

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

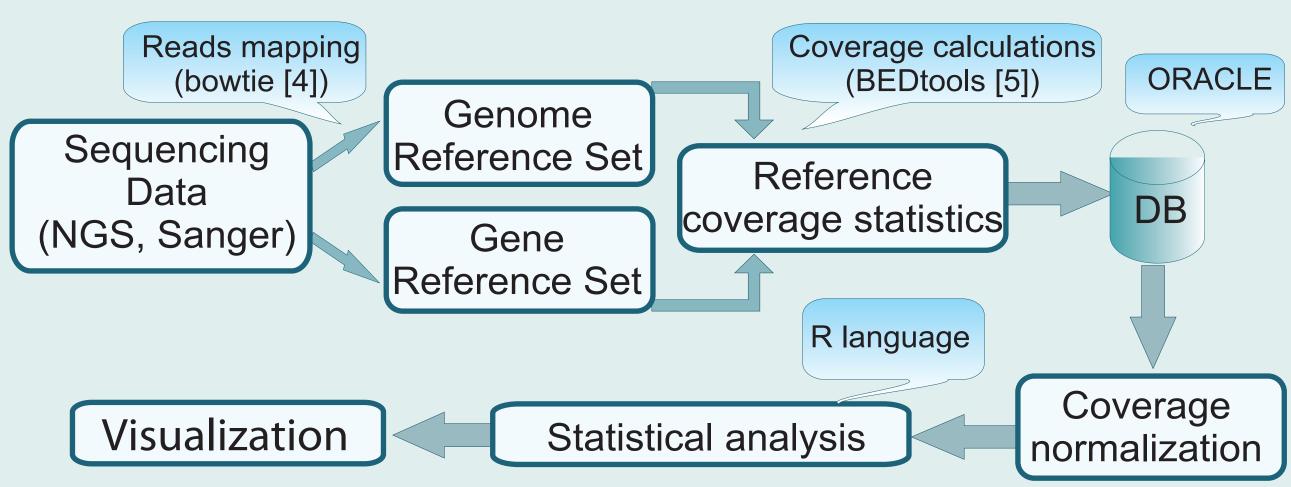
Anna S. Popenko<sup>1</sup>, Alexander V. Tyakht<sup>1</sup>, Ilya A. Altukhov<sup>1,2</sup>, Dmitry G. Alexeev<sup>1,2</sup>

<sup>1</sup> Research Institute of Physico-Chemical Medicine, Moscow, Russia

<sup>2</sup> Moscow Institute of Physics and Technology, Moscow, Russia a.s.popenko@niifhm.ru

## Pipeline workflow

http://malina.metagenome.ru



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.

### 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 '.

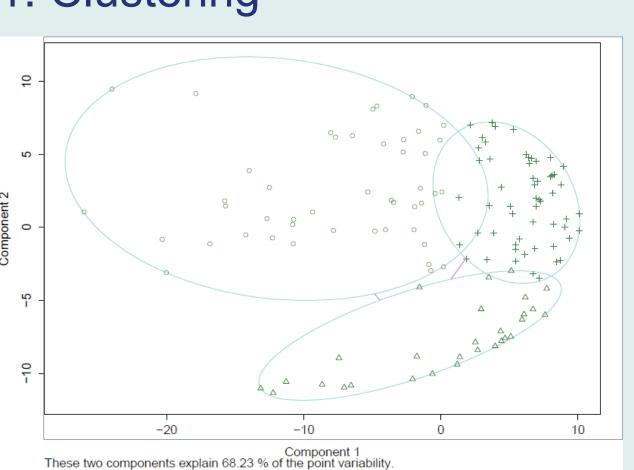
## MALINA interface

#### 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.

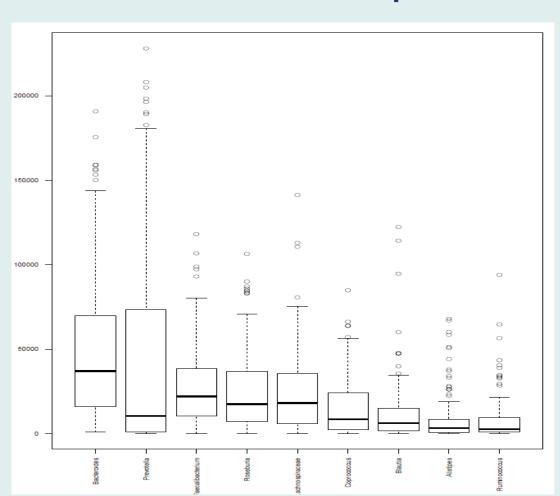
#### 1. Clustering



Metagenomic analysis results

Clusters are prodused using k-medoids algorithm

#### 2. Taxonomic composition



Top represented genera accounting for 80% abundance are displayed.

4. Principal Component Analysis

#### 2. Get started

malina.metagenome.ru								
Menu logout		Bind read files to read sets On server, your data is structured according to the hierarchy:						
Get Started		Groups of samples > Samples > Read sets						
My Files		A group of samples contains several metagenomic DNA samples that have some common feature (i.e. case subjects or control subjects) and are to be analyzed together as a class. Samples, in their turn, consist of read sets that are single (usually) or multiple (if several sequencing replicates per same DNA sample were done).						
My Samples My Results		Using left-side navigation tree, create groups of samples, then populate them with samples and the latter - with read sets. Then visit each leaf of tree (read set) to bind the uploaded files with reads to each read set.						
New analysis		Analysis After the files with reads are bound to corresponding read sets, you can proceed with statistical analysis and visualization. Go to Analysis page. The read sets can be aligned to two reference sequence sets: human gut microbial genome catalog (so called "HMP" here) and gene catalog (so called "BGI" here). Coverage of any of these reference sets can be selected for analysis. Select: - read sets you want to analyze - which reference to use (genes-BGI or genomes-HMP) - basis onto which to project your read sets - Show on plots with: which built-in read sets will be shown on plots together with your data After selection, press Analyze button. You will be notified of the results by e-mail.						
		Results  After the analysis is complete, its status on Results page changes to Complete and the files are ready to be downloaded as an archive (NUM,zip, where NUM is analysis ID). Number of files depends on						
		input parameters.						
		1) svod_genus_NUM.tsv / svod_cog_NUM.tsv: tables containing feature vectors (either on genera or COG clusters relative abundance) that were input to statistical analysis. Values are normalized base-pair counts for each read-set and each feature.						
		2) clustering_E_NUM.tsv / clustering_C_NUM.tsv: table with k-medoids clustering results both for your data (genera and COG, correspondingly) and additional read-sets, if chosen.						
		3) clusplot_E_NUM.pdf / clusplot_E_NUM.pdf: visualization of clustering results						
		4) boxplot_user_data_E_NUM.pdf / boxplot_user_data_C_NUM.pdf; boxplots for the most abundant genera/COGs that form 80% of normalized base pair coverage for user's dataset.						

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

	malina.metagenome.ru								
Mei	Menu 🕒 logout		Groups of samples		Samples		Reads sets		Read sets files
			Add new group of samples		O Add new sample		Add new reads set		File with reads for this reads set already loaded
	My Files		Description	Timedate	Description	Timedate	Description	Timedate	File name: MG_1165
			Anna	07.12.2011	test1	11.03.2012	test11	11.03.2012	Date added: 11.03.2012
	My Samples		try_bmc	23.12.2011					
	My Results		test	11.03.2012					
	New analysis								

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

## 3. Load your data

#### 4. Start analysis

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

Groups of samples with reads sets  Please note that this shows only those "G which files h  Group Anna Sample try_2: Reads set try_2_2:	roups of samples", "Samples" and "Reads sets" for have been added.   HMP BGI	Analysis settings  Use for basis:	MetaHIT 85 samples     Arumugam et al 37 samples	
	✓ HMP ☐ BGI	Use for basis:		
Sample try_1: Reads set try_1_1:  Group test	▼ HMP □ BGI	Show on plots with:	MetaHIT 85 samples Arumugam et al 37 samples	
Sample test1: Reads set test11:	☐ HMP ☐ BGI		✓ Russian samples  ☐ HMP human gut samples	
	Group test	Group test	Show on plots with:	Show on plots with:  Group test Sample test1: Reads set test11: HMP BGI  HMP human gut samples  HMP human gut samples

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

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

### 5. Download results

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

malina.metagenome.ru										
Menu	logout	Your Results								
C-1 Ch-1-1		Request ID	Use for basis	Reference data base	Reads sets	Show on plots with	Date	Status 🔺		
Get Started		58	Arumugam et al 37 samples	HMP	try_1_1,try_2_2	НМР	07.12.2011	$\checkmark$	<b>€</b>	
My Files		81	Arumugam et al 37 samples	HMP	try_1_1,try_2_2	HMP	09.12.2011	$\checkmark$	<b>₽</b>	
My Samples		162	MetaHIT 85 samples	HMP	Et_A,Et_B,Et_C,Et_CD1,Et_CD2,Et_D,Et_E,Et_F1-S,Et_F1-T,Et_F2-V,Et_F2-W,Et_F2	None	16.12.2011	$\checkmark$	<b>₽</b>	
My Results		181	MetaHIT 85 samples	HMP	Et_A,Et_B,Et_C,Et_CD1,Et_CD2,Et_D,Et_E,Et_F1-S,Et_F1-T,Et_F2-V,Et_F2-W,Et_F2	None	16.12.2011	$\checkmark$	<b>€</b>	
New analysis		202	Arumugam et al 37 samples	HMP	146	None	23.12.2011	$\checkmark$	<b>₽</b>	
		340	Arumugam et al 37 samples	HMP	try_2_2,try_1_1,Et_A,Et_B,Et_C,Et_CD1	None	22.02.2012	$\checkmark$	<b>₽</b>	
		369	MetaHIT 85 samples	HMP	Et_A,Et_C	Russian Samples	06.03.2012	$\checkmark$	<b>&amp;</b>	

## Reference

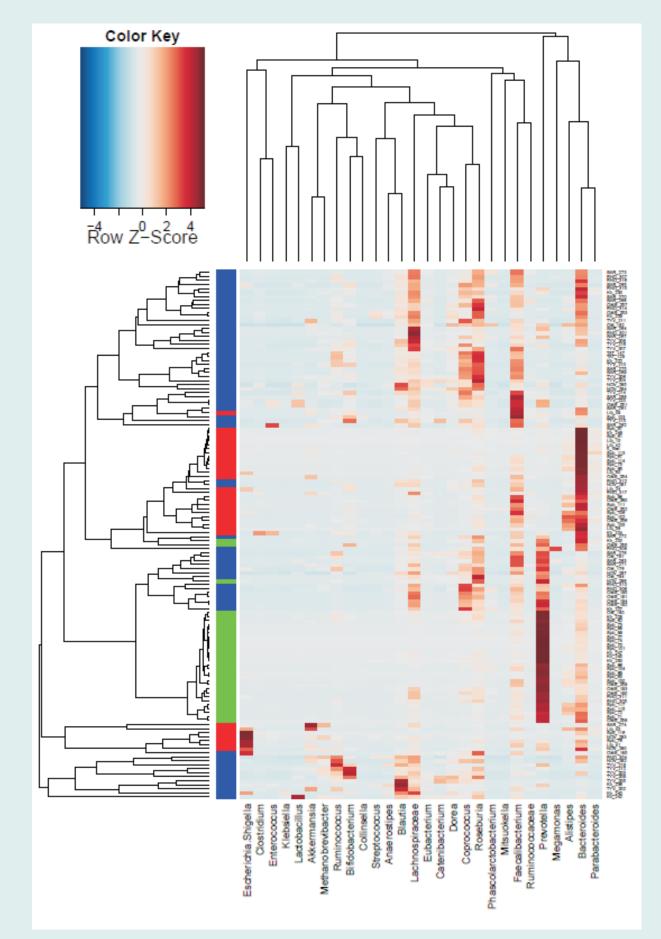
[3] http://www.hmpdacc.org

- [1] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, 'Enterotypes of the human gut microbiome', Nature 473,174–180
- [2] Qin J, Li R, R aes J, Arumugam M,'A human gut microbial gene catalogue established by metagenomic sequencing.',.Nature. 2010 Mar 4;464(7285):59-65
- [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

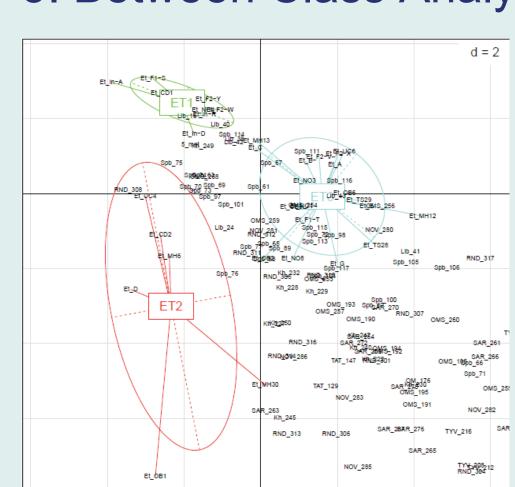
We thank Prof. Vadim Govorun who participated in study design and coordination. This work was supported by State Contracts 16.512.11.2111, 16.552.11.7034.

#### 3. Heatplot with clustering



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.

## 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).