MALT and MEGAN Tutorial

Dr. Laura Weyrich
Laura.weyrich@adelaide.edu.au
Microbiome Center KickStart Bioinformatics Workshop

Both developed by Daniel Huson *et al.* at the University of Tuebingen, Germany

MALT: http://ab.inf.uni-tuebingen.de/data/software/malt/download/welcome.html

MALT Download Page

This is the official download site for MALT (the MEGAN alignment tool).

MALT is a fast replacement for BLASTX, BLASTP and BLASTN, and provides both local and semi-global alignment capabilities. By default, M
addition, MALT can provide alignments in Text, Tab or SAM format.

MALT is an extension of MEGAN.

Program installers:

MALT_macos_0_3_8.dmg (64-bit, MacOS X)
MALT_windows-x64_0_3_8.exe (64-bit, Windows)
MALT_unix_0_3_8.sh (64-bit, Linux, Unix)

Manual
Release notes md5sums

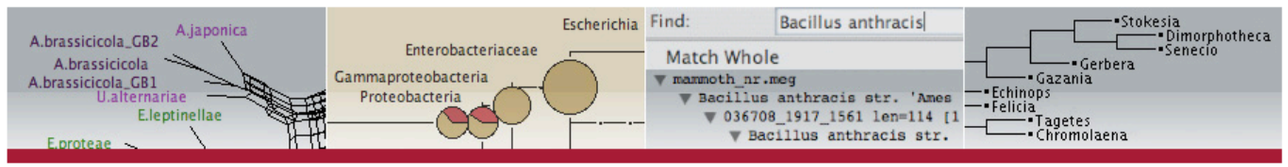
Auxiliary mapping files:

MALT is able to perform taxonomic and functional classification during alignment. This requires the use of the same mapping files that MEGAN download page.

MEGAN5: http://ab.inf.uni-tuebingen.de/software/megan5/

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT
Algorithms in Bioinformatics

Search
A A A
Uni A-Z
Login



HOME | NEWS | PEOPLE | RESEARCH | PUBLICATIONS | SOFTWARE | TALKS | TEACHING | THESES | ADDRESS

Home > Software > MEGAN5

ALGORITHMS IN BIOINFORMATICS

- SplitsTree4
- MEGAN6
- Dendroscope
- DIAMOND
- MEGANServer
- MALT
- Metasim
- **MEGAN5**
- MEGAN4
- SplitsTree3.2
- Crosslink
- Darwin Rocks!
- LOCAS
- NRPSpredictor
- microHARVESTER
- OSLay
- PAT
- PAUDA- High-throughput protein aligner
- 2dTiler
- CrossClassify

MEGAN5 - MEtaGenome ANalyzer

(Download here)

MEGAN5

Please use MEGAN6 - Huson et al, MEGAN Community Edition - Interactive exploration and analysis of large-scale microbiome sequencing data, to appear in: PLoS Computational Biology, 2016.

MEGAN5 was written by D. H. Huson.

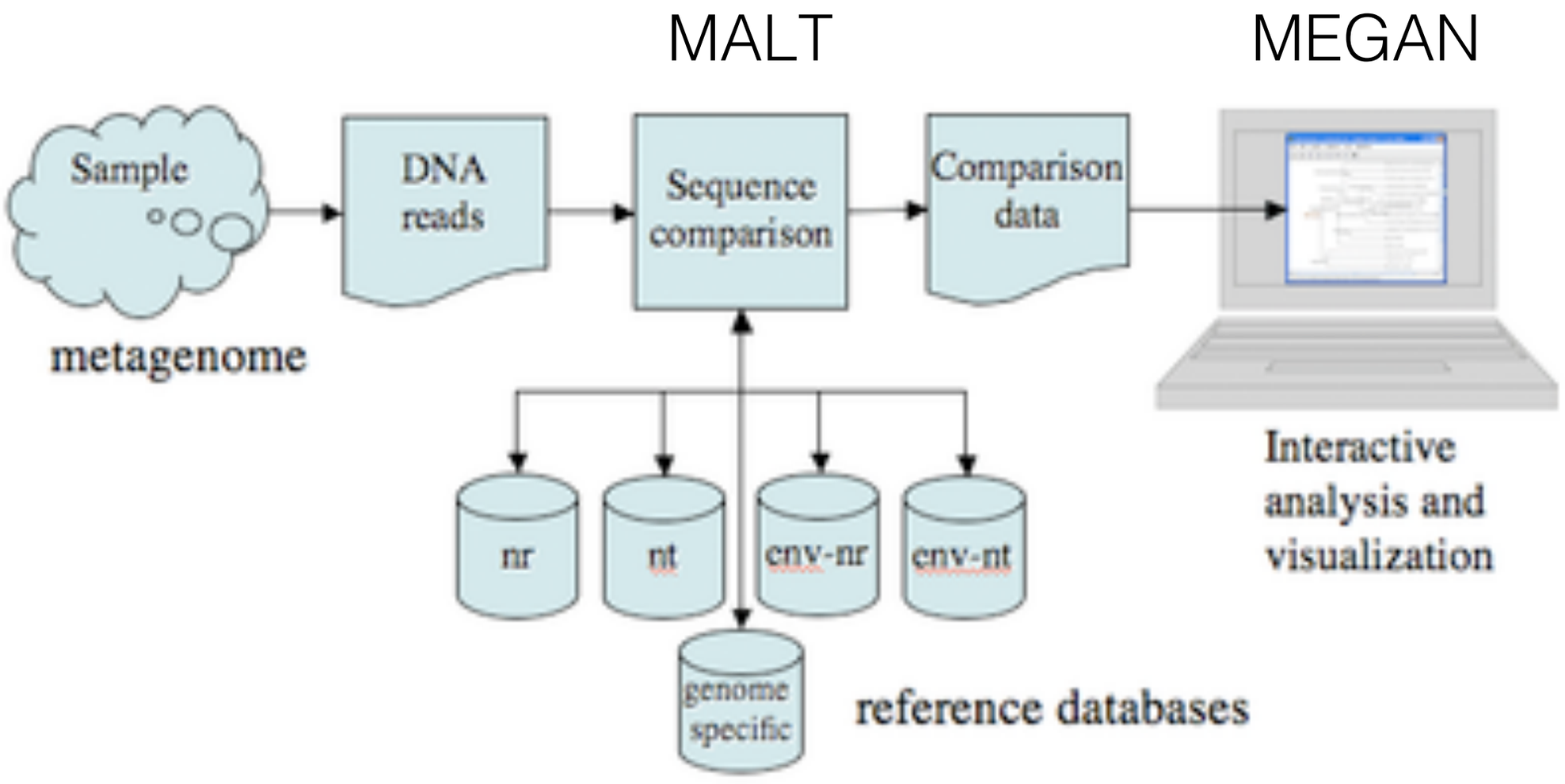
Introduction

In metagenomics, the aim is to understand the composition and operation of complex microbial consortia in environmental samples through sequencing and analysis of their DNA. Similarly, metatranscriptomics and metaproteomics target the RNA and proteins obtained from such samples. Technological advances in next-generation sequencing methods are fueling a rapid increase in the number and scope of environmental sequencing projects. In consequence, there is a dramatic increase in the volume of sequence data to be analyzed.

Basic computational questions

The first three basic computational tasks for such data are taxonomic analysis, functional analysis and comparative analysis. These are also known as the "who is out there?', "what are they doing?" and "how do they compare?" questions. They pose an immense conceptual and computational challenge, and there is a great need for new bioinformatics tools and methods to address them.

Workflow


MALT
MEGAN
Sample
DNA reads
Sequence comparison
Comparison data
metagenome
Interactive analysis and visualization
nr
nt
env-nr
env-nt
genome specific
reference databases

What is MALT?

- **M**EGAN **AL**ignment **T**ool (MALT)
- Nucleotide (MALTn) and protein alignment (MALTx)
- Taxonomic and functional analysis of metagenomes
- A *fast* replacement for BLASTX, BLASP and BLASTN
- Aligns metagenomic reads against a given database of reference sequences (any fasta file, e.g. NCBI nr, GenBank, Greengenes, Silva etc…)
- Uses hash (#) tables to rapidly find information

What is MALT?

Seed and Extend Algorithm

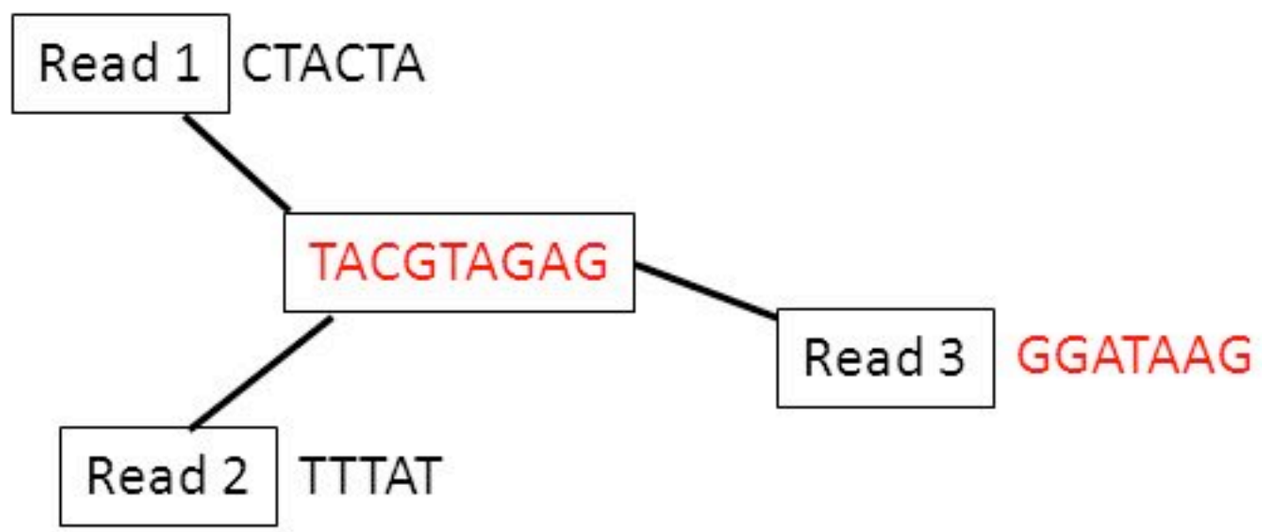
- Reduce reads into overlapping “K”mers

ACGTACGTAGAGGGATAAGATAGAGAGAG
ACGTACGTA AGGGATAAG
CGTACGTAG GGGATAAGA
GTACGTAGA GGATAAGAT
TACGTAGAG GATAAGATA

- Hash the kmers for rapid retrieval

```
for i in kmer_string:
        Hash long = (long << 5) + hash + int_value(i)
```

- Select identical hash hits, and extend read to find best match



What is MALT?

Figure 1 – Schematic overview of MALT. For each preprocessed metagenomic sequencing read, the algorithm generates all contained spaced seeds and looks them up in a hash table of spaced seeds representing the reference database. A banded alignment is calculated for each match of seeds. Once all alignments for a given read have been calculated taxonomic binning of the read is performed using the LCA algorithm.

Table 1 – Runtime comparison of BLAST, *lambda* and MALT on test datasets of different size. Runtimes are given in seconds.

number of reads	runtime BLAST	runtime lambda	runtime MALT
100.000	1.729	1.024	38
1.000.000	16.077	6.414	155
10.000.000	150.045	31.163	1.252

10 millions = 20 minutes
=
25X faster than lamba
120X faster than BLAST

Herbig *et al.,* MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. Biorxiv. 2016

What is MALT?

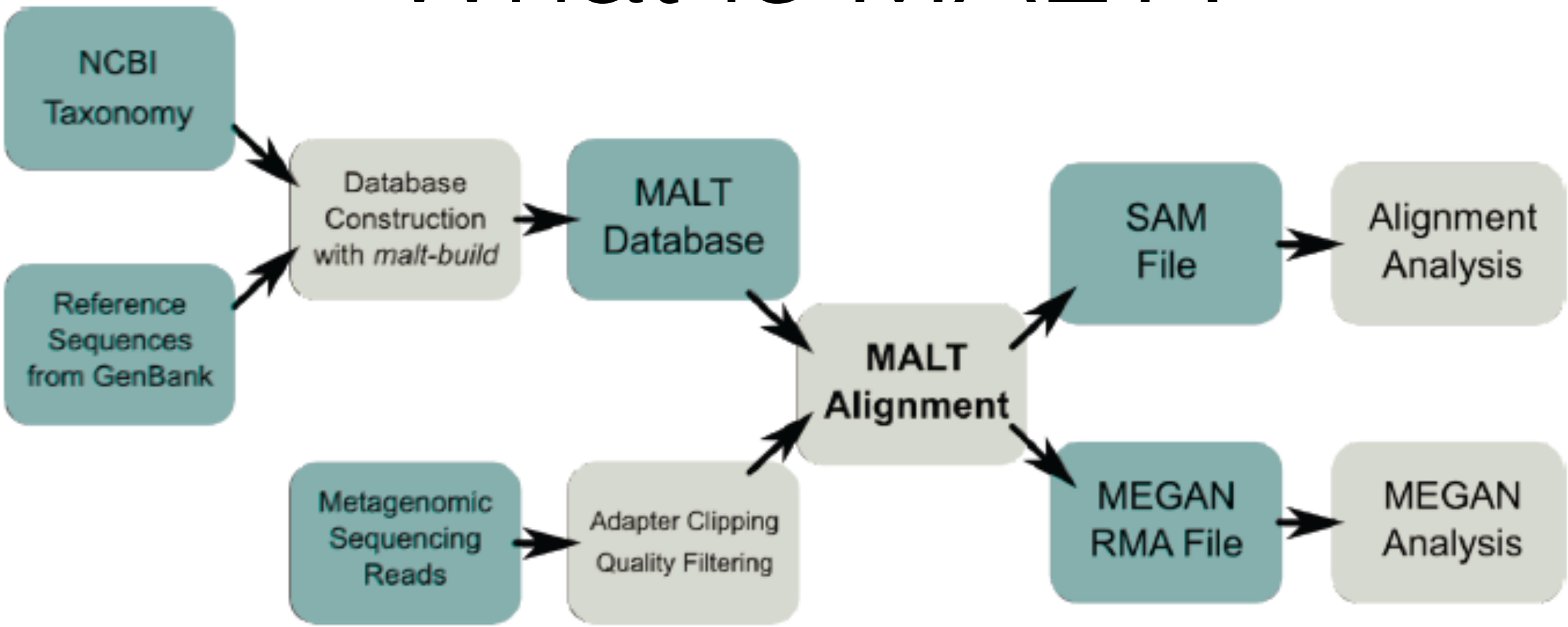


Figure 2 – Schematic overview of our MALT-based analysis workflow. A reference index is generated for all bacterial genomes from GenBank. Preprocessed metagenomic sequencing reads are aligned against the reference sequences. An RMA result file is produced for further analysis in MEGAN. Alignments are also stored in SAM format.

What does MALT use?

- Input: Fasta or fastq file AND indexed reference
  - Sequence reads (trimmed and collapsed)
  - Reference sequences
- Output: RMA (Read Map Alignment) format, or BLAST-Text, BLAST-Tab, SAM…
  - Gene alignments
  - Taxonomic information
  - Functional information

How to use it?

- Two main algorithms:
  - **MALT-build**: Used to build an index for the given database.

```
malt-build -i <References> -d <Index_Name> -L <MEGAN_license> [+auxiliary dbs]
```

  - **MALT-run**: Used to perform alignments and analyses

```
malt-run -i <input-files> -d <Index_Name> -m <alignment-mode> -L <MEGAN_license>
```

malt-build

malt-build -i <References> -d <Index_Name> -L <MEGAN_license> [+auxiliary dbs]

- Uses the input reference files and builds an indexed reference database
- Adds information from auxiliary mapping files:
  - -g2t : GI number to taxon-id mapping
  - -g2k: GI number to Kegg
  - -tre & -map: NCBI taxonomy for taxonomic analysis
  - -r2k, -r2c & r2s: refSeq to Kegg, COGs Db, Seed

malt-run

malt-run -i <input-files> -d <Index_Name> -m <alignment-mode> -[options]

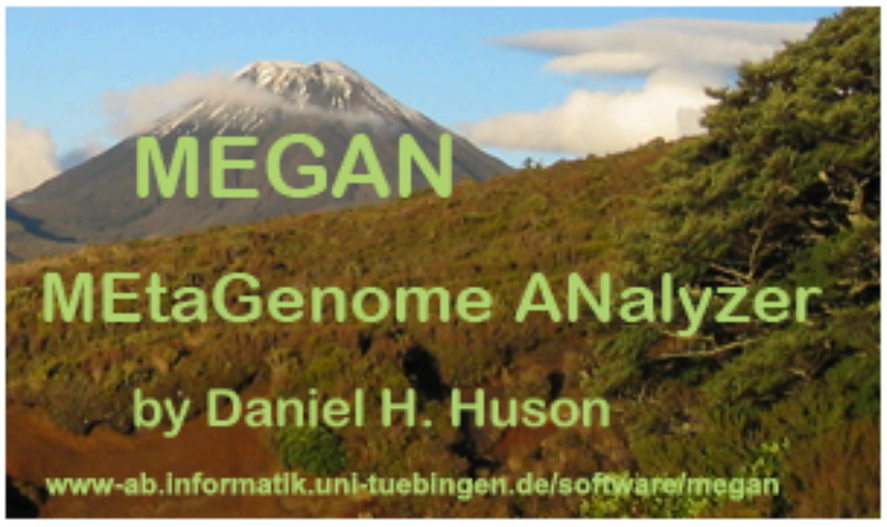
- Uses the indexed reference files and compares sequenced reads to reference using a seed and extend algorithm
- m: Unknown, BlastN, BlastP, BlastX, Classifier
- Takes one file or a directory containing FASTQ or FASTQ.GZ and a MALT-index
- Outputs RMA and SAM files

Why use MALT?

- Advantages
    - Speed
    - Flexibility
- Disadvantages
    - Large amounts of memory needed (250 GB RAM)
    - Command line only

MEGAN

- MEtaGenome ANalysis software
- GUI (and command line) based visualization and analysis program
- Point and click utilization makes it easy for beginners

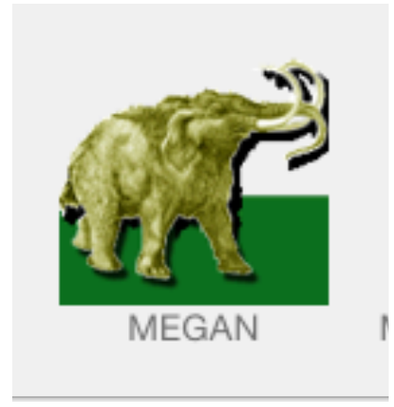


MEGAN

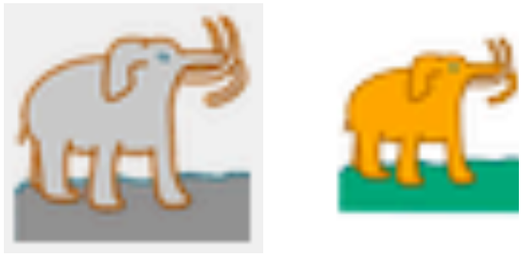
- Taxonomic analysis using the NCBI taxonomy or a customized taxonomy such as SILVA
- Functional analysis using InterPro2GO, SEED, eggNOG or KEGG
- Bar charts, word clouds, tree maps and many other charts
- PCoA, clustering and networks
- Supports metadata
- MEGAN parses many different types of input

Two versions available

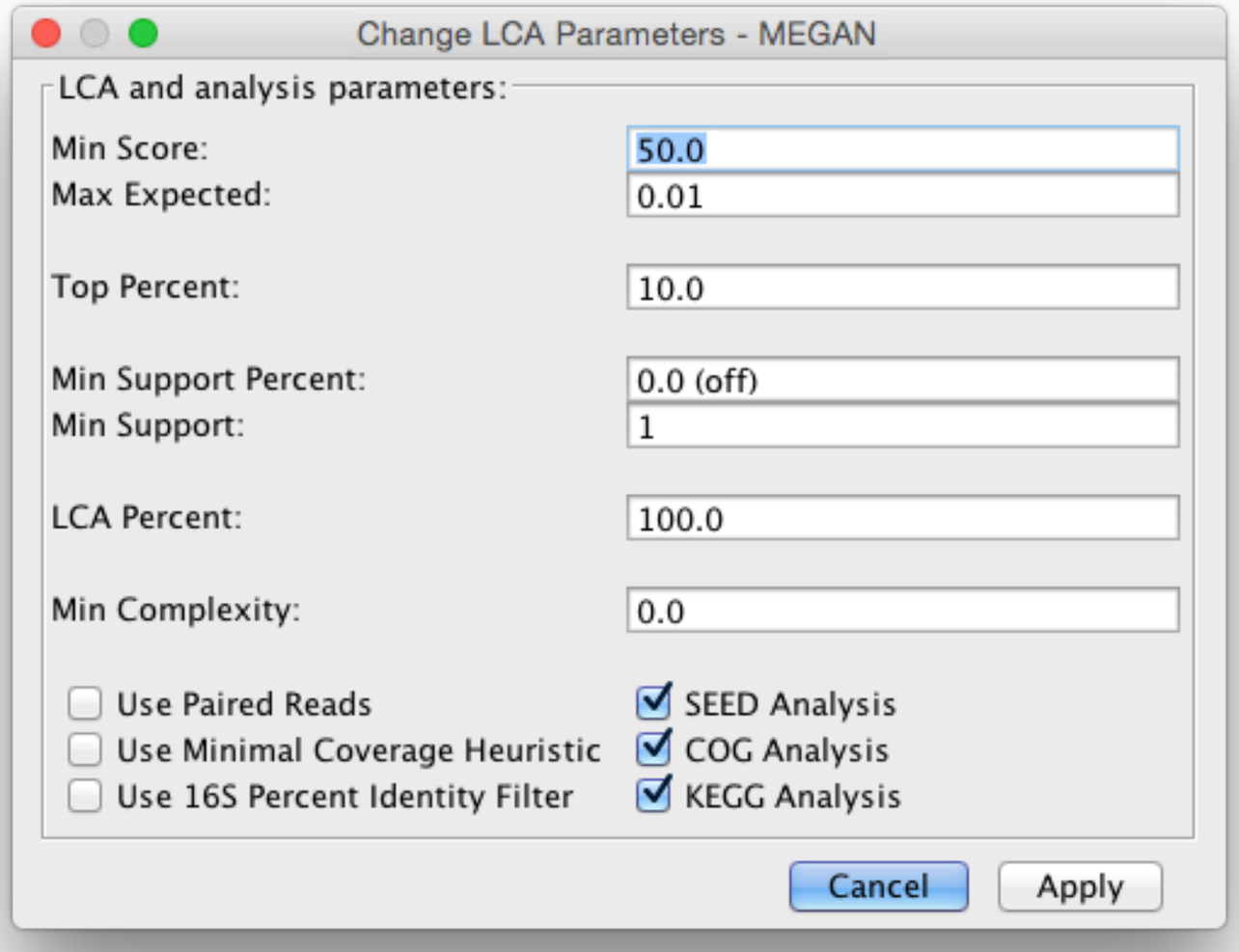
- ~~MEGAN5~~
  - ~~Stable version that is updated~~
  - ~~No background access~~



- MEGAN6CE
  - Open source
  - Community driven



LCA Parameters (least common ancestor)



Min score

Ignore all matches with bit score below this value

Max Percent

Ignore all matches with a expected value (e-value) above this

Top Percent

Must lie within this percentage of the best score to be considered

Min Support Percent

Minimum number of the total percent that a taxa must obtain to be considered

Min Support

Minimum number of reads a taxon must have

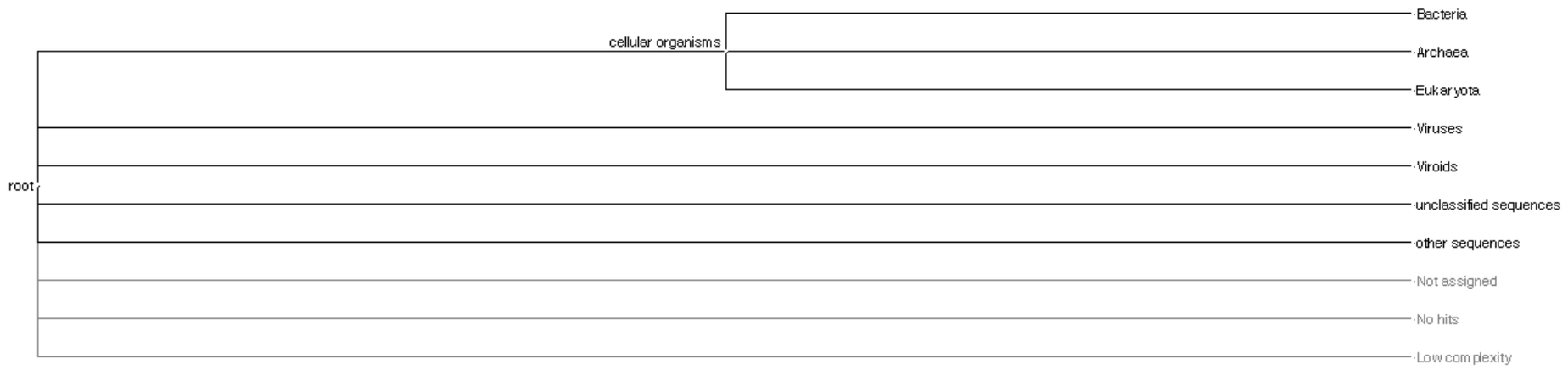
LCA Percent

Percent of matches that a placement of read must cover

Min Complexity

Minimum complexity before a read can be considered non-repetitive

Untitled - MEGAN (version 5.10.3, built 23 Apr 2015) - Licensed to user: Laura Weyrich, Academic institution: University of Adelaide, Email: laura.weyrich@adelaide.edu.au



Messages - MEGAN

```
Executing: new;
Induce Taxonomy tree, keeping 1266115 of 1266115 nodes
Induced taxonomy tree has 12 nodes
Executing: close;
Executing: close;
Executing: close;
Executing: close;
Executing: quit;
Induce Taxonomy tree, keeping 1266115 of 1266115 nodes
Induced taxonomy tree has 12 nodes
```

Messages

Taxa=12 | Taxonomy size=1,266,115, | 2.7 of 3.5G

Why use MEGAN?

- Advantages
    - LCA parameters
    - GUI interface
        - Speed and flexibility
- Disadvantages
    - Only visualization and no stats
    - GUI operations can be difficult to replicate
    - Limited to taxonomic/functional databases in MEGAN

What are we doing today?

- Demo MALT
    - Analyzing viral proteins present in an ancient calculus specimen.
- Demo MEGAN5
    - Working the basics
    - Filtering, analyzing, and comparing ancient and modern dental calculus samples through time.