eurofins | Genomics

Data Analysis Report: Metagenome Analysis v1.1

Project / Study: GATC-Demo

Project description: INVIEW METAGENOME EXPLORE

Date: February 27, 2018

Table of Contents

1 Analysis workflow

The schematic diagram of the data analysis steps that have been performed is shown in figure 1.

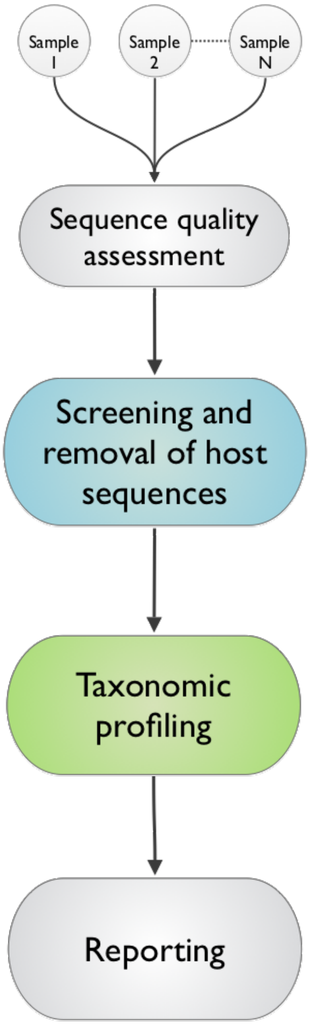

Sample
1
Sample
2
Sample
N
Sequence quality assessment
Screening and removal of host sequences
Taxonomic profiling
Reporting

Figure 1: Metagenome Analysis v1.1 Workflow

2 Samples Analysed

sample_1, sample_2, sample_3.

3 Reference Database

Table 1: Homo sapiens reference database.

Tag	Description
Name	Homo sapiens
Version	hg19
Source	UCSC
Size	3.137 GB
Sequences	23

Table 2: Taxonomic Profiling database composition.

Kingdom	Organisms	Sequences	Source
Archaea	199	213	NCBI Genomes (complete)
Bacteria	3,699	5,128	NCBI Genomes (complete)
Fungi	198	66,196	NCBI Genomes (complete + contigs)
Protozoa	77	226,377	NCBI Genomes (complete + contigs)
Virus	4,925	6,678	NCBI Genomes (complete)

4 Results

4.1 Sequence Quality Metrics

The base quality of each sequence read is inspected. Low quality calls are removed before proceeding with further processing. Using a sliding window approach, bases with low quality are removed from the 3' and 5' ends. Bases are removed if the average phred quality is below 15. Finally only mate pairs (forward and reverse read) were used for the next analysis step. The total amount of raw sequence data and the results of the quality filtering is collected and reported in the following table.

Table 3: Sequence quality metrics per sample

Sample	Total Reads	LQ Reads	Single Reads	HQ Reads
sample_1	10,000,000	40,060 (0.4%)	38,310 (0.4%)	9,921,630 (99.2%)
sample_2	10,000,000	22,081 (0.2%)	20,779 (0.2%)	9,957,140 (99.6%)
sample_3	10,000,000	188,146 (1.9%)	161,678 (1.6%)	9,650,176 (96.5%)

Total Reads: Total number of sequence reads analysed for each sample.

LQ Reads: Number (percentage) of low quality reads.

Single Reads: High quality reads without mates (2nd read). These are not included for further analysis.

HQ Reads: Number (percentage) of high quality reads used for further analysis.

4.2 Screening for host genome background

The sequence reads are mapped against a reference database of the host organism using Bowtie[1] with default parameters. The following table contains the number of reads mapped to the references for each sample. Accuracy of the reference and better quality of reads lead to a higher percentage of reads mapped to the reference. The details of the reference database used are mentioned in chapter 3, table 1.

Table 4: Mapped read metrics observed per sample.

Sample Name	HQ Reads	Mapped to hg19
sample_1	9,921,630	9,398 (0.1 %)
sample_2	9,957,140	175,124 (1.8 %)
sample_3	9,650,176	49,508 (0.5 %)

4.3 Taxonomic profiling

After screening and removing host sequence reads, non-host reads are subjected to taxonomic profiling algorithm. Taxonomic profiling is done using Kraken[2] and the Minikraken reference database. Kraken classifies reads by breaking each into overlapping k-mers. Each k-mer is mapped to the lowest common ancestor (LCA) of the genomes containing that k-mer in a precomputed reference database. For each read, a classification tree is found by pruning the taxonomy and only retaining taxa (including ancestors) associated with k-mers in that read. Each node is weighted by the number of k-mers mapped to the node, and the path from root to leaf with the highest sum of weights is used to classify the read. The final classified and unclassified reads are reported in table 5.

Table 5: Taxonomic Profiling metrics per sample.

Sample Name	Reads	Classified	Unclassified
sample_1	9,911,980	1,880,406 (18.97 %)	8,031,574 (81.03 %)
sample_2	9,778,356	1,509,674 (15.44 %)	8,268,682 (84.56 %)
sample_3	9,598,988	1,289,758 (13.44 %)	8,309,230 (86.56 %)

Table 6: Number of reads assigned to different kingdoms for sample_1, sample_2, sample_3.

Kingdom	sample_1		sample_2		sample_3	
Archaea	18,574	0.99 %	428	0.03 %	392	0.03 %
Bacteria	1,806,158	96.05 %	1,443,674	95.63 %	1,199,722	93.02 %
Eukaryota	1,100	0.06 %	3,224	0.21 %	1,606	0.12 %
Fungi	2,536	0.13 %	3,664	0.24 %	1,608	0.12 %
Viruses	1,306	0.07 %	1,068	0.07 %	246	0.02 %
Ambiguous	50,732	2.70 %	57,616	3.82 %	86,184	6.68 %

Ambigious: Reads which can not be assigned to one specific kingdom.

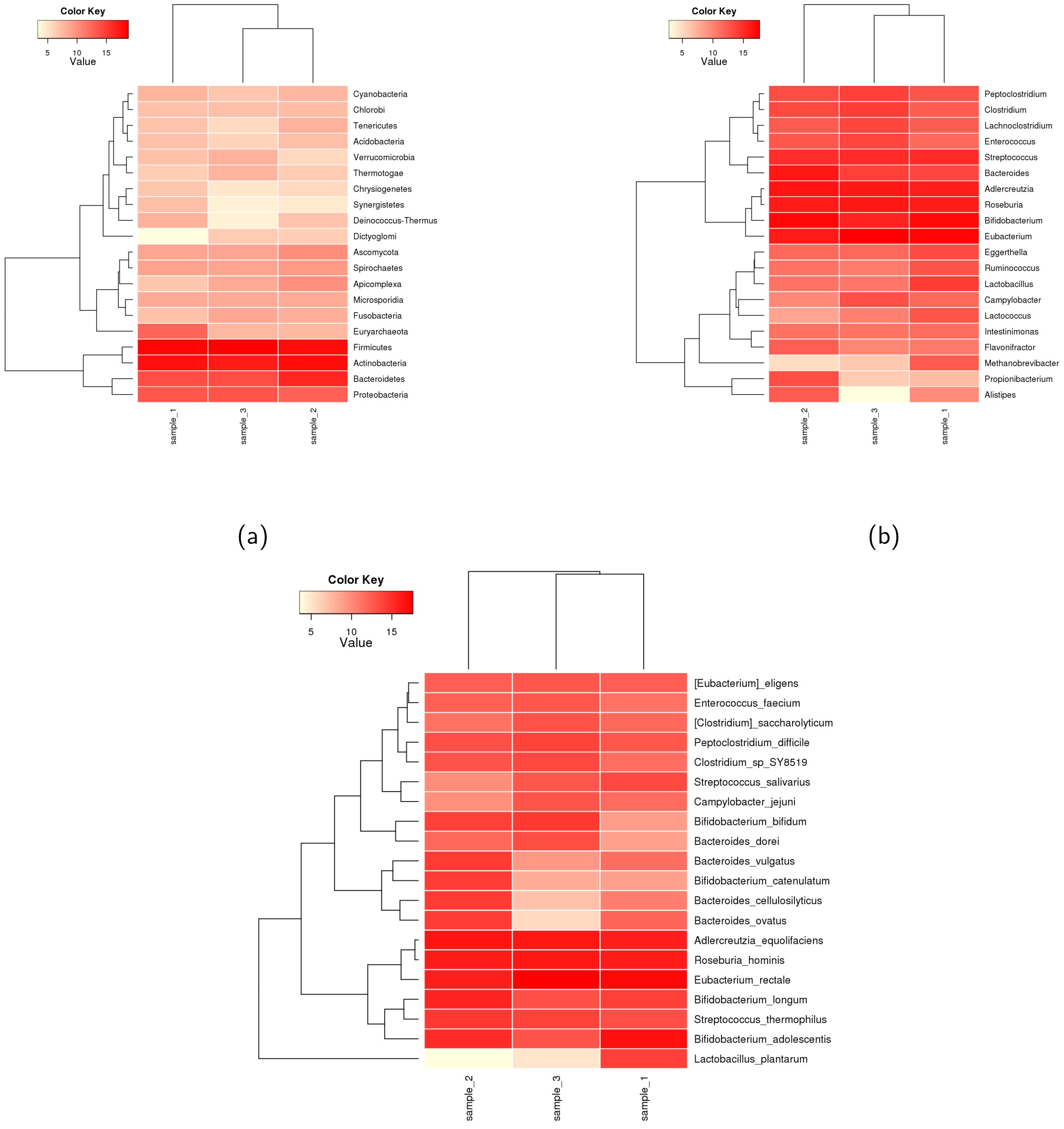
Eukaryota: Parasitic and non-parasitic Protozoa.

4.3.1 Taxa abundance

Abundance measured by the percentage of OTU assigned reads from various taxonomic levels is determined. The measured abundance levels are in OTU distribution tables (Taxa-level.composition.tsv). Heatmap and bar plots representing the taxonomic abundance at various levels are in OTU abundance heatmap (Taxa-level.rarefaction_heatmap.png) and OTU distribution plots (Taxa-level.barplot.png), respectively.
Read counts of input samples observed at various taxa levels (Phylum, Genus, and Species) are collected and normalized by using the rarefy function implemented in the Vegan bioconductor package[3] to compare species richness from all samples in the analysis run. Rarefied read counts enable better comparisons of OTU profiles between samples with different sample sizes. The final read counts in the tables (Taxa-level.composition.reads.tsv) contain normalized/rarefied read counts and NOT raw read counts.

Abundance measured by the percentage of OTU assigned reads from various taxonomic levels is determined and are used to generate heatmaps and bar plots at Phylum, Genus and Species levels.

The measured abundance levels are in OTU distribution tables (Taxa-level.composition.tsv). Heatmap and bar plots representing the taxonomic abundance at various levels are in OTU abundance heatmap (Taxa-level.rarefaction_heatmap.png) and OTU distribution plots (Taxa-level.barplot.png), respectively.

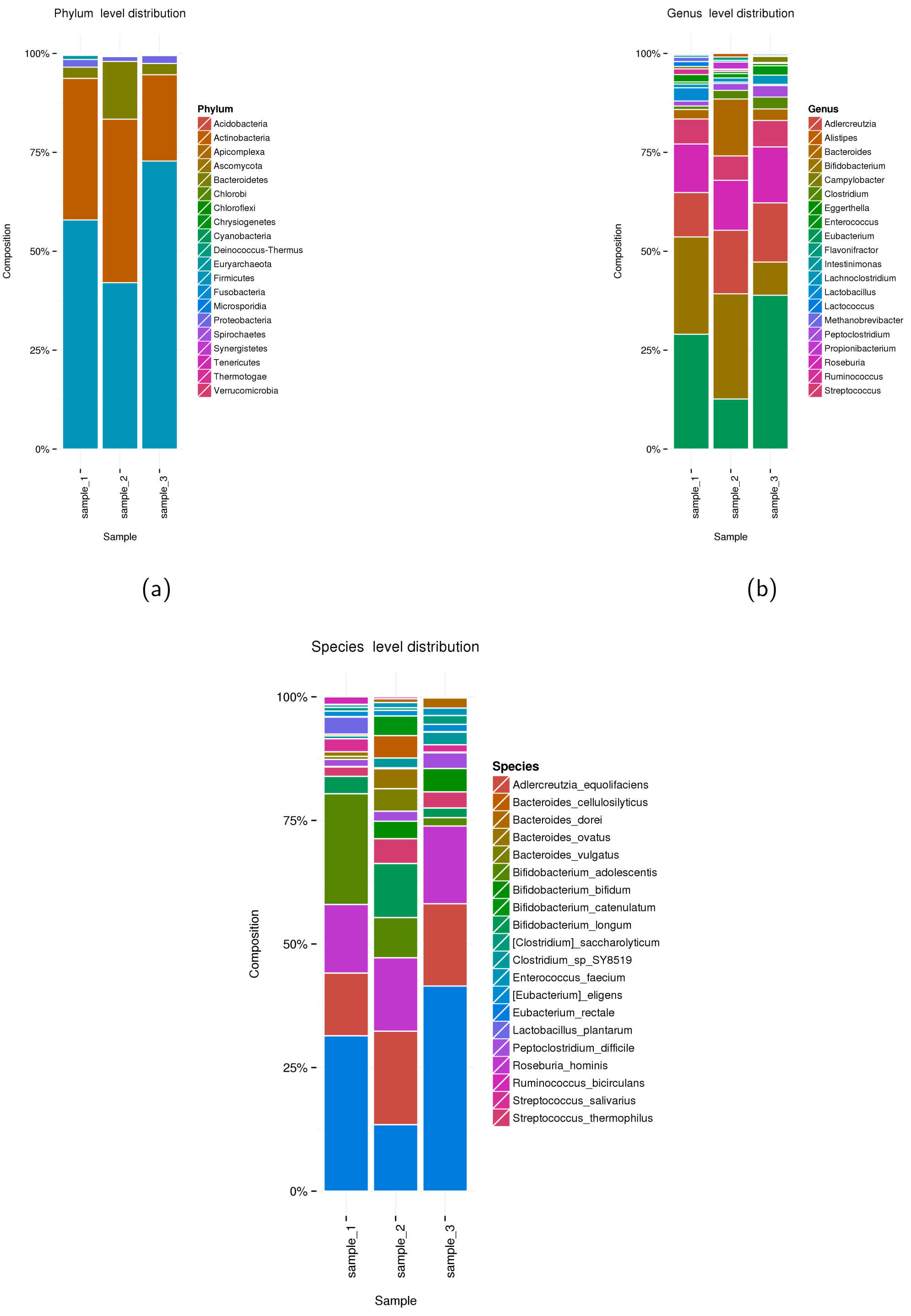


(a)

(b)

(c)

Figure 2: Heat map(s) showing the taxonomic abundance and their relation across the samples. Dendrograms determined by computing hierarchical clustering from the abundance levels shows the relationship between the species (left) and the samples (top). The abundance levels (number of reads associated with each taxa) are logarithmically transformed to base 2 for clarity. (a) Taxa-level: Phylum; (b) Taxa-level: Genus; (c) Taxa-level: Species



(a)

(b)

(c)

Figure 3: Bar plot(s) showing the taxonomic abundance across the samples. (a) Taxa-level: Phylum; (b) Taxa-level: Genus; (c) Taxa-level: Species

4.3.2 Species diversity

A diversity index is a quantitative measure that reflects how many different types (such as species) 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 all species are present at nearly the same level. For a given number of species, the value of a diversity index is maximized when all species are equally abundant.

The following diversity indices are computed using vegan[3] package in R.
Simpson refers to Simpson diversity index and has values ranging from 0 to 1. Values near 1 are simple environments and smaller values are diverse environments.
InvSimpson refers to inverse Simpson diversity and has values >0. A larger value means greater diversity.
Shannon refers to Shannon diversity index and has values >0. A higher value means greater diversity.
Alpha refers to Fischer's model of predicting species richness by computing alpha diversity and has values >0. A larger value means greater diversity.
Evenness refers to the distribution of individuals across species and is determined by Pielou's measure of species evenness. The index tends to 0 as the evenness decreases in simple environments (species-poor communities).
SpeciesNo refers to the absolute number of species found in each sample.

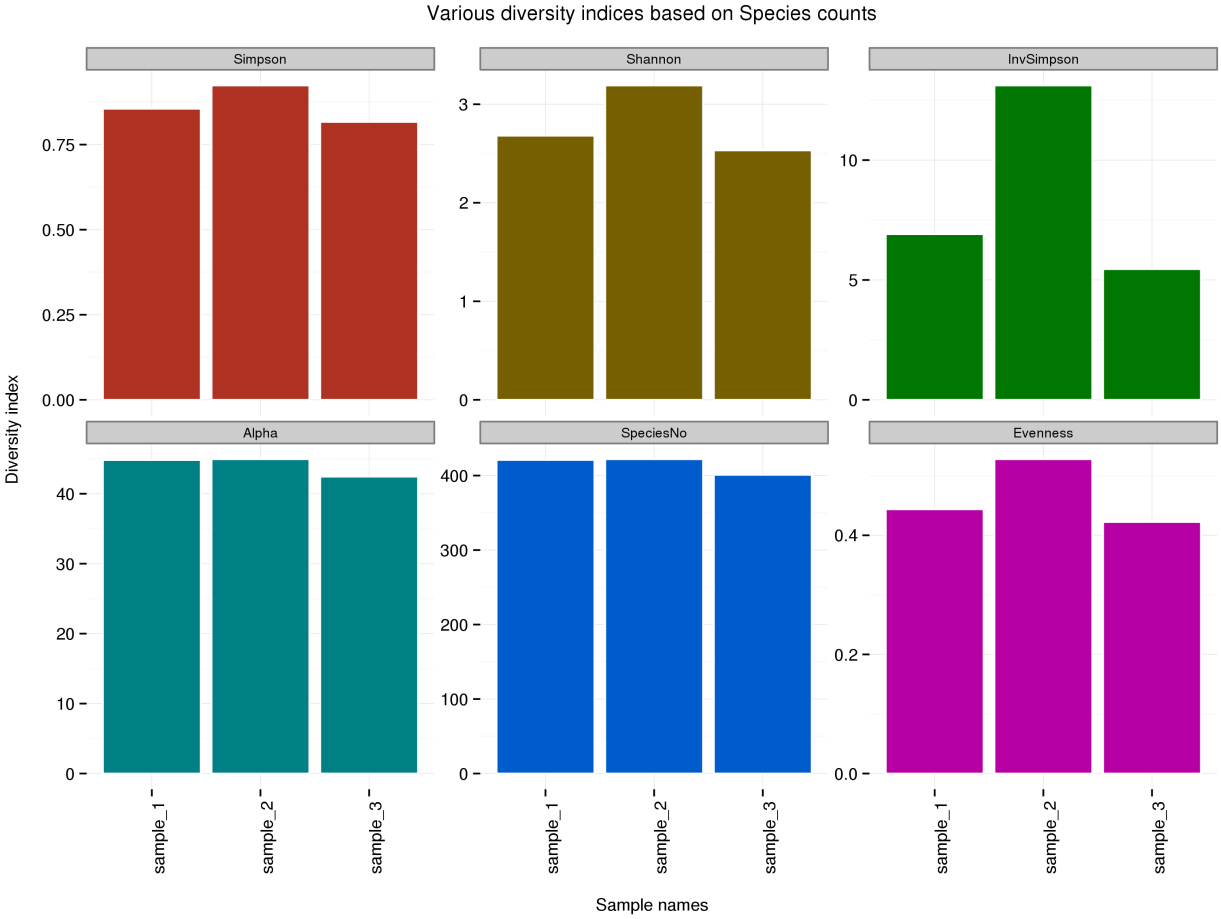


Figure 4: Various diversity indices computed based on the species counts found in each sample.

4.3.3 Rarefaction curves

Rarefaction allows the calculation of species richness for a given number of individual samples, based on the construction of rarefaction curves. This curve is a plot of the total number of distinct species found as a function of the number of sequences sampled. Sampling curves generally rise very quickly at first and then level off towards an asymptote as fewer new species are found in each sample. These rarefaction curves are calculated from the table of species abundance. The curves represent the average number of different species found for subsamples of the complete dataset.

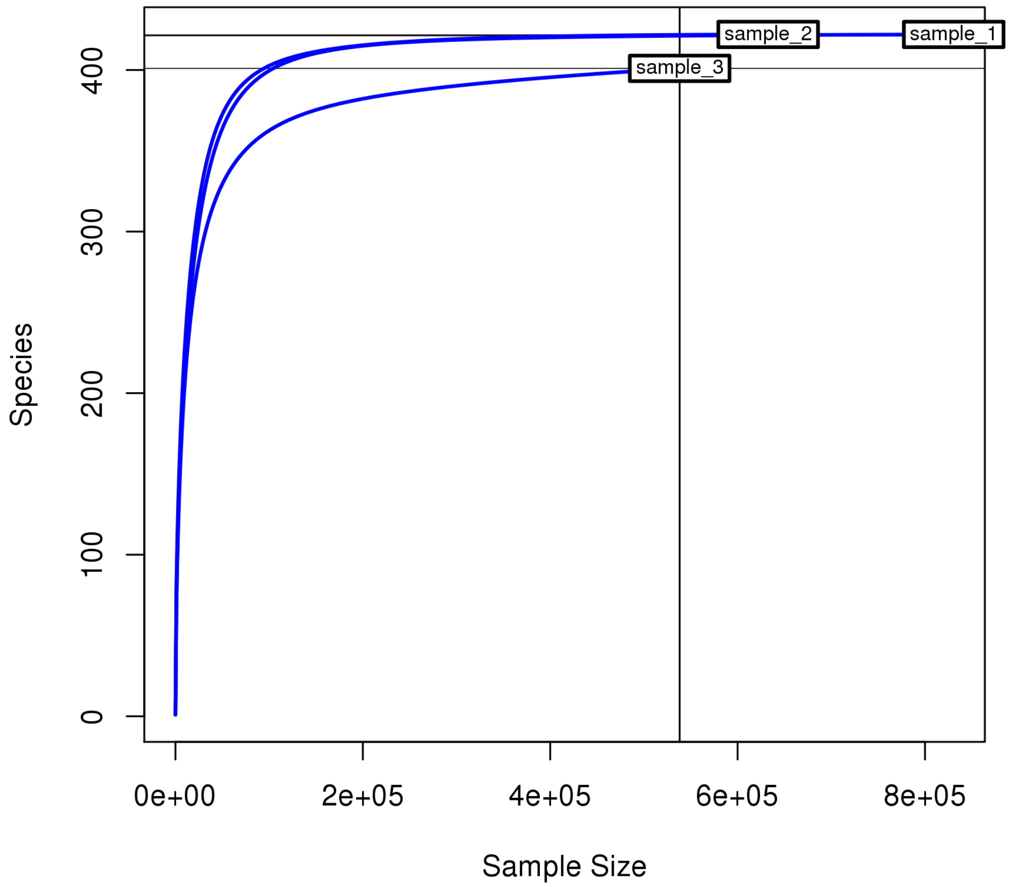


Figure 5: Rarefaction curve of annotated species richness.

4.3.4 Interactive plots

Taxonomic profiling results produced by Kraken[2] are used to generate interactive plots using Krona[4]. Krona is a visualization tool that allows intuitive exploration of relative abundances and confidences within the complex hierarchies of metagenomic classifications.

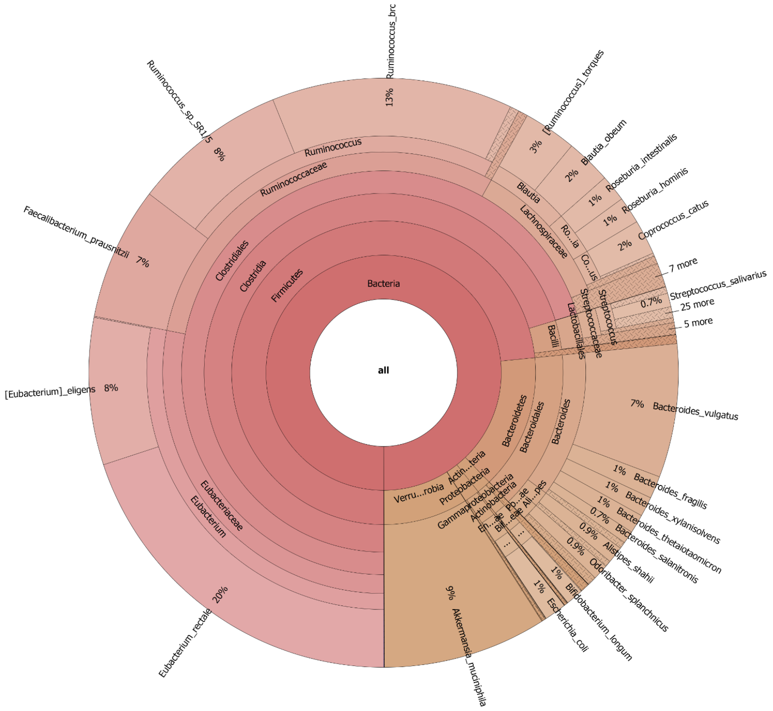


Figure 6: Example of an interactive plot generated by Krona (interactive_plots.html).

5 Deliverables

Table 7: List of delivered files, format and recommended programs to access the data.

File	Format	Program To Open File
All.interactive_plots.html	HTML	Web browser
Genus.barplot.png	PNG	Image viewer
Genus.composition.proportion.tsv	TSV	Spreadsheet Editor
Genus.composition.reads.tsv	TSV	Spreadsheet Editor
Genus.diversity_indices.png	PNG	Image viewer
Genus.diversity_indicies.tsv	TSV	Spreadsheet Editor
Genus.rarefaction_heatmap.log2scale.png	PNG	Image viewer
Genus.rarefaction_heatmap.png	PNG	Image viewer
Phylum.barplot.png	PNG	Image viewer
Phylum.composition.proportion.tsv	TSV	Spreadsheet Editor
Phylum.composition.reads.tsv	TSV	Spreadsheet Editor
Phylum.rarefaction_heatmap.log2scale.png	PNG	Image viewer
Phylum.rarefaction_heatmap.png	PNG	Image viewer
Species.barplot.png	PNG	Image viewer
Species.composition.proportion.tsv	TSV	Spreadsheet Editor
Species.composition.reads.tsv	TSV	Spreadsheet Editor
Species.diversity_indices.png	PNG	Image viewer

Table 7: List of delivered files, format and recommended programs to access the data.

File	Format	Program To Open File
Species.diversity_indicies.tsv	TSV	Spreadsheet Editor
Species.rarefaction_curve.png	PNG	Image viewer
Species.rarefaction_heatmap.log2scale.png	PNG	Image viewer
Species.rarefaction_heatmap.png	PNG	Image viewer
SAMPLE.alignment.bam	BAM	IGV, Tablet
SAMPLE.alignment.bam.bai	BAI	None
SAMPLE.unmapped.fastq	FASTQ	Text Editor

6 Formats

Table 8: References and descriptions of file format.

Format	Description
TSV	Tab separated table style text file. This can be imported into spreadsheet processing software like MS OFFICE Excel.
FASTQ[5]	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.
HTML	Standard markup language for creating web pages and web applications
BAM[6]	Compressed binary version of the Sequence Alignment/Mapping (SAM) format, a compact and index-able representation of nucleotide sequence alignments.
PNG	Figure or image in Portable Network Graphics format

7 FAQ

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 separator and comma as thousands separator. Make sure that you set both correctly.

Bibliography

[1] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25–10, March 2009.

[2] Derrick E. Wood and Steven L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46+, March 2014.

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

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

[5] 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.

[6] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[7] Picard. http://picard.sourceforge.net.

[8] 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.

[9] Marc Lohse, Anthony M. Bolger, Axel Nagel, Alisdair R. Fernie, John E. Lunn, Mark Stitt, and Björn Usadel. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, 40(W1):W622–W627, July 2012.

[10] Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, February 2015.

A Sequence Data Used

Table 9: Analysed samples (SE = single end, PE = paired end).

Sample	Read Type	File Name
sample_1	PE	GATC-Demo_sample_1_lib00007_1.fastq GATC-Demo_sample_1_lib00007_2.fastq
sample_2	PE	GATC-Demo_sample_2_lib00008_1.fastq GATC-Demo_sample_2_lib00008_2.fastq
sample_3	PE	GATC-Demo_sample_3_lib00009_1.fastq GATC-Demo_sample_3_lib00009_2.fastq

B Relevant Programs

Table 10: Name, version and description of relevant programs.

Program	Version	Description
Bowtie[1]	2.2.9	Bowtie is a ultrafast, memory-efficient short read aligner. It is based on Burrows-Wheeler transform algorithm.
Kraken[2]	0.10.6	Kraken is an ultrafast and highly accurate program for assigning taxonomic labels to metagenomic DNA sequences.
Krona[4]	2.5	Krona allows hierarchical data to be explored with zoomable pie charts.
Picard[7]	1.131	Picard is a java-based command-line utilities for processing SAM / BAM files.
R[8]	2.15.3	R is a programming language and environment for statistical computing.
Trimmomatic[9]	0.33	Trimmomatic performs a variety of useful trimming tasks for Illumina paired-end and single-end data.
sambamba[10]	0.6.6	Sambamba is a high performance modern robust and fast tool (and library), for working with SAM and BAM files.

C Filter Settings

Table 11: Filters used in postprocessing of taxonomic profiling results.

Filter	Value
Top OTUs to include in plots	20
Minimum read counts	50

Eurofins Genomics' products, services and applications reach the best quality and safety levels. They are carried out under strict QM and QA systems and comply with the following standards:

ISO 9001	Globally recognised as the standard quality management certification	GLP	The gold standard to conduct non-clinical safety studies
ISO 17025	Accredited analytical excellence	GCP	Pharmacogenomic services for clinical studies
ISO 13485	Oligonucleotides according to medical devices standard	cGMP	Products and testing according to pharma and biotech requirements

Eurofins Genomics • Anzinger Str. 7a • 85560 Ebersberg • Germany