What if we could look only for favorite genes but in many samples?

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

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

This material by ENIGMA-Ecosystems and Networks

Integrated with Genes and

(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 Biological & Environmental

DE-AC02-05CH11231

of Energy, Office of Science, Office

Research under contract number

Molecular Assemblies

Alexey E. Kazakov

Functional profiling

Comparative analysis

Gene-centric assembly

- Three reference datasets
- Normalization by sample size
  - Contig assembly by MEGAHIT

- Read trimming by Trimmomatic
- Normalization by gene length
- Read mapping by Bowtie

Manually curated reference proteins

• Utilizes DIAMOND, very fast BLASTX alternative

Normalization by average genome size

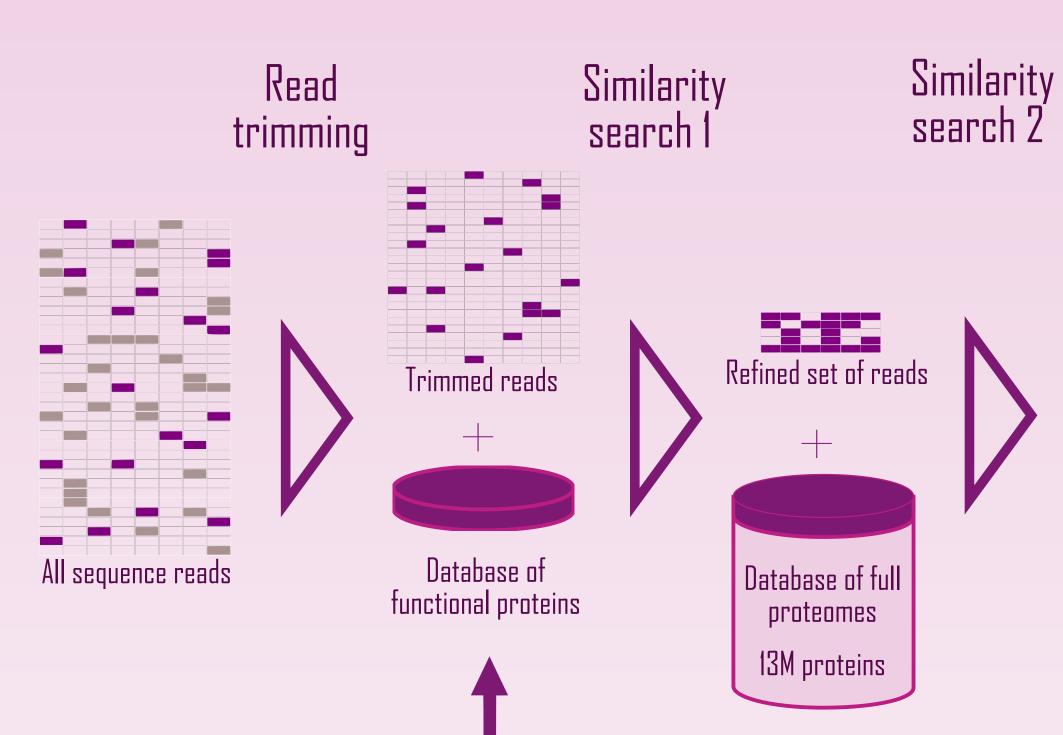
• TSV, XLSX output, interactive visualization

Gene prediction by Prodigal

### Kead 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)

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

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

Functional

profile

Taxonomic

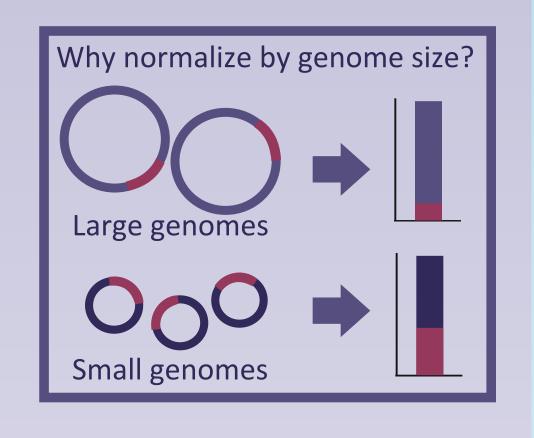
profile

 $N_f$ : number of reads mapped to function fEGL: effective gene length GE: genome equivalent

AGS: average genome size

Comparative

profile



Reports

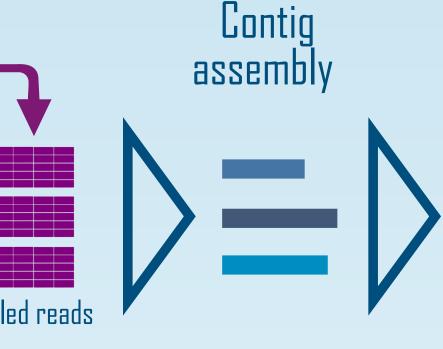
#### Code availability:

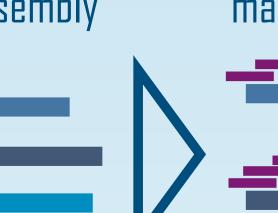
Functional profiling of predcted genes

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

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

## From reads to genes





Testing and benchmarking

Simulated metagenomic datasets (Lindgreen et al., 2015)







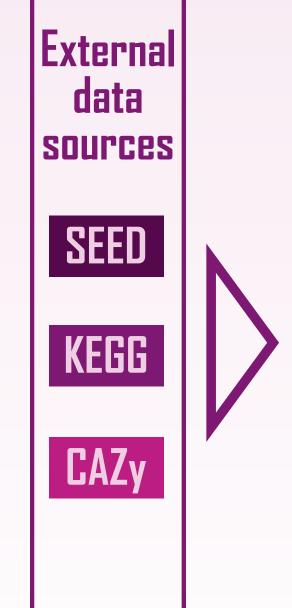
Taxonomic profile

Functional

# Calculation of average genome size

#### Subset of MicrobeCensus trimmed reads 1-2M reads Relative Average abundance of genome size marker genes Database of (Nayfach, Pollard, 2015) universal markers

# Reference protein datasets



Functional annotations

Protein

sequences

Gene coordinates

Dataset curation

Protein clustering

Gene co-localization

Universal single-copy markers 31 protein families,

> Carbohydrateactive enzymes 495 protein families, 325K proteins

78K proteins

Nitrogen cycle

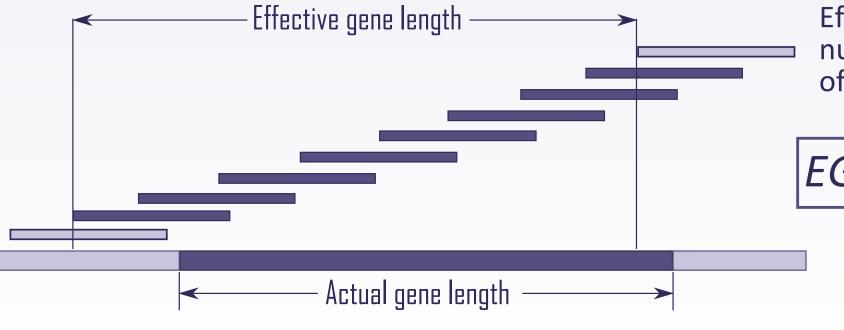
enzymes

57 protein families,

22K proteins

Mapped reads

# Effective vs. actual gene length



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

AL: minimal alignment length

RL: read length

 $EGL = GL + RL - 2*AL_{min} + 1$ GL: actual gene length

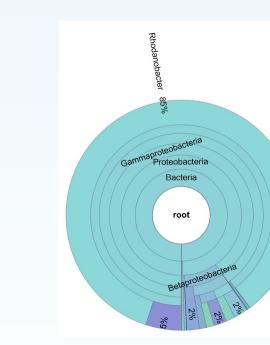
Proteobacteria dominate in FW106 sample but not in control

Fama successfully predicted

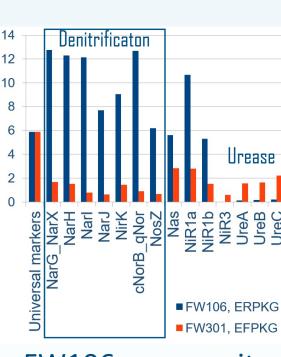
taxonomy structure with

universal marker genes

■ Thaumarchaeota



Rhodanobacter s a predominant organism in the contaminated sample



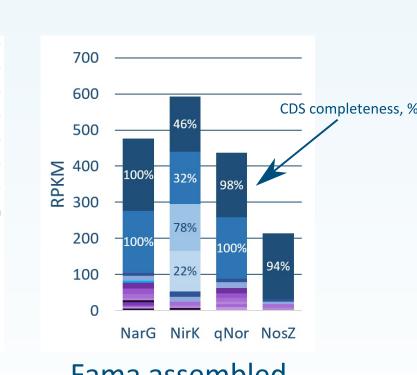
Fama outperforms other

tools using universal

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

marker proteins

FW106 community enriched with nitrate reduction and denitrification genes



Fama identified

simulated shift in

nitrogen fixation genes

Fama assembled denitrification genes from contaminated groundwater