Fama: a Computational Tool for Comparative Analysis of Shotgun Metagenomic Data

Alexey E. Kazakov
Pavel S. Novichkov
Lawrence Berkeley National Laboratory
Berkeley, CA, USA
aekazakov@lbl.gov

Functional profiling

- Three reference datasets
- Read trimming by Trimmomatic
- Manually curated reference proteins
- Utilizes DIAMOND, very fast BLASTX alternative

Comparative analysis

- Normalization by sample size
- Normalization by gene length
- Normalization by average genome size
- TSV, XLSX output, interactive visualization

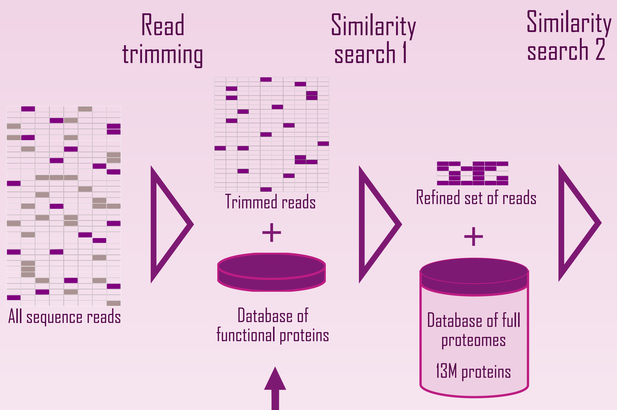
Gene-centric assembly

- Contig assembly by MEGAHIT
- Read mapping by Bowtie
- Gene prediction by Prodigal
- Functional profiling of predcted genes

This material by ENIGMA-Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research under contract number DE-AC02-05CH11231

Read mapping

Read mapping procedure selects sequence reads for functions of interest and calculates number of reads for each function.
For paired-end sequence libraries, it calculates number of sequenced fragments.



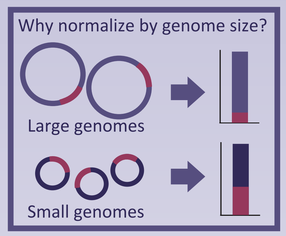
Normalization

ERPKG: number of reads per genome equivalent per kb of reference sequence (effective)

$$ERPKG_f = \frac{N_f}{EGL * GE}$$

$$GE = \frac{N_{trimmed\ reads}}{AGS}$$

N_f: number of reads mapped to function f
EGL: effective gene length
GE: genome equivalent
AGS: average genome size



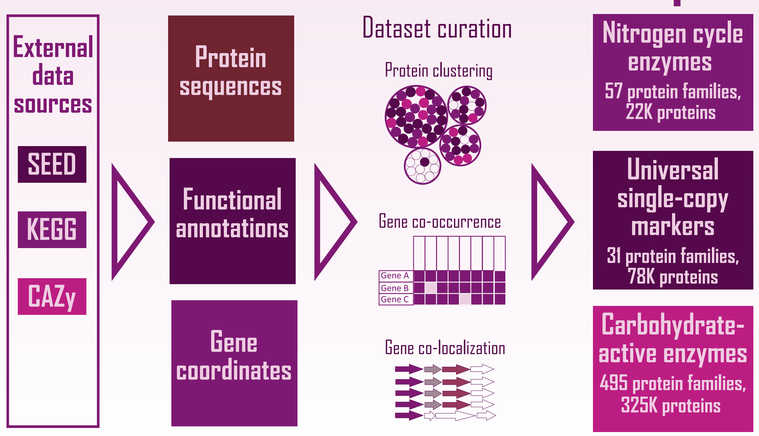
Code availability:

https://github.com/novichkov-lab/fama

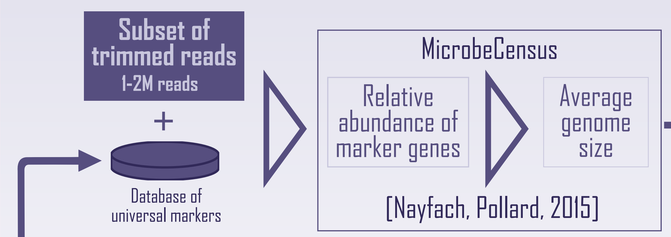
- Written in Python3
- Runs on OS Linux
- Developed and tested on Ubuntu 18.04

From reads to genes

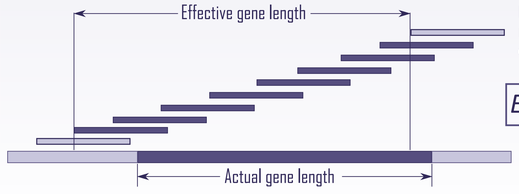
Reference protein datasets



Calculation of average genome size



Effective vs. actual gene length



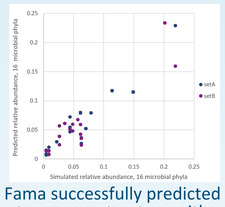
Effective gene length (EGL) is a number of possible start positions of reads mapped to the gene

$$EGL = GL + RL - 2*AL_{min} + 1$$

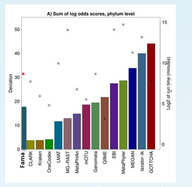
GL: actual gene length
RL: read length
AL: minimal alignment length

Testing and benchmarking

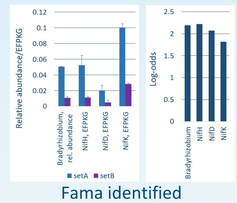
Simulated metagenomic datasets [Lindgreen et al., 2015]



Fama successfully predicted taxonomy structure with universal marker genes

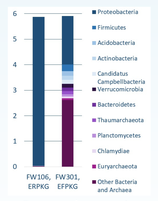


Fama outperforms other tools using universal marker proteins

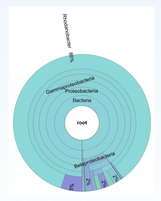


Fama identified simulated shift in nitrogen fixation genes

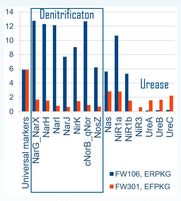
Nitrate-contaminated (FW106) and control (FW301) groundwater [Hemme et al., 2015]



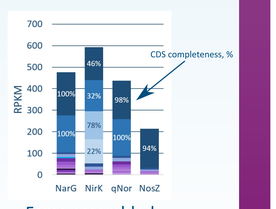
Proteobacteria dominate in FW106 sample but not in control



Rhodanobacter s a predominant organism in the contaminated sample



FW106 community enriched with nitrate reduction and denitrification genes



Fama assembled denitrification genes from contaminated groundwater