eurofins | Genomics

Data Analysis Report: Metagenome Analysis v1.1

Project / Study: GATC-Demo

Project description: INVIEW METAGENOME ADVANCE

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.

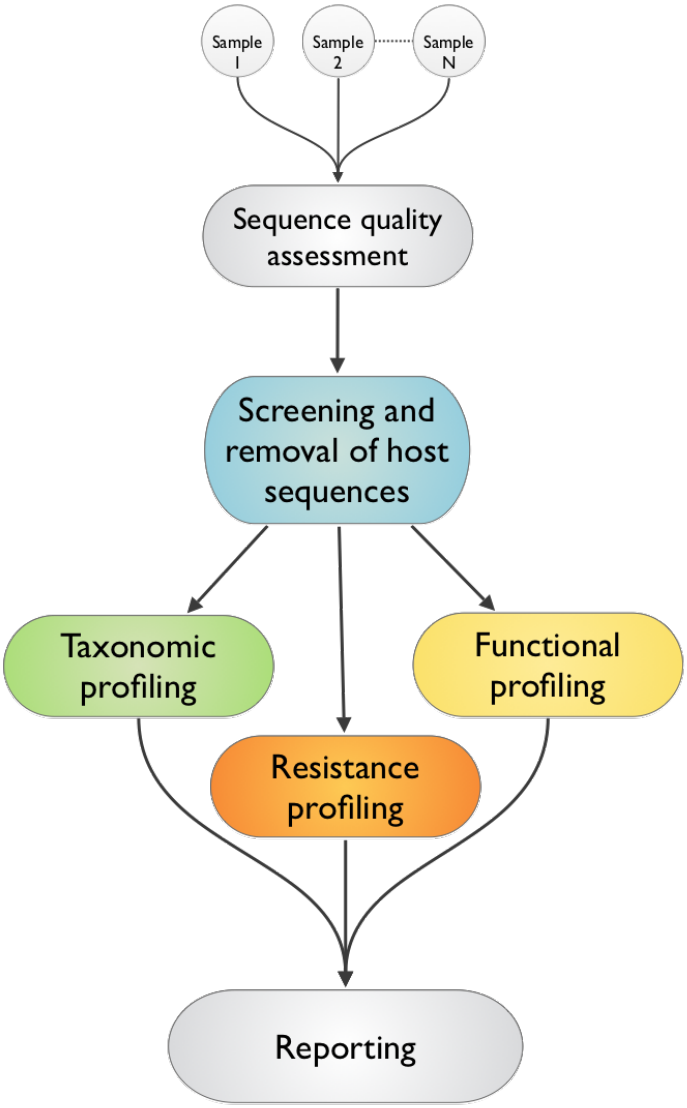


Figure 1: Metagenome Analysis v1.1 Workflow

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)

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
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[3] 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 6: 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[4] 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 7.

Table 7: 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 8: 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[5] 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.

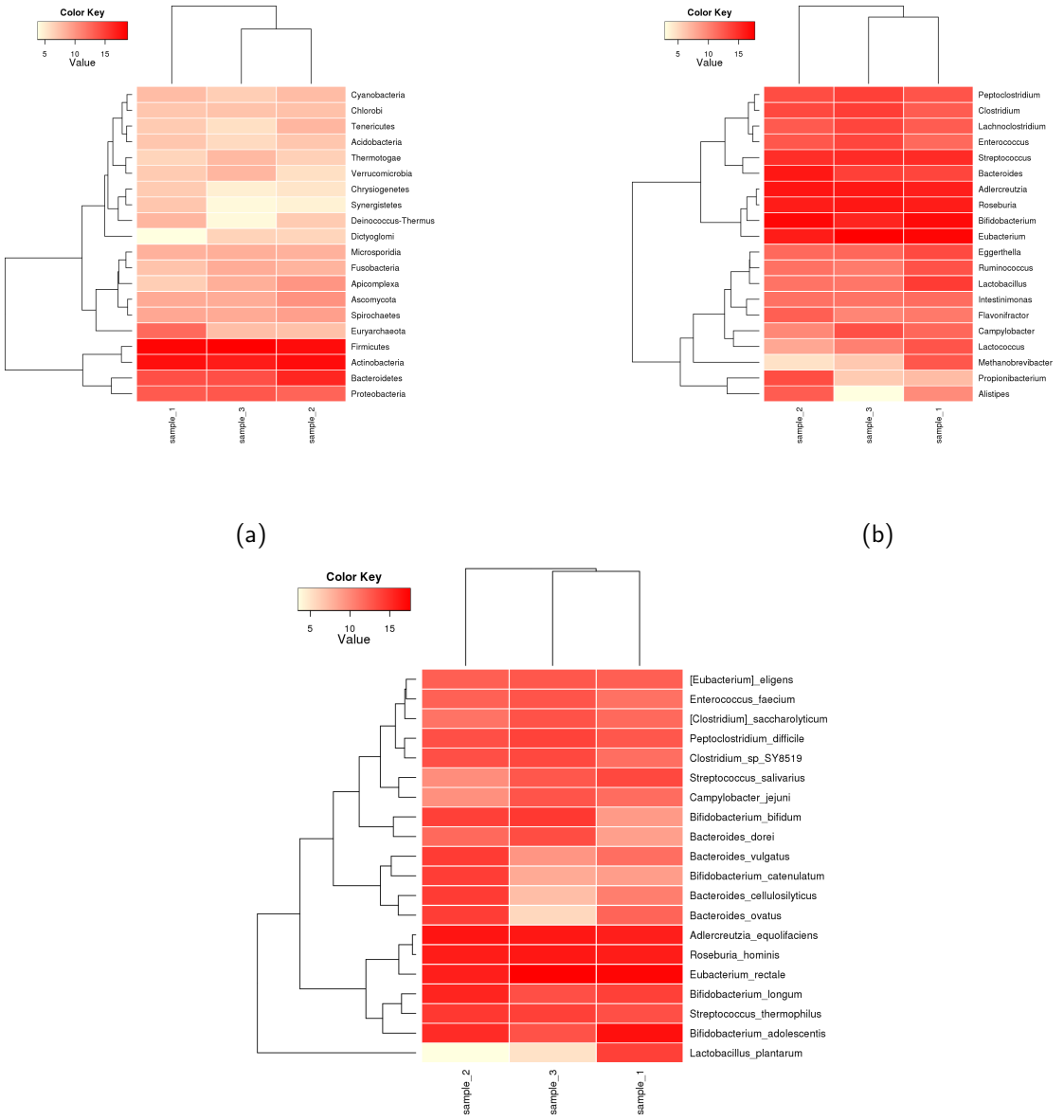
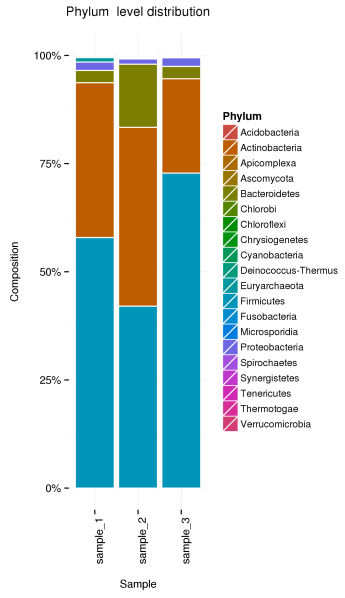
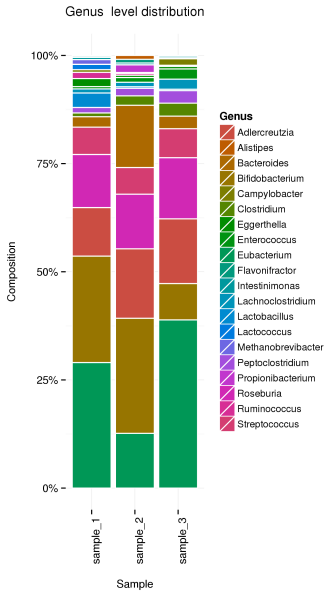


Figure 2: Heat map(s) showing the taxonomic abundance and their relation across the samples. Dendrograms determined by computing hierarchical clustering from the abundance levels shows the relationship between the species (left) and the samples (top). The abundance levels (number of reads associated 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[5] 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.

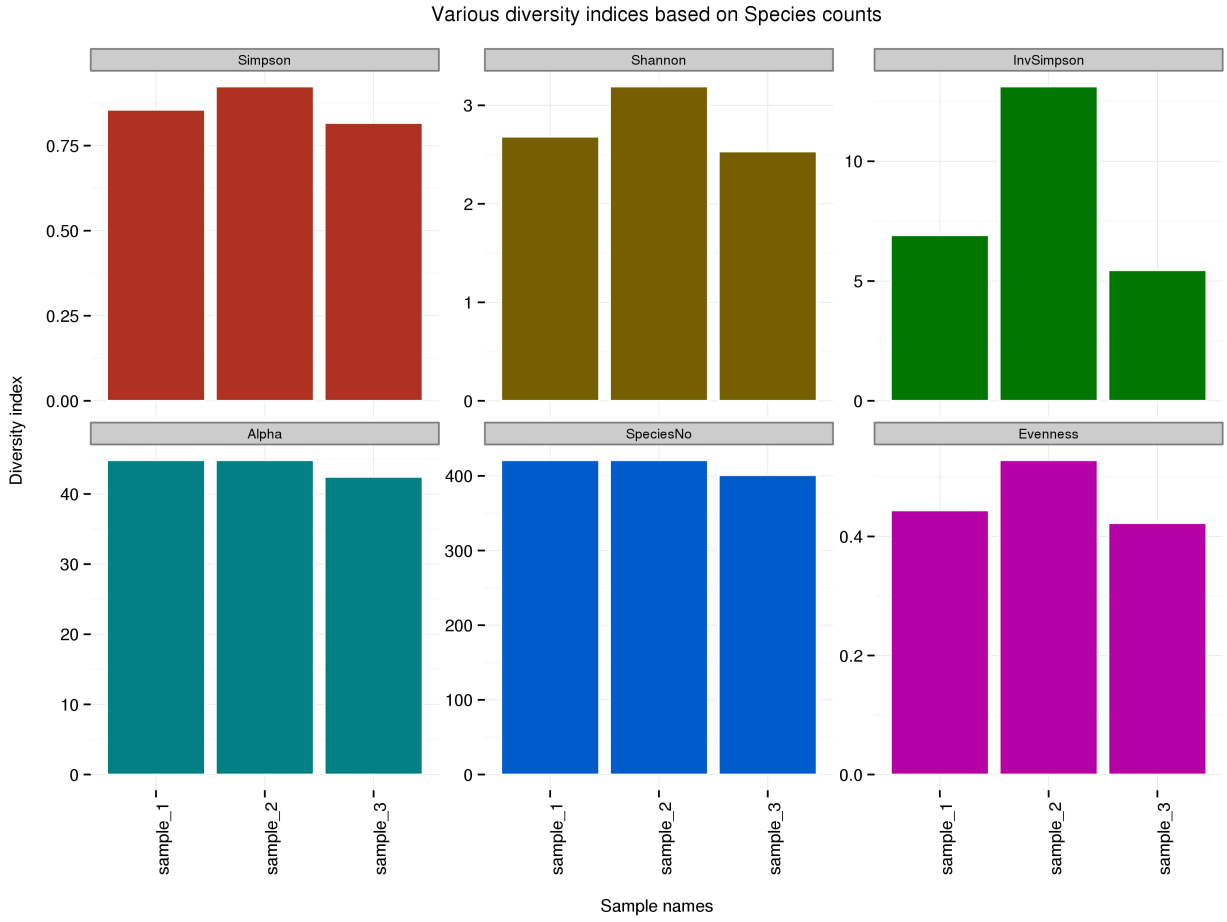


Figure 4: Various diversity indices computed based on the species counts found in each sample.

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.

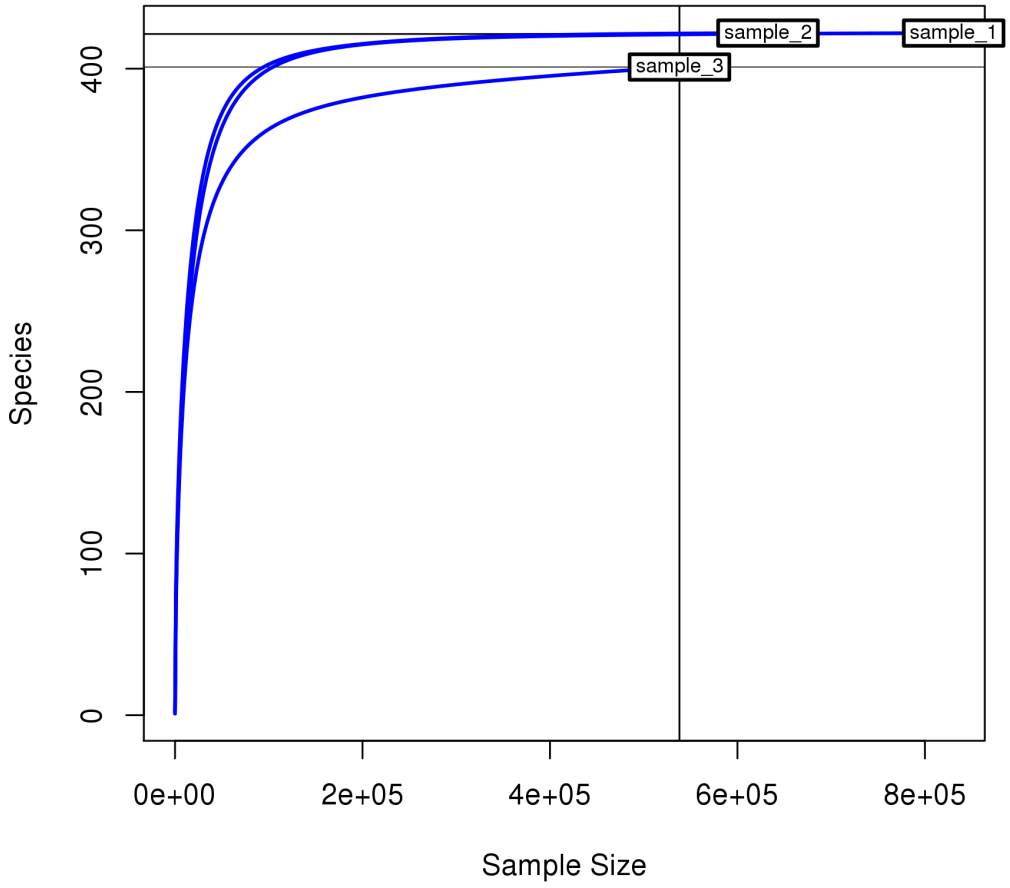


Figure 5: Rarefaction curve of annotated species richness.

4.3.4 Interactive plots

Taxonomic profiling results produced by Kraken[4] are used to generate interactive plots using Krona[6]. Krona is a visualization tool that allows intuitive exploration of relative abundances and confidences within the complex hierarchies of metagenomic classifications.

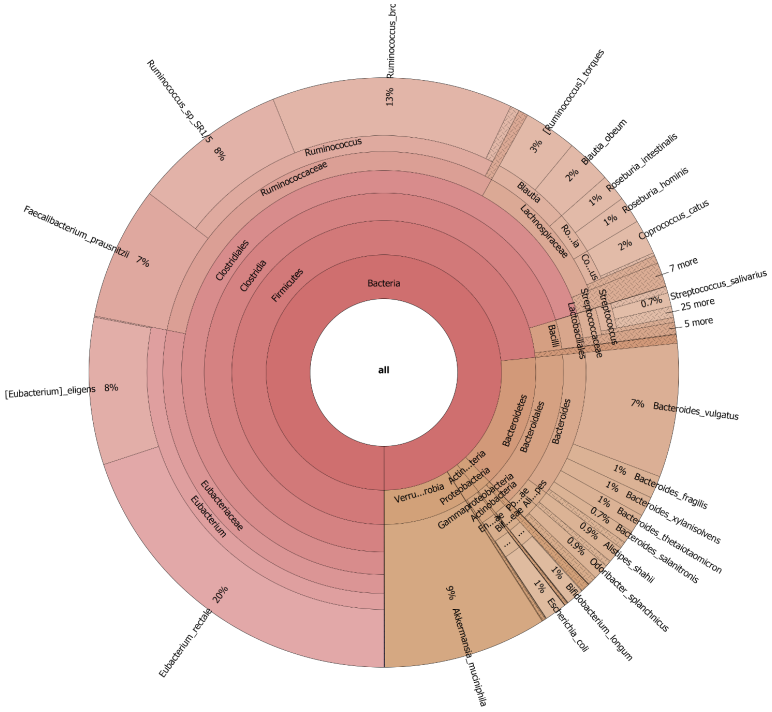


Figure 6: Example of an interactive plot generated by Krona (interactive_plots.html).

4.4 Functional profiling

Non-host sequence reads are mapped against a reference dataset - the integrated reference catalog (IGC[1]) using Bowtie[3] 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 9. 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 9: Functional Profiling metrics per sample

Sample Name	Reads	Mapped Reads
sample_1	9,911,980	8,138,998 (82.11%)
sample_2	9,778,356	8,213,899 (84.00%)
sample_3	9,598,988	7,893,725 (82.23%)

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 categories **Non-unique, Unique, Singletons, Cross-Contig** are calculated based on the number of reads mapping to entire reference.

Table 10: Read metrics for sample_1, sample_2, sample_3.

Read category	sample_1	sample_2	sample_3
Mapped	8,138,998	8,213,899	7,893,725
Unique	3,822,400 (46.96%)	3,610,224 (43.95%)	3,301,304 (41.82%)
Non-unique	4,316,598 (53.04%)	4,603,675 (56.05%)	4,592,421 (58.18%)
Singletons	964,804 (11.85%)	955,211 (11.63%)	950,357 (12.04%)
Cross-Contig	828,268 (10.18%)	807,682 (9.83%)	829,866 (10.51%)

IGC associated reads are consolidated based on the Kyoto Encyclopedia of Genes and Genomes (KEGG)[7] 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 11: Composition of top 20 functional categories for all sample(s) (KEGG_ANNOTATION.composition.top_hits.tsv

FUNCTION	sample_1	sample_2	sample_3
unknown	26.42	26.79	28.78
Carbohydrate Metabolism	8.78	8.77	7.97
Membrane Transport	8.19	8.25	8.26
Cellular Processes and Signaling	6.46	6.40	5.98
Metabolism	5.50	5.78	5.45
Amino Acid Metabolism	5.71	5.45	5.11
Replication and Repair	5.26	5.09	5.08
Genetic Information Processing	4.96	4.73	5.51
Poorly Characterized	4.39	4.46	4.01
Nucleotide Metabolism	3.45	3.48	3.32
Energy Metabolism	2.86	3.05	2.86
Enzyme Families	2.82	2.88	2.77
Transcription	2.42	2.62	2.65
Metabolism of Cofactors and Vitamins	2.25	2.23	2.16
Translation	2.22	2.06	1.94
Signal Transduction	1.56	1.58	1.55
Folding, Sorting and Degradation	1.66	1.45	1.37
Lipid Metabolism	1.18	1.16	1.23
Glycan Biosynthesis and Metabolism	1.18	1.11	0.96
Infectious Diseases	0.51	0.50	0.78

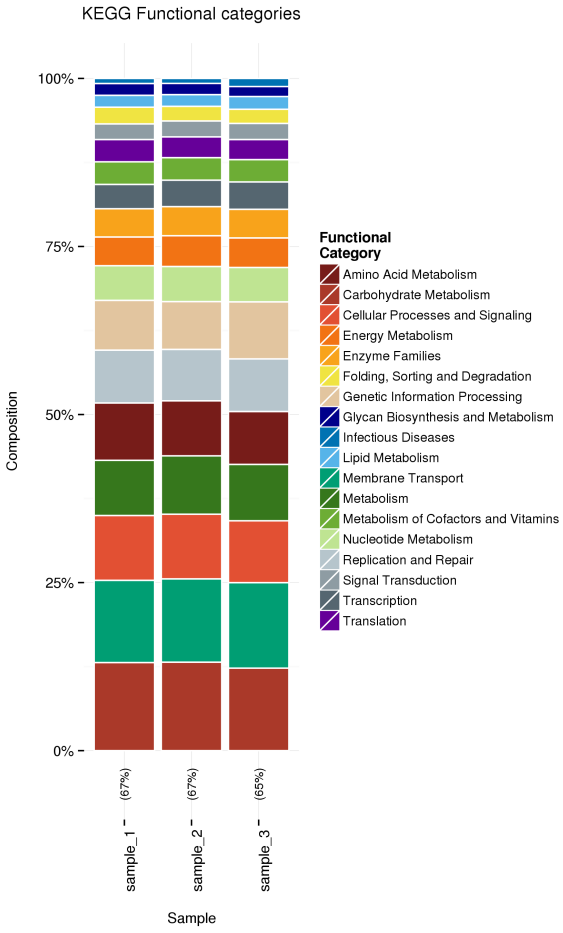


Figure 7: Bar plot showing the relative number of genes found in the most highly represented functional categories for all samples.

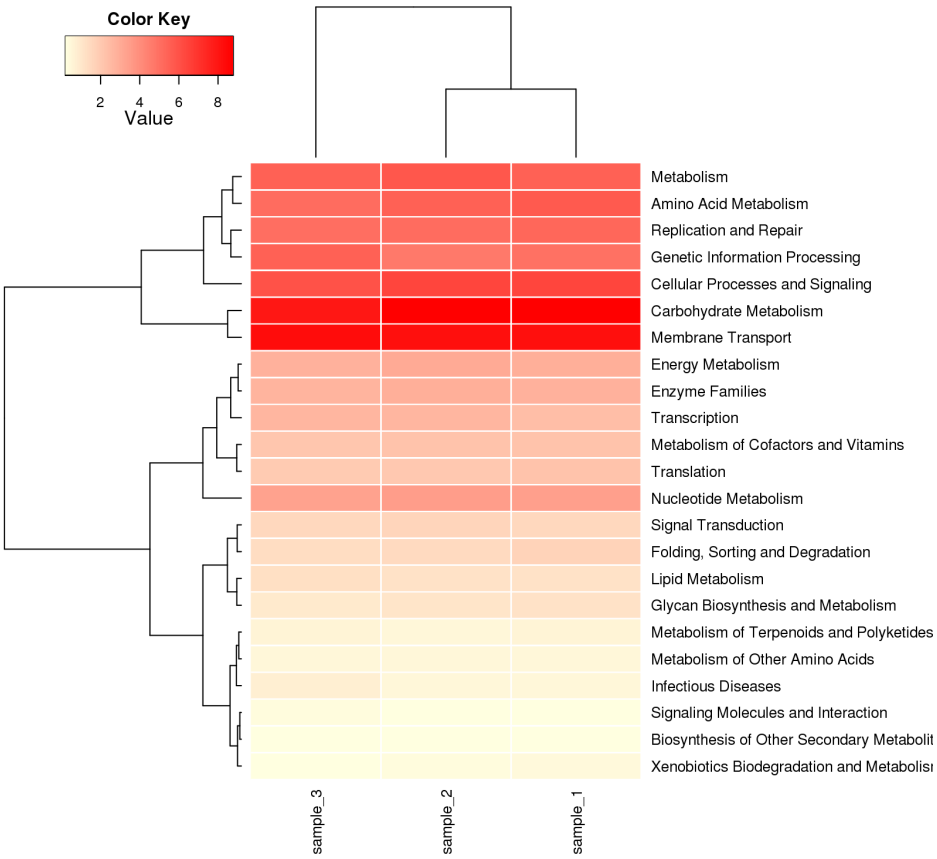


Figure 8: Heat map showing the frequency of the most highly represented functional categories and their relation across the samples. Dendrograms determined by computing hierarchical clustering from the frequencies shows the relationship between the various functional categories (left) and the samples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

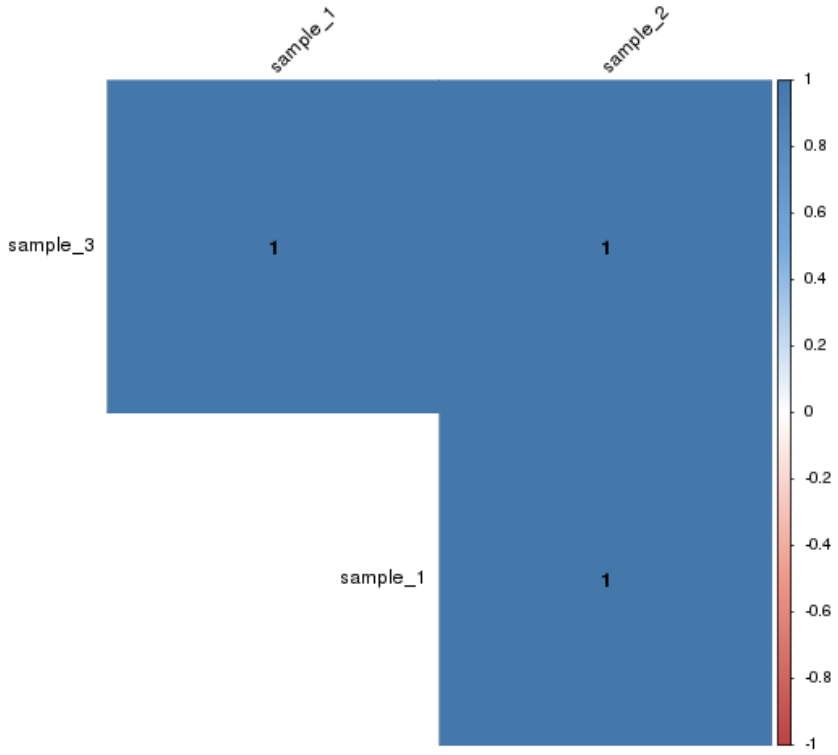


Figure 9: Correlation plot showing the relationship between the samples based on the identified functional profiles of the respective samples. Values close to +1 indicate a high degree of positive correlation between the sample pair, whereas values close to -1 indicate a high degree of negative correlation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

4.5 Resistance screening

Non-host sequence reads are mapped against a resistance gene dataset - the microbial virulence database (MvirDB[2]) using Bowtie[3] 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 12. 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 12: Resistance screening metrics per sample

Sample Name	Reads	Mapped Reads
sample_1	9,911,980	13,383 (0.14%)
sample_2	9,778,356	16,833 (0.17%)
sample_3	9,598,988	16,377 (0.17%)

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 categories **Non-unique, Unique, Singletons, Cross-Contig** are calculated based on the number of reads mapping to entire reference.

Table 13: Read metrics for sample_1, sample_2, sample_3.

Read category	sample_1	sample_2	sample_3
Mapped	13,383	16,833	16,377
Unique	3,382 (25.27%)	5,933 (35.25%)	2,163 (13.21%)
Non-unique	10,001 (74.73%)	10,900 (64.75%)	14,214 (86.79%)
Singletons	5,803 (43.36%)	6,447 (38.30%)	6,793 (41.48%)
Cross-Contig	134 (1.00%)	106 (0.63%)	156 (0.95%)

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

Table 14: Distribution of virulence factors for all sample(s) (VIRULENCE.reads.tsv)

Virulence_Factor_Type	sample_1	sample_2	sample_3
antibiotic resistance	1204	1832	5028
pathogenicity island	5058	7238	3782
virulence protein	84	90	62

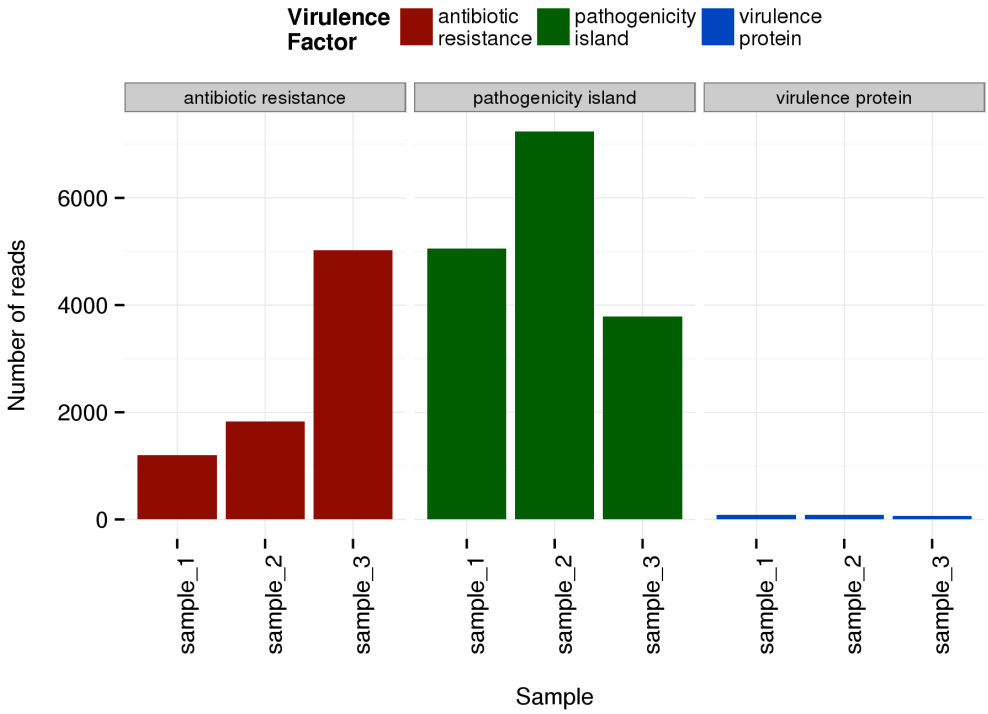


Figure 10: Sample-wise read distribution of various virulence factors.

Table 15: 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	sample_1	sample_2	sample_3
pathogenicity island	Blo57N	tRNA 44 (Asparagine)(...)	19.28	31.37	5.19
pathogenicity island	Sagn167L	tRNA 41 (Leucine) of(...)	17.48	12.40	16.30
pathogenicity island	Sagt164L	tRNA 44 (Leucine) of(...)	16.88	12.10	16.60
pathogenicity island	Blo17B	tRNA 10 (Initiator) (...)	6.06	8.80	1.50
antibiotic resistance	CAM12479|CAM12479	Tet(40) protein [unc(...)	0.11	0.26	9.81
antibiotic resistance	ZP_01996651|ZP_01996(...)	hypothetical protein(...)	1.23	2.32	2.16
antibiotic resistance	YP_600745|YP_600745	tetracycline resista(...)	0.49	0.26	3.48
pathogenicity island	NP_696649|NP_696649	hypothetical protein(...)	1.58	1.43	0.00
antibiotic resistance	EDP15418|EDP15418	hypothetical protein(...)	0.60	0.24	2.12
antibiotic resistance	AAV80411|AAV80411	TetO [Enterococcus f(...)	0.52	0.44	1.86
antibiotic resistance	RecName: Full=Tetrac(...)	RecName: Full=Tetrac(...)	0.66	0.32	1.50
antibiotic resistance	ZP_01072284|ZP_01072(...)	tetracycline resista(...)	0.25	0.40	1.82
antibiotic resistance	AAT12289|AAT12289	tetracycline resista(...)	0.41	0.44	1.60
antibiotic resistance	AAT27386|AAT27386	tetracycline resista(...)	0.57	0.36	1.47
antibiotic resistance	CAD20561|CAD20561	TetW protein [Mitsuo(...)	0.55	0.24	1.58
antibiotic resistance	AAY62597|AAY62597	TetW [Bifidobacteriu(...)	0.60	0.36	1.39
antibiotic resistance	TET.|O52836	RecName: Full=Tetrac(...)	0.44	0.54	1.37
antibiotic resistance	AAV80411|AAV80411	TetO [Enterococcus f(...)	0.41	0.32	1.60
antibiotic resistance	CAD13485|CAD13485	TetW protein [Rosebu(...)	0.52	0.28	1.43
antibiotic resistance	CAD20560|CAD20560	TetW protein [Butyri(...)	0.63	0.12	1.26
antibiotic resistance	EDP11831|EDP11831	hypothetical protein(...)	0.38	0.66	0.90
antibiotic resistance	TET.|P72533	RecName: Full=Tetrac(...)	0.49	0.16	1.28
antibiotic resistance	CAD20560|CAD20560	TetW protein [Butyri(...)	0.36	0.32	1.22
antibiotic resistance	EDO59803|EDO59803	hypothetical protein(...)	0.30	0.62	0.96
antibiotic resistance	ZP_02042851|ZP_02042(...)	hypothetical protein(...)	0.36	0.62	0.86

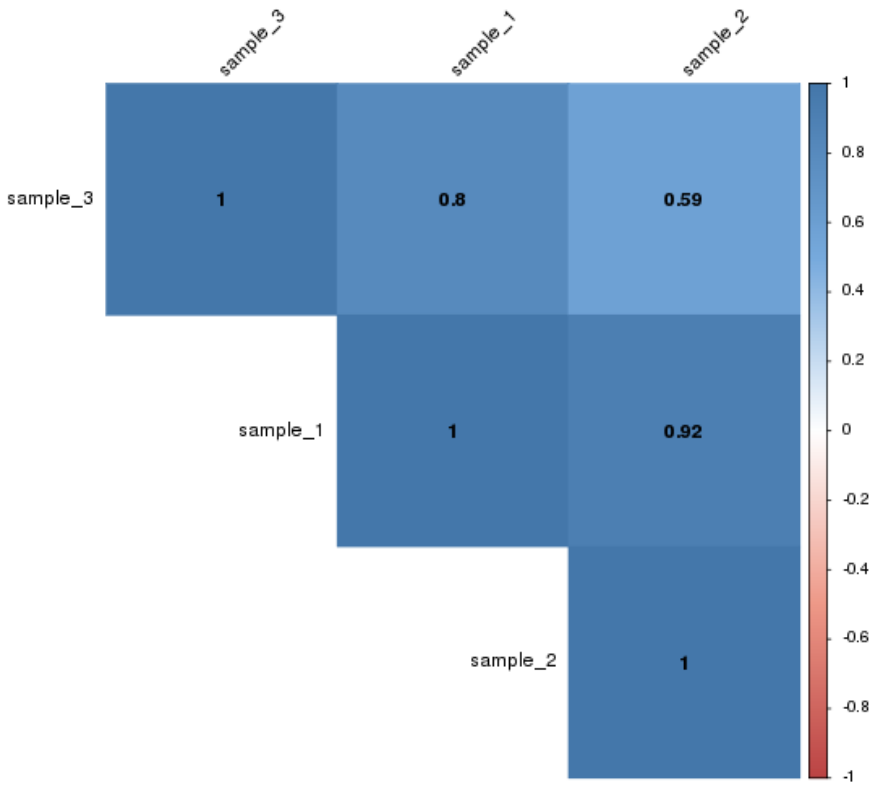


Figure 11: Correlation plot showing the relationship between the samples based on the identified functional profiles of the respective samples. Values close to +1 indicate a high degree of positive correlation between the sample pair in comparison whereas values close to -1 indicate a high degree of 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 16: 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
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
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
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 17: 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[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.
HTML	Standard markup language for creating web pages and web applications
BAM[9]	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] 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] 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.

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

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

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

[7] 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.

[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] 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.

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

[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] 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.

[13] 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.

[14] 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.

[15] 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 18: 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 19: Name, version and description of relevant programs.

Program	Version	Description
Bowtie[3]	2.2.9	Bowtie is a ultrafast, memory-efficient short read aligner. It is based on Burrows-Wheeler transform algorithm.
Kraken[4]	0.10.6	Kraken is an ultrafast and highly accurate program for assigning taxonomic labels to metagenomic DNA sequences.
Krona[6]	2.5	Krona allows hierarchical data to be explored with zoomable pie charts.
Picard[10]	1.131	Picard is a java-based command-line utilities for processing SAM / BAM files.
R[11]	2.15.3	R is a programming language and environment for statistical computing.
SAMTools[12]	0.1.18	SAMtools provide various utilities for manipulating alignments in the SAM format.
Trimmomatic[13]	0.33	Trimmomatic performs a variety of useful trimming tasks for Illumina paired-end and single-end data.
bamtools[14]	2.3.0	BamTools provides a small, but powerful suite of command-line utility programs for manipulating and querying BAM files for data.
sambamba[15]	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 20: Filters used in postprocessing of taxonomic profiling results.

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

Table 21: 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 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