MALINA — a Web-service for human gut microbiota whole-genome metagenomic reads analysis

Anna S. Popenko[1], Alexander V. Tyakht[1], Ilya A. Altukhov[1,2], Dmitry G. Alexeev[1,2]
[1] Research Institute of Physico-Chemical Medicine, Moscow, Russia
[2] Moscow Institute of Physics and Technology, Moscow, Russia
a.s.popenko@niifhm.ru

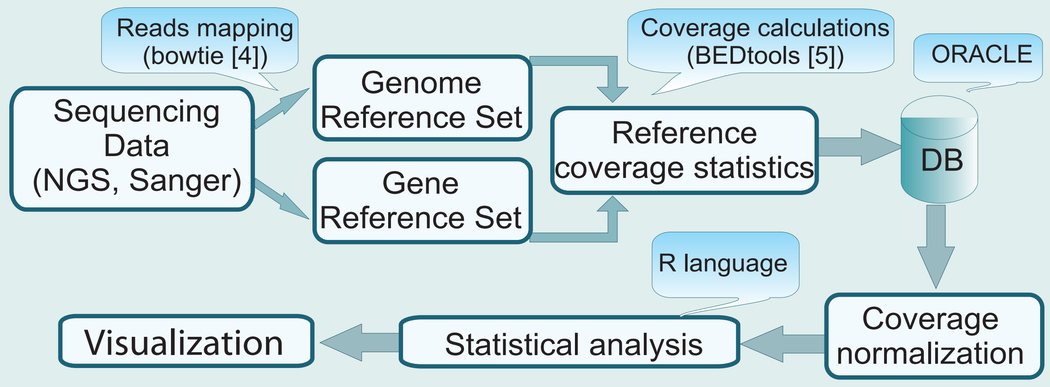
http://malina.metagenome.ru

Introduction

Huge volume of whole-genome metagenomic samples in repositories keeps growing at increasing pace. In community of researchers, there is a need for robust data analysis tools that allow efficient description of composition, classification and clustering coupled with comprehensive visualization of results Particularly, human gut microbiome is one of extensively studied topics in metagenomic research.We developed a software pipeline for bioinformatic analysis of metagenomic data obtained from human gut microbiota sequencing currently available as a free web service MALINA.

The pipeline was used in the workflow of Russian Metagenome Project . It included analysis of 132 human gut metagenomic samples from various groups of Russian population, including healthy citizen of large cities, country-side and isolated communities, and patients with diseases of heart and intestine. The samples were sequenced in short reads on ABI SOLiD. The research is described in poster 'Examining composition of Russian human gut microbiota by assesing relative abundance of functional and taxonomical units '.

Pipeline workflow



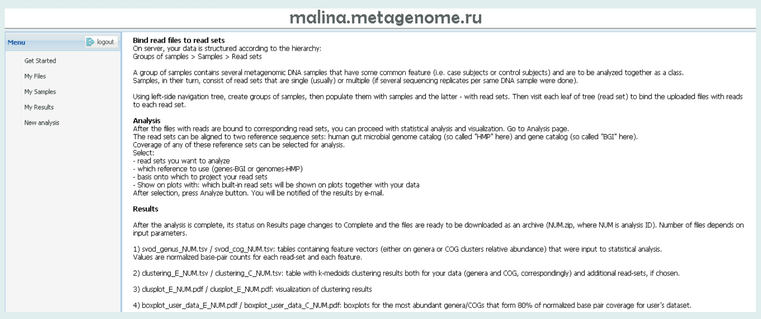
Web-service can process metagenomic short reads (including color-space SOLiD reads). Additionally, long reads can be analyzed (via pre-fragmentation). Two reference sets are catalogs of human gut microbial genomes (>400 genomes) and prevalent genes [2]. Genomes are aggregated at genus level, genes are grouped by COGs.

MALINA interface

1. Login

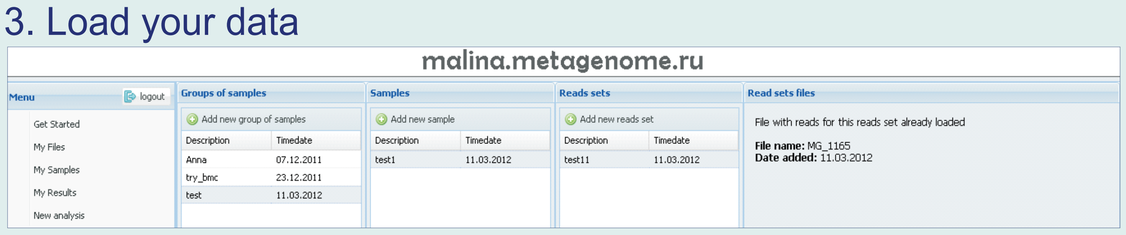
Demo access allows to analyze datasets openly to everyone. To analyze datacets privately and receive e-mail notifications about analysis statys, register your account.

2. Get started



Instructions on usage of service are provided, with examples of output

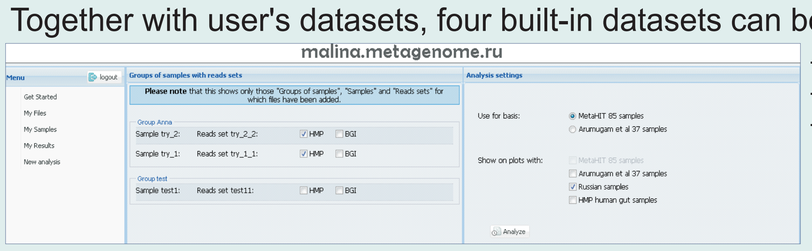
3. Load your data



Sessions of your account including datasets, analysis runs, results are stored and can accessed at any time via Web browser

4. Start analysis

Together with user's datasets, four built-in datasets can be processed:



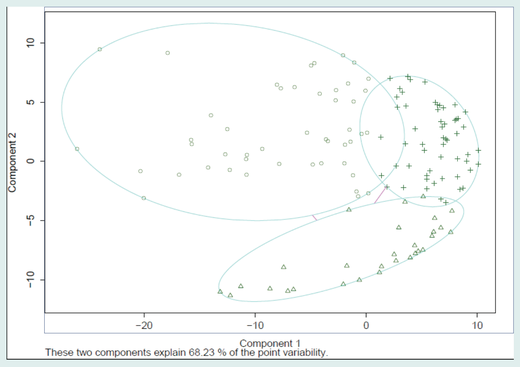
- 37 Sanger samples from [1]
-124 MetaHIT samples from [2]
- Novel 132 Russian Metagenome Project samples

5. Download results

After the analysis is complete, download archived results (PDF and text files).

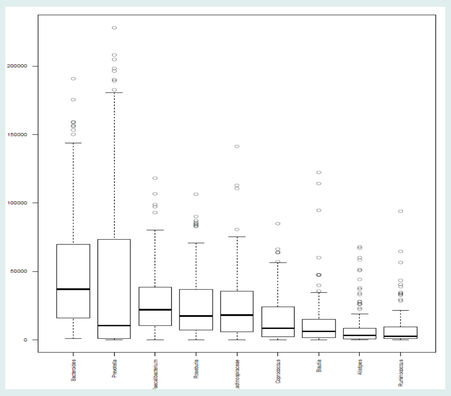
Metagenomic analysis results

1. Clustering



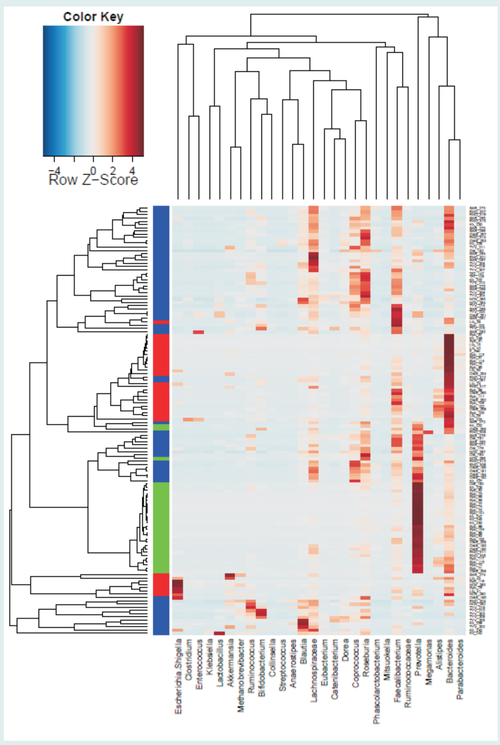
Clusters are produsd using k-medoids algorithm

2. Taxonomic composition



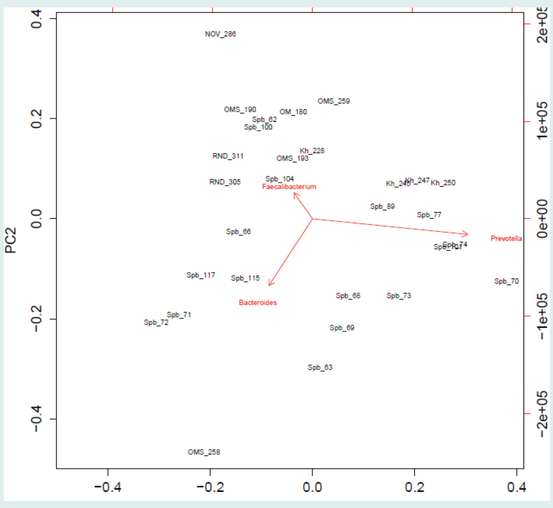
Top represented genera accounting for 80% abundance are displayed.

3. Heatplot with clustering



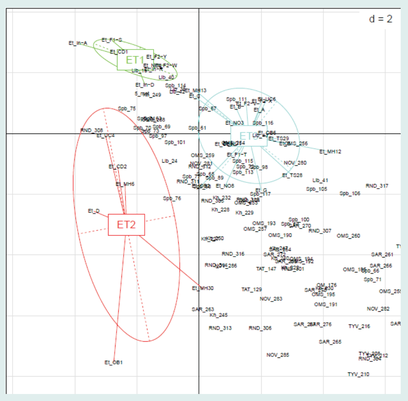
Heatplot is generated for most abundant genera, showing both k-methods clustering and hierarchical clustering using average method

4. Principal Component Analysis



PCA is performed for genera accounting for top represented genera accounting for 80% of abundance. Thus drivers are emphasized.

5. Between Class Analysis



BCA is performed on user data with optional inclusion of any of the four built-in datasets.

Reference

[1] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, 'Enterotypes of the human gut microbiome', Nature 473,174–180
[2] Qin J, Li R, Raes J, Arumugam M,'A human gut microbial gene catalogue established by metagenomic sequencing.',.Nature. 2010 Mar 4;464(7285):59-65
[3] http://www.hmpdacc.org
[4] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
[5] Quinlan AR and Hall IM, 2010. BEDT ools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842

Acknowledgements

We thank Prof. Vadim Govorun who participated in study design and coordination. This work was supported by State Contracts 16.512.11.2111, 16.552.11.7034.

Comparison with existing data

The pipeline was tested on metagenomic data obtained from existing studies [1, 2]. Analysis of short (Illumina) and long metagenomic reads (Sanger, pre-fragmented) resulted in relative abundance of bacterial genera that was significantly correlated with data from original studies (p-value < 0,05). Similar results were obtained for reconstruction of mock community (shot reads simulated from five gut microbial genomes mixed in different proportions).