MEGAN Tutorial

(Version 6 Ultimate Version)

What is MEGAN?
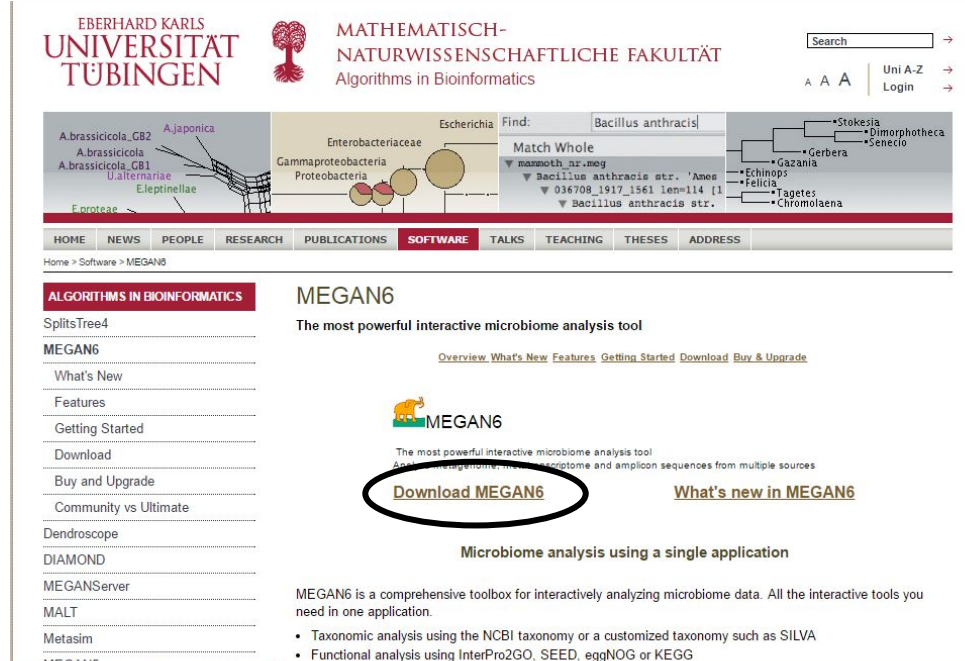
A comprehensive toolbox for analyzing microbiome data (metagenome, metatranscriptome, and amplicon sequences) using a GUI interface. Tools include:

· Taxonomic analysis using the NCBI database or a customized taxonomy (SILVA or greengenes)

· Functional analysis using InterPro2Go, SEED, eggnog or KEGG

· Bar charts, word clouds, trees and multiple other charts

· Distance matrices and PCoA plots and networks

· Supports metadata

· Ability to parse many types of input – Blast, Diamond (DAA), SAM, RDP, Silva, CSV, BIOM, zip/gz files

Why use?

1. Easy GUI interface for fast analyses with menus, toolbars, and graphics...no graphics required.

2. Very powerful. MEGAN6 allows you to work with hundreds of samples containing hundreds of millions of reads. Uses a Blast-like approach and compatible with DIAMOND

3. Comprehensive. Offers a large range of analysis tools and is under active development

4. Has an open source and paid version

Downloading Program

https://ab.inf.uni-tuebingen.de/software/megan6

Importing Data

- Input files can vary – Blast, Diamond (DAA), BIOM, Fasta, SAM, RDP, Silva, CSV, BIOM, zip/gz files

- Metadata friendly allowing for easy manipulation of data with downstream analyses

- Fasta files need to be imported with a blast search result file or a DAA file can be substituted using Diamond to search for blast results

Importing Data
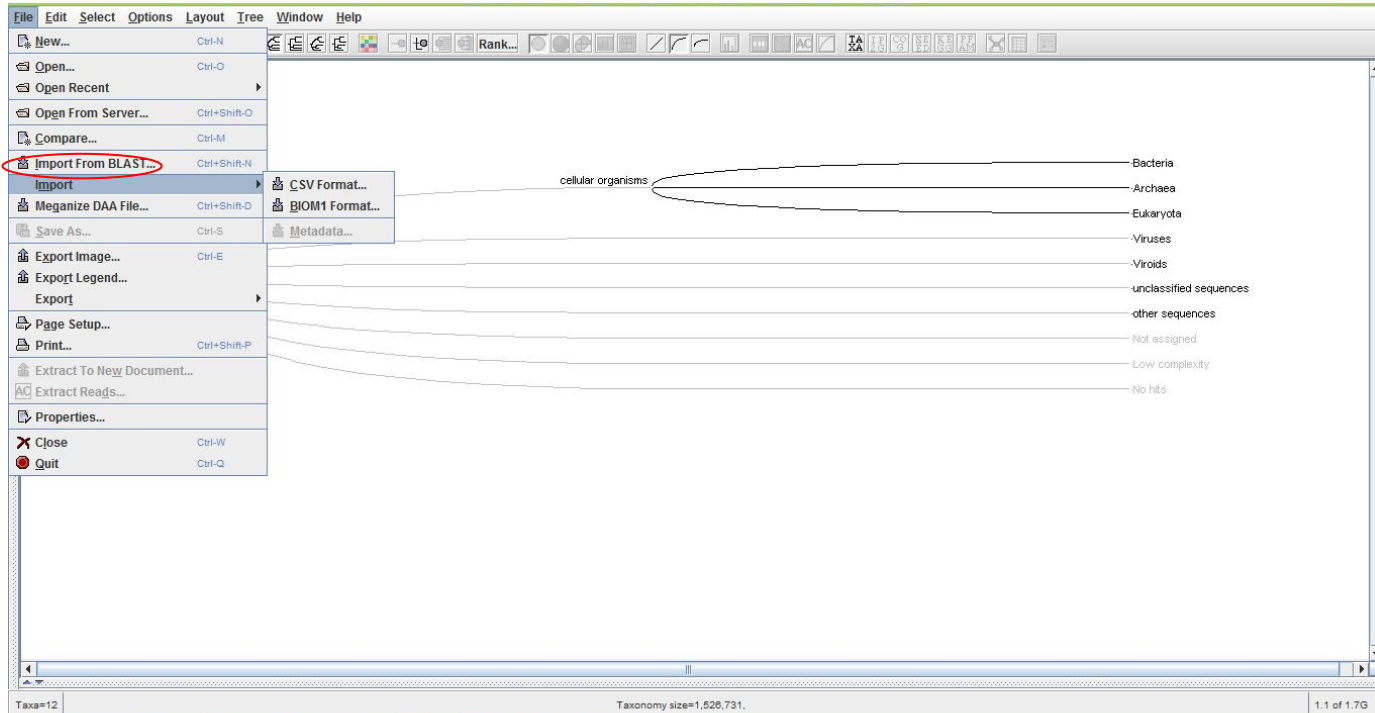
Importing Data from Blast

Blast Search File

Fasta File of your reads

New MEGAN friendly file name
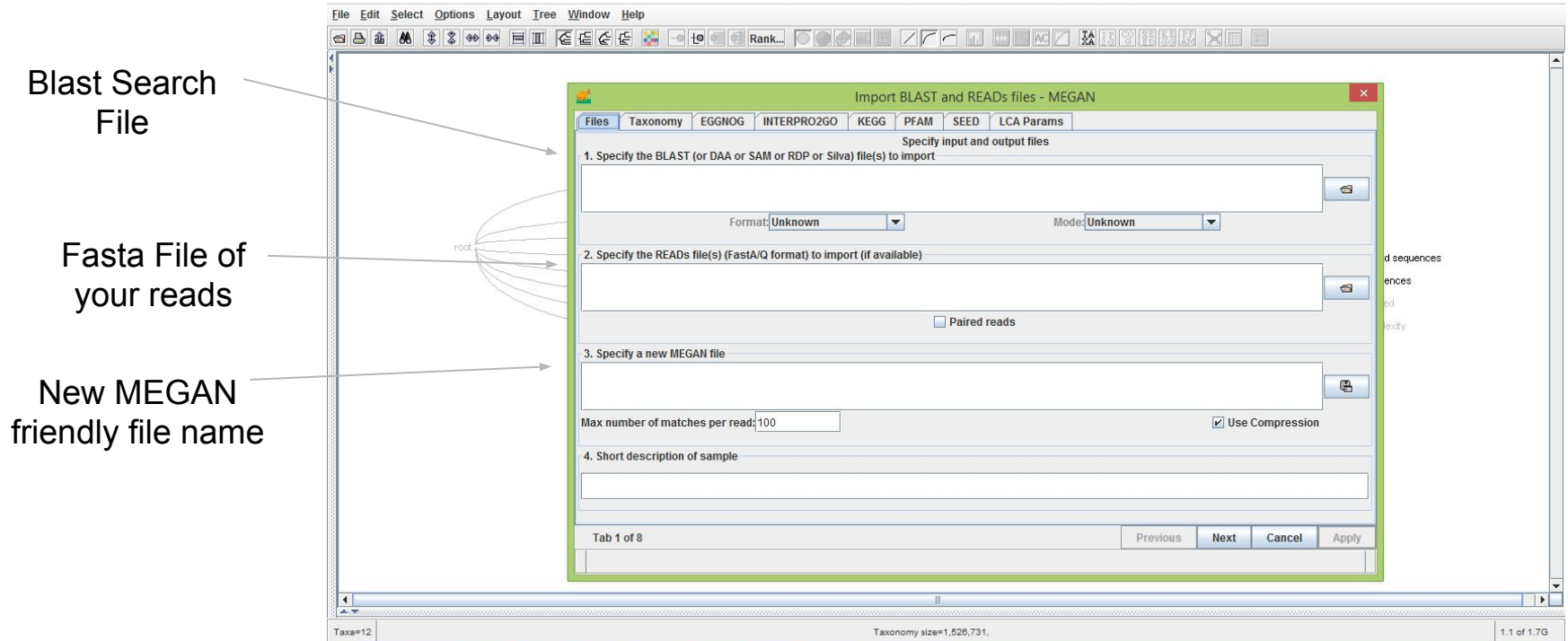
Lowest Common Ancestor (LCA)

- Algorithm for assigning taxonomy to the input reads
- Default parameters - takes the top 2 blast hits and compares the lowest common ancestor (on a phylogenetic tree) and assigns taxonomy. However these settings can be adjusted to help filter low quality results
  - Min Score: refers to the minimum BLAST score required to associate a read with a specific taxa
  - Top Percent: if two reads are found with the same taxonomy are within this percentage of each other, LCA will be calculated just using these two reads
  - Min Support Percent: refers to the minimum number of reads associated with a taxa before LCA is applied (singletons or low frequency reads)
  - Weighted LCA – gives weight to reads on the first round of LCA, then goes back a second time and uses these weighted nodes to assign the reads based on a 75% threshold
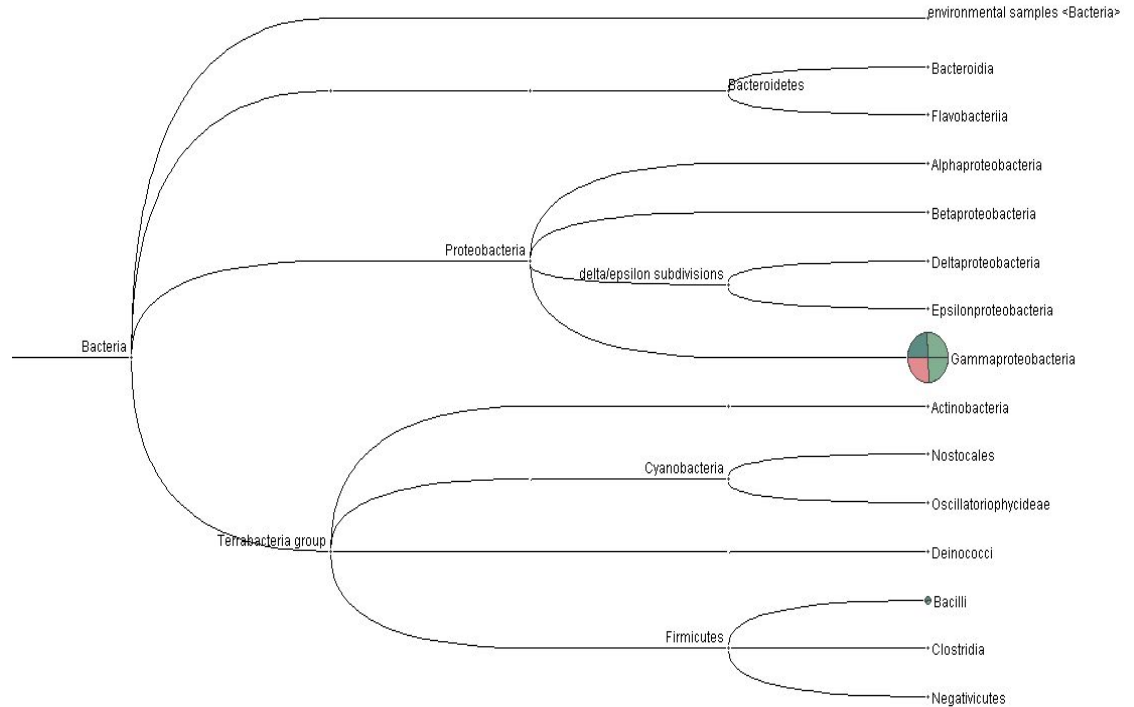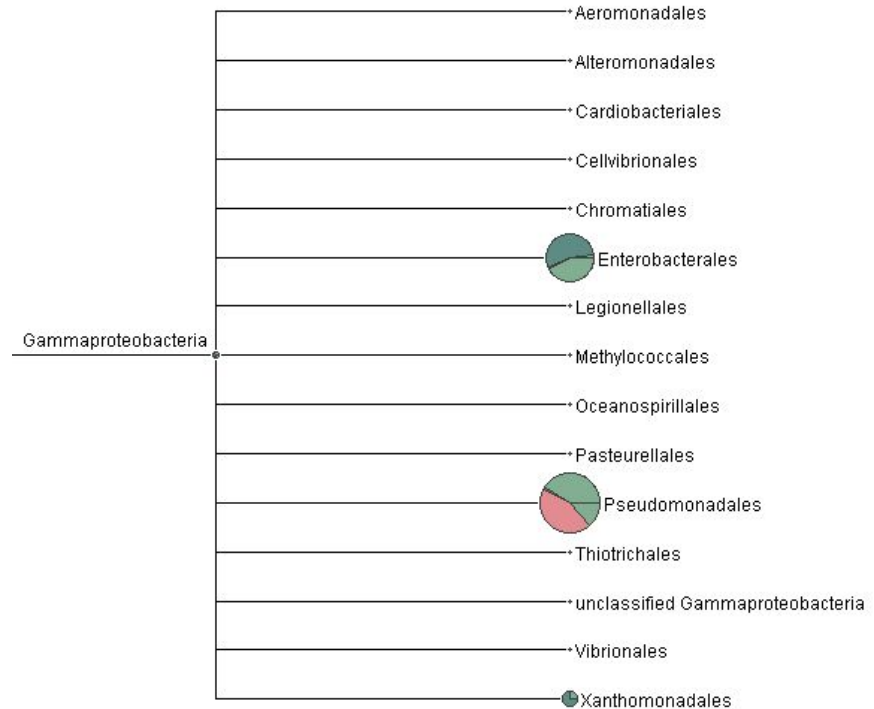
Taxonomic Chart

- MEGAN can also display taxonomic charts
- Gives a good visual overview of the taxonomy of the samples
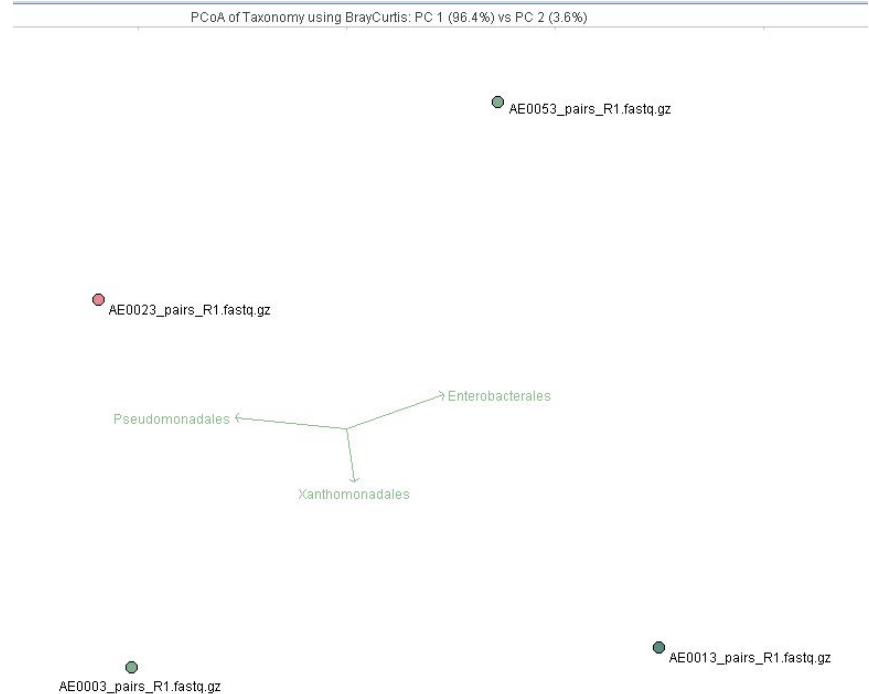 - Pie charts show relative abundances of the given species

Phylogenetic Tree

- MEGAN can also display phylogenetic trees by right clicking a node and selecting "uncollapse", or pressing Ctrl+U

Cluster Analysis

- Cluster Analysis Window allows PCoA plots to be generated
  - Can be manipulated in 3D
- Distance Matrices are also shown in the cluster analysis window
  - Distance formulas such as Euclidean, Bray-Curtis, and others may be used

PCoA of Taxonomy using BrayCurtis: PC 1 (96.4%) vs PC 2 (3.6%)

AE0053_pairs_R1.fastq.gz

AE0023_pairs_R1.fastq.gz

Pseudomonadales ← → Enterobacterales

Xanthomonadales

AE0003_pairs_R1.fastq.gz

AE0013_pairs_R1.fastq.gz

Microbrial Attributes
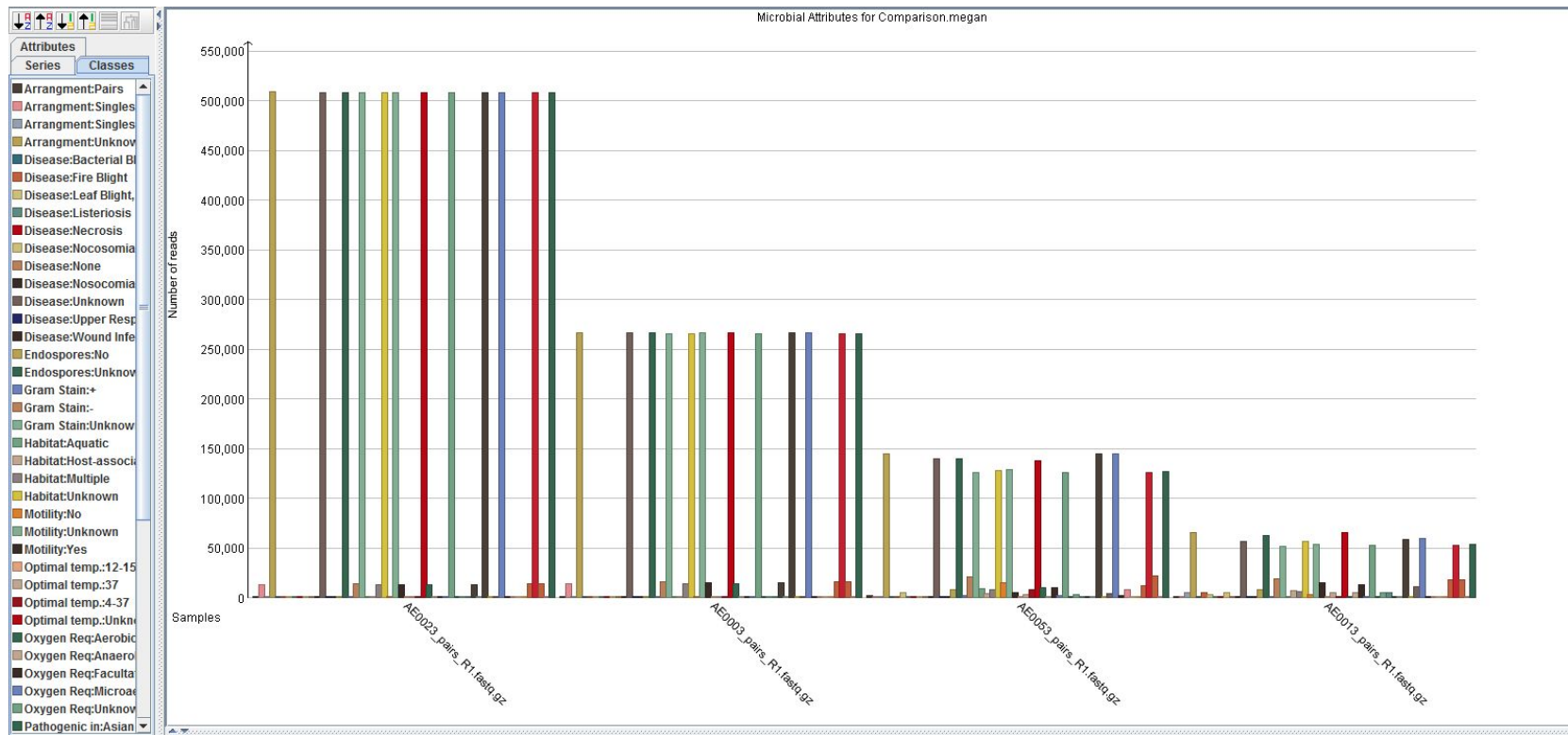
Microbrial Attributes

- Microbrial Attributes chart can be brought up under the "Window" menu
    - The charts may be sorted alphabetically, by amount of reads, or by source of the samples
   - Attributes can be toggled on and off to allow for easier visualization of attributes of interest

Rarefaction Plot

- MEGAN allows you to make rarefaction plots. This allows the user to randomly subsample at a set number of reads to make sure they are assigning all OTUs/Reads.
- Think of this as a species accumulation curve. Look for a point where your lines level out. This will be the point where you have enough reads to get all OTUs or taxonomic assignments.
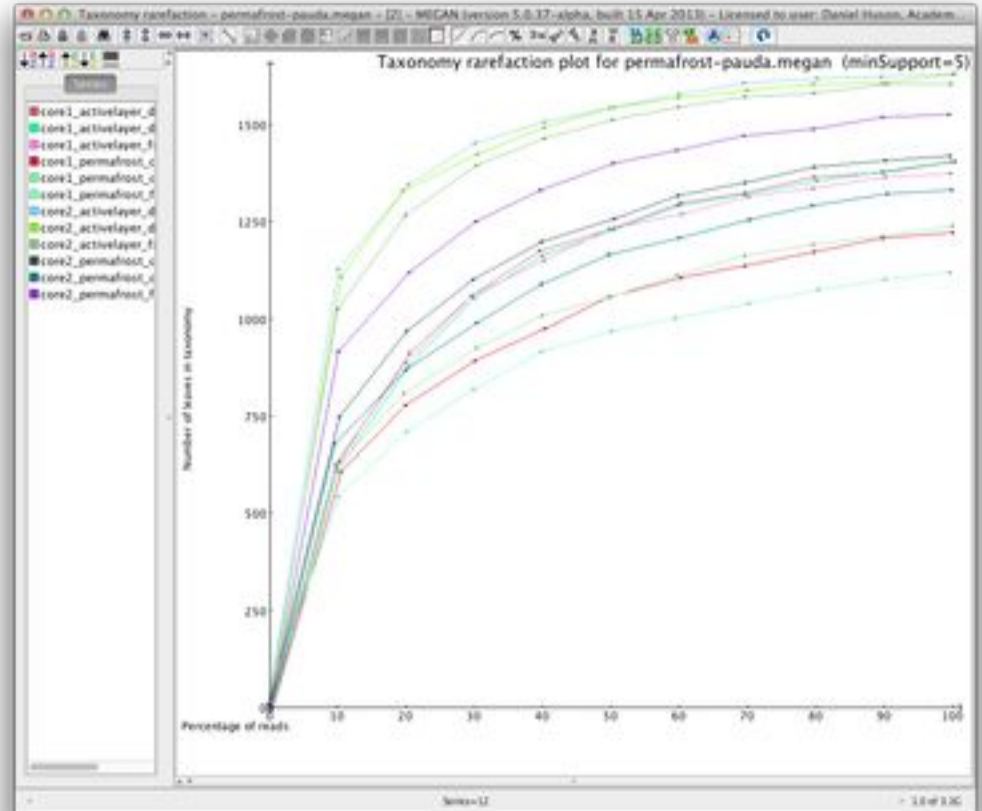
Functional Annotation

MEGAN allows the user to easily annotate their reads based on the following databases:

1. INTERPRO2GO – based on the Gene Ontology metagenomic goslim and InterPro families. Looks at protein families
2. COG – gene function identifier defined by clusters of orthologous groups
3. SEED – classification of gene function based on a collection of biologically defined subsystems
4. KEGG – A collection of metabolic pathways and other pathways
5. PFAM - A database collection of protein families

Functional Annotation

Left to Right: TAXA, INTERPRO2GO, COG, SEED, KEGG, PFAM