May 21, 2021

Functional and Taxonomic Characterization of sequence data using GhostKOALA

Helena Pound[1], Eric Gann[1], Steven W Wilhelm[1]

[1]The University of Tennessee, Knoxville

 dx.doi.org/10.17504/protocols.io.buvbnw2n

The Aquatic Microbial Ecology Research Group - AMERG (The Buchan, Zinser and Wilhelm labs)
Great Lakes Center for Fresh Waters and Human Health

Helena Pound
University of Tennessee, Knoxville

ABSTRACT

A method of functional and taxonomic annotation and expression of assembled sequence data using GhostKOALA.

DOI

dx.doi.org/10.17504/protocols.io.buvbnw2n

PROTOCOL CITATION

Helena Pound, Eric Gann, Steven W Wilhelm 2021. Functional and Taxonomic Characterization of sequence data using GhostKOALA . **protocols.io** https://dx.doi.org/10.17504/protocols.io.buvbnw2n

KEYWORDS

GhostKOALA, annotation, metatranscriptome, microbial ecology

CREATED

May 10, 2021

LAST MODIFIED

May 21, 2021

PROTOCOL INTEGER ID

49795

MATERIALS TEXT

Assembled sequences.

BEFORE STARTING

Before starting this protocol, you will need to quality control, trim, remove residual rRNA, and assemble your sequences. Given the many available tools and personal preferences for parameters, we have not covered that process in this protocol.

1 Assemble sequences from samples you wish to characterize. Final file should be in .fasta format.

Citation: Helena Pound, Eric Gann, Steven W Wilhelm (05/21/2021). Functional and Taxonomic Characterization of sequence data using GhostKOALA . https://dx.doi.org/10.17504/protocols.io.buvbnw2n

1.1 Users can assemble as they wish, but we established the following protocol if a user needs direction. Pound and Wilhelm 2021, https://dx.doi.org/10.17504/protocols.io.buvdnw26

2 Extract GFF gene prediction coding sequences from assembly using MetaGeneMark, exporting identified gene sequences as both proteins and nucleotides.

2.1 http://exon.gatech.edu/meta_gmhmmp.cgi

Wenhan Zhu, Alex Lomsadze and Mark Borodovsky. Ab initio gene identification in metagenomic. *Nucleic Acids Research* (2010) 38, e132.

John Besemer and Mark Borodovsky. Heuristic approach to deriving models for gene finding . *Nucleic Acids Research* (1999) 27, pp 3911-3920.

2.2 User may want to rename files as "assembly_protein/nucleotide" and save as .fasta files.

3 Upload the protein fasta file from MetaGeneMark to GhostKOALA, selecting the genus_prokaryotes+family_eukaryotes+viruses KEGG GENES database.

Kanehisa, M., Sato, Y., and Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J*ournal of Molecular Biology* (2016) 428, pp 726-731.

3.1 Note that an email will be sent that requires you to click a "submit" link to actually run GhostKOALA. Only one protein file can be run at a time from any given email address.

4 Download the detailed annotation and taxonomic descriptions. File extensions should be changed to .txt and can be manipulated in Microsoft excel.

5 The functional annotations and taxonomic annotations can be combined in a single excel sheet, taking care to ensure that the annotations for each gene align properly.

5.1 There will be some genes that aren't functionally annotated and some that aren't taxonomically annotated, so be aware that the annotations will probably not align perfectly when you first put them together. The countif function in excel can help identify genes that were not annotated in one category or another.

6 Annotated genes can then be sorted by organism or KEGG orthology.

7 The list of gene numbers from the KO/organism/whatever you are interested in can then be saved as a .txt file. This file should only contain the list of gene numbers, not any of the annotation information.

Citation: Helena Pound, Eric Gann, Steven W Wilhelm (05/21/2021). Functional and Taxonomic Characterization of sequence data using GhostKOALA . https://dx.doi.org/10.17504/protocols.io.buvbnw2n

8 A custom python script is then used to extract the subset list of nucleotide gene sequences from the nucleotide .fasta file exported from MetaGeneMark. **subset_fasta.py**

9 Reads can then be recruited back to gene sequences of interest to establish estimates of expression using the "Map reads to reference" tool in CLC Genomics Workbench.

9.1 We recommend using 90% identity and 90% length parameters, but can be adjusted for user preference.

10 Read counts can then be exported and the incorporated back into the GhostKOALA gene annotation file in order to evaluate gene expression by organism or KO number.

11 Many additional analyses can then be performed, including but not limited to, comparisons of conserved function/taxonomy between samples, differential expression between samples, phylogenetic analyses to resolve multiple genotypes of a single gene, etc.

Citation: Helena Pound, Eric Gann, Steven W Wilhelm (05/21/2021). Functional and Taxonomic Characterization of sequence data using GhostKOALA . https://dx.doi.org/10.17504/protocols.io.buvbnw2n