CeGaT

Research & Pharma Solutions

Shotgun Metagenomic Sequencing

Bioinformatic Note

Shotgun Metagenomic Sequencing

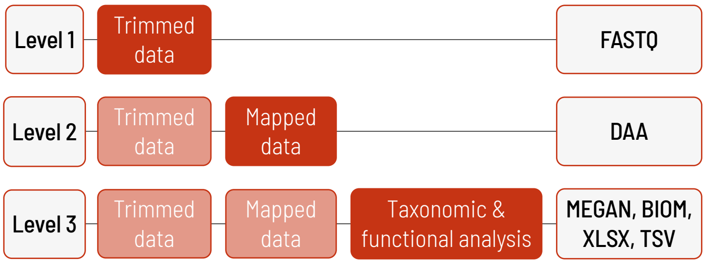
Recent research suggests a close relationship between the composition of the human microbiome and the occurrence of a variety of diseases, including the response towards pharmaceutical drugs. According to recent findings, this relationship has been particularly shown in the gut microbiome, but it can also be extended to other body sites, such as the skin, mouth, or nose. Furthermore, microorganisms are involved in many biochemical reactions in the environment and represent important constituents of environmental ecosystems. Understanding microbial functions in specific microbiomes or host-microbe relationships offers great potential for new therapeutic discoveries, especially for microbes, which are difficult to cultivate.

Shotgun metagenomic sequencing analyzes the complete DNA content of a sample and allows accurate detection of microbes (bacteria, archaea, fungi, protozoa, viruses, etc.) down to species level. Accordingly, functional genes encoding specific metabolic enzymes can be analyzed. Shotgun metagenomic sequencing is the best choice when microbiomes need to be thoroughly characterized, including accurate identification of microbial species and their functional repertoire.

Applications of shotgun metagenomic sequencing are diverse and include:

- Disease monitoring
- Microbial biomarker detection
- Drug development
- Characterization of environmental microbiomes
- Discovery of new microbial species (de novo assembly)

Different levels of bioinformatic data analysis are available:



With increasing bioinformatic levels, more data are delivered. All higher levels include the data from the lower levels, e.g., in Level 2, trimmed data and mapped data are provided. In addition to the data, and independent of the analysis level, a project report is generated.

Level 1

If you wish to analyze your data, we recommend Level 1 or 2. The default level for raw data is Level 1, where trimmed reads in FASTQ format are delivered. In this level, the sequencing data are demultiplexed and trimmed. This level is provided for every project, regardless of additional purchased bioinformatic analyses.

The additionally generated project report provides information for every sample about the laboratory protocol, including data about quality control of the starting material, library preparation, sequencing parameters, and the Q30 value of the sequencing. For the trimmed data, the number of sequenced fragments and bases is reported, and the sequence length, quality of the reads, and the GC content are illustrated in bar plots for all samples.

Level 2

In Level 2, the trimmed reads are mapped. The aligned reads are provided in DAA format. The DAA format is a binary file format. Thus, inspecting the files with a text editor is not possible. However, the files can be opened with MEGAN6 to inspect the alignment, adjust parameters, or do further analyses.

Level 3

In Level 3, the reads are assigned to the taxonomy and the KEGG database. Only taxa with relative sequence abundances above 0.01% are considered. Functional classification is carried out. The relative sequence abundances for taxonomic units and functions as well as diversity indices and Bray-Curtis dissimilarities are calculated.

In the project report, the number of reads that could be assigned to the taxonomy and the KEGG database are indicated in a table. An additional table indicates the percentages of reads that are assigned to bacteria, archaea, eukaryotes, and viruses. A small percentage of reads cannot be assigned to any taxonomic unit in our database. These reads are ignored in the downstream analysis.

Reads are assigned to units on all taxonomic levels, and the counts and relative sequence abundances can be found in respective Excel files. Each file contains several spreadsheets that indicate the distribution of reads on different taxonomic (species, genus, family, order, class, phylum). An excerpt of the species counts file of the taxonomy is given in table 1. Every project receives a unique S-number (SXXXX). In this example, the S-number is S1163. For every sample of the project, a column is included in the Excel file. The taxonomic identifier and the taxonomic levels are indicated in the columns, and the counts for every sample are given in the respective sample column. The file with the relative abundances is similar to the counts file, except for abundance values in the respective sample columns instead of count values.

Table 1 | Excerpt of the counts XLSX file.

taxid	level	super-kingdom	phylum	class	order	family	genus	species	S1163Nr200	S1163Nr201	S1163Nr202
2173	species	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	Methanobrevibacter	Methanobrevibacter smithii	1,516	6,699	7,056
446660	species	Bacteria	Actinobacteria	Coriobacteriia	Eggerthellales	Eggerthellaceae	Adlercreutzia	Adlercreutzia equolifaciens	461	285	1,228
394340	species	Bacteria	Actinobacteria	Coriobacteriia	Eggerthellales	Eggerthellaceae	Asaccharobacter	Asaccharobacter celatus	6,871	5,041	20,037

Similar to the counts and abundance file for all taxonomic levels, counts and abundance files for the functional levels based on KEGG are provided. Again, each file contains several spreadsheets that indicate the distribution of reads over the different KEGG levels (1-4). In table 2, an excerpt of the level 1 KEGG counts is shown. The relative abundance file for the functional stacked barplots of the relative abundances are provided for every taxonomic level (phylum, class, order, family, genus, and species) in PNG format (see figure 1).

Table 2 | Excerpt of the KEGG counts XLSX file.

Function	S1163Nr200	S1163Nr201	S1163Nr202
Metabolism	569,419	566,655	592,755
Genetic Information Processing	194,768	198,840	213,862
Environmental Information Processing	120,664	124,759	127,979
Cellular Processes	85,021	86,804	88,193
Organismal Systems	37,999	38,049	38,556
Human Diseases	66,502	65,208	72,186
Brite Hierarchies	795,478	811,907	846,364
Not Included in Pathway or Brite	163,226	165,293	166,325

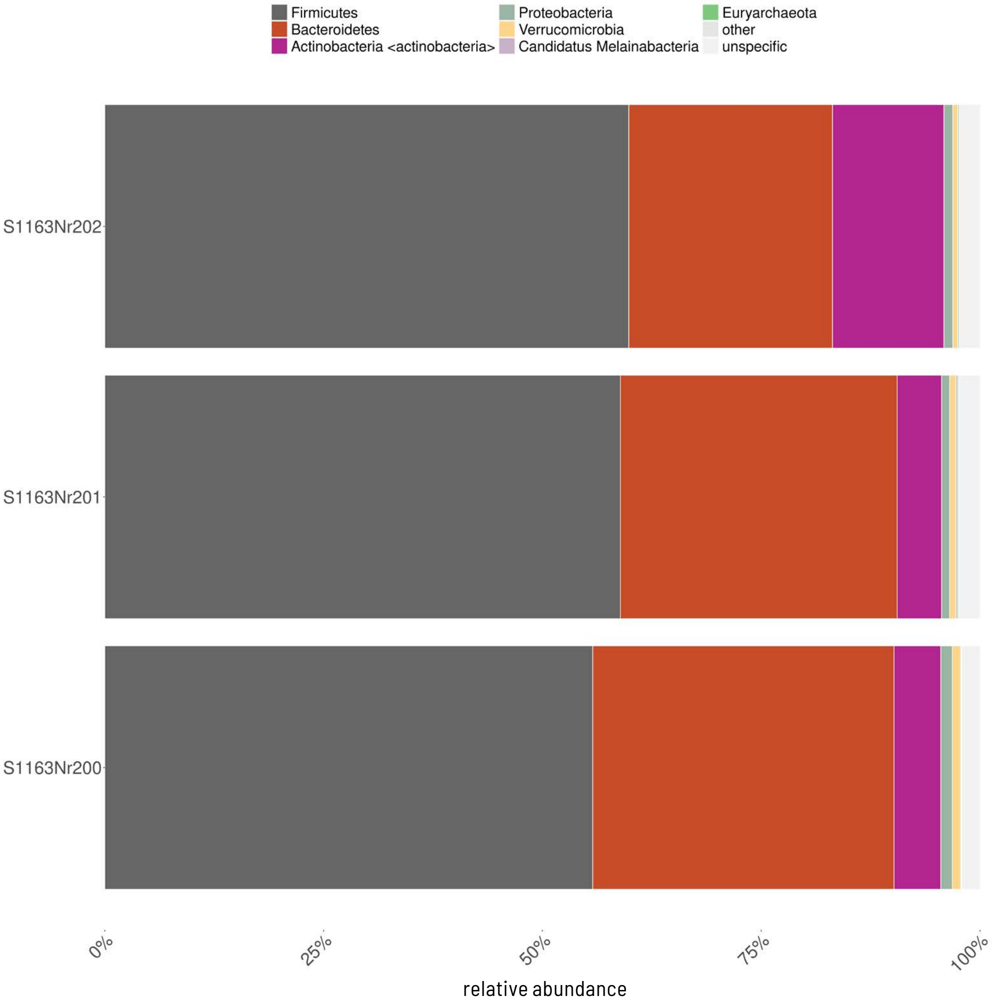


Figure 1 | Stacked barplot of relative abundances of the phyla present in each sample. Relative abundances were derived for all phyla present in any sample and the samples sorted by Bray-Curtis similarity. Reads marked as ambiguous cannot be assigned to a single phylum.

Data comparison files are provided in MEGAN and BIOM format. The MEGAN file can be opened with MEGAN6 for interactive taxonomic and functional exploration of the dataset, generation of graphs, data export, and further analyses. The BIOM format was developed for easy transfer of microbiome data between different tools and pipelines. Therefore, the BIOM file can, for example, be imported into MetaPhlAn. We provide one BIOM file on the phylum level and one on the species level. Additional files on other taxonomic levels can be exported from the MEGAN file using MEGAN6.

Diversity measures for all taxonomic levels are provided in TSV format. These diversity measures include the Shannon-Wiener Index (indicated as Shannon index in table 3), the Evenness, and the Richness. The Richness is the total number of species present in each sample. The Shannon-Wiener Index is a measure of the Richness and the relative sequence abundance (see figure 2).

To evaluate the similarity between the samples, Bray-Curtis distances were calculated using the relative sequence abundances of the detected species (>0.01%) and a principal coordinates analysis (PCoA) conducted to assign each sample a location in the 2-dimensional space (see figure 3). Using this analysis, groups of very similar samples can be identified as well as outliers.

The index increases with the number of species but is highest if all species occur with the same relative sequence abundance (e.g., all species occur exactly once). This also means that there is no maximum value for the Shannon-Wiener Index as it is only limited by the number of species in the reference. In contrast, the Evenness of a sample is between 0 and 1. If all species have the same abundance, the Evenness is close to 1, but is low if there are big differences between the abundances of the species in a sample. As an example, the diversity measures on species level are indicated in table 3.

Table 3 | Diversity measures on species level.

Sample	Shannon Index	Evenness	Richness
S1163Nr200	4.06	0.73	262
S1163Nr201	4.06	0.73	265
S1163Nr202	3.96	0.72	252

The results of the diversity measures are additionally visualized in the project report: The Shannon-Wiener Index is plotted against the Richness and the Evenness is color-coded. The visualization is shown in figure 2.

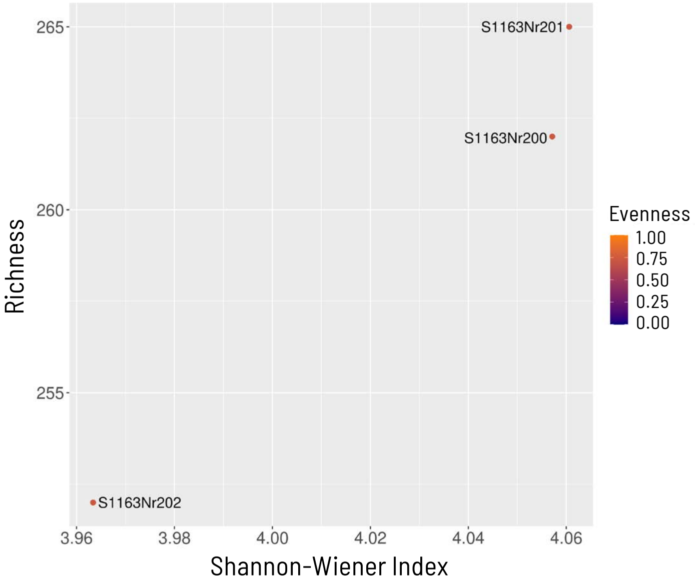


Figure 2 | Diversity of all samples. The Shannon-Wiener Index (based on the natural logarithm) and the Richness were calculated for each sample on the species level and plotted against each other. Higher indices indicate that more species were present in a sample and that their relative abundance is more even. The color of the dots represents the species' Evenness in each sample.

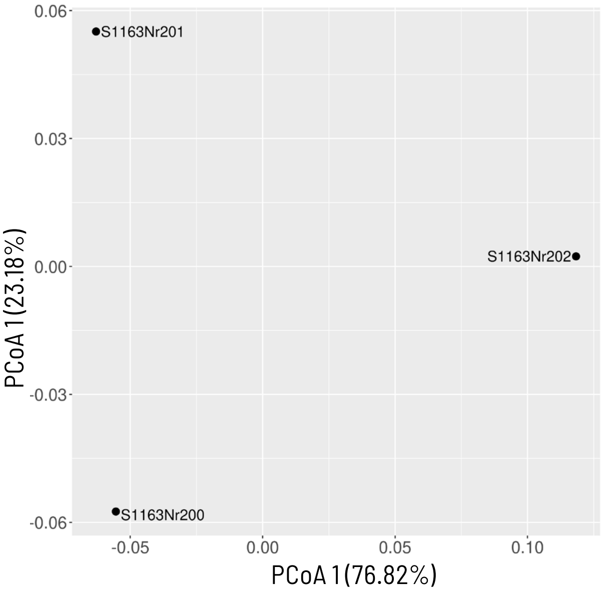


Figure 3 | PCoA of all samples. Plotted are the two main axes that explain the most variation in the data. If the relative sequence abundances in two samples are very similar, they are plotted very close to each other.

In addition to the diversity measures, a functional analysis is performed. This functional analysis of the samples reveals the pathways and genes present in the detected species. However, this annotation is only based on the presence and absence of the genes in the DNA of the detected species and does not imply that these genes are actually expressed. We include counts for the top four levels in the KEGG database, and the plot shows the top five functions of all four levels. Functions on KEGG Level 1 are very broad, while Level 4 is very specific.

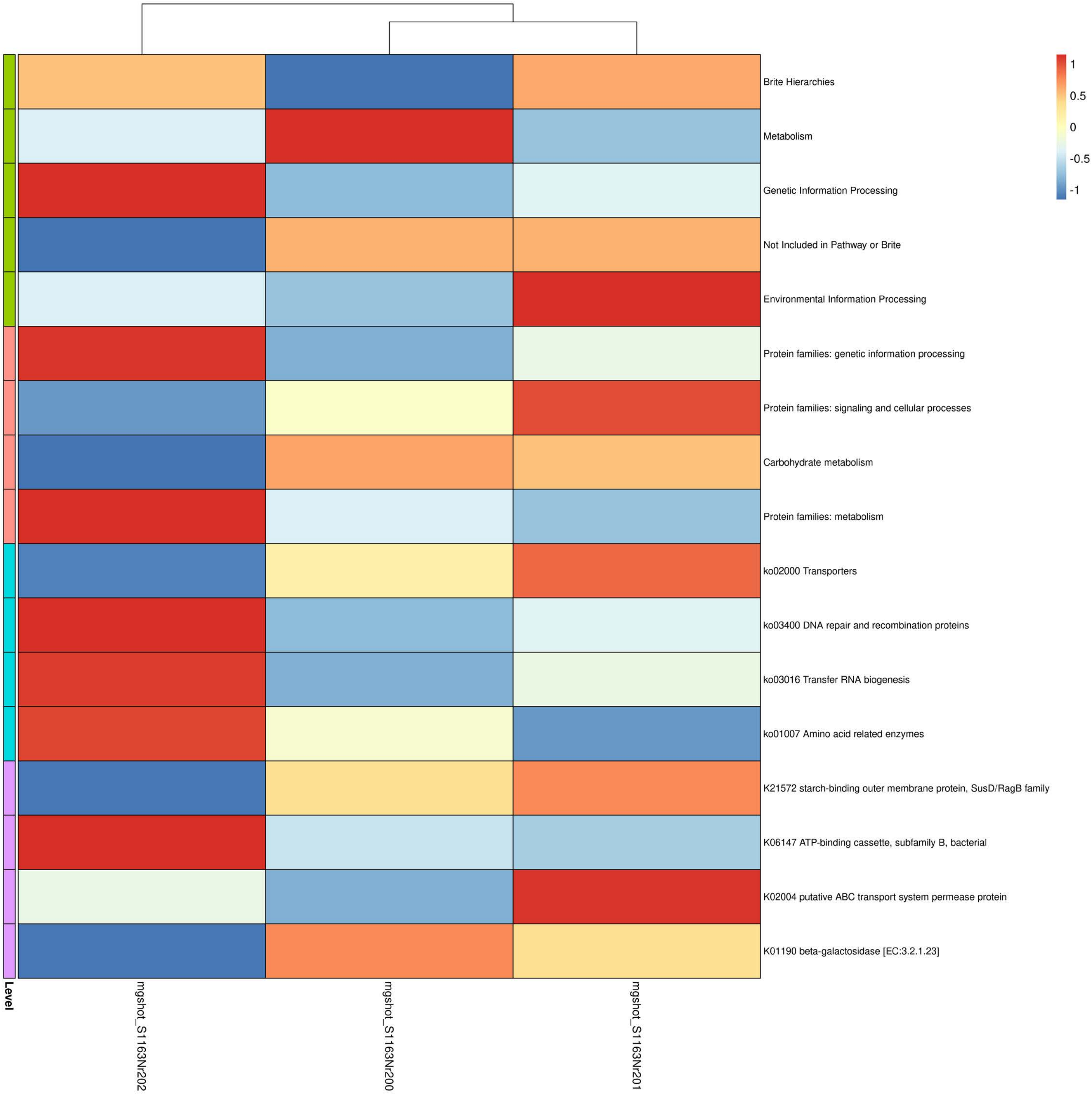


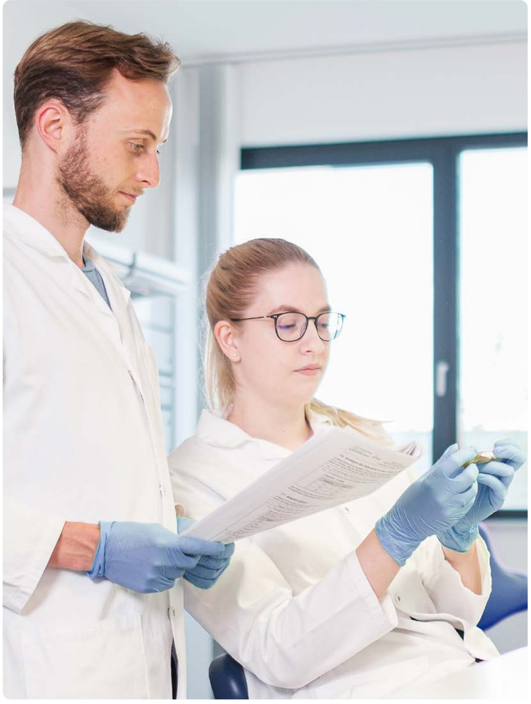
Figure 4 | The five most abundant KEGG functions at the first four levels (color coded on the left) of the KEGG database. Sequence abundances were row scaled, positive values (red colors) depict high abundances while negative values (blue colors) depict low abundances. Samples were sorted using the complete clustering of the distances derived from the displayed functions.

About Us

CeGaT was founded in 2009 in Tübingen, Germany. Our scientists are specialized in next-generation sequencing (NGS) for genetic diagnostics, and we also provide a variety of sequencing services for research purposes and pharma solutions. Our sequencing service portfolio is complemented by analyses suited for microbiome, immunology, and translational oncology studies.

Our dedicated project management team of scientists and bio-informaticians works closely with you to develop the best strategy to realize your project. Depending on its scope, we select the most suitable library preparation and conditions on our sequencing platforms.

We would be pleased to provide you with our excellent service.
Contact us today to start planning your next project.



For more details please visit
www.cegat.com/rps

CeGaT GmbH
Research & Pharma Solutions
Paul-Ehrlich-Str. 23
72076 Tübingen
Germany

Phone: +49 7071 56544-333
Fax: +49 7071 56544-56
Email: rps@cegat.com

CAP ACCREDITED
COLLEGE of AMERICAN PATHOLOGISTS
CLIA CERTIFIED ID: 99D2130225

DAkkS
Deutsche Akkreditierungsstelle
D-PL-13206-01-00
Accredited by DAkkS according to
DIN EN ISO/IEC 17025:2018

2023/12