eurofins | Genomics

Data Analysis Report: Metagenome Analysis v2.3

Project / Study: EF-Demo

Project description: INVIEW METAGENOME ADVANCE

Date: May 27, 2021

Table of Contents

1 Analysis workflow

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

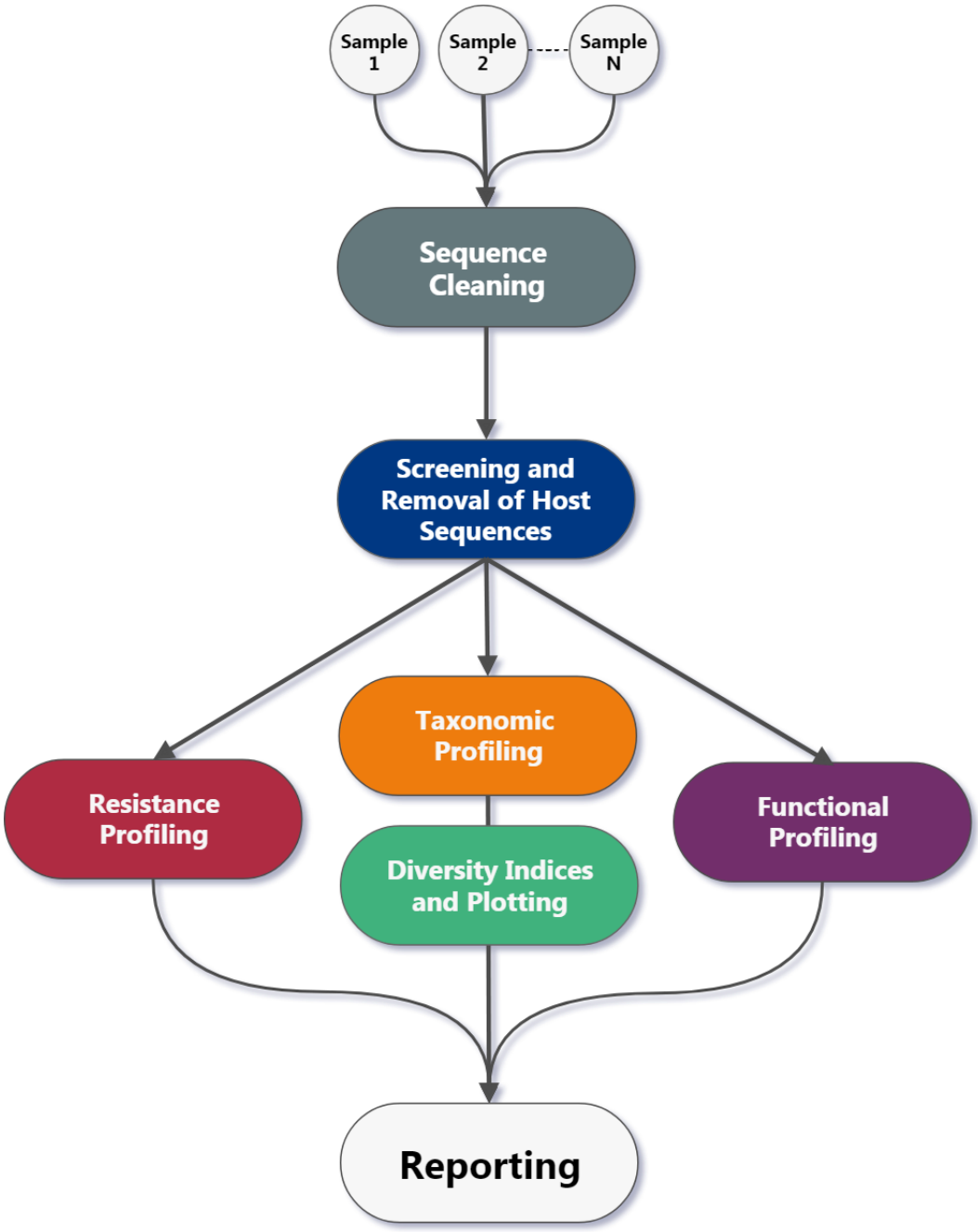

Figure 1: Metagenome Analysis v2.3 Workflow

2 Samples Analysed

Sample2, sample1.

3 Reference Database

Table 1: Taxonomic Profiling database composition (Metaphlan2).

Kingdom	Organisms	Sequences	Source
Archaea	292	311	NCBI Genomes (complete)
Bacteria	6,084	11,585	NCBI Genomes (complete)
Fungi	277	45,331	NCBI Genomes (complete + contigs)
Virus	10,212	13,701	NCBI Genomes (complete)

Table 2: Taxonomic Profiling database composition (KrakenUniq).

Kingdom	Organisms	Sequences	Source
Archaea	258	6,323	NCBI Genomes (complete)
Bacteria	5,550	303,145	NCBI Genomes (complete)
Fungi	245	171,704	NCBI Genomes (complete + contigs)
Protozoa	75	363,606	NCBI Genomes (complete + contigs)
Virus	9,201	11,876	NCBI Genomes (complete)

Table 3: IGC database (Integrated Gene Catalog of the human gut microbiome) [1].

Tag	Description
Name	IGC
Release	Mar. 2014
Genes (Million)	9.88
% Complete ORFs	57.74 %
Total length (Mbp)	7,436
Average length (bp)	753
N50 (bp)	1,035
N90 (bp)	384
Max length (bp)	88,230
Min length (bp)	100
% annotated on Phylum level	21.30 %
% annotated on Genus level	16.30 %
% annotated on KEGG	42.10 %
% annotated on eggNOG	60.40 %

Table 4: Mvir database of known toxins, virulence factors, and antibiotic resistance genes [2].

Tag	Description
Name	MvirDB
Release	Dec. 2015
Total sequences	26,373
Longest (bp)	198,867
Smallest (bp)	17
Mean (bp)	1,188
Median (bp)	798

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 5: Sequence quality metrics per sample

Sample	Total Reads	LQ Reads	Single Reads	HQ Reads
Sample2	44,215,290	1,317,449 (3.0%)	673,865 (1.5%)	42,223,976 (95.5%)
sample1	42,714,088	1,583,424 (3.7%)	814,862 (1.9%)	40,315,802 (94.4%)

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 Taxonomic profiling

After screening and removing host sequence reads, non-host reads are subjected to taxonomic profiling algorithm. Taxonomic profiling is done using Metaphlan2[3].

Uclassified reads are than subjected to KrakenUniq[4]. Kraken[5] 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. KrakenUniq computes the number of unique k-mers observed for each taxon, which allows to filter more false positives. Filters applied are listed in table **??**. The final classified, unclassified and filter passed reads are reported in table 6.

Table 6: Taxonomic Profiling metrics per sample.

Sample Name	Reads	Classified	Unclassified
Sample2	42,223,976	14,802,779 (35.06 %)	27,421,197 (64.94 %)
sample1	40,315,802	16,378,043 (40.62 %)	23,937,759 (59.38 %)

Table 7: Number of reads assigned to different kingdoms for Sample2, sample1.

Kingdom	Sample2		sample1	
Archaea	4,826	0.03 %	6,444	0.04 %
Bacteria	14,731,766	99.52 %	16,292,751	99.48 %
Eukaryota	1,592	0.01 %	1,304	0.01 %
Fungi	6,414	0.04 %	4,778	0.03 %
Viruses	33,764	0.23 %	47,996	0.29 %
Ambiguous	24,416	0.16 %	24,768	0.15 %

Ambigious: Reads which can not be assigned to one specific kingdom.
Eukaryota: Parasitic and non-parasitic Protozoa.

4.2.1 Taxa abundance

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[6] 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.normalized.tsv) contain normalized / rarefied read counts. The corresponding raw read counts are in Taxa-level.composition.reads.raw.tsv.

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.
Possible reasons for missing plots -

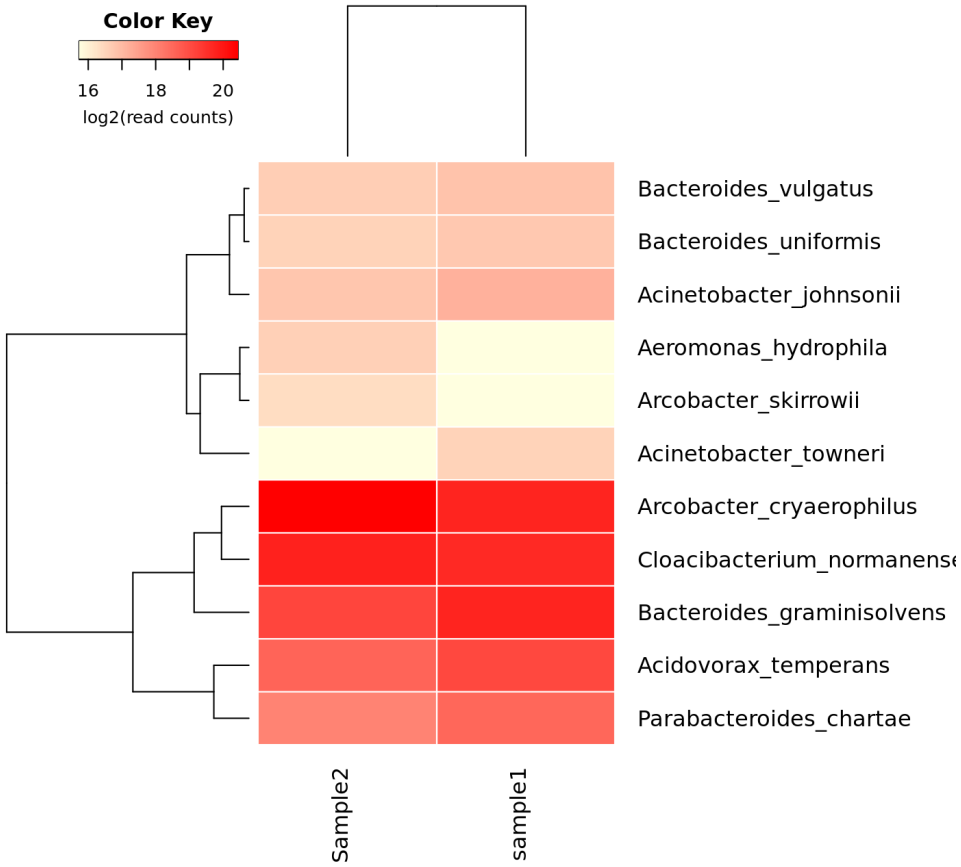

Figure 2: Heat map(s) showing the taxonomic abundance and their relation across the samples. Dendrograms determined by computing hierarchical clustering from the abundance levels shows the relationship between the species (left) and the samples (top). The abundance levels (number of reads associated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

i) **only one sample in the analysis**

ii) **missing data**

iii) **selected parameters are too stringent**

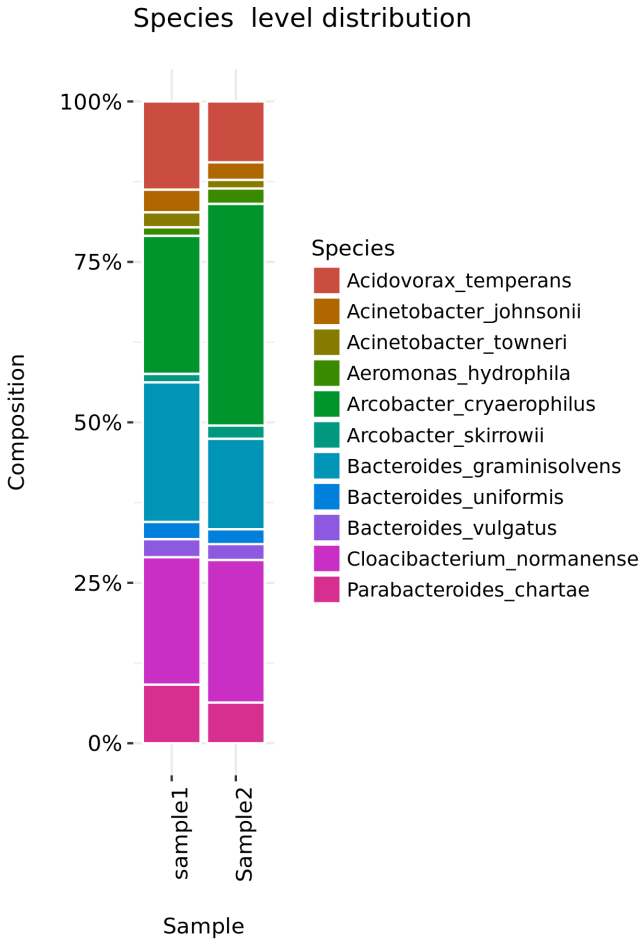

Figure 3: Bar plot(s) showing the taxonomic abundance across the samples. Taxa-level: Species

Possible reasons for missing plots -

i) **only one sample in the analysis**

ii) **missing data**

iii) **selected parameters are too stringent**

4.2.2 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[6] 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.

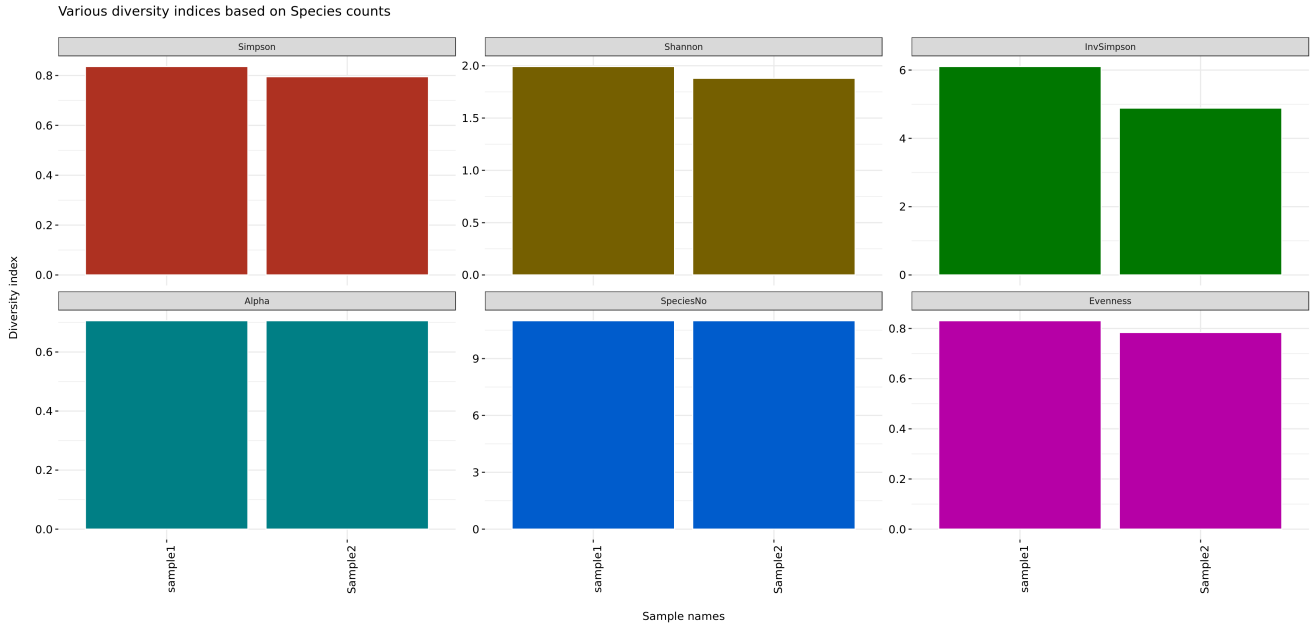

Figure 4: Various diversity indices computed based on the species counts found in each sample.

4.2.3 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.

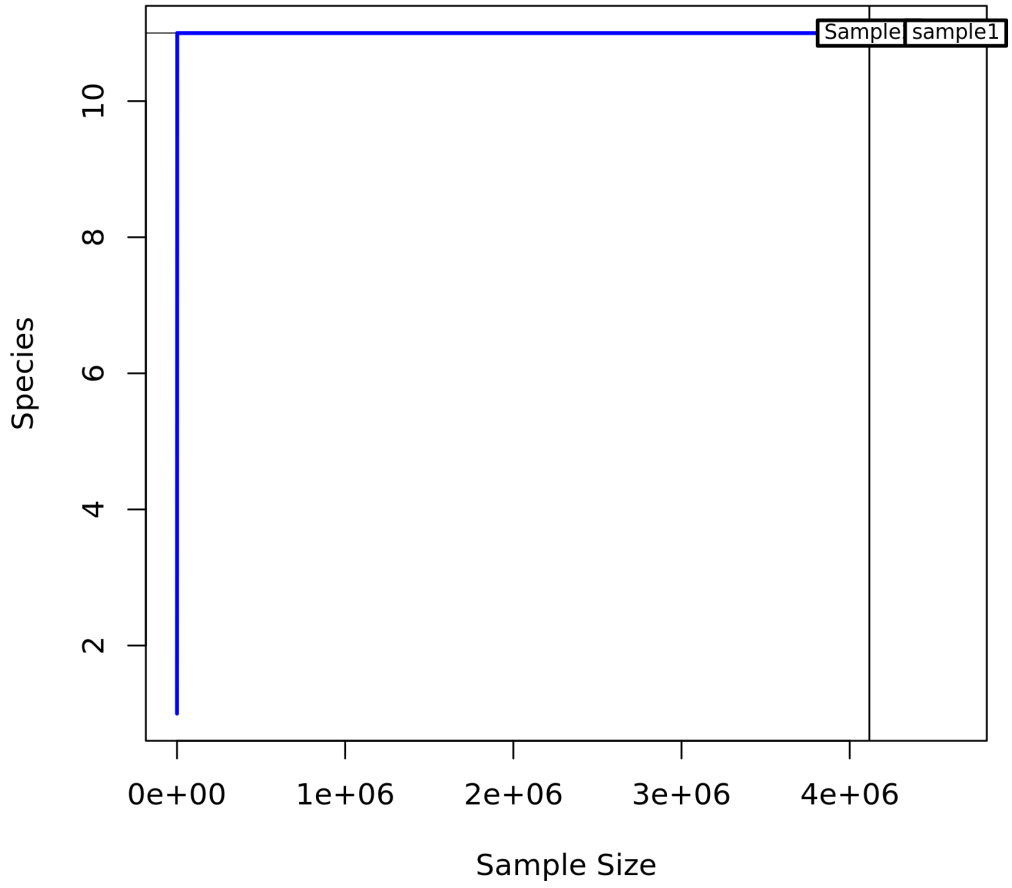

Figure 5: Rarefaction curve of annotated species richness.

4.2.4 Interactive plots

Taxonomic profiling results produced by KrakenUniq[4] are used to generate interactive plots using Krona[7]. Krona is a visualization tool that allows intuitive exploration of relative abundances and confidences within the complex hierarchies of metagenomic classifications.

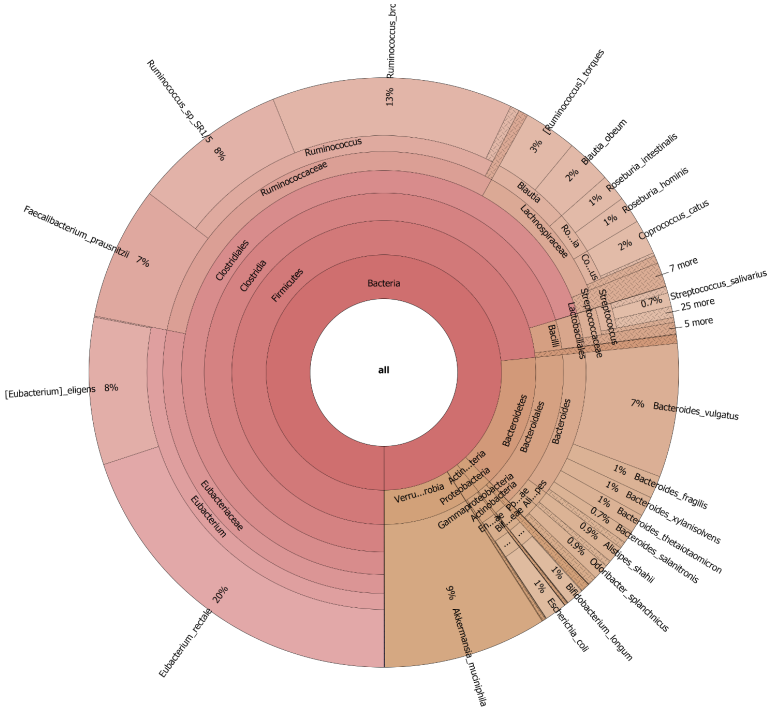

Figure 6: Example of an interactive plot generated by Krona (interactive_plots.html).

4.3 Functional profiling

Non-host sequence reads are mapped against a reference dataset - the integrated reference catalog (IGC[1]) using Bowtie[8] with default parameters. IGC contains high quality reference genes identified in the human microbiome project (http://commonfund.nih.gov/hmp/overview).
Reads that could be associated to IGC gene sets are recorded in table 8. IGC associated reads are further filtered to include only reads that could be placed uniquely and have both reads in a pair. High quality IGC associated reads are annotated, consolidated and reported.

Table 8: Functional Profiling metrics per sample

Sample Name	Reads	Mapped Reads
Sample2	41,741,414	4,073,777 (9.76%)
sample1	39,805,788	5,812,462 (14.60%)

The alignment classification table includes the following read categories:

- Mapped: Reads mapped to reference.
- Unique: Reads mapped to exactly one site on the reference.
- Non-unique: Reads mapped to more than one site on the reference.
- Singletons: Reads with itself mapped and its mate unmapped.
- Cross-Contig: Reads with the other end mapped to a different site.

Percentage of reads in category **Unique** is calculated based on the number of reads mapping to entire reference.

Table 9: Read metrics for Sample2, sample1.

Read category	Sample2	sample1
Mapped	4,073,777	5,812,462
Unique	2,405,703 (59.05%)	3,205,435 (55.15%)
Non-unique	1,668,074 (40.95%)	2,607,027 (44.85%)
Singletons	660,863 (16.22%)	793,612 (13.65%)
Cross-Contig	362,176 (8.89%)	364,946 (6.28%)

IGC associated reads are consolidated based on the Kyoto Encyclopedia of Genes and Genomes (KEGG)[9] functional annotations. KEGG is a database resource for understanding high-level functions and utilities of a biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput technologies.

The composition of various functional categories for each sample is summarized in the following table and figures.

Table 10: Composition of top 20 functional categories for all sample(s) (KEGG_ANNOTATION.composition.top_hits.tsv

FUNCTION	Sample2	sample1
unknown	33.40	33.48
Carbohydrate Metabolism	8.03	8.02
Cellular Processes and Signaling	6.53	6.46
Genetic Information Processing	5.57	5.48
Membrane Transport	5.44	5.49
Poorly Characterized	5.39	5.37
Metabolism	5.22	5.20
Replication and Repair	4.69	4.69
Amino Acid Metabolism	4.64	4.64
Nucleotide Metabolism	3.06	3.05
Enzyme Families	2.58	2.57
Energy Metabolism	2.37	2.35
Translation	2.04	2.09
Transcription	1.90	1.90
Folding, Sorting and Degradation	1.76	1.76
Metabolism of Cofactors and Vitamins	1.68	1.70
Glycan Biosynthesis and Metabolism	1.34	1.31
Signal Transduction	1.25	1.25
Lipid Metabolism	1.11	1.10
Metabolism of Terpenoids and Polyketides	0.52	0.53

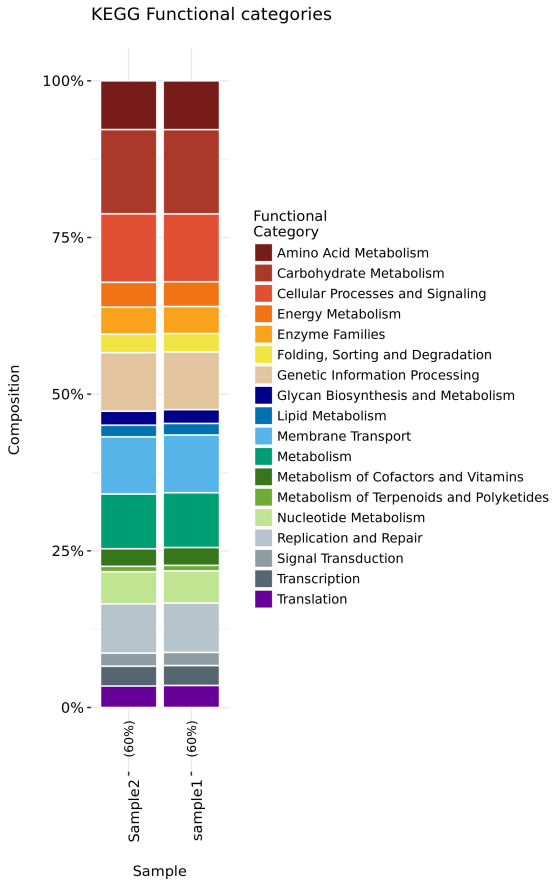

Figure 7: Bar plot showing the relative number of genes found in the most highly represented functional categories for all samples.

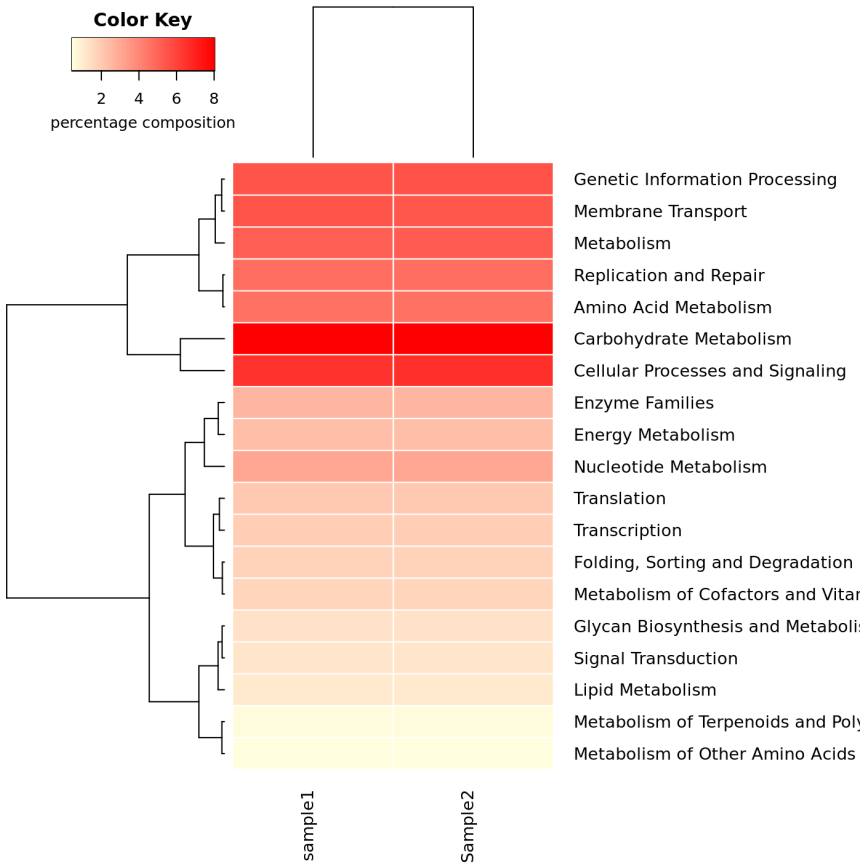

Figure 8: Heat map showing the frequency of the most highly represented functional categories and their relation across the samples. Dendrograms determined by computing hierarchical clustering from the frequencies shows the relationship between the various functional categories (left) and the samples (top).

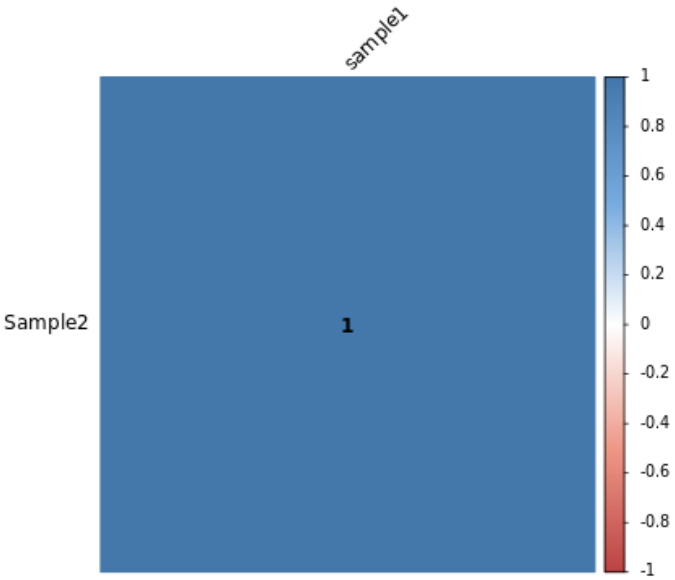

Figure 9: Correlation plot showing the relationship between the samples based on the identified functional profiles of the respective samples. Values close to +1 indicate a high degree of positive correlation between the sample pair, whereas values close to -1 indicate a high degree of negative correlation between the sample pair in comparison. Values close to zero indicate poor correlation of either kind, and 0 indicates no correlation at all.

4.4 Resistance screening

Non-host sequence reads are mapped against a resistance gene dataset - the microbial virulence database (MvirDB[2]) using Bowtie[8] with default parameters. MvirDB is a collection of genes known to have virulence properties like antibiotic resistance, pathogenicity island, resistance protein and transcription factors http://nar.oxfordjournals.org/content/35/suppl_1/D391.full.
Reads mapping to MvirDB are recorded in table 11. Virulence associated reads are further filtered to include only reads that could be placed uniquely and have both reads of a pair. High quality virulence associated reads are annotated, consolidated and reported.

Table 11: Resistance screening metrics per sample

Sample Name	Reads	Mapped Reads
Sample2	41,741,414	78,238 (0.19%)
sample1	39,805,788	89,000 (0.22%)

The alignment classification table includes the following read categories:

- Mapped: Reads mapped to reference.
- Unique: Reads mapped to exactly one site on the reference.
- Non-unique: Reads mapped to more than one site on the reference.
- Singletons: Reads with itself mapped and its mate unmapped.
- Cross-Contig: Reads with the other end mapped to a different site.

Percentage of reads in category **Unique** is calculated based on the number of reads mapping to entire reference.

Table 12: Read metrics for Sample2, sample1.

Read category	Sample2	sample1
Mapped	78,238	89,000
Unique	41,021 (52.43%)	42,840 (48.13%)
Non-unique	37,217 (47.57%)	46,160 (51.87%)
Singletons	19,496 (24.92%)	18,064 (20.30%)
Cross-Contig	2,326 (2.97%)	2,116 (2.38%)

Read distribution on various virulence factors for each sample is summarized in the following table and figure.

Table 13: Distribution of virulence factors for all sample(s) (VIRULENCE.reads.tsv)

Virulence_Factor_Type	Sample2	sample1
antibiotic resistance	13258	15028
pathogenicity island	18480	20384
transcription factor	246	364
virulence protein	1370	1538

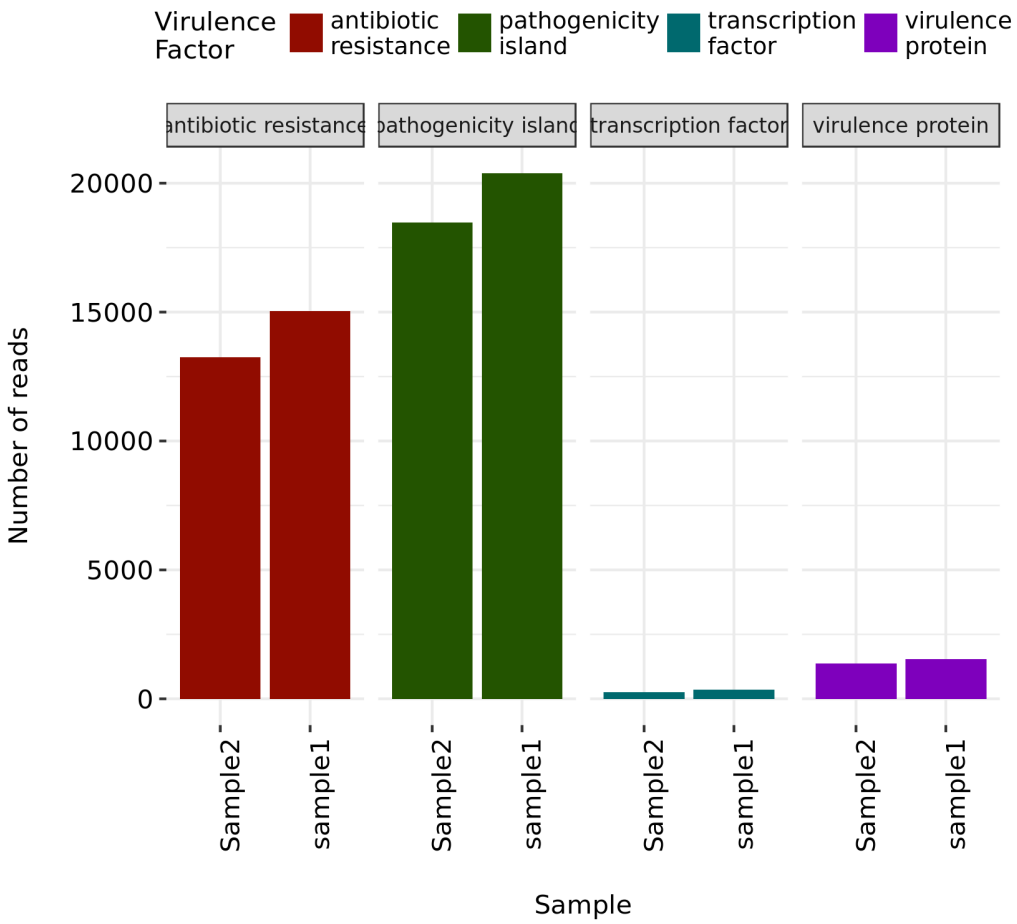

Figure 10: Sample-wise read distribution of various virulence factors.

Table 14: Relative composition of 25 most abundantly represented virulence factors for all sample(s) (VIRULENCE.annotation.filtered.report.tsv)

Virulence_Factor_Type	Gene_Names	Short_Description	Sample2	sample1
pathogenicity island	Xfa67G	tRNA 32 (Glycine) of(...)	7.54	3.86
antibiotic resistance	PSEMT	Pseudomonas aerugino(...)	6.42	4.57
pathogenicity island	Eco34T	tRNA 5 (Threonine) o(...)	4.67	4.00
antibiotic resistance	AAM70497|AAM70497|AF(...)	transposase TnpA [Es(...)	4.29	3.96
pathogenicity island	Sagn167L	tRNA 41 (Leucine) of(...)	3.50	4.35
pathogenicity island	Sagt164L	tRNA 44 (Leucine) of(...)	3.42	4.13
pathogenicity island	Eco21R	tRNA 8 (Arginine) of(...)	3.43	2.62
antibiotic resistance	AAG14402|AAG14402|AF(...)	putative transposase(...)	3.02	2.64
virulence protein	AAL08440|transposase(...)	transposase TnpA [Sh(...)	2.21	2.03
antibiotic resistance	CAG25423|CAG25423	truncated transposas(...)	2.27	1.96
pathogenicity island	NP_414793|NP_414793	IS5 transposase and (...)	2.37	1.82
pathogenicity island	NP_415084|NP_415084	IS5 transposase and (...)	2.31	1.76
antibiotic resistance	AAR25034|AAR25034	TnpA [Escherichia co(...)	1.90	1.78
antibiotic resistance	CAA51175|CAA51175	transposase [Klebsie(...)	1.32	1.06
antibiotic resistance	AAA22911|AAA22911	putative [Bacteroide(...)	0.80	1.39
pathogenicity island	Bth90F	tRNA 40 (Phenylalani(...)	0.82	1.03
pathogenicity island	Ecoc48X	tRNA 50 (tmRNA) of C(...)	0.75	1.04
pathogenicity island	Ecoc105F	tRNA 61 (Phenylalani(...)	0.62	0.88
pathogenicity island	Stic134F	tRNA 73 (Phenylalani(...)	0.61	0.78
antibiotic resistance	AAG54073|AAG54073	transposase [Escheri(...)	0.44	0.74
pathogenicity island	Ecoc114S	tRNA 25 (Serine) of (...)	0.45	0.69
antibiotic resistance	AAR18978|AAR18978	transposase [Klebsie(...)	0.47	0.64
pathogenicity island	Bth60K	tRNA 36 (Lysine) of (...)	0.51	0.60
antibiotic resistance	AAL02126|AAL02126	transposase [Escheri(...)	0.38	0.71
transcription factor	SubName: Full=RteB p(...)	SubName: Full=RteB p(...)	0.45	0.54

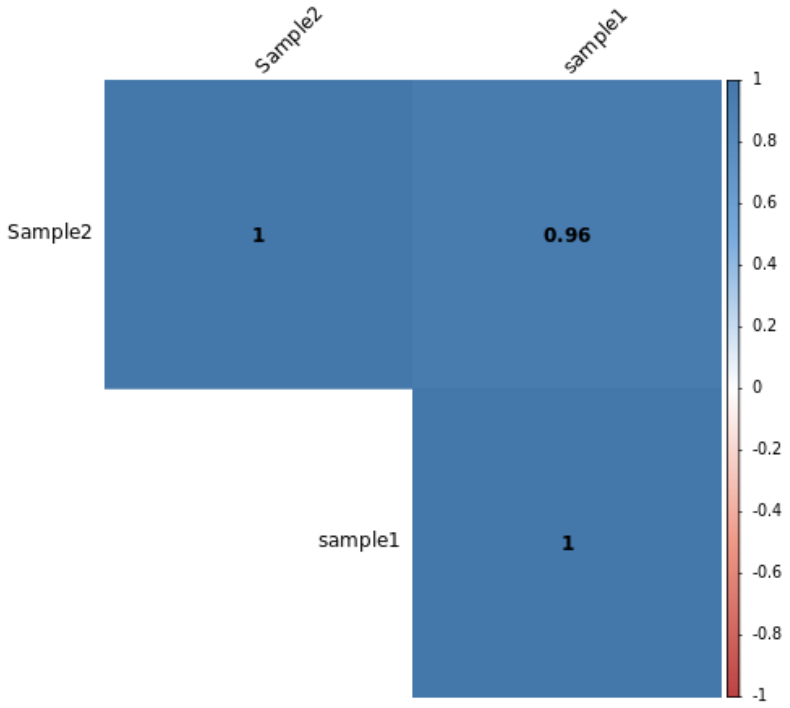

Figure 11: Correlation plot showing the relationship between the samples based on the identified functional profiles of the respective samples. Values close to +1 indicate a high degree of positive correlation between the sample pair in comparison whereas values close to -1 indicate a high degree of negative correlation between the sample pair. Values close to zero indicate poor correlation of either kind, and 0 indicates no correlation at all.

5 Deliverables

Table 15: 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
KEGG_ANNOTATION.barplot.png	PNG	Image viewer
KEGG_ANNOTATION.composition.filtered.tsv	TSV	Spreadsheet Editor
KEGG_ANNOTATION.composition.top_hits.tsv	TSV	Spreadsheet Editor
KEGG_ANNOTATION.composition.tsv	TSV	Spreadsheet Editor
KEGG_ANNOTATION.correlation.png	PNG	Image viewer
KEGG_ANNOTATION.heatmap.png	PNG	Image viewer
KEGG_ANNOTATION.reads.tsv	TSV	Spreadsheet Editor
KEGG_ANNOTATION.tilemap.labels.png	PNG	Image viewer
KEGG_ANNOTATION.tilemap.png	PNG	Image viewer
Species.barplot.png	PNG	Image viewer
Species.composition.proportion.tsv	TSV	Spreadsheet Editor
Species.composition.reads.normalized.tsv	TSV	Spreadsheet Editor
Species.composition.reads.raw.tsv	TSV	Spreadsheet Editor
Species.diversity_indices.png	PNG	Image viewer
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
VIRULENCE.annotation.filtered.tsv	TSV	Spreadsheet Editor
VIRULENCE.barplot.log10.png	PNG	Image viewer
VIRULENCE.barplot.png	PNG	Image viewer
VIRULENCE.correlation.png	PNG	Image viewer
VIRULENCE.reads.tsv	TSV	Spreadsheet Editor

6 Formats

Table 16: References and descriptions of file format.

Format	Description
HTML	Standard markup language for creating web pages and web applications
PNG	Figure or image in Portable Network Graphics format
TSV	Tab separated table style text file. This can be imported into spreadsheet processing software like MS OFFICE Excel.

7 FAQ

Q: How can I open a CSV, TSV, or VCF file in Excel?
A: You can open CSV, TSV, VCF, or any other text file using Excel. Please follow this procedure:

i) Start Excel
ii) Click on the "File" menu button in the top left corner
iii) Click on the "Open" menu button in the left menu pane
iv) Click on the dropdown-menu in the bottom right corner of the small window that opens. Initially, it should show "All Excel files (*.xls; *.xlsx)".
v) Select the topmost entry "All files (*.*)"
vi) Navigate to the directory with the text files. They should be visible now.
vii) Open the files and click through the appearing "Text Import Wizard" dialog (Next, Next, Done).

Depending on the content of the text file you want to import, you might want to change some settings in the "Text Import Wizard" dialog. Most often, you want to change the decimal separator. The provided text files use the dot as decimal separator and comma as thousands separator. Make sure that you set both correctly. To do this, click on the "Advanced" button in pane 3 of the "Text Import Wizard" dialog. You can find additional information in this article at the Microsoft Office support site.

Q: How can I view alignments and variants?
A: A convenient tool to view alignments and variant data is the *Integrative Genomics Viewer* (IGV) for Unix, MS Windows, and MacOS X. It can be downloaded and installed locally, or can be run as web-application.

- Before loading alignments or variant data into IGV, the reference genome FASTA file has to be loaded via the *Genomes -> Load Genome from File* menu. Make sure that you load the same reference genome FASTA file that was used during mapping.
- To load alignments into IGV select the BAM files via the *File -> Load from File* menu. Please note that you need to zoom-in to about 30kb to see alignments. You can set this visibility range threshold and other displaying and filtering options via the *View -> Preferences -> Alignments* menu, or the right-click context menu.
- To load variant data into IGV select the VCF files via the *File -> Load from File* menu. IGV can color mismatch bases and InDel positions. Use the right-click context menu to configure this and other displaying and filtering options. Not all mismatch positions in alignments might have been considered significant by the variant analysis tool and therefore might not be contained in the variant tracks.
- Please visit the IGV online manual to get more information about loading genomes, viewing alignments, and viewing variants.

8 Bibliography

[1] Junhua Li, Huijue Jia, Xianghang Cai, Huanzi Zhong, Qiang Feng, Shinichi Sunagawa, Manimozhiyan Arumugam, Jens Roat R. Kultima, Edi Prifti, Trine Nielsen, Agnieszka Sierakowska S. Juncker, Chaysavanh Manichanh, Bing Chen, Wenwei Zhang, Florence Levenez, Juan Wang, Xun Xu, Liang Xiao, Suisha Liang, Dongya Zhang, Zhaoxi Zhang, Weineng Chen, Hailong Zhao, Jumana Yousuf Y. Al-Aama, Sherif Edris, Huanming Yang, Jian Wang, Torben Hansen, Henrik Bjørn B. Nielsen, Søren Brunak, Karsten Kristiansen, Francisco Guarner, Oluf Pedersen, Joel Doré, S. Dusko Ehrlich, MetaHIT Consortium, Peer Bork, Jun Wang, and MetaHIT Consortium. An integrated catalog of reference genes in the human gut microbiome. *Nature biotechnology*, 32(8):834–841, August 2014.

[2] C. E. Zhou, J. Smith, M. Lam, A. Zemla, M. D. Dyer, and T. Slezak. MvirDB-a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Research*, 35(suppl 1):D391–D394, January 2007.

[3] Duy Tin Truong, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12:902 EP –, Sep 2015. Correspondence.

[4] F. P. Breitwieser, D. N. Baker, and S. L. Salzberg. Krakenuniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biology*, 19(1):198, Nov 2018.

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

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

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

[8] 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.

[9] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research*, 27(1):29–34, January 1999.

[10] Derek Barnett, Erik Garrison, Aaron Quinlan, Michael Strömberg, and Gabor Marth. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12):btr174–1692, April 2011.

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

[12] 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.

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

[14] 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.

[15] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15):2114–2120, August 2014.

A Sequence Data Used

Naming convention for FASTQ files:
<project-id>_<sample-id>_<lib-id>_<run-id>_<lane-no>_<read-no>.fastq.gz

<project-id> the unique identifier of this project.

<sample-id> the sample name as provided by the customer.

<lib-id> a unique identifier of the sequencing library created in the lab. Multiple sequencing libraries may have been created from the same sample material, depending e.g. on project setup.

<run-id> a unique identifier of the sequencing run that created this file.

<lane-no> a number specifying the lane of the sequencing device used for sequencing.

<read-no> either _1 or _2. For paired-end runs, these numbers identify the associated forward and reverse read files (mate pairs).

Table 17: Analysed samples.

No.	Sample	File Name
1	Sample2	EF-Demo_Sample2_lib12346_1234_1_1.fastq.gz.gz
		EF-Demo_Sample2_lib12346_1234_1_2.fastq.gz.gz
2	sample1	EF-Demo_sample1_lib12345_1234_1_1.fastq.gz.gz
		EF-Demo_sample1_lib12345_1234_1_2.fastq.gz.gz

B Relevant Programs

Table 18: Name, version and description of relevant programs.

Program	Version	Description
bamtools[10]	2.3.0	BamTools provides a small, but powerful suite of command-line utility programs for manipulating and querying BAM files for data.
Bowtie[8]	2.3.3.1	Bowtie is a ultrafast, memory-efficient short read aligner. It is based on Burrows-Wheeler transform algorithm.
KrakenUniq[5, 4]	0.5.3	Kraken is an ultrafast and highly accurate program for assigning taxonomic labels to metagenomic DNA sequences. KrakenUniq adds some additional functionality - most notably a unique k-mer count using the HyperLogLog algorithm.
Krona[7]	2.5	Krona allows hierarchical data to be explored with zoomable pie charts.
Metaphlan[3]	2.9.20	MetaPhlAn2 is a computational tool for profiling the composition of microbial communities (Bacteria, Archaea, Eukaryotes and Viruses) from metagenomic shotgun sequencing data with species level resolution.
Picard[11]	1.131	Picard is a java-based command-line utilities for processing SAM / BAM files.
R[12]	3.2.4	R is a programming language and environment for statistical computing.
sambamba[13]	0.6.6	Sambamba is a high performance modern robust and fast tool (and library), for working with SAM and BAM files.
SAMTools[14]	0.1.18	SAMtools provide various utilities for manipulating alignments in the SAM format.
Trimmomatic[15]	0.33	Trimmomatic performs a variety of useful trimming tasks for Illumina paired-end and single-end data.

C Filter Settings

Table 19: Filters used in postprocessing of taxonomic profiling results.

Filter	Value
Top OTUs to include in plots	20
Minimum read count proportion	0.01

Table 20: Filters used in postprocessing of functional profiling, resistance screening results.

Filter	Value
Top hits to include in plots	20.00
Minimum composition across samples	0.50
Exclude categories from plots	unknown,Poorly Characterized

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

Eurofins Genomics Europe Sequencing GmbH • Jakob-Stadler-Platz 7 • 78467 Constance • Germany