Research &
Pharma Solutions

Full-Length 16S Sequencing

Bioinformatic Note

Full-Length 16 S Sequencing

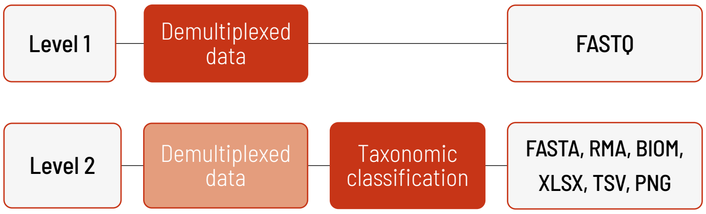
The 16S ribosomal RNA (rRNA) gene is approximately 1.5 kb long and contains several conserved and hypervariable regions (V1-V9) that vary between different bacteria. These hypervariable sequences can be used to identify and characterize microbial diversity. Therefore, the 16S rRNA gene is a common marker to characterize microbial communities in various specimens.

Using the PacBio single molecule real-time (SMRT) technology, we accurately sequence the full-length 16S rRNA gene, covering all variable regions with an average HiFi read length of about 1.5 kb. The extraordinary accuracy and length of PacBio HiFi reads generated using circular consensus sequencing (CCS) mode allows microbial taxa detection at high resolution.

Applications of full-length 16S sequencing are diverse and include:

- Characterization of different microbial communities
- Microbial biomarker detection
- Disease monitoring
- Drug development

Different levels of bioinformatic data analysis are available:



With increasing bioinformatics level, more data are delivered. The higher level includes the data from the lower levels. Thus, in Level 2, the taxonomic classification as well as the demultiplexed 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 yourself, we recommend Level 1. At this level, demultiplexing of the sequencing reads is performed. HiFi reads (CCS reads with a predicted accuracy ≥ Q20) are extracted and converted to FASTQ-format. The quality of the FASTQ files is analyzed. 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, and sequencing parameters. The number of reads, bases, and the mean HiFi read length are reported for the demultiplexed data. The read length, the average base quality per read, and the GC content per read are additionally illustrated in bar plots.

Level 2

In Level 2, the primers are trimmed. Reads are filtered for reads containing Ns and low-quality bases before quality trimming, denoising, error correction, and removal of chimeras. The resulting FASTA files are then mapped against a custom database. The resulting files in RMA format can be opened with MEGAN6 for alignment inspection, adjustment parameters, and further analyses. Thus, a FASTA and an RMA file are provided for every sample.

Taxonomic placement is performed. Only taxa with relative sequence abundances above 0.001% are considered. Reads are assigned to units on the following taxonomic levels:

Species → Genus → Family → Order → Class → Phylum

Relative sequence abundances for taxonomic units, diversity indices and Bray-Curtis dissimilarities, are calculated.

In the project report, a list indicates the number of HiFi reads, trimmed reads, filtered reads, denoised reads, nonchimeric reads, and assigned reads. Additionally, the percentage of the assigned reads is listed.

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. In addition to the files in Excel format, the counts and relative abundances are also provided in a TSV file. As TSV files do not have spreadsheets, the information is stored in one large table. The additional column "tax" (taxonomy) indicates which taxonomic level the respective row provides.

Stacked bar plots of the relative abundances are provided for every taxonomic level (phylum, class, order, family, genus, and species) in PNG format. One example is given in figure 1.

Table 1 | Excerpt of the counts XLSX file.

taxid	level	superkingdom	phylum	class	order	family	genus	species	S1163Nr221	S1163Nr225	S1163Nr213
1680	species	Bacteria	Actinobacteria	Actinomycetia	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	Bifidobacterium adolescentis	40	60	43
1683	species	Bacteria	Actinobacteria	Actinomycetia	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	Bifidobacterium angulatum	7	5	6

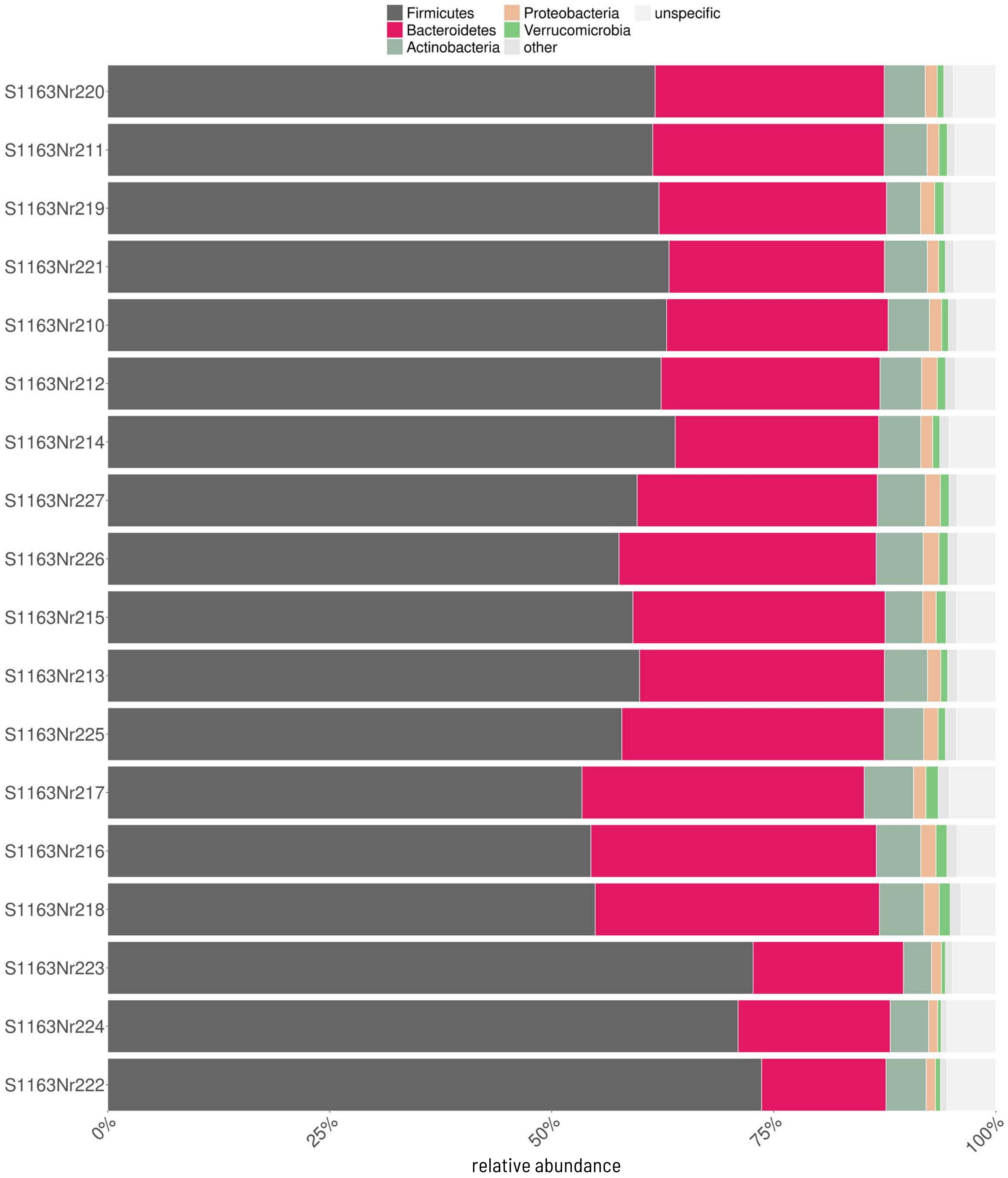


Figure 1 | Stacked bar plot of relative abundances of the phyla present in each sample. Relative abundances were derived for all phyla present in any sample, and the samples were sorted according to their Bray-Curtis similarity. Reads marked as unspecific cannot be assigned to a single phylum.

Table 2 | Diversity measures on species level.

sample	Shannon Index	Evenness	Richness
S1163Nr210	3.76	0.80	111
S1163Nr211	3.72	0.78	115
S1163Nr212	3.76	0.80	110
S1163Nr213	3.69	0.79	109
S1163Nr214	3.86	0.82	112
S1163Nr215	3.70	0.79	108
S1163Nr216	3.67	0.77	117
S1163Nr217	3.60	0.77	104
S1163Nr218	3.65	0.78	106
S1163Nr219	3.72	0.79	112
S1163Nr220	3.73	0.79	111
S1163Nr221	3.74	0.78	117
S1163Nr222	3.79	0.80	117
S1163Nr223	3.75	0.79	113
S1163Nr224	3.79	0.79	119
S1163Nr225	3.66	0.78	111
S1163Nr226	3.70	0.78	111
S1163Nr227	3.73	0.79	114

Data comparison files are provided in MEGAN and BIOM format. The MEGAN file can be opened with MEGAN6 for interactive taxonomic 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 the Shannon index in table 2), Evenness, and 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. 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 significant differences between the abundances of the species in a sample. For example, the diversity measures on the species level are indicated in table 2.

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.

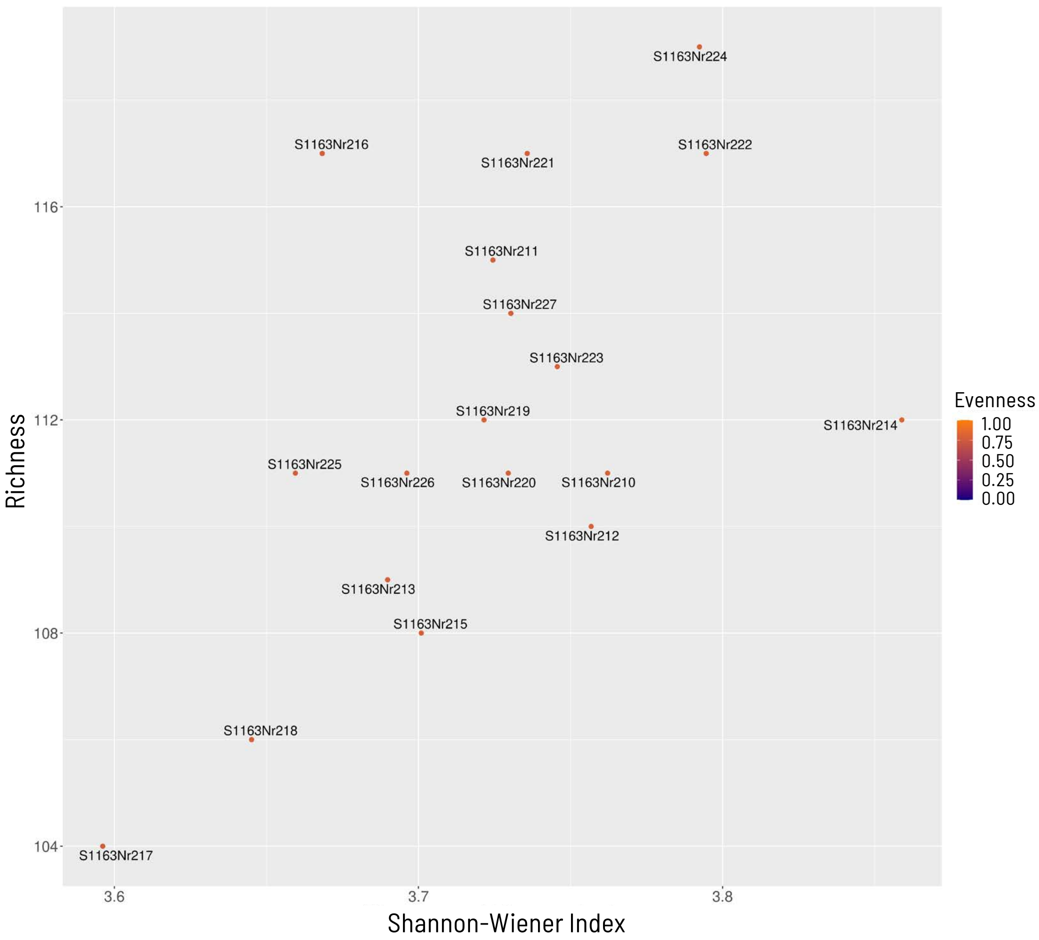


Figure 2 | Diversity of all samples. The Shannon-Wiener Index (based on the natural logarithm) and the Richness were calculated for each sample on the species level and plotted against each other. Higher indices indicate that more species were present in a sample and that their relative abundance is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

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. Using this analysis, groups of very similar samples can be identified as well as outliers (see figure 3).

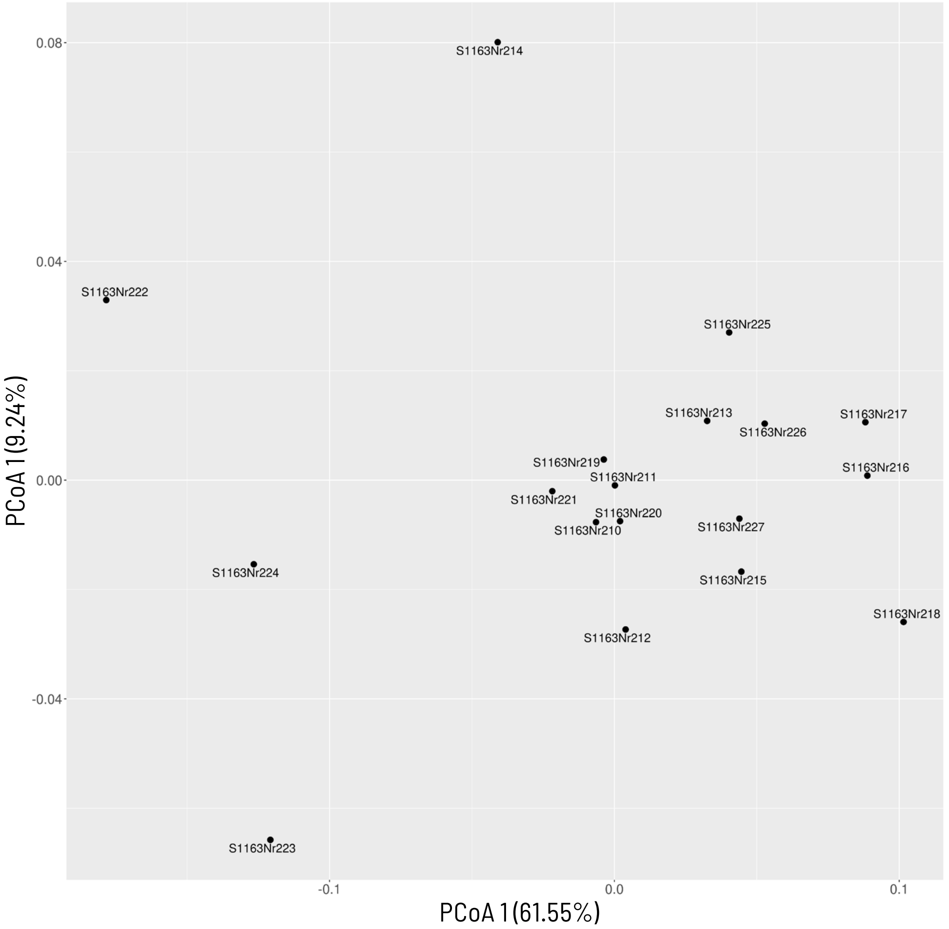


Figure 3 | PCoA of all samples. Plotted are the two main axes that explain the most variation in the data. If the relative sequence abundances in two samples are very similar, they are plotted very close to each other.

To further compare the samples, the 15 most abundant species across all samples were extracted and the samples clustered based on these abundances. Reads mapping to other than the 15 most abundant species are combined and labelled 'other'. Reads that map to more than one species are labeled as 'unspecific' (see figure 4).

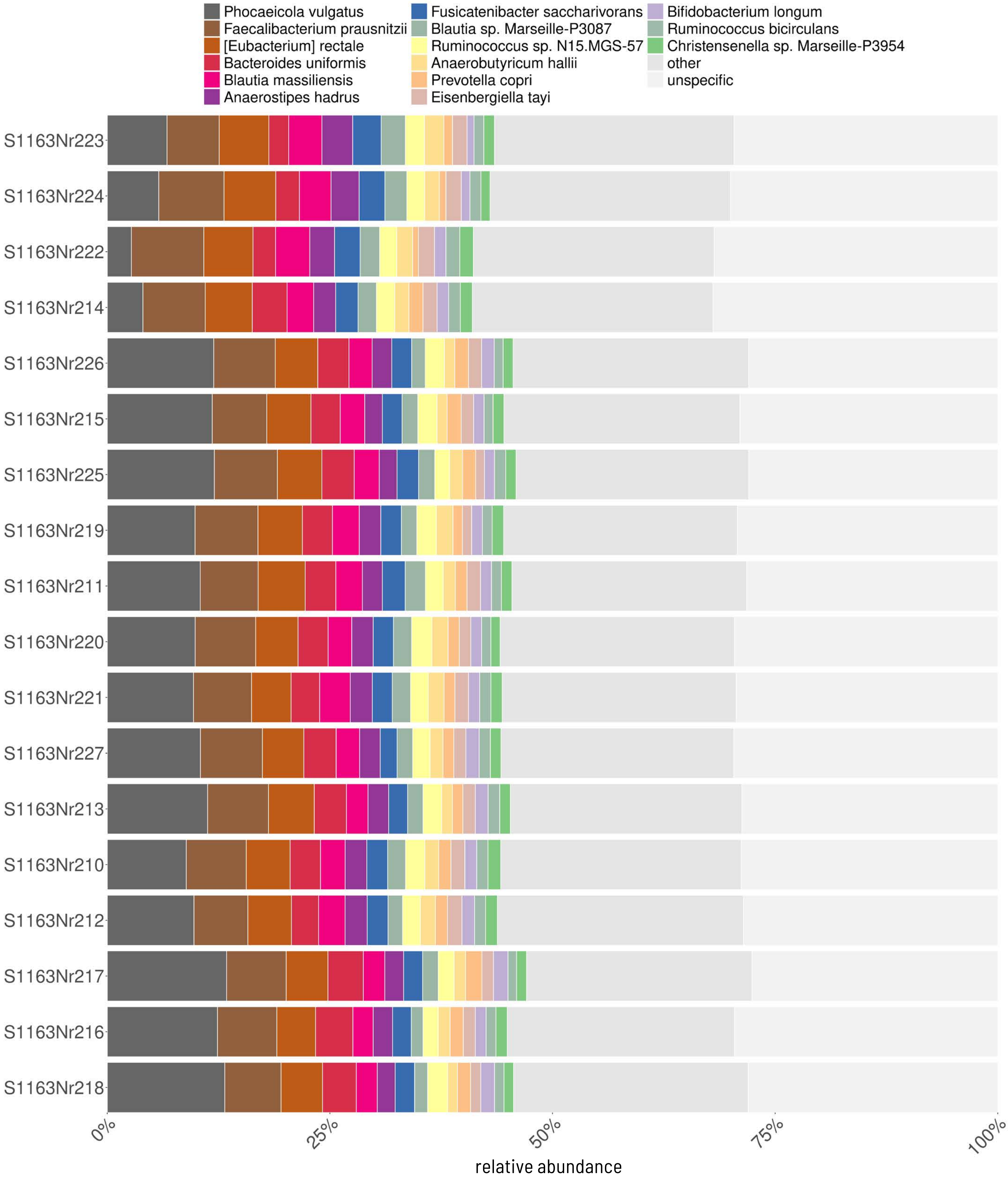


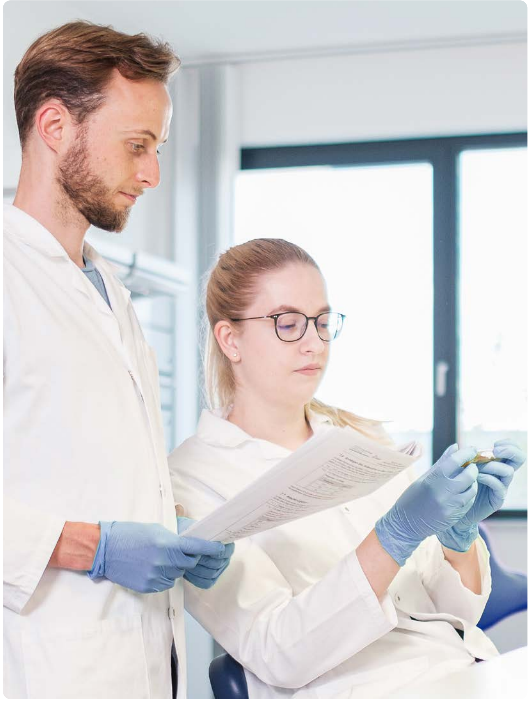
Figure 4 | Stacked barplot of the relative sequence abundances (based on the number of assigned reads) of microbial species with a relative sequence abundance >0.01% in at least one sample. Colors representing the 15 most abundant species are depicted at the top. Counts of all other species were summed up and are shown as 'other'. Reads that map to more than one species are shown as 'unspecific'.

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