Data Analysis Report

GATC Microbiome Profiling v3.6

Project(s): NG-13133

Project description: Project 1

October 4, 2017

Table of Contents

1 Introduction

The analysis of genes common to or ubiquitous amongst various organisms like bacterial 16S rRNA or fungal ITS is a time- and cost-effective method to characterise microbial diversity in complex samples. Amplification and high-throughput sequencing of the hypervariable regions of these genes is therefore a commonly used method for studying phylogeny and taxonomy. It is particularly suitable for analysing diverse samples and unculturable microorganisms and is therefore usable for various industrial, agricultural, medical and environmental applications.
This amplicon-based method has been optimised regarding study design and bioinformatics processing to provide a ready to use solution for researchers who are seeking to characterise microbiomes from various sources and samples which are usually difficult to study.

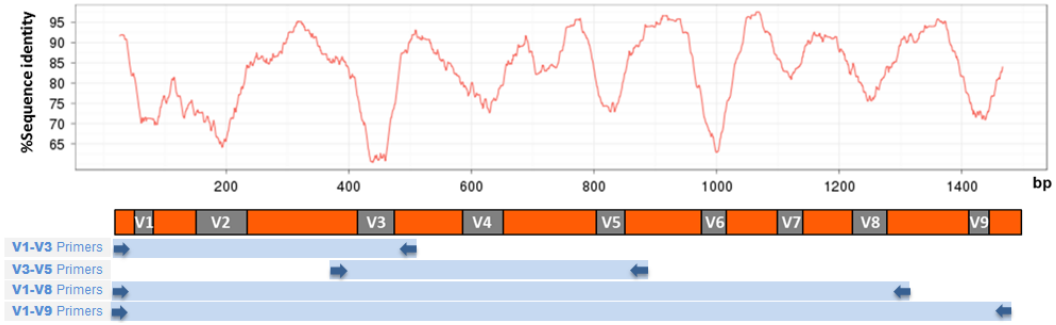

Figure 1: Schematic overview of the 16S rRNA gene. The sequence identity of the 16S rRNA gene of more than 6,000 bacteria compared to consensus sequence is shown. Dips indicate hypervariable regions. Hypervariable regions (V1-V9) are shown in grey and the conserved regions in orange.

2 Samples

Table 1: 16S Primers used.

Variable Region	Primer	Sequence	Product size[1]
V3-V5[1]	357F	CCTACGGGAGGCAGCAG	570 bp
	926R	CCGTCAATTCMTTTRAGT	

Table 2: Analysed samples.

Sample	File Name
Project_1_B1_t144	NG-13133_Project_1_B1_t144_lib196227_5593_2_1.fastq NG-13133_Project_1_B1_t144_lib196227_5593_2_2.fastq

[1] excluding primer lengths

Table 2: Analysed samples.

Sample	File Name
Project_1_B1_t48	NG-13133_Project_1_B1_t48_lib196224_5593_2_1.fastq NG-13133_Project_1_B1_t48_lib196224_5593_2_2.fastq
Project_1_G1_t144	NG-13133_Project_1_G1_t144_lib196228_5593_2_1.fastq NG-13133_Project_1_G1_t144_lib196228_5593_2_2.fastq
Project_1_G1_t48	NG-13133_Project_1_G1_t48_lib196225_5593_2_1.fastq NG-13133_Project_1_G1_t48_lib196225_5593_2_2.fastq
Project_1_G2_t144	NG-13133_Project_1_G2_t144_lib196229_5593_2_1.fastq NG-13133_Project_1_G2_t144_lib196229_5593_2_2.fastq
Project_1_G2_t48	NG-13133_Project_1_G2_t48_lib196226_5593_2_1.fastq NG-13133_Project_1_G2_t48_lib196226_5593_2_2.fastq
Project_1_t0	NG-13133_Project_1_t0_lib196223_5593_2_1.fastq NG-13133_Project_1_t0_lib196223_5593_2_2.fastq

3 Analysis Summary

3.1 Workflow

The schematic diagram of data analysis performed is displayed in the following graphic.

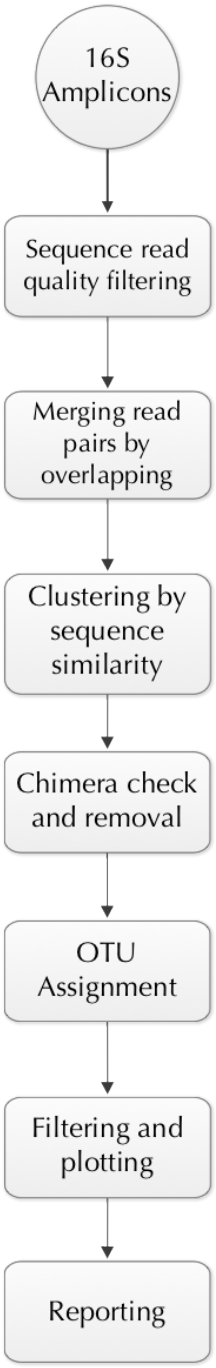

Figure 2: Microbiome Profiling Workflow

3.2 Merging read pairs by overlapping

In case of paired end sequencing where amplicons are sequenced in both the directions, the resulting read pairs are merged based on overlapping bases using FLASh[2] with maximum mismatch density of 0.25. Merging read pairs extends the read length to reflect the amplicon length which increases the possibility and accuracy of OTU assignment during the downstream processing.

3.3 Clustering by sequence similarity

In any given sample, there is an uneven representation of the microbiome biota which results in uneven amplification and sequence coverage. In order to reduce the computational time incurring for further downstream processing, the sequence data is compressed by performing sequence clustering based on 99% similarity accounting for PCR and sequencing errors (<1%). To achieve this, cd-hit[3], a clustering program is used. At high sequencing depth each original template is sequenced multiple times. Therefore singletons, clusters containing only one sequence, are removed from further analysis.

3.4 Chimera check and removal

PCR is an essential step in generating the amplicons from DNA samples. Due to the high similarity of different 16S rRNA, the possibility that small amounts of chimeric PCR products are generated is high. Therefore, the clustered data is checked for chimeras and the corresponding clusters are removed from further analysis. Chimera check is performed with UCHIME[4] using a full length, good quality, and non-chimeric 16S rRNA gene reference database.

3.5 OTU assignment

Non-chimeric, unique clusters are then subjected to BLASTn[5] analysis using non-redundant 16S rRNA reference sequences with an E-value cutoff of 1e-06.Reference 16S rRNA sequences are obtained from Ribosomal Database Project[6] (RDP Release 11 updated on September, 2016). Only good quality and unique 16S rRNA sequences which have a taxonomic assignment are considered and used as a reference database to assign operational taxonomic unit (OTU) status to the clusters. Taxonomic classification is based on NCBI Taxonomy[7] - http://www.ncbi.nlm.nih.gov/taxonomy. The number of sequences and length characteristics of the reference database used are described in table 3.

Table 3: Number of sequences and length characteristics for the reference database.

Total Sequences	Biggest	Smallest	Mean
11,795	1,768 bp	1,200 bp	1,461 bp

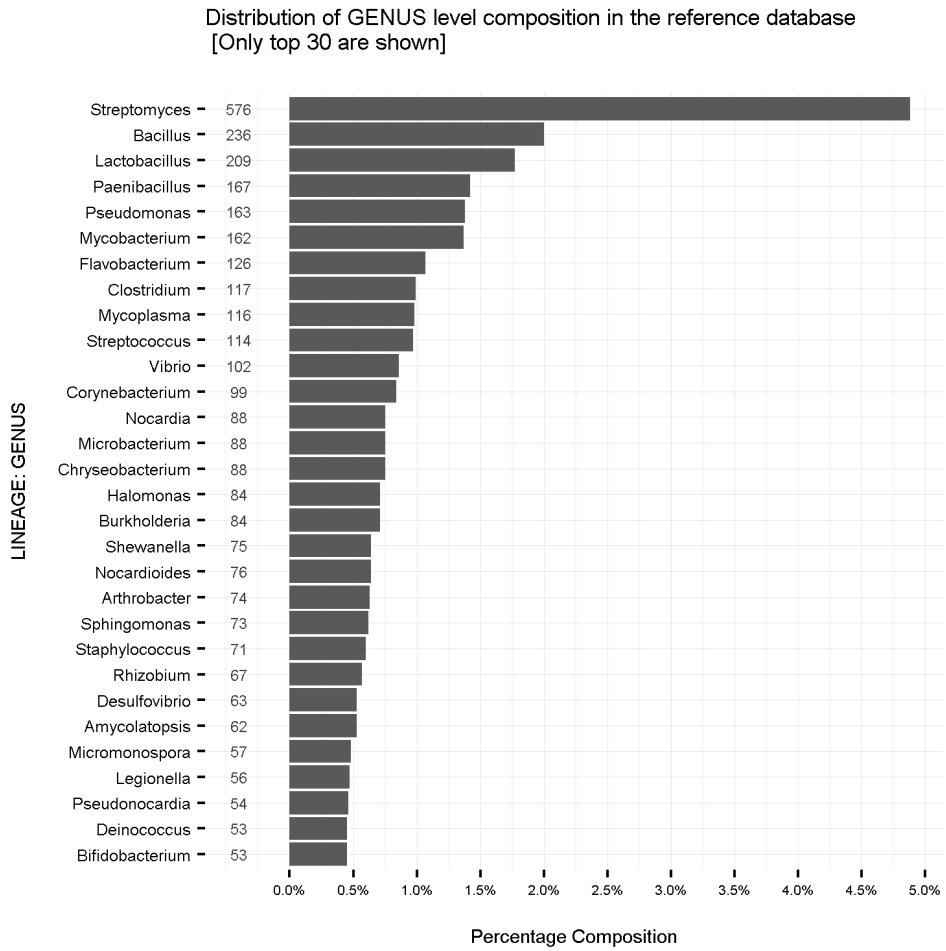

Figure 3: GENUS Distribution plot for the reference database.

3.6 Filtering and plotting

Except E-value cutoff (1e-06), no other thresholds were used during the BLAST analysis. All the hits to reference 16S rRNA database are considered and specific filters are applied to the hits to remove false positives. Further, best hit per cluster and multiple hits per cluster were analysed separately to determine the discriminatory power of the clusters with respect to the assigned OTUs. The various thresholds applied are reported in table 29. Finally, classification of OTU clusters and size of OTUs (number of reads within one cluster) are consolidated to compute relative abundancies (percentage composition).

4 Results

4.1 Read statistics

The number of reads and clusters which passed every computational step are listed in the read statistic tables. It gives an overview of the subset of reads that carry the discriminatory phylogenetic signal.

4.2 OTU tables

The complete taxonomy of the OTU clusters is provided in the following file. The various levels reported based on taxonomic lineage are mentioned in table 30.

- Sample.OTU.list.tsv
- Sample.OTU.list.high_abundant.tsv

Genus specific OTU tables are provided summarizing the read level measurements at genus and species level, respectively. In addition, the percentage abundance is determined for each OTU by considering all the OTUs in the sample and is used as a criterion (>0.5%) to select highly abundant OTUs in the sample.

- Sample.GENUS.ReadCounts.tsv
- Sample.GENUS.ReadCounts.high_abundant.tsv

4.3 OTU distribution plots

The distribution of highly abundant (>0.5% representation in the sample) OTUs at GENUS level are given in the sample specific OTU distribution plot(s). Such plots are helpful to capture the prominent OTUs present in each sample.

4.4 Sequences

In addition to the OTU tables, various sequence files are delivered. The following table lists the files and description.

Table 4: Sequence files delivered.

Description	File
Non-chimeric and unique clusters	Sample.non_chimeric_clusters.fasta
Discarded clusters / singletons	Sample.discarded_clusters.fasta
Clusters with no OTU assignment	Sample.no_hits.fasta
OTU assigned clusters (multiple filter failed)	Sample.multiple_hits.min_similarity.97.00.failed.fasta
OTU assigned clusters (best filter failed)	Sample.best_hits.min_similarity.97.00.failed.fasta

Table 5: QC report table for sample Project_1_B1_t144.

Description	Read pairs	% Read pairs
Total	3,298,680	100.0
Cleaned	1,053,233	31.9
Cleaned (orphan)	1,155,581	35.0
Merged by overlapping	984,872	29.9
Clustered by similarity	652,381	19.8
Chimeric	224,732	6.8
Final high quality	427,649	13.0

Table 6: OTU assignment table for sample Project_1_B1_t144.

Description	Read pairs	% Read pairs
High quality	427,649	100.0
OTU assigned	427,649	100.0
Filter passed OTUs (best hit only)	398,853	93.3
Filter passed OTUs (multiple hits)	390,026	91.2

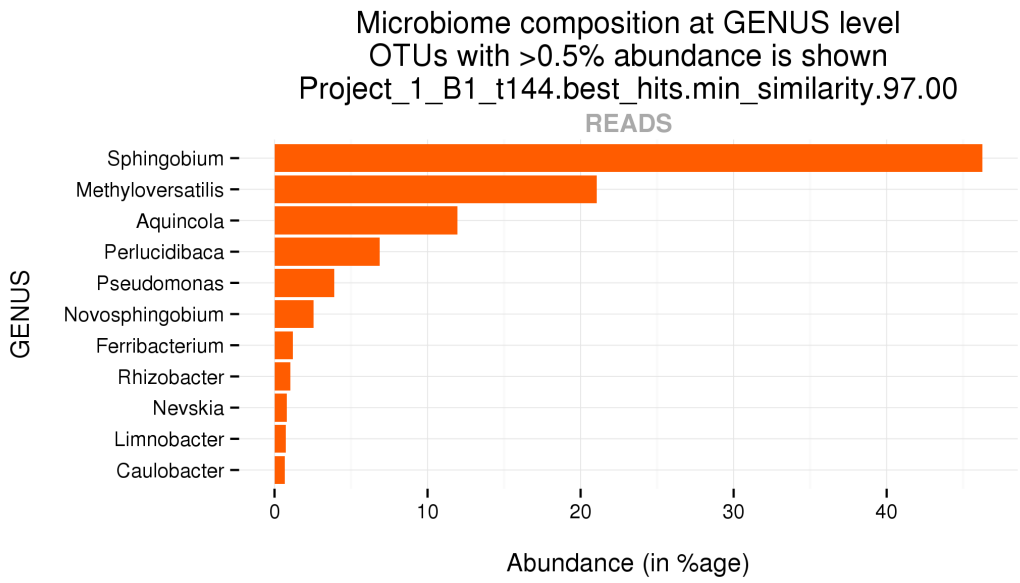

Figure 4: Distribution plot for sample Project_1_B1_t144.

Table 7: OTU table for sample Project_1_B1_t144 (file: Project_1_B1_t144.Project_1_B1_t144.best_hits.min_similarity.97.00.OTU.list.high_abundant.tsv).

% ABUNDANCE	TAXA_ID	KINGDOM	PHYLUM	CLASS	ORDER	FAMILY	GENUS	SPECIES
42.18	121428	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	Sphingobium xenophag...
20.94	1000565	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Methyloversatilis	Methyloversatilis un...
11.96	391953	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Aquincola	Aquincola_tertiaric...
6.88	392589	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Perlucidibaca	Perlucidibaca piscin...
3.59	432308	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	Sphingobium rhizovic...
2.64	674054	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas_baetica
2.35	48936	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Novosphingobium	Novosphingobium subt...
1.19	76259	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Ferribacterium	Ferribacterium limne...
0.98	1395940	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Rhizobacter	Rhizobacter_sp._PL...
0.8	64002	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Sinobacteraceae	Nevskia	Nevskia ramosa
0.74	131080	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Limnobacter	Limnobacter thiooxid...

Table 8: QC report table for sample Project_1_B1_t48.

Description	Read pairs	% Read pairs
Total	2,074,420	100.0
Cleaned	663,370	32.0
Cleaned (orphan)	733,942	35.4
Merged by overlapping	631,004	30.4
Clustered by similarity	380,155	18.3
Chimeric	205,775	9.9
Final high quality	174,380	8.4

Table 9: OTU assignment table for sample Project_1_B1_t48.

Description	Read pairs	% Read pairs
High quality	174,380	100.0
OTU assigned	174,380	100.0
Filter passed OTUs (best hit only)	128,677	73.8
Filter passed OTUs (multiple hits)	124,183	71.2

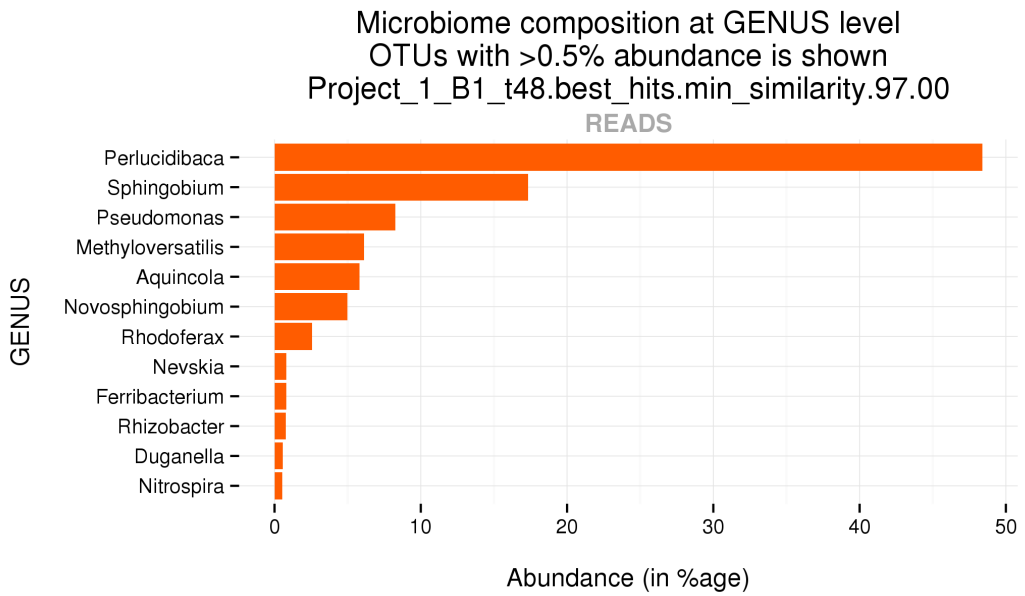

Figure 5: Distribution plot for sample Project_1_B1_t48.

Table 10: OTU table for sample Project_1_B1_t48 (file: Project_1_B1_t48.Project_1_B1_t48.best_hits.min_similarity.97.00.OTU.list.high_abundant.tsv).

% ABUNDANCE	TAXA_ID	KINGDOM	PHYLUM	CLASS	ORDER	FAMILY	GENUS	SPECIES
48.39	392589	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Perlucidibaca	Perlucidibaca piscin...
15.91	121428	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	Sphingobium xenophag...
6.11	1000565	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Methyloversatilis	Methyloversatilis un...
5.82	391953	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Aquincola	Aquincola_tertiaric...
5.34	674054	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas_baetica
4.7	48936	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Novosphingobium	Novosphingobium subt...
2.5	614083	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Rhodoferax	Rhodoferax saiden-bac...
1.14	432308	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	Sphingobium rhizovic...
1.07	1148509	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas sp.
0.82	64002	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Sinobacteraceae	Nevskia	Nevskia ramosa
0.8	76259	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Ferribacterium	Ferribacterium limne...
0.72	1395940	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Rhizobacter	Rhizobacter_sp._PL...
0.57	762836	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	Duganella	Duganella phyllospha...
0.52	42253	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira moscovien...

Table 11: QC report table for sample Project_1_G1_t144.

Description	Read pairs	% Read pairs
Total	2,739,959	100.0
Cleaned	676,892	24.7
Cleaned (orphan)	1,007,542	36.8
Merged by overlapping	625,255	22.8
Clustered by similarity	375,955	13.7
Chimeric	113,215	4.1
Final high quality	262,740	9.6

Table 12: OTU assignment table for sample Project_1_G1_t144.

Description	Read pairs	% Read pairs
High quality	262,740	100.0
OTU assigned	262,740	100.0
Filter passed OTUs (best hit only)	246,591	93.9
Filter passed OTUs (multiple hits)	245,020	93.3

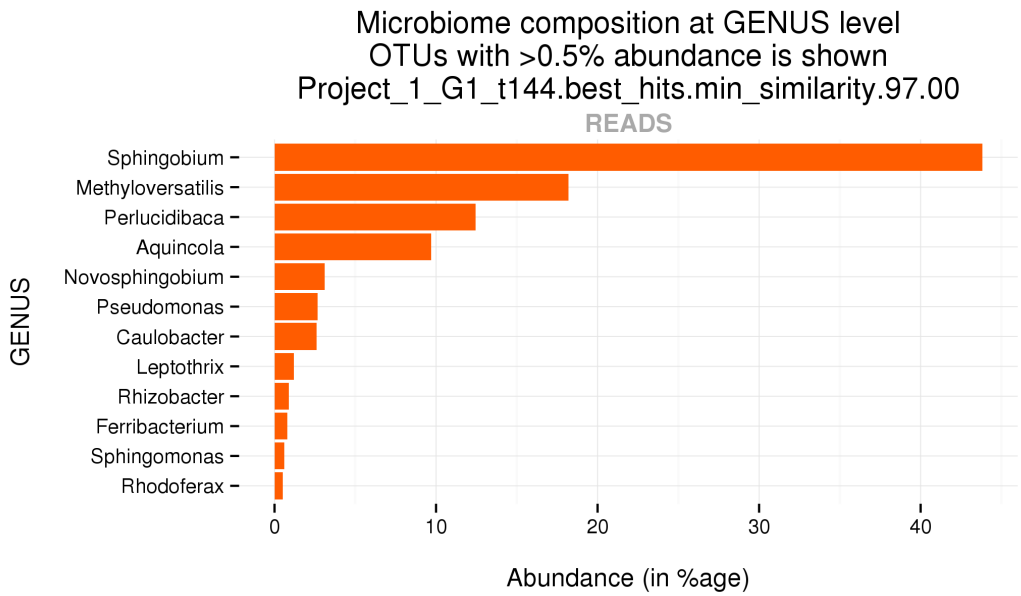

Figure 6: Distribution plot for sample Project_1_G1_t144.

Table 13: OTU table for sample Project_1_G1_t144 (file: Project_1_G1_t144.Project_1_G1_t144.best_hits.min_similarity.97.00.OTU.list.high_abundant.tsv).

% ABUNDANCE	TAXA_ID	KINGDOM	PHYLUM	CLASS	ORDER	FAMILY	GENUS	SPECIES
40.38	121428	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	Sphingobium xenophag...
18.19	1000565	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Methyloversatilis	Methyloversatilis un...
12.45	392589	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Perlucidibaca	Perlucidibaca piscin...
9.71	391953	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Aquincola	Aquincola_tertiaric...
2.85	432308	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	Sphingobium rhizovic...
2.64	48936	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Novosphingobium	Novosphingobium subt...
1.97	674054	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas_baetica
1.4	1380046	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Caulobacter	Caulobacter profundu...
1.19	89	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Leptothrix	Leptothrix_discopho...
0.85	1395940	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Rhizobacter	Rhizobacter_sp._PL...
0.78	76259	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Ferribacterium	Ferribacterium limne...
0.72	88688	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Caulobacter	Caulobacter segnis

Table 14: QC report table for sample Project_1_G1_t48.

Description	Read pairs	% Read pairs
Total	2,689,623	100.0
Cleaned	693,362	25.8
Cleaned (orphan)	1,024,545	38.1
Merged by overlapping	651,147	24.2
Clustered by similarity	357,922	13.3
Chimeric	144,566	5.4
Final high quality	213,356	7.9

Table 15: OTU assignment table for sample Project_1_G1_t48.

Description	Read pairs	% Read pairs
High quality	213,356	100.0
OTU assigned	213,356	100.0
Filter passed OTUs (best hit only)	94,686	44.4
Filter passed OTUs (multiple hits)	93,605	43.9

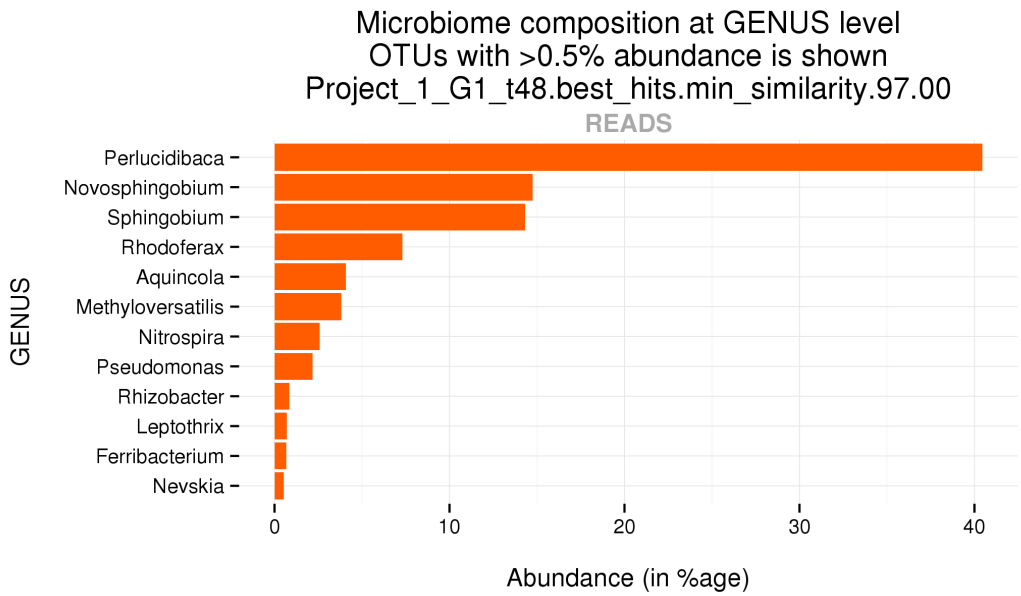

Figure 7: Distribution plot for sample Project_1_G1_t48.

Table 16: OTU table for sample Project_1_G1_t48 (file: Project_1_G1_t48.Project_1_G1_t48.best_hits.min_similarity.97.00.OTU.list.high_abundant.tsv).

% ABUNDANCE	TAXA_ID	KINGDOM	PHYLUM	CLASS	ORDER	FAMILY	GENUS	SPECIES
40.45	392589	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Perlucidibaca	Perlucidibaca piscin...
14.02	48936	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Novosphingobium	Novosphingobium subt...
13.01	121428	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	Sphingobium xenophag...
6.95	614083	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Rhodoferax	Rhodoferax saiden-bac...
4.08	391953	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Aquincola	Aquincola_tertiaric...
3.8	1000565	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Methyloversatilis	Methyloversatilis un...
2.58	42253	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira moscovien...
0.9	674054	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas_baetica
0.87	432308	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	Sphingobium rhizovic...
0.77	1395940	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Rhizobacter	Rhizobacter_sp._PL...
0.69	89	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Leptothrix	Leptothrix_discopho...
0.67	76259	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Ferribacterium	Ferribacterium limne...
0.57	1148509	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas sp.
0.52	64002	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Sinobacteraceae	Nevskia	Nevskia ramosa

Table 17: QC report table for sample Project_1_G2_t144.

Description	Read pairs	% Read pairs
Total	2,172,180	100.0
Cleaned	645,765	29.7
Cleaned (orphan)	760,912	35.0
Merged by overlapping	607,912	28.0
Clustered by similarity	375,941	17.3
Chimeric	160,755	7.4
Final high quality	215,186	9.9

Table 18: OTU assignment table for sample Project_1_G2_t144.

Description	Read pairs	% Read pairs
High quality	215,186	100.0
OTU assigned	215,186	100.0
Filter passed OTUs (best hit only)	196,521	91.3
Filter passed OTUs (multiple hits)	193,481	89.9

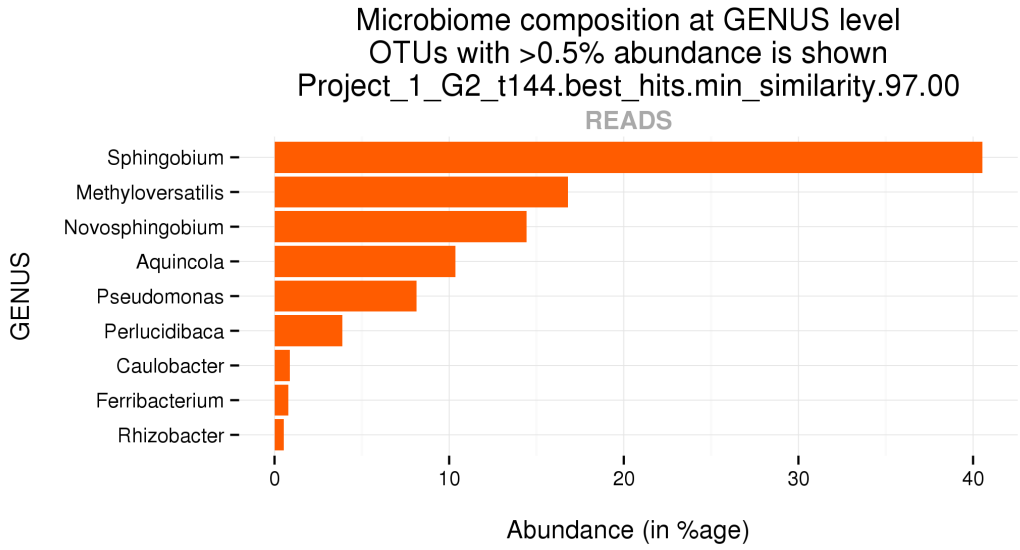

Figure 8: Distribution plot for sample Project_1_G2_t144.

Table 19: OTU table for sample Project_1_G2_t144 (file: Project_1_G2_t144.Project_1_G2_t144.best_hits.min_similarity.97.00.OTU.list.high_abundant.tsv).

% ABUNDANCE	TAXA_ID	KINGDOM	PHYLUM	CLASS	ORDER	FAMILY	GENUS	SPECIES
37.21	121428	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	Sphingobium xenophag...
16.78	1000565	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Methyloversatilis	Methyloversatilis un...
13.64	48936	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Novosphingobium	Novosphingobium subt...
10.36	391953	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Aquincola	Aquincola_tertiaric...
6.07	674054	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas_baetica
3.87	392589	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Perlucidibaca	Perlucidibaca piscin...
2.42	432308	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	Sphingobium rhizovic...
0.8	76259	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Ferribacterium	Ferribacterium limne...
0.62	1007511	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	Sphingobium limnetic...
0.59	494916	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas brassica...
0.58	1148509	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas sp.
0.51	1395940	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Rhizobacter	Rhizobacter_sp._PL...

Table 20: QC report table for sample Project_1_G2_t48.

Description	Read pairs	% Read pairs
Total	2,557,593	100.0
Cleaned	830,144	32.5
Cleaned (orphan)	891,956	34.9
Merged by overlapping	791,669	31.0
Clustered by similarity	488,314	19.1
Chimeric	117,144	4.6
Final high quality	371,170	14.5

Table 21: OTU assignment table for sample Project_1_G2_t48.

Description	Read pairs	% Read pairs
High quality	371,170	100.0
OTU assigned	371,170	100.0
Filter passed OTUs (best hit only)	311,256	83.9
Filter passed OTUs (multiple hits)	115,664	31.2

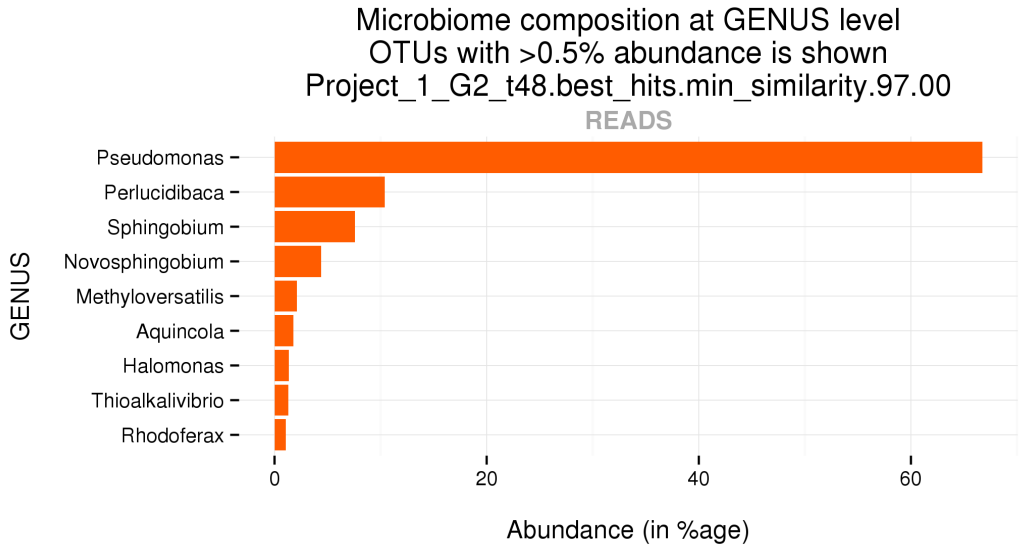

Figure 9: Distribution plot for sample Project_1_G2_t48.

Table 22: OTU table for sample Project_1_G2_t48 (file: Project_1_G2_t48.Project_1_G2_t48.best_hits.min_similarity.97.00.OTU.list.high_abundant.tsv).

% ABUNDANCE	TAXA_ID	KINGDOM	PHYLUM	CLASS	ORDER	FAMILY	GENUS	SPECIES
61.13	556533	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas_ benzeni...
10.37	392589	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Perlucidibaca	Perlucidibaca piscin...
6.98	121428	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium	Sphingobium xenophag...
4.24	48936	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Novosphingobium	Novosphingobium subt...
3.19	674054	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas_baetica
2.09	1000565	Bacteria	Proteobacteria	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Methyloversatilis	Methyloversatilis un...
1.8	391953	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Aquincola	Aquincola_tertiaric...
1.3	419597	Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	Halomonas	Halomonas shengliens...
1.3	396588	Bacteria	Proteobacteria	Gammaproteobacteria	Chromatiales	Ectothiorhodospiraceae	Thioalkalivibrio	Thioalkalivibrio sul...
1.03	614083	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Rhodoferax	Rhodoferax saiden-bac...
0.8	494916	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas bras-sica...
0.72	132476	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	Pseudomonas kilo-nens...

Table 23: QC report table for sample Project_1_t0.

Description	Read pairs	% Read pairs
Total	2,210,881	100.0
Cleaned	500,333	22.6
Cleaned (orphan)	843,775	38.2
Merged by overlapping	459,111	20.8
Clustered by similarity	226,821	10.3
Chimeric	468	0.0
Final high quality	226,353	10.2

Table 24: OTU assignment table for sample Project_1_t0.

Description	Read pairs	% Read pairs
High quality	226,353	100.0
OTU assigned	226,351	100.0
Filter passed OTUs (best hit only)	30,167	13.3
Filter passed OTUs (multiple hits)	29,938	13.2

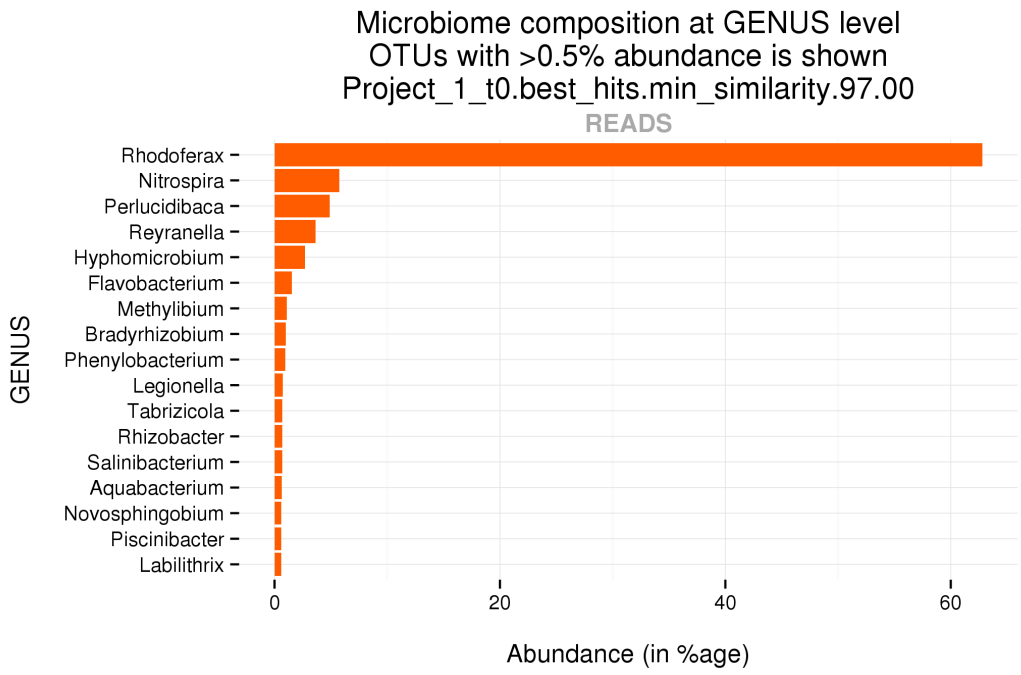

Figure 10: Distribution plot for sample Project_1_t0.

Table 25: OTU table for sample Project_1_t0 (file: Project_1_t0.Project_1_t0.best_hits.min_similarity.97.00.OTU.list.high_abundant.tsv).

% ABUNDANCE	TAXA_ID	KINGDOM	PHYLUM	CLASS	ORDER	FAMILY	GENUS	SPECIES
62.38	614083	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Rhodoferax	Rhodoferax saidenbac...
5.77	42253	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira moscovien...
4.89	392589	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Perlucidibaca	Perlucidibaca piscin...
2.92	1230389	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	unclassified Rhodospirillales	Reyranella	Reyranella soli
1.57	83	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	Hyphomicrobium vulga...
1.09	420662	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Methylibium	Methylibium_petrole...
0.78	1114874	Bacteria	Bacteroidetes	Flavobacteriia	Flavobacteriales	Flavobacteriaceae	Flavobacterium	Flavobacterium_pisc...
0.72	1205680	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	unclassified Rhodospirillales	Reyranella	Reyranella massilien...
0.69	909926	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Tabrizicola	Tabrizicola aquatica
0.67	386302	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	Salinibacterium	Salinibacterium xinj...
0.65	582840	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium	Hyphomicrobium facil...
0.61	392597	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Piscinibacter	Piscinibacter_aquat...
0.6	1179767	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Bradyrhizobium	Bradyrhizobium_ganz...
0.58	1391654	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales	Labilitrichaceae	Labilithrix	Labilithrix_luteola
0.54	1395940	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Unknown	Rhizobacter	Rhizobacter_sp._PL...

5 Deliverables

Table 26: List of deliverable files, format and recommended programs to access.

File	Format	Open file with
Sample.best_hits.min_similarity.97.00.OTU.list.tsv	TSV	Spreadsheet editor
Sample.best_hits.min_similarity.97.00.OTU.list.high_abundant.tsv	TSV	Spreadsheet editor
Sample.best_hits.min_similarity.97.00.GENUS.ReadCounts.tsv	TSV	Spreadsheet editor
Sample.best_hits.min_similarity.97.00.GENUS.ReadCounts.high_abundant.tsv	TSV	Spreadsheet editor
Sample.best_hits.min_similarity.97.00.OTU.distribution.png	PNG	Image viewer
Sample.best_hits.min_similarity.97.00.failed.fasta	FASTA	Text editor
Sample.multiple_hits.min_similarity.97.00.OTU.list.tsv	TSV	Spreadsheet editor
Sample.multiple_hits.min_similarity.97.00.OTU.list.high_abundant.tsv	TSV	Spreadsheet editor
Sample.multiple_hits.min_similarity.97.00.GENUS.ReadCounts.tsv	TSV	Spreadsheet editor
Sample.multiple_hits.min_similarity.97.00.GENUS.ReadCounts.high_abundant.tsv	TSV	Spreadsheet editor
Sample.multiple_hits.min_similarity.97.00.OTU.distribution.png	PNG	Image viewer
Sample.multiple_hits.min_similarity.97.00.failed.fasta	FASTA	Text editor
Sample.min_similarity.97.00.otu_statistics.tsv	TSV	Spreadsheet editor
Sample.non_chimeric_clusters.fasta	FASTA	Text editor
Sample.discarded_clusters.fasta	FASTA	Text editor
Sample.no_hits.fasta	FASTA	Text editor
Microbiome_Profiling_Report.pdf	PDF	PDF reader

6 Formats

Table 27: References and descriptions of file formats

Format	Description
FASTQ[8]	Text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are encoded with a single ASCII character for brevity.
FASTA	Text-based format for storing a biological sequence (usually nucleotide sequence).
TSV	Tab seperated table style text file. Can be imported into spreadsheet processing software like MS OFFICE Excel.
PNG	Visual representation in Portable Network Graphics format.

7 Software Tools

Table 28: Name, Version, Reference and Description of relevant programs

Program	Version	Description
CD-HIT[3]	4.6	CD-HIT is a very widely used program for clustering and comparing protein or nucleotide sequences.
Cutadapt[9]	1.10	Cutadapt removes adapter sequences from DNA high-throughput sequencing data.
FLASh[2]	1.2.11	Fast Length Adjustment of Short reads (FLASh) is a very fast and accurate software tool to merge paired-end reads from next-generation sequencing experiments.
Krona[10]	2.5	Krona allows hierarchical data to be explored with zoomable pie charts.
NCBI BLAST+[5]	2.4.0+	The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.
R[11]	2.15.3	R is a programming language and environment for statistical computing.
Sickle[12]	1.33	Sickle is a tool that uses sliding windows along with quality and length thresholds to determine when quality is sufficiently low to trim the 3'-end of reads.
UCHIME[4]	4.2.40	UCHIME is an algorithm for detecting chimeric sequences, which are sequences formed from two or more biological sequences joined together during PCR.

8 Tables

Table 29: Description of filters used.

Name	Thresholds
% Identity	$\geq$ 97.00
E-value	$\leq$ 1e-06
% Alignment coverage	$\geq$ 95.00
Min. query length	428
% bitscore threshold for multiple hits	10
Max. hits to consider for multiple hits	50
% abundance	$>$ 0.5

Table 30: Description of Taxonomy.

No.	Name	Description	Example
1	READ.COUNTS	Number of sequences hitting the same OTU	8726
2	PERCENT.COMPOSITION.READS	Percentage of SEQUENCES hitting the same OTU and is calculated based on all the OTUs observed in the sample	21.09
3	CLUSTER.COUNTS	Number of CLUSTERS hitting the same OTU	499
4	PERCENT.COMPOSITION.CLUSTERS	Percentage of CLUSTERS hitting the same OTU and is calculated based on all the OTU CLUSTERs observed in the sample	11.12
5	GI_ID	NCBI GenBank ID	X80725
6	TAXA_ID	NCBI Taxon ID	866789
7	KINGDOM	Name of the kingdom	Bacteria
8	PHYLUM	Name of the phylum	Proteobacteria
9	CLASS	Name of the class	Gammaproteobacteria
10	ORDER	Name of the order	Enterobacteriales
11	FAMILY	Name of the family	Enterobacteriaceae
12	GENUS	Name of the genus	Escherichia
13	SPECIES	Name of the species	Escherichia coli DSM 30083

9 FAQ

Q: What is the necessary coverage for microbiome analysis?
A: The required sequencing depth mainly depends on the complexity of the sample (number, genome size and representation of individual species) and the aim of the project. If you expect your sample to contain only a few different bacteria, a low coverage is sufficient; with many different bacteria expected, a higher coverage is needed. In case of doubt we recommend determining the required depth of sequencing through performing a pilot on a sub-set of samples.

Q: Which organisms can be detected?
A: Phylogenic characterisation and analysis of microbial communities can be performed for various sample types and organisms. We have tested and demonstrated the utility of this approach for the identification and description of complex and non-complex food/industrial, environmental and medical samples. The focused sequencing of hypervariable regions enables the detection of bacteria present at extremely low frequency.

Q: Down to which taxonomic level can the microbiome be sequenced?
A: Usually the microbiome of a given sample can be resolved down to the genus level with a high degree of certainty. However, related organisms (e.g. belonging to the same genus) may have identical or very similar 16S rRNA genes and therefore, the species cannot be resolved. If the identification of closely related bacteria is of interest, sequencing of further 16S hypervariable regions and/or other genes can be performed.

Q: What is the difference between 'best_hits' and 'multiple_hits'?
A: Both refer to the same BLAST search. While the 'best_hits' summary only takes the first entry of the BLAST hits, the 'multiple_hits' summary utilizes all the hits for the statistics. An additional filter is applied in the 'multiple_hits' evaluation: if a sequence gets more than 50 or 250 hits (depending on the size of the database) these hits are discarded because the informative value is questionable.

Q: How can I open a TSV file in Excel?
A: Start Excel and click File -> Open and select the TSV file you want to open. Next an assistant dialog should show up. Make sure that you select tab as separator. Set the format of all rows without numbers to text. The TSV files use the dot as decimal mark and comma as thousands separator. Make sure that you set both correctly.

Bibliography

[1] A framework for human microbiome research. *Nature*, 486(7402):215–221, June 2012.

[2] Tanja Magoč and Steven L. Salzberg. FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies. *Bioinformatics*, 27(21):2957–2963, September 2011.

[3] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, July 2006.

[4] Robert C. Edgar, Brian J. Haas, Jose C. Clemente, Christopher Quince, and Rob Knight. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200, August 2011.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, October 1990.

[6] J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic acids research*, 37(Database issue):D141–D145, January 2009.

[7] Scott Federhen. The NCBI Taxonomy database. *Nucleic acids research*, 40(Database issue):D136–D143, January 2012.

[8] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, 2010.

[9] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011.

[10] Brian Ondov, Nicholas Bergman, and Adam Phillippy. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12(1):385+, 2011.

[11] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[12] Sickle. https://github.com/najoshi/sickle.

GATC Biotech AG
European Genome and Diagnostics Center
Jakob-Stadler-Platz 7
78467 Konstanz

www.gatc-biotech.com