Data Analysis Report

GATC Microbiome Profiling (Combined Analysis) v3.6

Project(s): NG-13133

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.

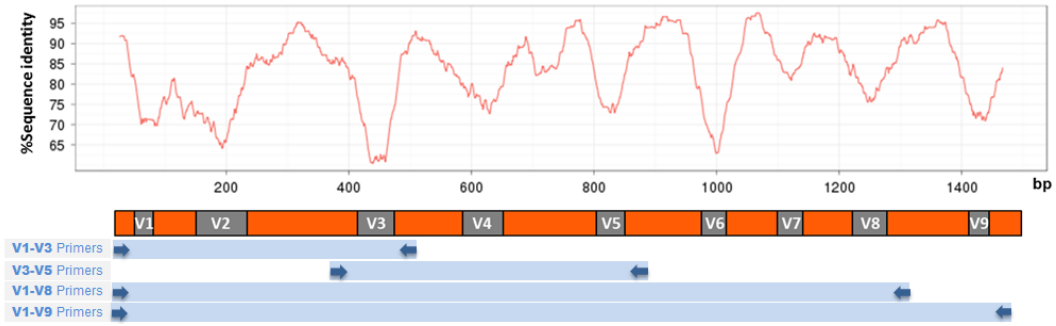

Figure 1: Schematic overview of the 16S rRNA gene. The sequence identity of the 16S rRNA gene of more than 6,000 bacteria compared to consensus sequence is shown. Dips indicate hypervariable regions. Hypervariable regions (V1-V9) are shown in grey and the conserved regions in orange.

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
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

[1] excluding primer lengths

Table 2: Analysed samples.

Sample	File Name
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.

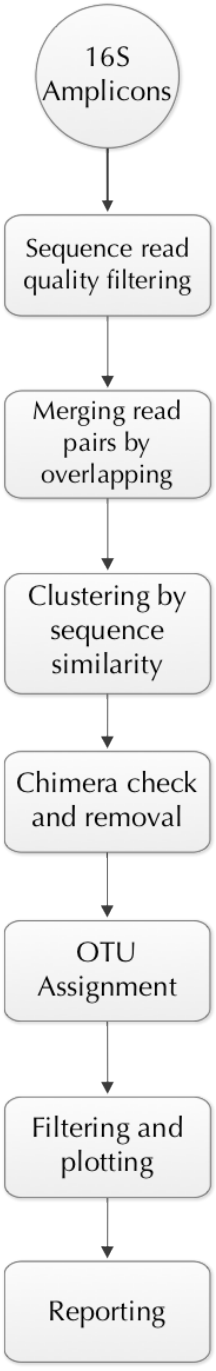

Figure 2: Microbiome Profiling Workflow

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

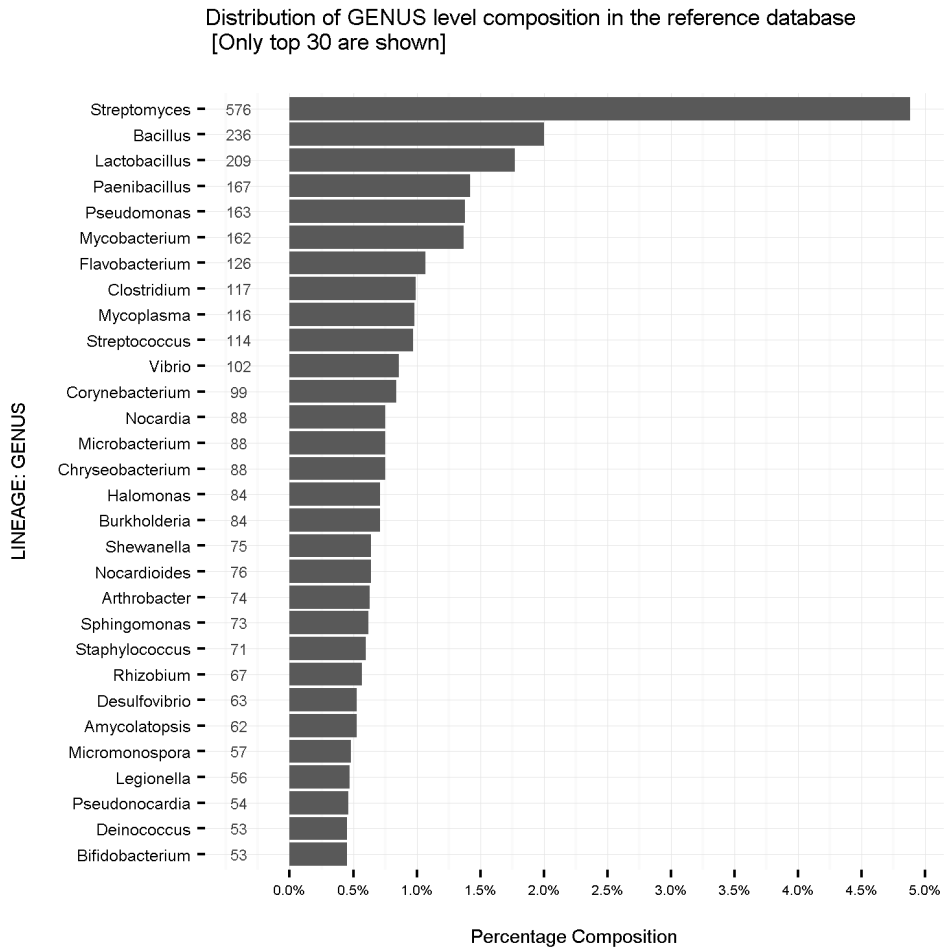

Figure 3: GENUS Distribution plot for the reference database.

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 19. Finally, classification of OTU clusters and size of OTUs (number of reads within one cluster) are consolidated to compute relative abundancies (percentage composition).

3.7 Microbial Diversity Indices

3.7.1 Overview

A diversity index is a quantitative measure that reflects how many different types (such as species) there are in a dataset, and simultaneously takes into account how evenly the basic entities (such as individuals) are distributed among those types. The value of a diversity index increases both when the number of species increases and when evenness increases. For a given number of species, the value of a diversity index is maximized when all species are equally abundant.

Although there are many indices available to measure the species diversity, the most popular ones used in sequence based OTU profiling are - Shannon index and Simpson index.

3.7.2 Shannon Index

The idea behind this index is that diversity of a community is similar to the amount of information contained in a sampled environment. The Shannon index increases as both the richness and the evenness of the community increase. Since the index incorporates both components of biodiversity (richness and evenness) it provides a simplistic summary of species diversity. On the other hand, it makes it difficult to compare communities that differ greatly in richness. In order to overcome this limitation, a second index - Simpson index is used for comparative studies, combining a direct estimate of species richness (the total number of species in the community - Shannon Index) with some measure of dominance or evenness (Simpson index) [8].

3.7.3 Simpson Index

It measures the probability of any two individuals randomly selected from a sample will belong to the same species. Simpson index gives the probability of any two individuals drawn from noticeably large community belonging to same species. Simpson index can be used to measure the species evenness, richness and diversity [9].

For general information about various diversity indices see [10].

4 Results

4.1 Read statistics

Table 4: Read statistics and OTU assignment

Sample	Total (Read pairs)	Cleaned	Merged by overlapping	Clustered by similarity	Chimeric	High quality	OTU assigned	Filter passed OTUs (best hit only)
Project_1_B1_t144	3,298,680	1,053,233	984,872	652,381	224,732	427,649	427,649	403,717
Project_1_B1_t48	2,074,420	663,370	631,004	380,155	205,775	174,380	174,380	132,722
Project_1_G1_t144	2,739,959	676,892	625,255	375,955	113,215	262,740	262,740	249,152
Project_1_G1_t48	2,689,623	693,362	651,147	357,922	144,566	213,356	213,356	103,558
Project_1_G2_t144	2,172,180	645,765	607,912	375,941	160,755	215,186	215,186	199,666
Project_1_G2_t48	2,557,593	830,144	791,669	488,314	117,144	371,170	371,170	316,700
Project_1_t0	2,210,881	500,333	459,111	226,821	468	226,353	226,351	45,741

4.2 Diversity index tables

Both Shannon and Simpson indices are computed using the R package Vegan[11]. The results are found in the diversity index tables (Taxa-level.diversity_index_tables.tsv) and a graphical representation of the indices are in diversity plots (Taxa-level.diversity.png).

4.3 OTU abundance tables

Abundance measured by the percentage of OTU assigned reads from various taxonomic level was computed. The measured abundance levels are in OTU distribution tables (Taxa-level.combined.table.tsv) and the bar plots representing the abundance levels at various taxonomic level are in OTU distribution plots (Taxa-level.OTU.distribution.combined.png).

4.4 Phylum

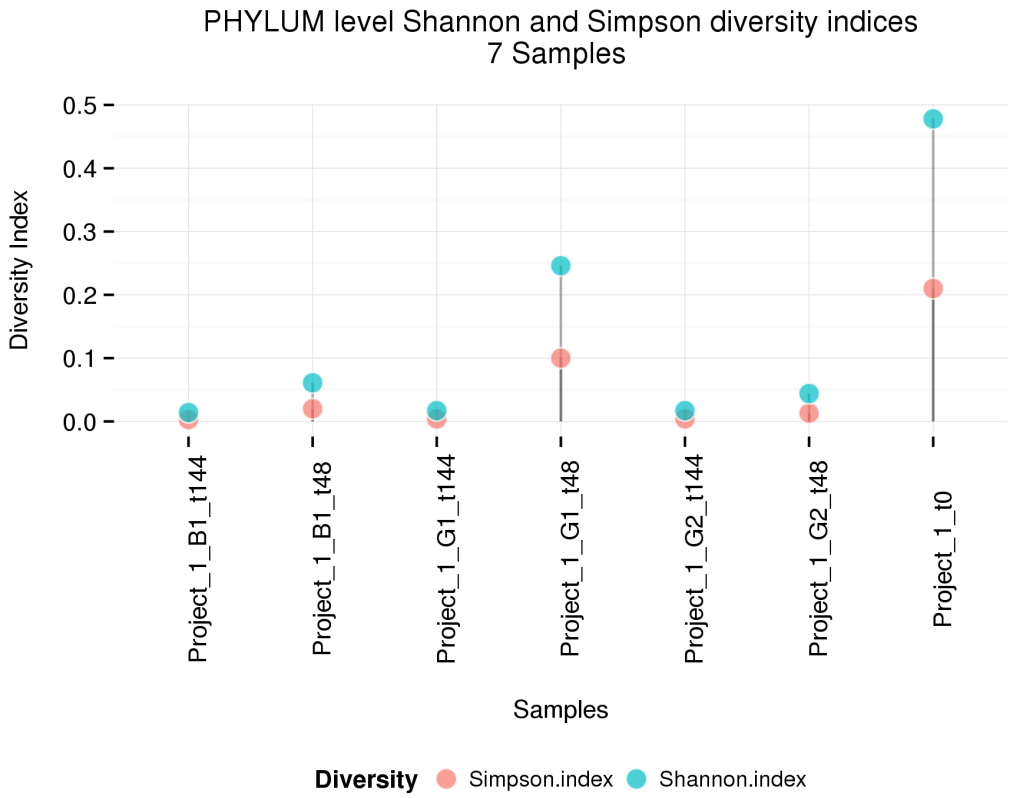

Figure 4: Phylum diversity indices (file: PHYLUM.diversity.png)

Table 5: Phylum diversity indices table (file: PHYLUM.diversity_index_tables.tsv)

Sample	Simpson.index	Shannon.index	OTUs
Project_1_B1_t144	0.003	0.014	3
Project_1_B1_t48	0.02	0.061	3
Project_1_G1_t144	0.004	0.017	3
Project_1_G1_t48	0.1	0.246	5
Project_1_G2_t144	0.004	0.017	3
Project_1_G2_t48	0.013	0.044	4
Project_1_t0	0.21	0.478	7

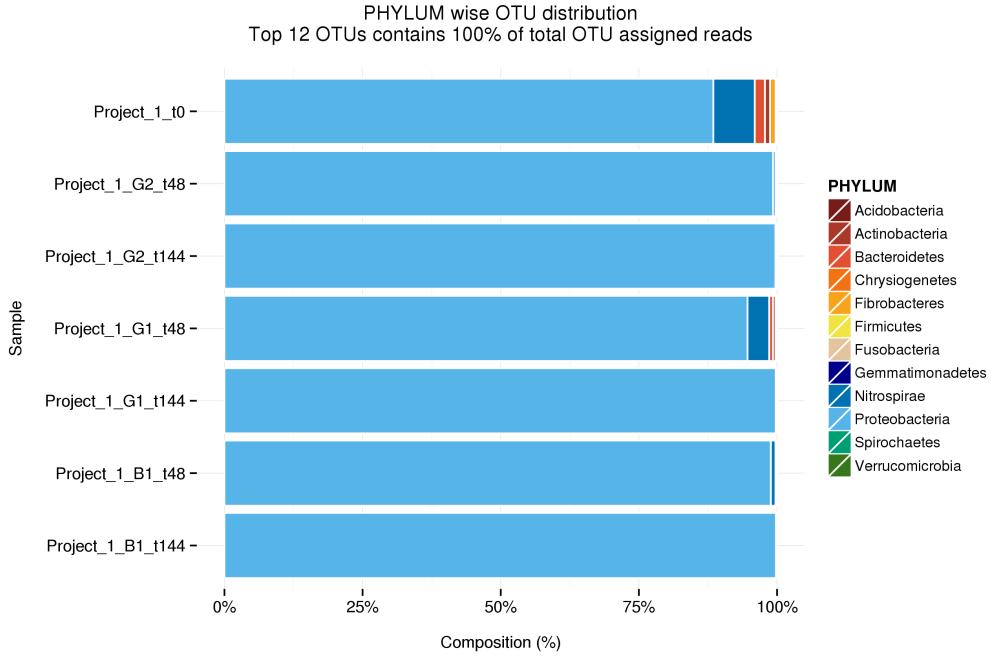

Figure 5: Phylum distribution plot (file: PHYLUM.OTU.distribution.combined.png)

Table 6: Phylum distribution table (file: PHYLUM.OTU.combined.table.percent.top.12.tsv)

PHYLUM	Project_1_B1_t144	Project_1_B1_t48	Project_1_G1_t144	Project_1_G1_t48	Project_1_G2_t144	Project_1_G2_t48	Project_1_t0
Proteobacteria	99.81	98.89	99.76	94.66	99.71	99.28	88.48
Nitrospirae	0.11	0.84	0.08	3.92	0.11	0.42	7.49
Bacteroidetes	0.06	0.17	0.14	0.67	0.1	0.1	1.85
Actinobacteria	0.01	0.04	0.01	0.47	0.01	0.03	0.93
Fibrobacteres	0	0.05	0.01	0.05	0.01	0.02	1
Firmicutes	0	0	0	0.15	0.05	0.11	0.02
Verrucomicrobia	0	0	0	0.01	0	0	0.12
Spirochaetes	0	0	0	0.02	0	0	0.05
Gemmatimonadetes	0	0	0	0.01	0	0	0.06
Chrysiogenetes	0	0	0	0	0	0.03	0
Acidobacteria	0	0	0	0.02	0	0	0
Fusobacteria	0	0	0	0	0	0	0

4.5 Class

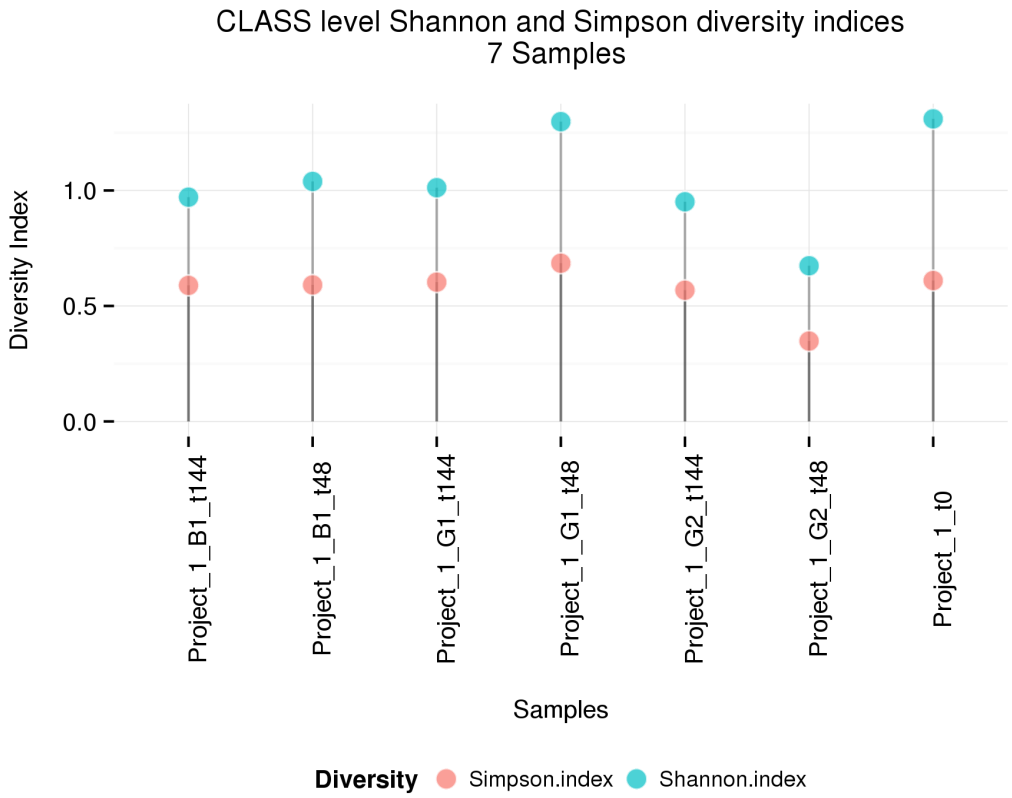

Figure 6: Class diversity indices (file: CLASS.diversity.png)

Table 7: Class diversity indices table (file: CLASS.diversity_index_tables.tsv)

Sample	Simpson.index	Shannon.index	OTUs
Project_1_B1_t144	0.589	0.971	4
Project_1_B1_t48	0.591	1.039	6
Project_1_G1_t144	0.603	1.012	6
Project_1_G1_t48	0.685	1.298	9
Project_1_G2_t144	0.568	0.951	4
Project_1_G2_t48	0.348	0.674	7
Project_1_t0	0.61	1.31	11

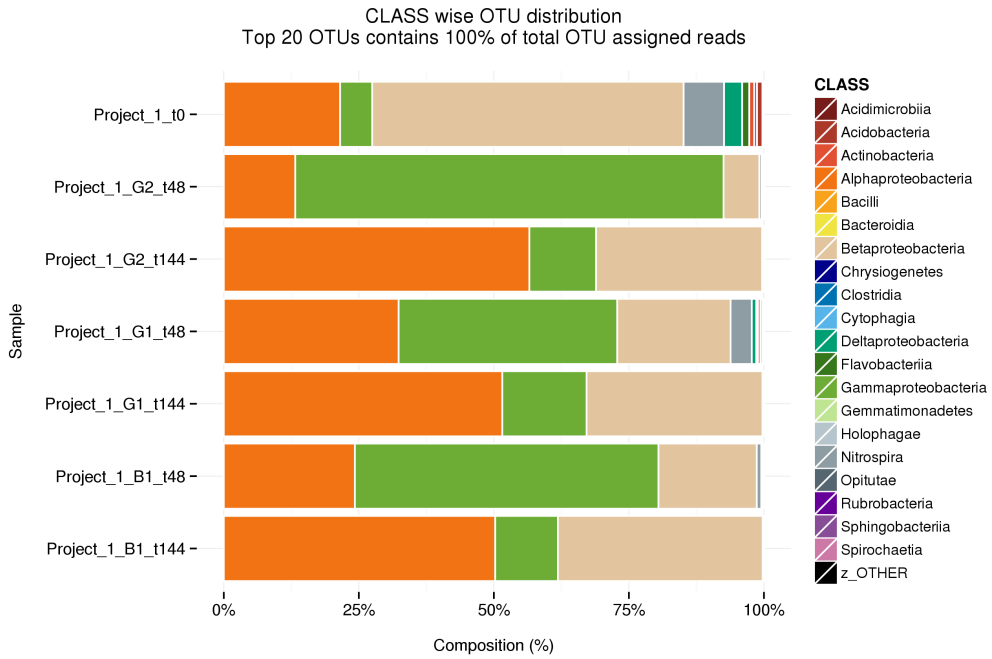

Figure 7: Class distribution plot (file: CLASS.OTU.distribution.combined.png)

Table 8: Class distribution table (file: CLASS.OTU.combined.table.percent.top.20.tsv)

CLASS	Project_1_B1_t144	Project_1_B1_t48	Project_1_G1_t144	Project_1_G1_t48	Project_1_G2_t144	Project_1_G2_t48	Project_1_t0
Alphaproteobacteria	50.25	24.27	51.56	32.37	56.58	13.23	21.53
Gammaproteobacteria	11.6	56.21	15.6	40.47	12.35	79.3	5.94
Betaproteobacteria	37.93	18.2	32.57	20.99	30.74	6.63	57.66
Nitrospira	0.11	0.84	0.08	3.92	0.11	0.42	7.49
Deltaproteobacteria	0.04	0.21	0.03	0.83	0.04	0.12	3.35
Flavobacteriia	0.04	0.13	0.06	0.31	0.05	0.07	1.32
Actinobacteria	0	0.04	0.01	0.45	0.01	0.03	0.91
Acidobacteria	0	0.05	0.01	0.05	0.01	0.02	1
Sphingobacteriia	0.02	0.04	0.07	0.36	0.04	0.02	0.52
Clostridia	0	0	0	0.14	0.03	0.1	0.01
Opitutae	0	0	0	0.01	0	0	0.12
Bacilli	0	0	0	0.02	0.02	0.01	0.02
Spirochaetia	0	0	0	0.02	0	0	0.05
Gemmatimonadetes	0	0	0	0.01	0	0	0.06
Chrysiogenetes	0	0	0	0	0	0.03	0
Holophagae	0	0	0	0.02	0	0	0
Bacteroidia	0	0	0	0	0.01	0	0
Cytophagia	0	0	0	0	0	0	0.01
Acidimicrobiia	0	0	0	0.01	0	0	0
Rubrobacteria	0	0	0	0.01	0	0	0

4.6 Order

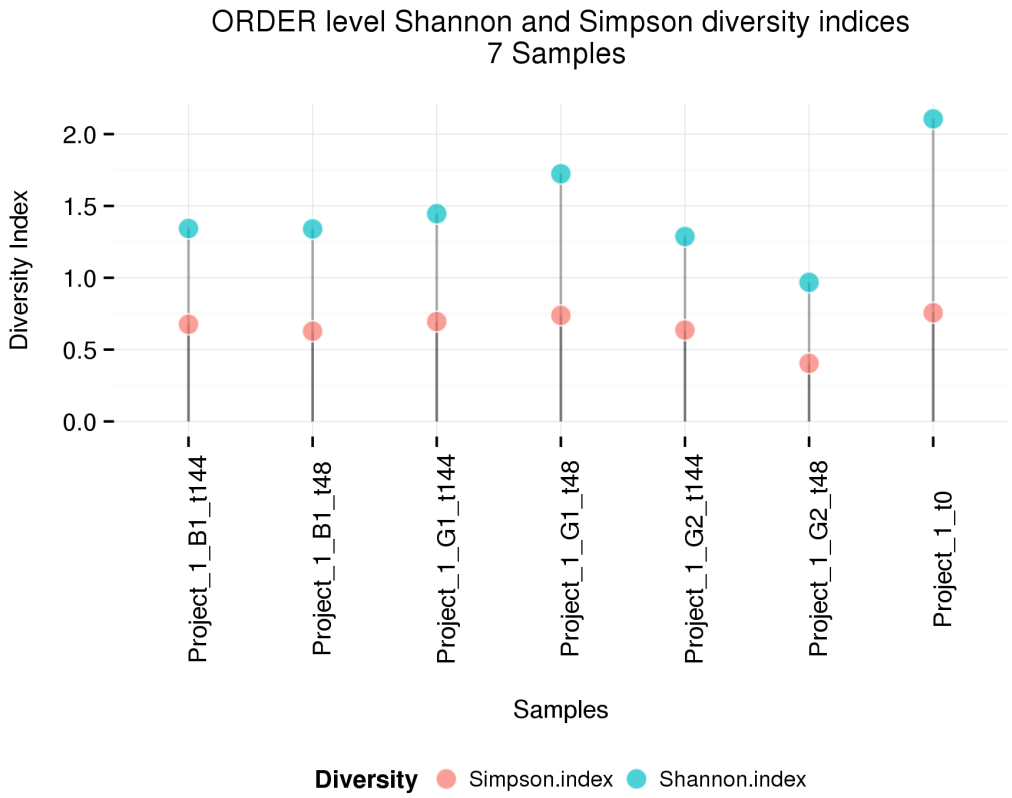

Figure 8: Order diversity indices (file: ORDER.diversity.png)

Table 9: Order diversity indices table (file: ORDER.diversity_index_tables.tsv)

Sample	Simpson.index	Shannon.index	OTUs
Project_1_B1_t144	0.676	1.343	10
Project_1_B1_t48	0.628	1.34	17
Project_1_G1_t144	0.695	1.446	15
Project_1_G1_t48	0.738	1.724	24
Project_1_G2_t144	0.636	1.287	12
Project_1_G2_t48	0.404	0.968	20
Project_1_t0	0.756	2.106	30

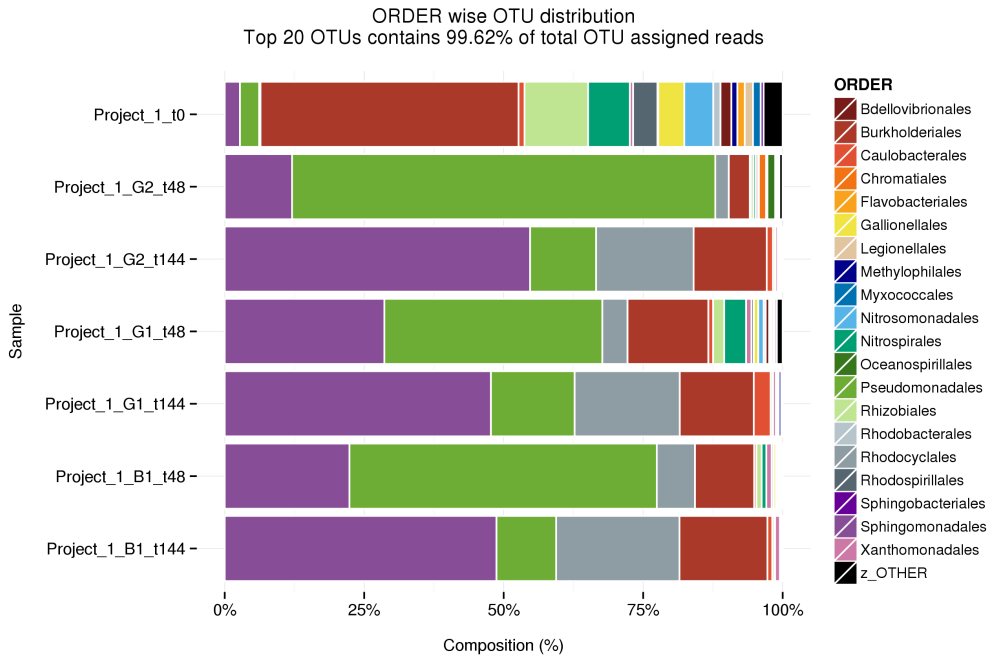

Figure 9: Order distribution plot (file: ORDER.OTU.distribution.combined.png)

Table 10: Order distribution table (file: ORDER.OTU.combined.table.percent.top.20.tsv)

ORDER	Project_1_B1_t144	Project_1_B1_t48	Project_1_G1_t144	Project_1_G1_t48	Project_1_G2_t144	Project_1_G2_t48	Project_1_t0
Sphingomonadales	48.72	22.36	47.72	28.62	54.73	12.07	2.74
Pseudomonadales	10.7	55.09	15	39.07	11.84	75.85	3.43
Burkholderiales	15.76	10.62	13.27	14.49	13.1	3.76	46.24
Rhodocyclales	22.1	6.84	18.84	4.51	17.48	2.42	0.26
Rhizobiales	0.38	0.94	0.36	1.99	0.35	0.48	11.4
Nitrospirales	0.11	0.84	0.08	3.92	0.11	0.42	7.49
Caulobacterales	0.88	0.4	3.02	0.85	1.11	0.19	1.07
Nitrosomonadales	0.02	0.26	0.01	0.97	0.01	0.13	5.19
Gallionellales	0.02	0.41	0.02	0.77	0.06	0.24	4.66
Rhodospirillales	0.16	0.34	0.18	0.43	0.2	0.16	4.37
Xanthomonadales	0.87	0.97	0.5	0.96	0.42	0.39	0.58
Bdellovibrionales	0.03	0.15	0.02	0.63	0.03	0.08	1.94
Rhodobacterales	0.08	0.14	0.22	0.36	0.15	0.29	1.3
Legionellales	0.02	0.11	0.02	0.35	0.02	0.05	1.47
Flavobacteriales	0.04	0.13	0.06	0.31	0.05	0.07	1.32
Methylophilales	0.03	0.03	0.43	0.14	0.08	0.06	1.09
Myxococcales	0.01	0.06	0.01	0.21	0.01	0.04	1.39
Chromatiales	0.01	0.03	0.07	0.02	0.05	1.29	0.14
Oceanospirillales	0	0	0	0	0.01	1.39	0.01
unclassified Acidobacteria	0	0.05	0.01	0.05	0.01	0.02	1

4.7 Family

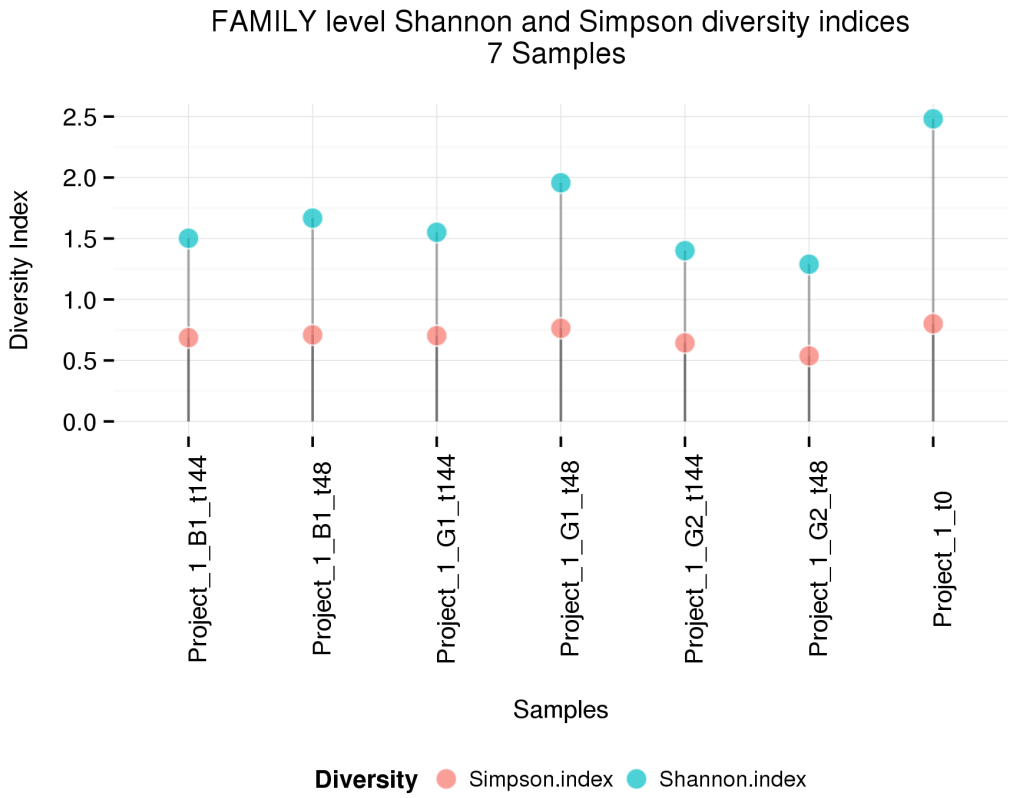

Figure 10: Family diversity indices (file: FAMILY.diversity.png)

Table 11: Family diversity indices table (file: FAMILY.diversity_index_tables.tsv)

Sample	Simpson.index	Shannon.index	OTUs
Project_1_B1_t144	0.687	1.502	16
Project_1_B1_t48	0.71	1.667	25
Project_1_G1_t144	0.703	1.551	18
Project_1_G1_t48	0.764	1.957	36
Project_1_G2_t144	0.644	1.4	17
Project_1_G2_t48	0.537	1.289	25
Project_1_t0	0.801	2.481	47

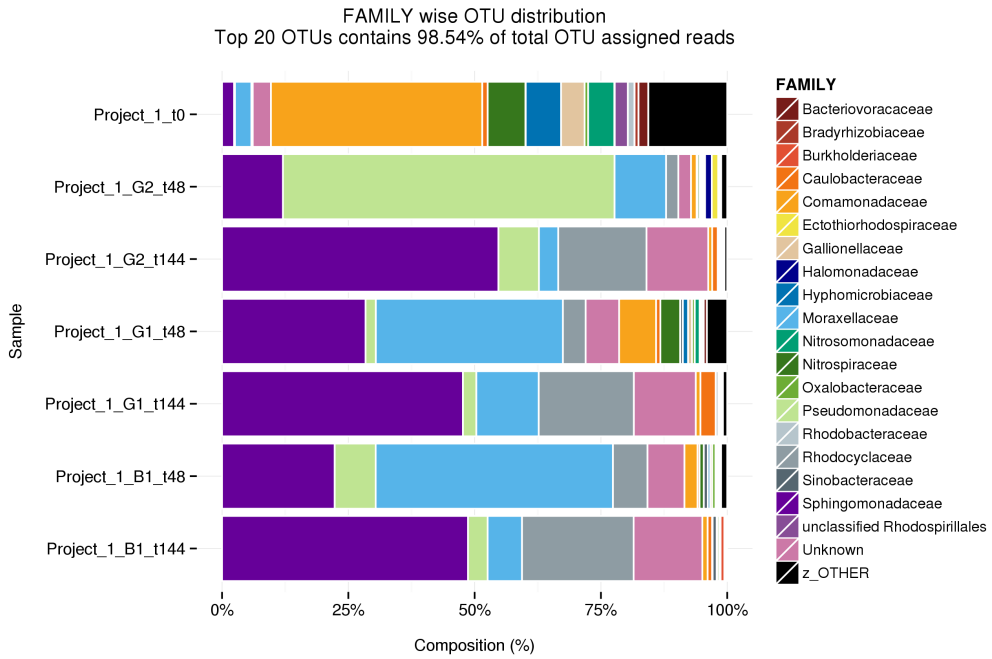

Figure 11: Family distribution plot (file: FAMILY.OTU.distribution.combined.png)

Table 12: Family distribution table (file: FAMILY.OTU.combined.table.percent.top.20.tsv)

FAMILY	Project_1_B1_t144	Project_1_B1_t48	Project_1_G1_t144	Project_1_G1_t48	Project_1_G2_t144	Project_1_G2_t48	Project_1_t0
Sphingomonadaceae	48.69	22.3	47.69	28.4	54.71	12.05	2.39
Moraxellaceae	6.81	46.97	12.35	37.03	3.82	10.2	3.26
Pseudomonadaceae	3.89	8.12	2.65	2.04	8.02	65.65	0.17
Rhodocyclaceae	22.1	6.84	18.84	4.51	17.48	2.42	0.26
Unknown	13.6	7.32	12.26	6.62	12.22	2.51	3.61
Comamonadaceae	1.05	2.59	0.92	7.32	0.78	1.12	41.84
Nitrospiraceae	0.11	0.84	0.08	3.92	0.11	0.42	7.49
Hyphomicrobiaceae	0.1	0.46	0.1	1.04	0.12	0.23	6.99
Caulobacteraceae	0.88	0.4	3.02	0.85	1.11	0.19	1.07
Nitrosomonadaceae	0.02	0.26	0.01	0.97	0.01	0.13	5.19
Gallionellaceae	0.02	0.41	0.02	0.77	0.06	0.24	4.66
unclassified Rhodospirillales	0.16	0.31	0.18	0.3	0.19	0.15	2.59
Sinobacteraceae	0.8	0.8	0.43	0.53	0.34	0.34	0.05
Bacteriovoracaceae	0.03	0.15	0.02	0.63	0.03	0.08	1.94
Oxalobacteraceae	0.41	0.66	0.14	0.54	0.13	0.12	0.7
unclassified Rhizobiales	0.02	0.12	0.02	0.37	0.04	0.09	2.03
Rhodobacteraceae	0.08	0.14	0.22	0.36	0.15	0.29	1.3
Legionellaceae	0.02	0.11	0.02	0.35	0.02	0.05	1.46
Flavobacteriaceae	0.04	0.13	0.06	0.31	0.05	0.07	1.3
Methylophilaceae	0.03	0.03	0.43	0.14	0.08	0.06	1.09

4.8 Genus

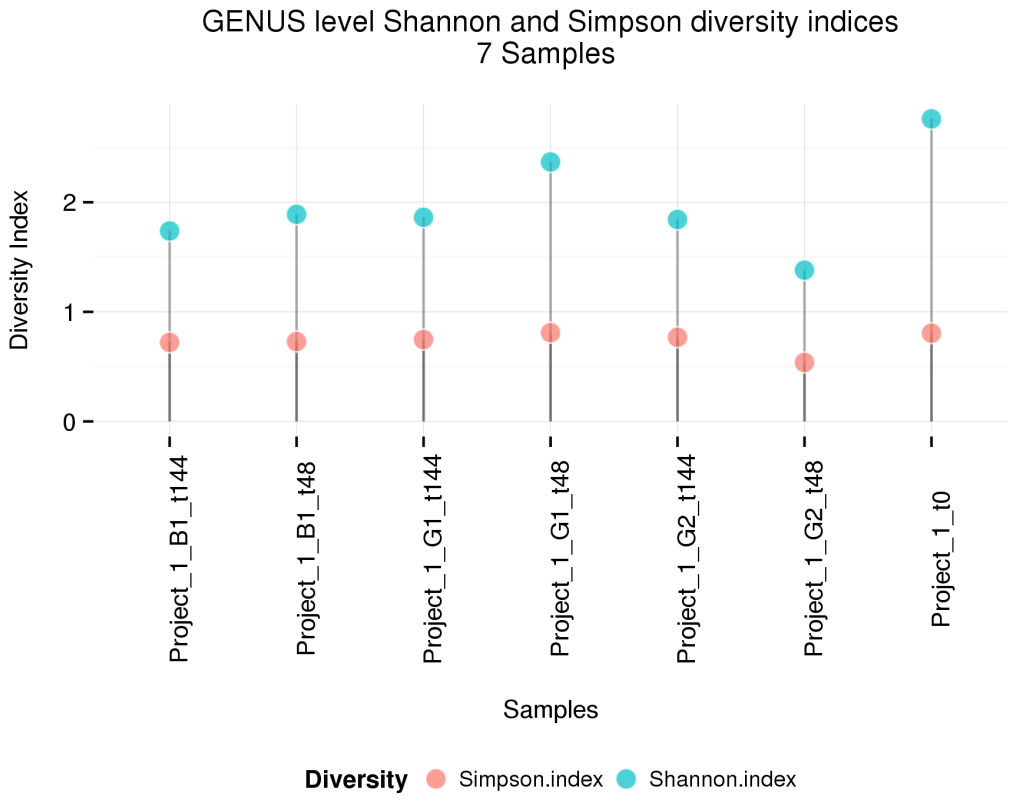

Figure 12: Genus diversity indices (file: GENUS.diversity.png)

Table 13: Genus diversity indices table (file: GENUS.diversity_index_tables.tsv)

Sample	Simpson.index	Shannon.index	OTUs
Project_1_B1_t144	0.721	1.738	27
Project_1_B1_t48	0.729	1.89	40
Project_1_G1_t144	0.747	1.863	33
Project_1_G1_t48	0.81	2.368	61
Project_1_G2_t144	0.768	1.843	31
Project_1_G2_t48	0.538	1.38	33
Project_1_t0	0.805	2.761	88

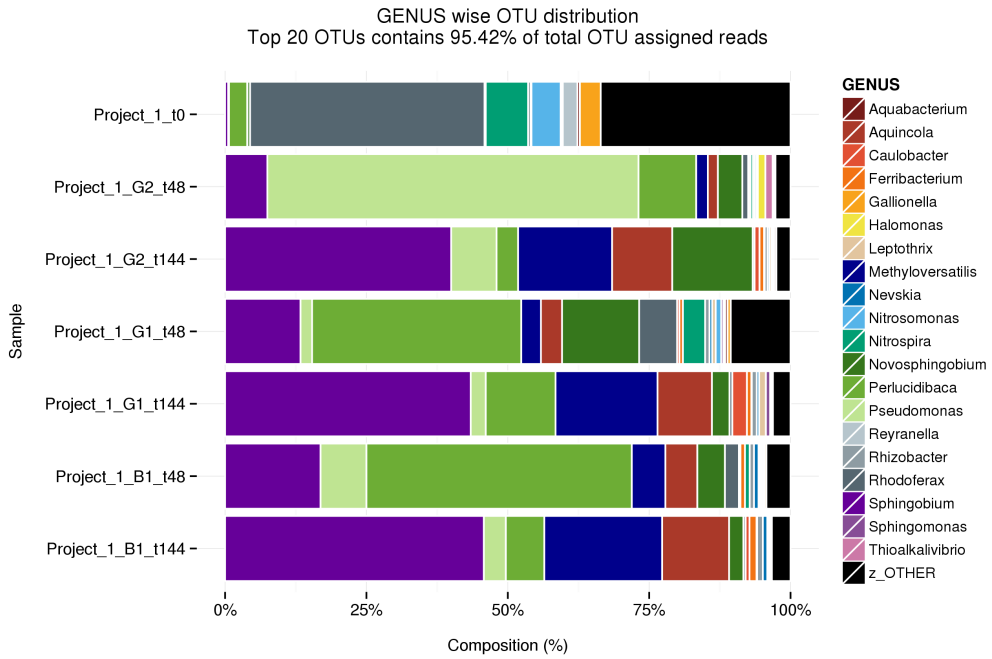

Figure 13: Genus distribution plot (file: GENUS.OTU.distribution.combined.png)

Table 14: Genus distribution table (file: GENUS.OTU.combined.table.percent.top.20.tsv)

GENUS	Project_1_B1_t144	Project_1_B1_t48	Project_1_G1_t144	Project_1_G1_t48	Project_1_G2_t144	Project_1_G2_t48	Project_1_t0
Sphingobium	45.77	16.89	43.45	13.35	39.99	7.48	0.54
Perlucidibaca	6.8	46.93	12.34	36.99	3.81	10.19	3.22
Pseudomonas	3.88	8.12	2.65	2.04	8.02	65.65	0.16
Methyloversatilis	20.84	5.93	18.04	3.48	16.62	2.06	0.03
Rhodoferax	0.39	2.48	0.5	6.72	0.33	1.04	41.45
Aquincola	11.84	5.66	9.63	3.74	10.64	1.78	0.02
Novosphingobium	2.54	4.88	3.1	13.63	14.23	4.32	0.46
Nitrospira	0.11	0.84	0.08	3.92	0.11	0.42	7.49
Nitrosomonas	0.02	0.26	0.01	0.97	0.01	0.13	5.18
Caulobacter	0.68	0.27	2.58	0.41	0.88	0.12	0.23
Gallionella	0.02	0.33	0.02	0.55	0.02	0.19	3.69
Rhizobacter	1.01	0.76	0.87	0.78	0.51	0.32	0.47
Ferribacterium	1.24	0.8	0.78	0.61	0.8	0.32	0
Reyranella	0.16	0.31	0.18	0.3	0.19	0.15	2.59
Pedomicrobium	0.03	0.19	0.03	0.35	0.04	0.09	3.07
Hyphomicrobium	0.05	0.21	0.05	0.47	0.05	0.1	2.72
Nevskia	0.79	0.8	0.43	0.47	0.34	0.33	0.01
Leptothrix	0.23	0.16	1.18	0.63	0.39	0.09	0.1
Bacteriovorax	0.03	0.15	0.02	0.62	0.03	0.08	1.76
Methylibium	0.25	0.38	0.16	0.36	0.19	0.13	0.72

4.9 Species

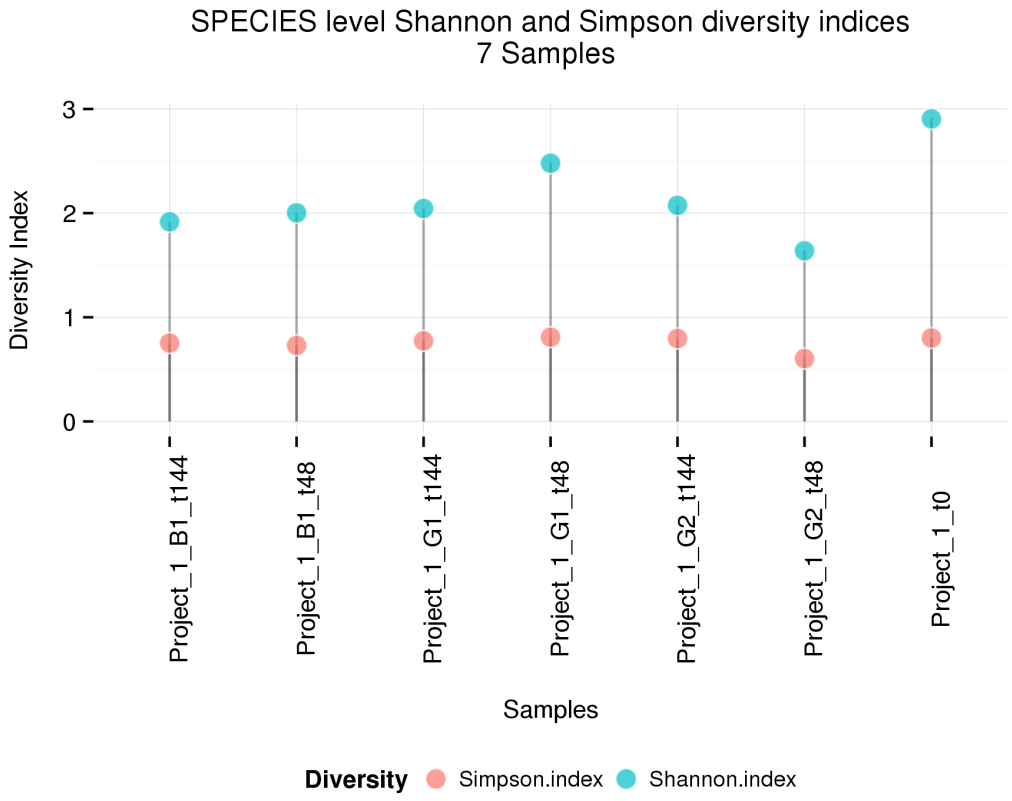

Figure 14: Species diversity indices (file: SPECIES.diversity.png)

Table 15: Species diversity indices table (file: SPECIES.diversity_index_tables.tsv)

Sample	Simpson.index	Shannon.index	OTUs
Project_1_B1_t144	0.753	1.917	36
Project_1_B1_t48	0.73	2.003	51
Project_1_G1_t144	0.773	2.044	46
Project_1_G1_t48	0.809	2.479	91
Project_1_G2_t144	0.796	2.075	45
Project_1_G2_t48	0.603	1.638	40
Project_1_t0	0.801	2.905	121

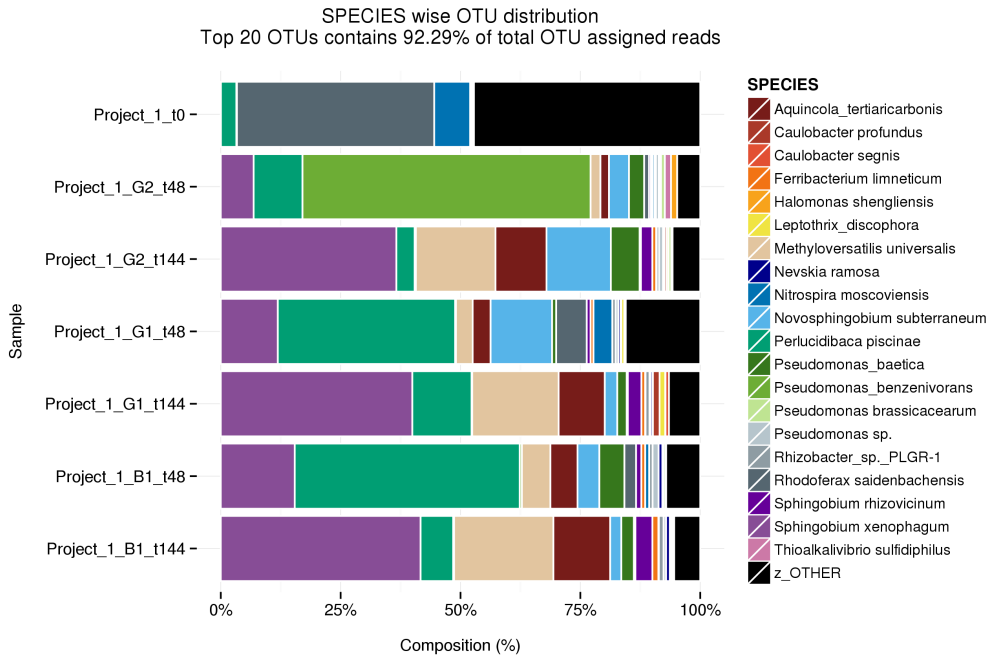

Figure 15: Species distribution plot (file: SPECIES.OTU.distribution.combined.png)

Table 16: Species distribution table (file: SPECIES.OTU.combined.table.percent.top.20.tsv)

SPECIES	Project_1_B1_t144	Project_1_B1_t48	Project_1_G1_t144	Project_1_G1_t48	Project_1_G2_t144	Project_1_G2_t48	Project_1_t0
Sphingobium xenophagum	41.67	15.43	39.96	11.89	36.63	6.86	0.03
Perlucidibaca piscinae	6.8	46.93	12.34	36.99	3.81	10.19	3.22
Methyloversatilis universalis	20.71	5.93	18.02	3.48	16.59	2.06	0.03
Pseudomonas_benzenivorans	0.19	0.46	0.15	0.19	0.27	60.08	0
Rhodoferax saidenbachensis	0.32	2.43	0.24	6.38	0.28	1.01	41.14
Aquincola_tertiaricarbonis	11.84	5.66	9.63	3.74	10.64	1.78	0.02
Novosphingobium subterraneum	2.32	4.56	2.61	12.82	13.43	4.17	0.08
Pseudomonas_baetica	2.62	5.24	1.97	0.86	6	3.17	0
Nitrospira moscoviensis	0.11	0.84	0.08	3.92	0.11	0.42	7.49
Sphingobium rhizovicinum	3.54	1.1	2.82	0.79	2.38	0.4	0
Gallionella capsiferriformans	0.02	0.33	0.02	0.55	0.02	0.19	3.69
Ferribacterium limneticum	1.24	0.8	0.78	0.61	0.8	0.32	0
Rhizobacter_sp._PLGR-1	0.98	0.7	0.84	0.7	0.5	0.3	0.38
Pseudomonas sp.	0.53	1.27	0.27	0.54	0.81	0.59	0.02
Nitrosomonas oligotropha	0.01	0.15	0.01	0.69	0.01	0.08	2.83
Pedomicrobium americanum	0.03	0.15	0.03	0.3	0.04	0.07	2.75
Reyranella soli	0.14	0.26	0.16	0.25	0.16	0.12	2.11
Nevskia ramosa	0.79	0.79	0.43	0.47	0.34	0.33	0.01
Leptothrix_discophora	0.23	0.16	1.18	0.63	0.39	0.09	0.1
Bacteriovorax_stolpii	0.03	0.15	0.02	0.62	0.03	0.08	1.76

5 Deliverables

Table 17: List of deliverable files, format and recommended programs to access.

File	Format	Program To Open File
Taxa-level.diversity_index_tables.tsv	TSV	Spreadsheet editor
Taxa-level.diversity.png	PNG	Image viewer
Taxa-level.combined.table.percent.top.X.tsv	TSV	Spreadsheet editor
Taxa-level.combined.table.percent.tsv	TSV	Spreadsheet editor
Taxa-level.combined.table.tsv	TSV	Spreadsheet editor
Taxa-level.OTU.distribution.combined.png	PNG	Image viewer

6 Formats

Table 18: List of deliverable files, format and recommended programs to access.

Format	Description
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 Tables

Table 19: 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 20: 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

8 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] Shannon diversity index. http://en.wikipedia.org/wiki/Diversity_index#Shannon_index.

[9] Simpson diversity index. http://en.wikipedia.org/wiki/Diversity_index#Simpson_index.

[10] Diversity index. http://en.wikipedia.org/wiki/Diversity_index.

[11] Ecological Diversity Indices and Rarefaction Species Richness (R package Vegan). http://cc.oulu.fi/~jarioksa/softhelp/vegan/html/diversity.html.

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

www.gatc-biotech.com