Title

CloVR-Metagenomics (orfs): Microbial community functional and taxonomic characterization from metagenomic shotgun sequences – standard operating procedure v. 1.0

James Robert White, Cesar Arze, Kevin Galens, Malcolm Matalka, Stephen Mekosh, David R. Riley, Mahesh Vangala, Owen White, Samuel V. Angiuoli & W. Florian Fricke

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

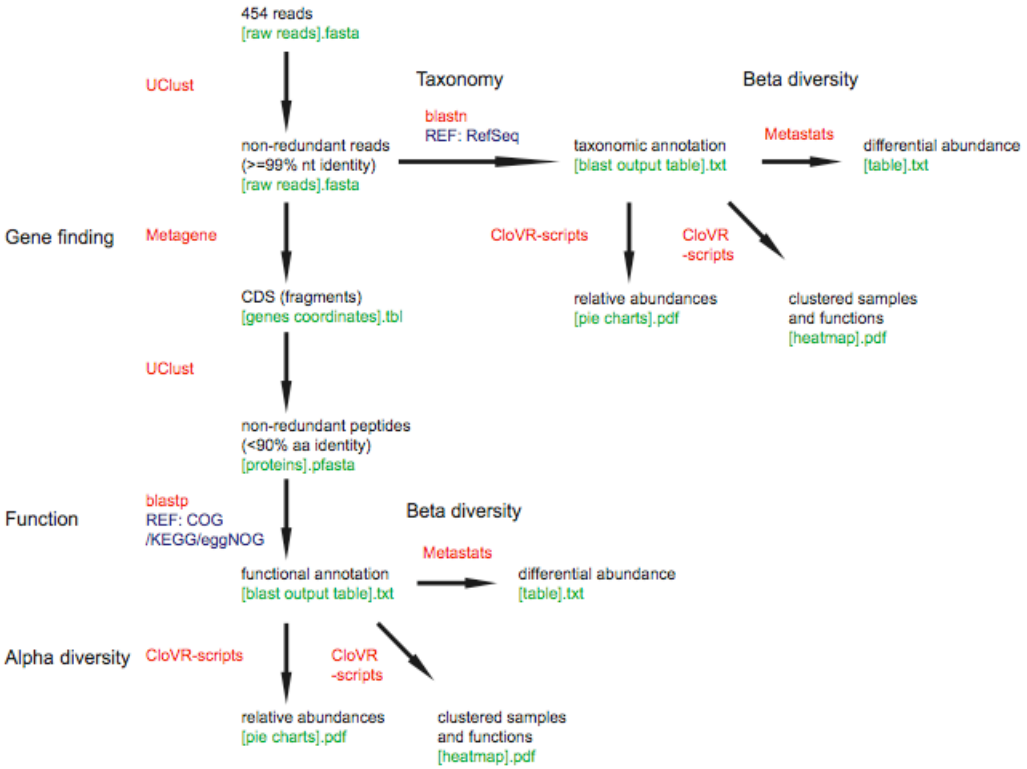
Abstract

The CloVR-Metagenomics (orfs) pipeline is an alternative protocol to the default CloVR-Metagenomics pipeline and employs several well-known tools and protocols for the analysis of metagenomic shotgun sequence datasets:

A) UCLUST – a C++-based software package for clustering redundant DNA and peptide sequences and removing artificial 454 replicate reads [1];
B) MetaGene predicts protein-coding sequences from partial gene fragments [2];
C) BLASTP and BLASTN for functional and taxonomic assignment of sequences, respectively [3];
D) Additional statistical and graphical evaluation within the pipeline includes Metastats [4] and custom R scripts implemented in CloVR.

The CloVR-Metagenomics (orfs) pipeline accepts as input multiple fasta files (1 sample per file) and a corresponding tab-delimited metadata file for comparative analysis.

Figure



Software

Step	Program	Version	Weblink	Reference
Clustering of redundant sequences and replicate removal	UCLUST	1.1.579q	http://www.drive5.com/usearch	[1]
ORF prediction on representative reads	MetaGene	1.0	http://metagene.cb.k.u-tokyo.ac.jp/metagene.html	[2]
Functional classification of putative peptides	BLASTP	2.2.21	http://blast.ncbi.nlm.nih.gov	[3]
Taxonomic classification of DNA sequences	BLASTN	2.2.21	http://blast.ncbi.nlm.nih.gov	[3]
Differential abundance detection	Metastats	1.0	http://metastats.cbcb.umd.edu/	[4]
Statistical evaluation	R	2.10.1-2	http://www.r-project.org/	

Reference data

Database	Data	Version	Weblink	Reference
COG	Clusters of orthologous proteins	1.0	http://www.ncbi.nlm.nih.gov/COG/	[5]
RefSeq	Finished bacterial and archaeal genomes with full taxonomic resolution	6/21/10	www.ncbi.nlm.nih.gov/refseq/	[6]
eggNOG	Functionally annotation orthologous proteins	2.0	http://eggnog.embl.de/	[7,8]
KEGG genes	Functionally annotated protein sequences	55.0/ 09-14	www.genome.jp/kegg/	[9,10]
NCBI NR	Protein sequences		ftp://ftp.ncbi.nlm.nih.gov/blast/db/	

Pipeline input

Data	Suffix	Description
Multiple fasta files	.fasta	Trimmed and binned shotgun 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 downsteam analysis
Translated ORFs	.concat.fsa	Fasta file of translated orfs predicted by Metagene
BLAST hits	.raw	BLASTN or BLASTP results with “-m 8” output format to reference datasets
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	Differential taxonomic or functional groups between pre-defined groups as given in the metadata input file
Skiff clustering	.pdf	Heatmap two-way clustering based on different feature abundances

Pie charts	.pdf	Pie charts describing feature 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 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
3. The header line begins with: #File<tab>
4. There are no duplicate header fields or file name
5. No header fields or corresponding entries contain invalid characters (alphanumeric and underscore only allowed)

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

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 [11], 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 [12] and is removed from further analysis. Taxonomic and functional annotations made to representative members are later propagated to all non-replicate sequences.

C. ORF prediction and clustering

C1. MetaGene ORF calling

The MetaGene component of this protocol analyzes the set of unique DNA sequences produced from UCLUST and determines which ORFs most likely exist on each fragment. Multiple ORFs may be called from a single read. These ORFs are subsequently translated to peptide sequences in fasta format.

C2. Peptide clustering with UCLUST

To improve the execution of downstream BLASTP searches, predicted peptide sequences are clustered with UCLUST using a similarity cutoff of 90%. Functional annotations assigned to a representative member of a cluster are later propagated to all members of that cluster.

D. Taxonomic assignment of DNA sequences

All representative DNA sequences from clusters are searched against the RefSeq database of finished prokaryotic genomes using BLASTN with an e-value threshold of 1e-5, “-b 1” and “-m 8” output. Each sequence is assigned to the taxonomy of the best-BLAST-hit.

E. Functional assignment of translated ORFs

All representative peptide sequences from UCLUST (section C2) are searched against the COG database of orthologous gene groups using BLASTP with an e-value threshold of 1e-5, “-b 1” and “-m 8” output. 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.

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

F.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 [4]. BLASTN results are processed to detect different taxonomic groups at multiple levels (phylum, class, order, family, genus, species), while BLASTP results are parsed for differential 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 do a comparison of 1 sample vs. 2 samples.

F.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.

F.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.

F.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. 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. Noguchi H, Park J, Takagi T (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res 34: 5623-5630.

3. 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.

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

5. 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.

6. 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.

7. 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.

8. 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.

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

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

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

12. 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.