16S/18S/ITS data analysis

Report v.1.1

Library preparation and sequencing

There are two amplification steps in the library workflow: an initial PCR amplification using locus specific PCR primers and a subsequent amplification that integrates relevant flow-cell binding domains and unique indices (NexteraXT Index Kit, FC-131-1001/FC-131-1002).
This method is used to amplify i) the variable V3 and V4 regions of the 16S rRNA gene aiming to characterize bacterial community compositions ii) ITS region for classification of fungal communities or iii) 18S rRNA gene to allow discrimination of other eukaryotic organisms.
The amplification is due to the following target sequences **16S-341F** 5'-CCTACGGGNGGCWGCAG-3' and **16S-805R** 5'-GACTACHVGGGTATCTAATCC-3' for the 16S locus; **ITS1** 5'- TCCGTAGGTGAACCTGCGG-3' and **ITS4** 5'-TCCTCCGCTTATTGATATGC-3' (White *et al*., 1990) for ITS locus; **18S-566F** 5'- CAGCAGCCGCGGTAATTCC-3' and **18S-1200R** 5'- CCCGTGTTGAGTCAAATTAAGC-3' (Hadziavdic *et al.*, 2014) for 18S locus.
Libraries are sequenced in a MiSeq run in paired end with 300-bp read length or HiSeq2500 with 250bp read length depending on the experiment type.

Standard bioinformatics analysis

IGAtech set up an internal pipeline to analyze 16S/18S/ITS amplicon sequences.

Description

Reads are de-multiplexed based on Illumina indexing system. Where amplicon length is permissive with the respective sequencing length, 3'-ends of pairs are overlapped to generate consensus pseudo-reads, while the remainders are maintained as separated pairs; for 16S sequencing we enforce retaining of only overlapping reads. After, a clipping routine is applied to remove low-quality bases at 3'tails. Reads are then retained if they maintain a minimum length of 200 bp. Any primer sequence at 5'-ends is removed and not accounted during the process.

Following the QIIME pipelines, the USEARCH algorithm (version 8.1.1756, 32-bit) allows following steps: chimera filtering, grouping of replicate sequences; sorting sequences per decreasing abundance and OTU identification. The Operational Taxonomic Unit (OTU) picking aims to group query sequences into clusters, represented by centroids. Each centroid shares a level of similarity with their member sequences. Two methods of analysis can be adopted depending on the need of each projects and the number of samples. Open-reference algorithm is used as default approach unless differently enquired. All reads are used in the analysis if they maintain a minimum length of 200bp after removal of primer sequences and low-quality bases. Paired reads with permissive overlap at their 3'-ends are merged to single fragment and used as such to improve assignment accuracy. Reads that do not support overlap are maintained in the pool for downstream processing.

In a "closed-reference" analysis query sequences not sharing similarity with a centroid (in such case a centroid of the properly selected database) will be discarded. In an "open-reference" analysis, sequences not matching any reference sequence of the database will constitute a novel OTU and the most abundant and long read in each OTU is selected as the representative sequence. In a "closed-reference" strategy, fragments will be aligned to a reference database and only matches with a minimum identity of 94% will be retained and subjected to further classification; database sequences will be maintained as representative sequences of OTUs. In a "open-reference" analysis OTUs are built *de novo* with a clustering threshold set at 97%, with sequences that passed a pre-filter step for minimum identity of 90% with any sequence present in the reference database. OTUs in "open-reference" analysis are generated with a minimum of 2 sequenced fragments.

Rarefaction curves end-points and normalization of counts for diversity analysis are set to 50% of the target sequencing coverage (i.e. for 100,000 fragments a cutoff of 50,000 fragments is applied). The cutoff can be modified accordingly to the sequencing yields. Samples not satisfying the count threshold will not be included in standard alpha- and beta-diversity estimators. The total count is retained for taxonomic abundances estimation and used accordingly for ad-hoc statistical testing of taxonomic abundance when enquired.

The RDP classifier and Reference database are used to assign taxonomy with a minimum confidence threshold of 0.50.

Reference databases:

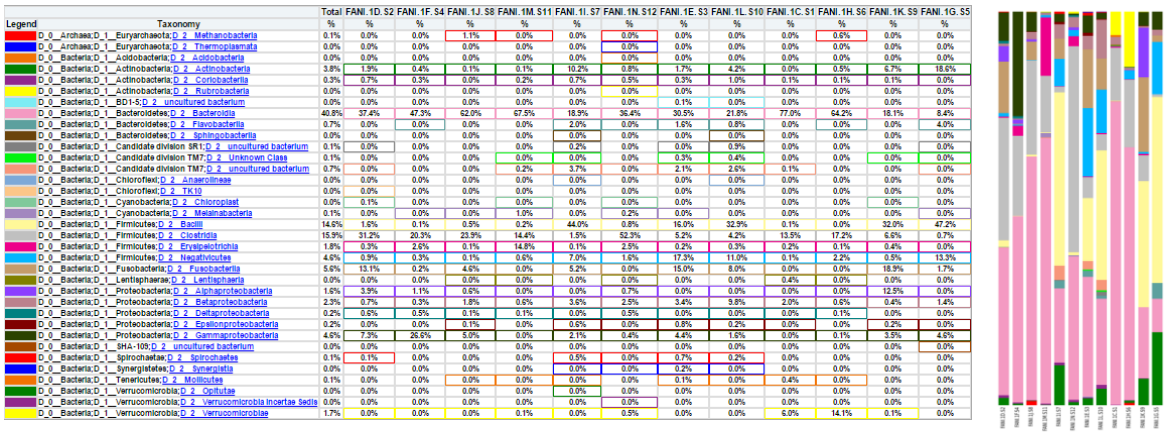
- 16S: modified GreenGene database (version 2013_8)
- 18S: internal database
- ITS: UNITE database

Files for customer

The "**summary_[pipeline_label]"** folder contains:

- The "**absolute_tax-count_tables_no-singletons"** folder contains '.biom' and '.txt' files reports absolute values (i.e read counts) for each single taxonomy label by each classification level (from phylum [L2] to species [L7]) after the removal of singleton sequences (OTU made by a single read). While all '.txt' files can be displayed by Excel or any other text editor, '.biom' files (version 2.1, with HDF5 format) are intended to be used with other software. It is not intended for visualization.

- The "**relative_tax-count_tables_no-singletons_plots"** folder which contains **bar_charts.html** (Figure 1, can be displayed in a Web browser; i.e.Chrome, Internet Explorer, Firefox) providing a graphical representation of how many reads (on a % basis) from each sample were assigned to a category in each taxonomic level. To simplify the comparison, the bar charts provide mouse-over information about taxonomies and relative abundance (maintaining the same color through different samples). All labels are resumed into the samples table below the chart. All **.biom/.txt** files indicated by L[2-7] suffix will contain relative quantities at taxonomy level (i.e. all OTU assigned to the same taxonomic lineage are collapsed at the given taxonomical level). Singletons (OTU made from a single sequence) are not accounted in this analysis.



Legend	Taxonomy	Total %	FANI.1D.S2 %	FANI.1F.S4 %	FANI.1J.S8 %	FANI.1M.S11 %	FANI.1I.S7 %	FANI.1N.S12 %	FANI.1E.S3 %	FANI.1L.S10 %	FANI.1C.S1 %	FANI.1H.S6 %	FANI.1K.S9 %	FANI.1G.S5 %
	D_0__Archaea;D_1__Euryarchaeota;D_2__Methanobacteria	0.1%	0.0%	0.0%	1.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%	0.0%	0.0%
	D_0__Archaea;D_1__Euryarchaeota;D_2__Thermoplasmata	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Acidobacteria;D_2__Acidobacteria	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria	3.8%	1.9%	0.4%	0.1%	0.1%	10.2%	0.8%	1.7%	4.2%	0.0%	0.5%	6.7%	18.6%
	D_0__Bacteria;D_1__Actinobacteria;D_2__Coriobacteriia	0.3%	0.7%	0.3%	0.0%	0.2%	0.7%	0.5%	0.3%	1.0%	0.1%	0.1%	0.1%	0.0%
	D_0__Bacteria;D_1__Actinobacteria;D_2__Rubrobacteria	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__BD1-5;D_2__uncultured bacterium	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia	40.8%	37.4%	47.3%	62.0%	67.5%	18.9%	36.4%	30.5%	21.8%	77.0%	64.2%	18.1%	8.4%
	D_0__Bacteria;D_1__Bacteroidetes;D_2__Flavobacteriia	0.7%	0.0%	0.0%	0.0%	0.0%	2.0%	0.0%	1.6%	0.8%	0.0%	0.0%	0.0%	4.0%
	D_0__Bacteria;D_1__Bacteroidetes;D_2__Sphingobacteriia	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Candidate division SR1;D_2__uncultured bacterium	0.1%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Candidate division TM7;D_2__Unknown Class	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.4%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Candidate division TM7;D_2__uncultured bacterium	0.7%	0.0%	0.0%	0.0%	0.2%	3.7%	0.0%	2.1%	2.6%	0.1%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Chloroflexi;D_2__Anaerolineae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Chloroflexi;D_2__TK10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Cyanobacteria;D_2__Chloroplast	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Cyanobacteria;D_2__Melainabacteria	0.1%	0.0%	0.0%	0.0%	1.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Firmicutes;D_2__Bacilli	14.6%	1.6%	0.1%	0.5%	0.2%	44.0%	0.8%	16.0%	32.9%	0.1%	0.0%	32.0%	47.2%
	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia	15.9%	31.2%	20.3%	23.9%	14.4%	1.5%	52.3%	5.2%	4.2%	13.5%	17.2%	6.6%	0.7%
	D_0__Bacteria;D_1__Firmicutes;D_2__Erysipelotrichia	1.8%	0.3%	2.6%	0.1%	14.8%	0.1%	2.5%	0.2%	0.3%	0.2%	0.1%	0.4%	0.0%
	D_0__Bacteria;D_1__Firmicutes;D_2__Negativicutes	4.6%	0.9%	0.3%	0.1%	0.6%	7.0%	1.6%	17.3%	11.0%	0.1%	2.2%	0.5%	13.3%
	D_0__Bacteria;D_1__Fusobacteria;D_2__Fusobacteriia	5.6%	13.1%	0.2%	4.6%	0.0%	5.2%	0.0%	15.0%	8.0%	0.0%	0.0%	18.9%	1.7%
	D_0__Bacteria;D_1__Lentisphaerae;D_2__Lentisphaeria	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Proteobacteria;D_2__Alphaproteobacteria	1.5%	3.9%	1.1%	0.5%	0.0%	0.0%	0.7%	0.0%	0.0%	0.0%	0.0%	12.5%	0.0%
	D_0__Bacteria;D_1__Proteobacteria;D_2__Betaproteobacteria	2.3%	0.7%	0.3%	1.8%	0.6%	3.6%	2.5%	3.4%	5.8%	2.0%	0.6%	0.4%	1.4%
	D_0__Bacteria;D_1__Proteobacteria;D_2__Deltaproteobacteria	0.2%	0.6%	0.5%	0.1%	0.1%	0.0%	0.5%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%
	D_0__Bacteria;D_1__Proteobacteria;D_2__Epsilonproteobacteria	0.2%	0.0%	0.0%	0.1%	0.0%	0.6%	0.0%	0.8%	0.2%	0.0%	0.0%	0.2%	0.0%
	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria	4.6%	7.3%	26.6%	5.0%	0.0%	2.1%	0.4%	4.4%	1.6%	0.0%	0.1%	3.5%	4.6%
	D_0__Bacteria;D_1__SHA-109;D_2__uncultured bacterium	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Spirochaetae;D_2__Spirochaetes	0.1%	0.1%	0.0%	0.0%	0.0%	0.5%	0.0%	0.7%	0.2%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Synergistetes;D_2__Synergistia	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Tenericutes;D_2__Mollicutes	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.4%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Verrucomicrobia;D_2__Opitutae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Verrucomicrobia;D_2__Verrucomicrobia Incertae Sedis	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	D_0__Bacteria;D_1__Verrucomicrobia;D_2__Verrucomicrobiae	1.7%	0.0%	0.0%	0.0%	0.1%	0.0%	0.5%	0.0%	0.0%	6.0%	14.1%	0.1%	0.0%

Figure 1. Taxonomic abundances are displayed as bar charts along with percentage of reads count per sample. Data can be browsed on html file and are stored in a "tsv" and a "BIOM" file.

- The **"qiime2_export_viz"** folder contains the OTU table '**otu_ table_no-singletons.qza**' and taxonomy assignment '**seqs_chimeras_filtered_rep_set_tax_assignments.qza**' and their visualization production '**taxa_barplots_no-singletons.qzv**' in the Qiime2 formats. **Qzv** file can be uploaded in https://view.qiime2.org/ to access interactive visualization and coloring of bar charts.

- The "**config_files"** folder contains parameters and configuration data used for the analysis pipeline.

- The **Core_diversity** folder contains all data about diversity.

  Data is stored in two main folders '**arare_max*#'** and '**bdiv_even*#'** sub-folders (where # is a minimum number of fragments used for the analysis set as cutoff, see above).

The first folder contains all rarefaction datasets (α-diversity): raw permutation tables, rarefaction curves, box plots for each metadata factor along with significance testing (t test) for each of the diversity indexes used (Shannon, Chao1, doubles, Simpson, singles, observed, observed_OTUs, goods_coverage). Once the rarefaction_plots.html is open, select the diversity metric (i.e. observed_otus) and the category (i.e. sample ID) to reveal the plot (Figure 2).

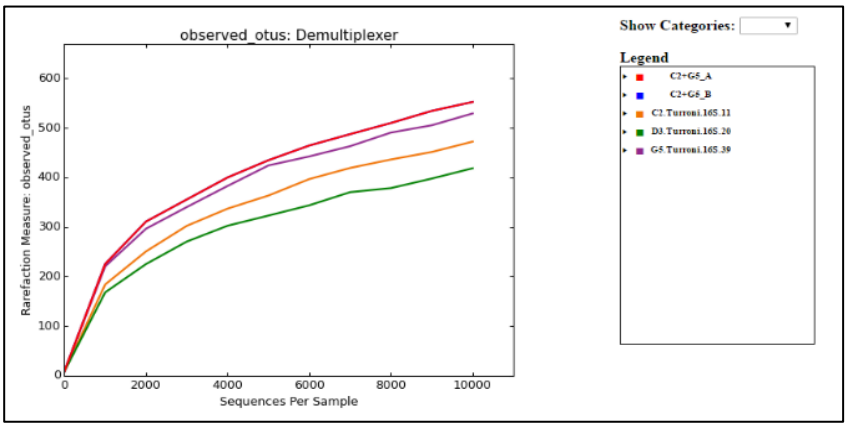


Figure 2. Alpha rarefaction is used to assess microbiome diversity in each sample. It also allows to estimate sequencing depth required to obtain exhaustive microbiome sampling.

The second folder contains distance matrix and principal coordinates based on Bray-Curtis estimation. All possible metadata grouping of distances pools are pair-wise tested for differences by a Student's t test; also, a non-parametric p-value is provided by 999 Monte Carlo permutations. Box plots of comparison are provided along. Distance input files for Emperor plot (based on β-diversity) are bray_curtis_dm.txt (distance matrix) and bray_curtis_pc.txt (principal components). To view the PCoA analysis in three dimensions with EMPEROR (Figure 3), open index.html in its folder. To the right side of the screen, the colors tab allows you to easily visualize 'clustering' by metadata category. The 3-D visualization software allows you to rotate the axes to see the data from different perspectives. In the "options" tab user can export the final image to a SVG file.

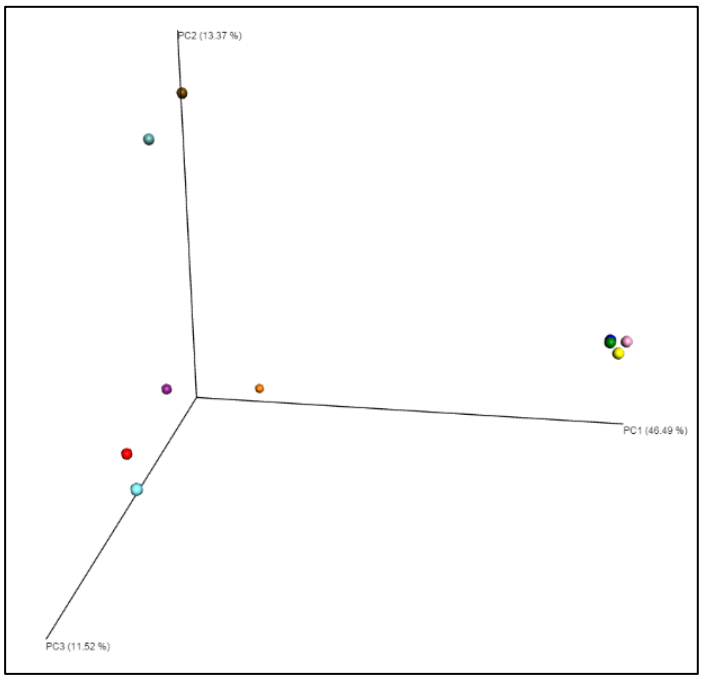


Figure 4: Alpha rarefaction is used to assess microbiome diversity in each sample. It also allows to estimate sequencing depth required to obtain exhaustive microbiome sampling.

The **index.html on the top directory** (can be displayed in a Web browser; i.e. Chrome, Safari, Firefox; **do not** use Internet Explorer), provide ease of access to all the analyses mentioned above.

- *Run summary data* section of the index.html main table list the **biom_table_summary.txt** that is a summary table of utilized reads per sample (after any overlap or filtering routine) and biom files normalized to *# reads (**table_mc*#.biom.gz** / **table_even*#.biom.gz**) used respectively for α-diversity and β-diversity estimations.
- *Alpha diversity results* section of the index.html main table allow to view the rarefaction plot and the printable data and bar charts of diversity analysis for each paired metric-metadata category.
- *Group significance results* section (this section of the table is generated only when metadata are available) of the index.html main table contains **group_significance_*.txt** in which group of samples (clustered by available metadata factors) are compared at single OTU level for significance of enrichment in its abundance for some group (Kruskal-Wallis test). These .txt files are available into core_diversity folder.
- *Beta diversity results* section of the index.html main table contains **Stats.txt** and **Distances.pdf** that represent distributions of inter-sample diversity between all possible pairs of groups.

- *In front of specific enquires we can perform statistical testing for taxonomic enrichment across groups using the fitZIG function implemented in the metagenomeSeq package.*

- The **otu_table** .tsv (derived from its biom counterpart **out_table_sample-metadata.biom**) is a tab-delimited file (can be displayed by Excel) contains all the raw fragments count at OTU level for each sample and its relative taxonomic attribution. Biom file also contains attached sample metadata. The "no-singletons.biom" file is the same type of table but reduced from OTUs made up from a single sequence.

- The **seqs_chimeras_filtered.fna.gz is** the effective original dataset used for OTU picking procedure (reads are already overlapped, quality trimmed, primer-removed and chimera-filtered); FASTA, gzip-compressed. Use open source software like 7-zip to unpack.

- The **seqs_chimeras_filtered_otus.txt** describes all the sequences belonging to a specific OTU.

- The **seqs_chimeras_filtered_rep_set.fasta** is a FASTA-formatted file containing a representative sequence for each OTU present in the analysis.

- The **seqs_chimeras_filtered_rep_set_tax_assignments.txt** contains the final taxonomic assignment for each OTU and it relative score provided by RDP.