CloVR-Metagenomics: Functional and taxonomic microbial community characterization from metagenomic whole-genome shotgun (WGS) sequences – standard operating procedure, version 1.0

James Robert White, Cesar Arze, Malcolm Matalka, the CloVR team, Samuel V. Angiuoli & W. Florian Fricke

The Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

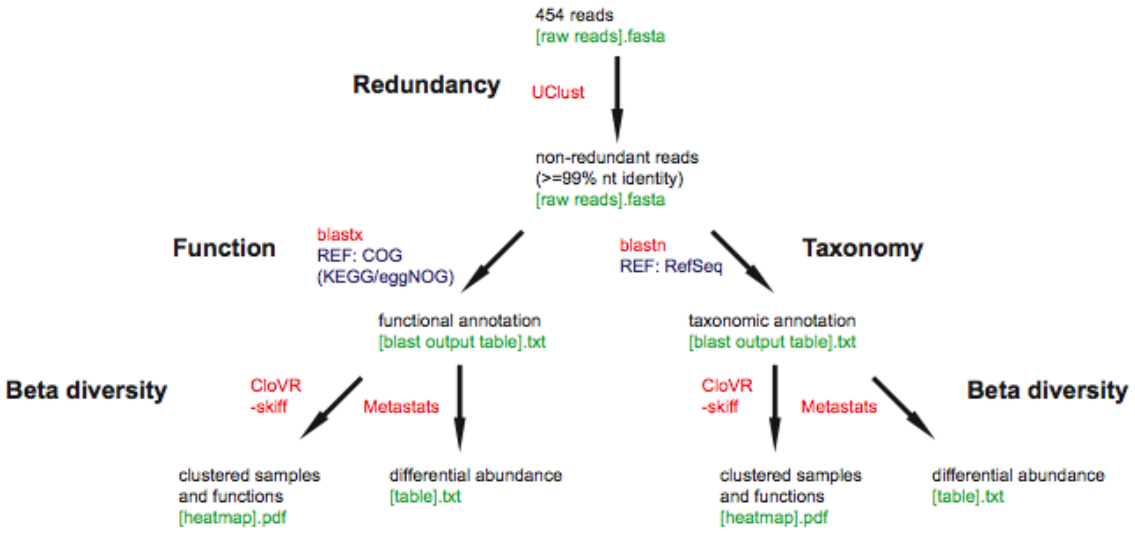
Abstract

The CloVR-Metagenomics pipeline employs several well-known tools and protocols for the analysis of metagenomic whole-genome shotgun (WGS) sequence datasets:

A) UCLUST [1] – a C++-based software package for clustering redundant DNA sequences and removing artificial 454 replicates [1];
B) BLASTX and BLASTN [2] for functional and taxonomic assignment of sequences, respectively;
C) Metastats [3] and custom R scripts to generate additional statistical and graphical evaluation.

The CloVR-Metagenomics pipeline accepts as input multiple fasta files (1 sample per file) and a corresponding tab-delimited metadata file that specifies features associated with the samples, which are used for comparative analysis. This protocol is available in CloVR beta version 0.5 and 0.6.

Overview



Software

Step	Program	Version	Weblink	Reference
Clustering of redundant sequences (replicate removal)	UCLUST	1.1.579q	http://www.drive5.com/usearch	[1]
Functional classification of DNA sequences	BLASTX	2.2.21	http://blast.ncbi.nlm.nih.gov	[2]
Taxonomic classification of DNA sequences	BLASTN	2.2.21	http://blast.ncbi.nlm.nih.gov	[2]
Differential abundance detection	Metastats	1.0	http://metastats.cbcb.umd.edu/	[3]
Statistical evaluation	R	2.10.1-2	http://www.r-project.org/	

Reference data

Database	Data	Version	Weblink	Reference
COG	Functionally annotated protein sequences (Clusters of Orthologous Genes)	1.0	http://www.ncbi.nlm.nih.gov/COG/	[4]
RefSeq	Taxonomically annotated bacterial and archaeal genomes	6/21/10	www.ncbi.nlm.nih.gov/refseq/	[5]
eggNOG	Functionally annotated orthologous proteins	2.0	http://eggnog.embl.de/	[6,7]
KEGG genes	Functionally annotated proteins	55.0/ 09-14	www.genome.jp/kegg/	[8,9]
NCBI NR	Non-redundant Proteins		ftp://ftp.ncbi.nlm.nih.gov/blast/db/	

Pipeline input

Data	Suffix	Description
Multiple fasta	.fasta	Metagenomic WGS sequences (1 file per sample)
Metadata	.txt	Sample-associated features (see section A for details)

Pipeline output

Data	Suffix	Description
UCLUST clusters	.clstr	List of clusters created to reduce redundant analysis
Replicate sequences	.txt	List of artificial 454 replicates removed from downstream analysis
BLAST hits	.raw	BLASTN or BLASTX to reference datasets results table ("-m 8" format)
Taxonomic assignments	.tsv	Table (tab-delimited) displaying taxonomic assignment counts for each sample
Functional assignments	.tsv	Table (tab-delimited) displaying functional assignment counts for each sample
Metastats output	.csv	Differentially abundant taxonomic or functional assignment groups (as pre-defined in Metadata input)
Skiff clustering	.pdf	Heatmap and two-way clustering based on taxonomic and functional assignment abundances
Pie charts	.pdf	Pie charts describing assignment abundances for up to 12 samples (not performed if >12 samples are given)
Stacked histograms	.pdf	Stacked histograms displaying relative abundances for up to

50 samples and 25 features (not performed if beyond these thresholds)

A. Requirements for Pipeline Input

To run the full CloVR-Metagenomics analysis track, two different inputs have to be provided by the user: a set of fasta-formatted sequence files and a tab-delimited metadata file in the .txt format. The metadata file provides sample-associated information with the following formatting requirements:

```
#File       SampleName    ph    Description
A.fasta     sampleA      high    control
B.fasta     sampleB      high    sick
C.fasta     sampleC      low     treated
D.fasta     sampleD      low     treated
```

where:

1. All entries are tab-delimited.
2. All entries in every column are defined (no empty fields).
3. The header line begins with: #File<tab>.
4. There are no duplicate header fields or file names.
5. No header fields or corresponding entries contain invalid characters (only alphanumeric and underscore characters allowed)

Pairwise comparisons: To utilize the Metastats statistical methodology for the detection of taxonomic and functional assignments with differential abundance, the associated header field must end with "_p", (e.g. "Treatment_p", or "ph_p"). Otherwise Metastats will skip pairwise analysis.

B. Sequence clustering and artificial replicate removal with UCLUST

To reduce redundant database searches downstream, the UCLUST component of CloVR-Metagenomics first clusters all DNA sequences using a stringent 99% identity threshold. Similar to the procedure in [10], any non-representative sequence in a cluster that shares a prefix of length 8 with the representative (and whose length is within 10 bp of the representative's length) is determined to be an artificial 454 pyrosequencing replicate [11] and is removed from further analysis. Taxonomic and functional annotations made to representative members are later propagated to all non-replicate sequences.

C. Taxonomic assignment of DNA sequences

All representative DNA sequences from clusters are searched against the RefSeq database of finished prokaryotic genomes (by default) using BLASTN with the following options: "-e 1.0e-5" (e-value threshold), "-b 1" (number of alignments to show) and "-m 8" (tabular output). Each sequence is assigned to the taxonomy of the best-BLAST-hit.

D. Functional assignment of DNA sequences

All representative DNA sequences from UCLUST (section B) are searched against the COG database of orthologous gene groups (by default) using BLASTX with the same options as in section C ("-e 1.0e-5 -b 1-m 8"). Alternatively, the user may opt to employ the KEGG genes, eggNOG or NCBI NR databases for functional annotation. Each sequence is assigned to the function of the best BLAST hit of the respective database.

E. Additional beta diversity analysis using Metastats and the R statistical package

E.1. Detection of differentially abundant features

The program Metastats uses count data from annotated sequences to compare two populations in order to detect differentially abundant features [3]. BLASTN results are processed to detect different taxonomic groups at multiple levels (phylum, class, order, family, genus), while BLASTX results are parsed for differentially abundant functional groups. Metastats produces a tab-delimited table displaying the mean relative abundance of a feature, variance and standard error together with a p value and q value to describe significance of the detected variations (see project website: http://metastats.cbcb.umd.edu/). Note Metastats can run analyses of 1 sample vs. 1 sample, or *N* samples vs. *M* samples, where *N* and *M* are greater than 1. It cannot perform a comparison of 1 sample vs. 2 samples.

E.2. Unsupervised sample clustering

Custom R scripts are used to normalize taxonomic or functional counts and subsequently calculate Euclidean-based distance matrices for samples and features. Complete-linkage (furthest neighbor) clustering is employed to create dendrograms of samples and taxa in the .pdf format. The R packages *RColorBrewer* and *gplots* are utilized.

E.3. Pie chart visualization

Custom R scripts are used to form pie charts displaying proportions of sequences assigned to specific functional and taxonomic levels for up to 12 samples. Outputs are in .pdf format. For more than 12 samples this function is not performed, as the visual comparison for the user would be cumbersome.

E.4. Stacked histogram visualization

Custom R scripts are used to form stacked histograms displaying proportions of sequences assigned to specific functional and taxonomic levels for up to 50 samples and 25 features. Graphical outputs are in .pdf format. For more than 50 samples or 25 features this function is not performed, as the visual comparison for the user would be difficult.

References

1. Edgar RC Search and clustering orders of magnitude faster than BLAST. Bioinformatics.

2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

3. White JR, Nagarajan N, Pop M (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol 5: e1000352.

4. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.

5. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61-65.

6. Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, et al. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res 38: D190-195.

7. Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, et al. (2008) eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res 36: D250-254.

8. Aoki-Kinoshita KF, Kanehisa M (2007) Gene annotation and pathway mapping in KEGG. Methods Mol Biol 396: 71-91.

9. Kanehisa M (2002) The KEGG database. Novartis Found Symp 247: 91-101; discussion 101-103, 119-128, 244-152.

10. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. Isme J 3: 1314-1317.

11. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, et al. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods 6: 639-641.