends the resulting classifications and indices to the end of the diamond file. Meganizing a diamond file takes much less time than generating an RMA file and reduces the number of files that are created.

- Initial database: NCBI nr + NCBI taxonomy
- Alignment of the reads on the database: Diamond
- Exploration of the data (MEGAN CE):
  - Taxonomic binning: LCA, lowest common ancestor
  - Functional analysis: mapping to KEGG, SEED, EggNOG and InterPro2GO
Lowest common ancestor with MEGAN (2007) 
origin of the LCA
Usage

• **diamond**: local installation
  https://ab.inf.uni-tuebingen.de/software/diamond

• **MEGAN CE (Community Edition)**: local installation
  https://github.com/danielhuson/megan-ce
  http://ab.inf.uni-tuebingen.de/software/megan6/
  ◦ JAVA graphical interface
  ◦ multiple viewers: KEGG Viewer (pathways), EggNog viewer,
  ◦ comparative analysis, rarefaction,…

• **MEGAN Ultimate Edition**
EBI Metagenomics
EBI Metagenomics

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


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

Mitchell AL1, Scheremetiew M1, Denise H1, Potter S1, Tarkowska A1, Qureshi M1, Salazar GA1, Pesseat S1, Boland MA1, Hunter FMI1, Ten Hoopen P1, Alako B1, Amid C1, Wilkinson Dj2, Curtis TP3, Cochrane G1, Finn RD1.
EBI Metagenomics pipeline

- cleaning and trimming of the short reads
- taxonomic analysis: Qiime, and then Mapseq on 16S reads
- functional analysis: InterPro + InterProScan + InterPro2GO
Data submission : ENA Webin

- data are 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."
**DISCUSSION**

As the field of metagenomics develops, data volumes grow, and sequence processing and analysis algorithms mature, it is important that analysis pipelines evolve to keep pace. EBI Metagenomics began life several years ago as a small resource, largely devoted to processing Roche 454 metagenomic data. It has grown to become one of the largest metagenomic repositories in the world, supporting the analysis of a range of study types, generated with a variety of different sequencing technologies, from a range of different biomes. The next stage in its development has been to extend and improve its processing and analyses to keep up with progress in the field. At the same time, it has looked to extend access to captured contextual metadata and analysis results, to become more useful to the research community.

To this end, the last two years have seen EBI Metagenomics pipeline changes aimed at updating tools and broadening the scope of analyses. These have included new components to enable eukaryotic taxonomic analysis and better representation of functional annotation using updated reference databases. At the same time, extensive data indexing to support web-based searches and a new RESTful API serving contextual metadata and results have been developed to offer powerful entry points for browsing, searching and discovery of data from both a manual and programmatic perspective.

Amor a fundamental change is the shift towards provision of assembly of both user-submitted and publicly available metagenomic datasets. This represents a new and exciting development for the resource, which will provide the opportunity to compare the differences between raw reads and assembly analysis outputs (a comparison that will require tracking of raw reads to contigs to maintain abundance counts).

Assembly also opens the door to more in-depth functional annotation. For example, provision of full length protein sequences potentially allows their annotation using the complete set of InterPro, some of which are excluded from the current EBI Metagenomics pipeline as they do not perform well at annotating sequence fragments. The ability to annotate at a very deep functional level (e.g. classifying sequences into specific protein subfamilies and identifying precise enzymatic functions) in turn allows more sophisticated analyses, such as reconstruction of precise metabolic pathways within a microbial community at very high resolution. Furthermore, assembly also unlocks the possibility of taxonomic binning, genome reconstruction and annotation, which brings the potential for identification of new organisms.

Such developments will provide new analysis types and entry points to the resource. The first of these is already...
Usage

- **webserver**: [https://www.ebi.ac.uk/metagenomics](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

• 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
Which tool is the best? With which parameters?

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

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

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


"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."