iTag amplicon sequencing for taxonomic identification at JGI

May 2016 Adam R. Rivers

The DOE Joint Genome institute sequences approximately 10,000 amplicon samples per year 3 using primer sets that target prokaryotic, eukaryotic or fungal organisms. This document describes the sequencing and bioinformatic processing of the samples. Information on the laboratory methods is available on the Protocols section of the JGI website.

Sequencing

JGI receives iTag samples in 96 well plates. Plates are quantified and individual libraries are amplified with single barcode primers (Figure 1) according the sequencing standard operating protocol. Samples are pooled at up to 184 samples per sequencing run and sequenced on an Illumina MiSeq sequencer in 2x300 run mode.

Analysis

All data from the sequencer is demultiplexed and stored in JGI's archiving and metadata organizer system (JAMO). Read data is then processed through JGI's centralized rolling quality control system. This verifies that there were no sequencing issues and removes known contaminant reads using the kmer filter in bbduk.

Quality controlled reads are then processed by **iTagger** (Tremblay *et al.*, 2015). The current version of iTagger (2.0) described in this document differs substantially from original version described in the cited paper. iTagger 2.0 processes sequencing amplicon data in three major steps: read clustering, the taxonomic assignment of operational taxonomic units (OTUs), and the analysis and summarization of ecological data. The iTagger program relies heavily on clustering and classification tools in the USEARCH software suite (Edgar, 2010) and ecological analysis scripts in QIIME (Caporaso *et al.*, 2010). Itagger also records the OTU's from each project in a central database, creating clusters of centroids that can be queried to identify other samples with matching OTU's. This feature allows comparison across JGI's large amplicon sample collection. iTagger was designed to allow JGI to use the best amplicon analysis tools in a production environment. The itagger program is open source and licensed under the Perl license (individual components have their own licenses and Usearch is not open source) The source code for iTagger is available on Bitbucket: http://bitbucket.org/berkeleylab/jgi_itagger. A summary of the amplicon sequencing process is provided in Table 1. The exact methods and parameters used for each itags sequencing run are available in the method.txt and congig.ini files placed in every iTag project directory on the JGI genome portal.

Amplicon primer sets at JGI

Archaeal and Bacterial 16S rRNA V4-V5
515F-Y GTGYCAGCMGCCGCGGTAA
926R CCGYCAATTYMTTTRAGTTT
(Parada *et al.*, 2015)

Eukaryotic 18S rRNA V4
565F CCAGCASCYGCGGTAATTCC
948R ACTTTCGTTCTTGATYRA
(Stoeck *et al.*, 2010)

Fungal ITS2
ITS9F GAACGCAGCRAAIIGYGA
ITS4R TCCTCCGCTTATTGATATGC
(Menkis *et al.*, 2012; White *et al.*, 1990)

Figure 1. The regions of the amplicon sequencing primers used for hybridization with target genes.

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

Sequencing and QC		• Pool up to 184 samples • Sequence 2x300bp on an Illumina MiSeq (200K sequences per sample) • De-multiplex samples • Filter contaminants and trim adapters (Bbtools)
iTagger	Read preparation	• Merge read pairs (Usearch) • Match primers (Userach) • Remove reads with high expected errors • Dereplicate count and sort reads • Create seqobs file
	Clustering	• Remove samples with insufficient numbers of sequences • Combine seqobs files from all samples in a project • Sort by decreasing abundance • Cluster iteratively at 99%, 98%, 97% identity (Usearch cluster_otus and Usearch_global, chimera checking performed here too)

	Classification	• Classify centroids taxonomically using Usearch utax
		• One of three reference databases are used are used for annotation:
		  ○ 16S – Silva SSU, quality filtered, trimmed, V4-V5
		  ○ 18S – Silva LSU, quality filtered, trimmed, V4 (Quast *et al.*, 2013)
		  ○ ITS – Unite fungal database, ITS2 (Kõljalg *et al.*, 2013)
		• Generate Biom files (.json)

	Summary analysis	• Align centroids (Mafft) (Katoh *et al.*, 2002)
		• Build phylogenetic tree (FastTree 2) (Price *et al.*, 2010)
		• QIIME v1.91 core diversity analysis is run. This script creates:
		  ○ Alpha rarefaction plots
		  ○ Beta diversity and PCoA plots
		  ○ Graphical taxonomic summaries
		  ○ Lists of OTU's enriched in environmental conditions

	Linking to other projects	• OTU's from the project are entered into a database of OTU's from other projects • The centroids from all projects are clustered to identify centroids and samples from other projects that match

Table 1 The iTag amplicon sequencing and analysis process at JGI.

References

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**:335–336.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**:2460–1.

Katoh K, Misawa K, Kuma K, Miyata T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**:3059–3066.

Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, *et al.* (2013). Towards a unified paradigm for sequence-based identification of fungi. *Mol Ecol* **22**:5271–7.

Menkis A, Burokienė D, Gaitnieks T, Uotila A, Johannesson H, Rosling A, *et al.* (2012). Occurrence and impact of the root-rot biocontrol agent Phlebiopsis gigantea on soil fungal communities in Picea abies forests of northern Europe. *FEMS Microbiol Ecol* **81**:438–45.

Parada A, Needham DM, Fuhrman JA. (2015). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time-series and global field samples. *Environ Microbiol*. doi:10.1111/1462-2920.13023.

Price MN, Dehal PS, Arkin AP. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**:e9490.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**:D590–6.

Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner H-W, *et al.* (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* **19 Suppl 1**:21–31.

Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, *et al.* (2015). Primer and platform effects on 16S rRNA tag sequencing. *Front Microbiol* **6**:771.

White TJ, Bruns T, Lee S, Taylor JW, others. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In:*PCR protocols: a guide to methods and applications*, Vol. 18, San Diego, pp. 315–322.