Research &
Pharma Solutions

Towards a Highly Accurate Microbiome Analysis

Tech Note

Towards a Highly Accurate Microbiome Analysis

Recent scientific publications impressively demonstrate the urgent need for standardized microbiome analysis workflows that accurately detect all microbes in a sample and thus significantly improve comparability between studies. Although the two most commonly applied analysis techniques, 16S rRNA gene analysis and shotgun metagenomics, have been optimized and applied for years, so far, a scientific "best practice" consensus has not been found. Accordingly, data quality, analysis accuracy and repeatability strongly depend on the individual wet lab protocol and analysis workflow chosen by the scientist.

Metagenomic Profiling (shotgun metagenomics)

This TechNote addresses key quality parameters of three commonly used sequencing library protocols for shotgun metagenomics. Protocol A: Illumina Nextera DNA Flex (input: 100 ng), Protocol B: Illumina Nextera XT DNA (input: 1 ng), Protocol C: Down-scaled Illumina Nextera XT DNA (input: 0.1 ng). In a recent publication, Hillmann et al. (2018) recommended the down-scaled Nextera XT DNA approach as a cost-efficient protocol for shotgun metagenomics.

Sample and analysis details

To test how well the protocols can detect the theoretical composition of a microbial community standard, we performed three library preparations for each protocol using the same standard. We used a commercially available microbial community standard that consists of DNA from ten microbial species (five gram-positive bacteria, three gram-negative bacteria, two fungi) at a known composition. The standard was designed to explicitly cover a wide variety of genome sizes and GC contents and thus mimics the conditions in more complex communities. Also, we investigated the repeatability and detectable diversity of all protocols using a complex microbial DNA pool, generated from a fecal sample containing stool from ten individuals.

All sequencing libraries were sequenced on the Illumina NovaSeq with a read length of 2x100 bp and a sequencing depth of 10 million clusters. Raw data of the standard samples were analyzed by mapping against a database containing only the ten microbial species theoretically available in the standard. For the analysis of the complex microbial DNA pool from stool samples, we used CeGaT's standard bioinformatics pipeline. This pipeline is based on mapping all reads against an extensive database with thousands of species (NCBI RefSeq).

Results

As shown in figure 1, all protocols were able to capture the microbial species present in the standard. However, regarding the correct representation of the community composition (i.e. the theoretical relative sequence abundance of the detected microbial species) the protocols revealed differences. The theoretical composition was most accurately determined with protocol A (figure 1 and table 1) as it showed the lowest Bray-Curtis dissimilarity values and highest correlation coefficients (Pearson-R and Lin's Concordance Correlation Coefficient).

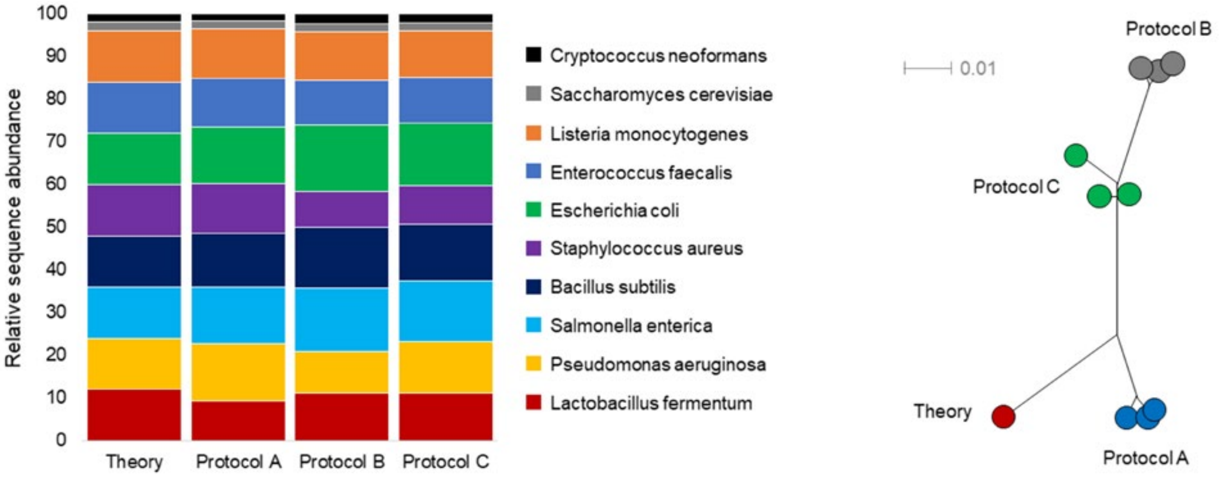


Figure 1: *Comparison of the results from three different protocols with the theoretical composition in the microbial community standard (Theory). Left: Relative sequence abundances in Theory and Protocols A, B, C. Bars of the protocols A, B, C represent means of three library preparations (n=3). Right: Neighbor-joining tree showing the similarity (Bray-Curtis) between Protocols A, B, C, and the theoretical values (Theory).*

	Pearson-R	Lin's CCC
Protocol A	0.960 ± 0.003	0.957 ± 0.006
Protocol B	0.876 ± 0.006	0.870 ± 0.010
Protocol C	0.929 ± 0.010	0.927 ± 0.015

Table 1: *Correlation coefficients (Pearson-R and Lin's CCC) originating from correlating the theoretical relative sequence abundances with the relative sequence abundances determined in the protocols A, B, C. Values represent means ± standard deviations of three library preparations (n=3).*

To determine the repeatability and detectable diversity of the different protocols, we performed three library preparations for each protocol using a complex microbial DNA pool from stool samples. As shown in figure 2, all protocols were able to detect the most common gut microbiome representatives (various Firmicutes and Bacteroidetes members) and covered a wide range of different microbes.

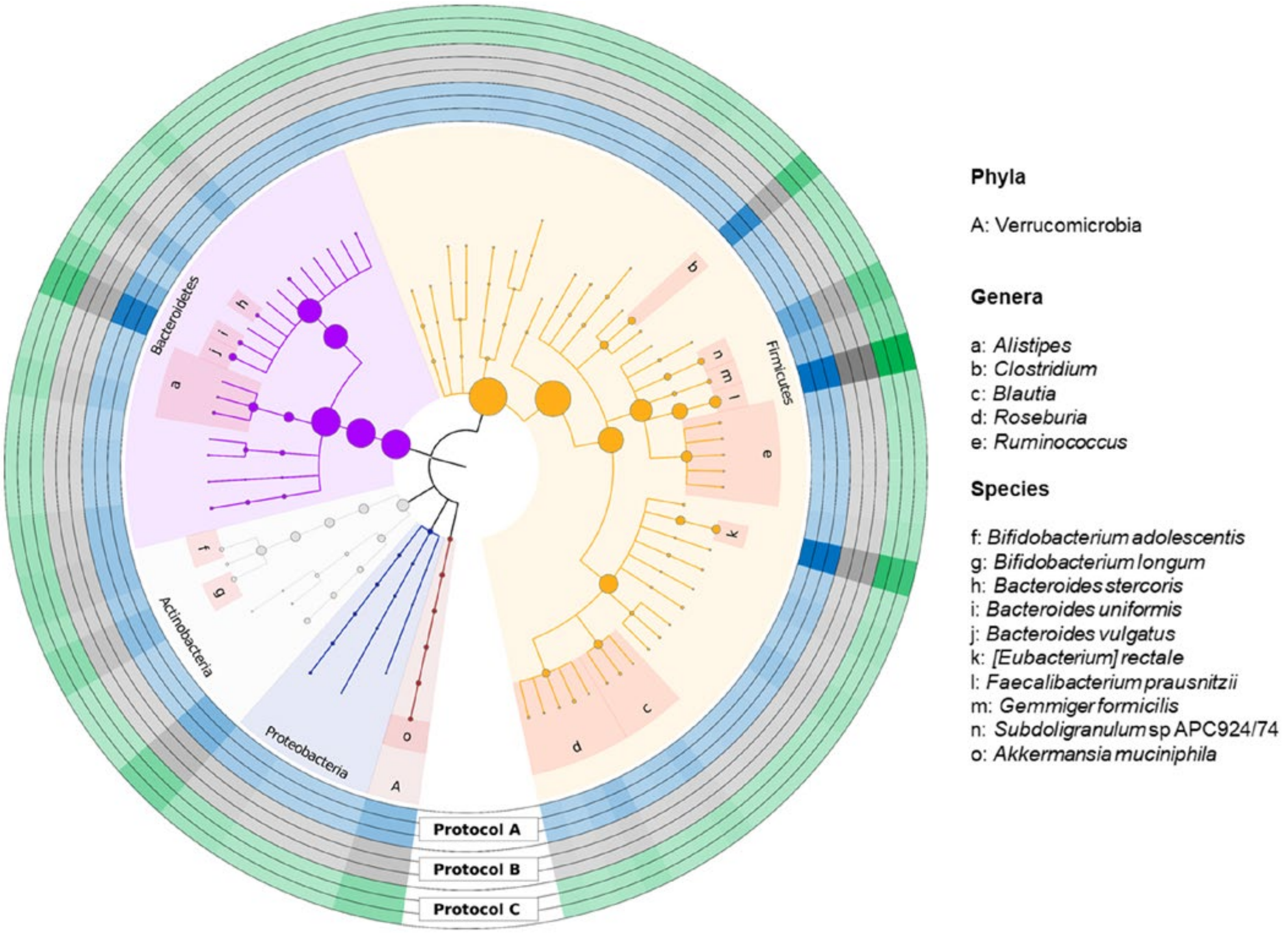
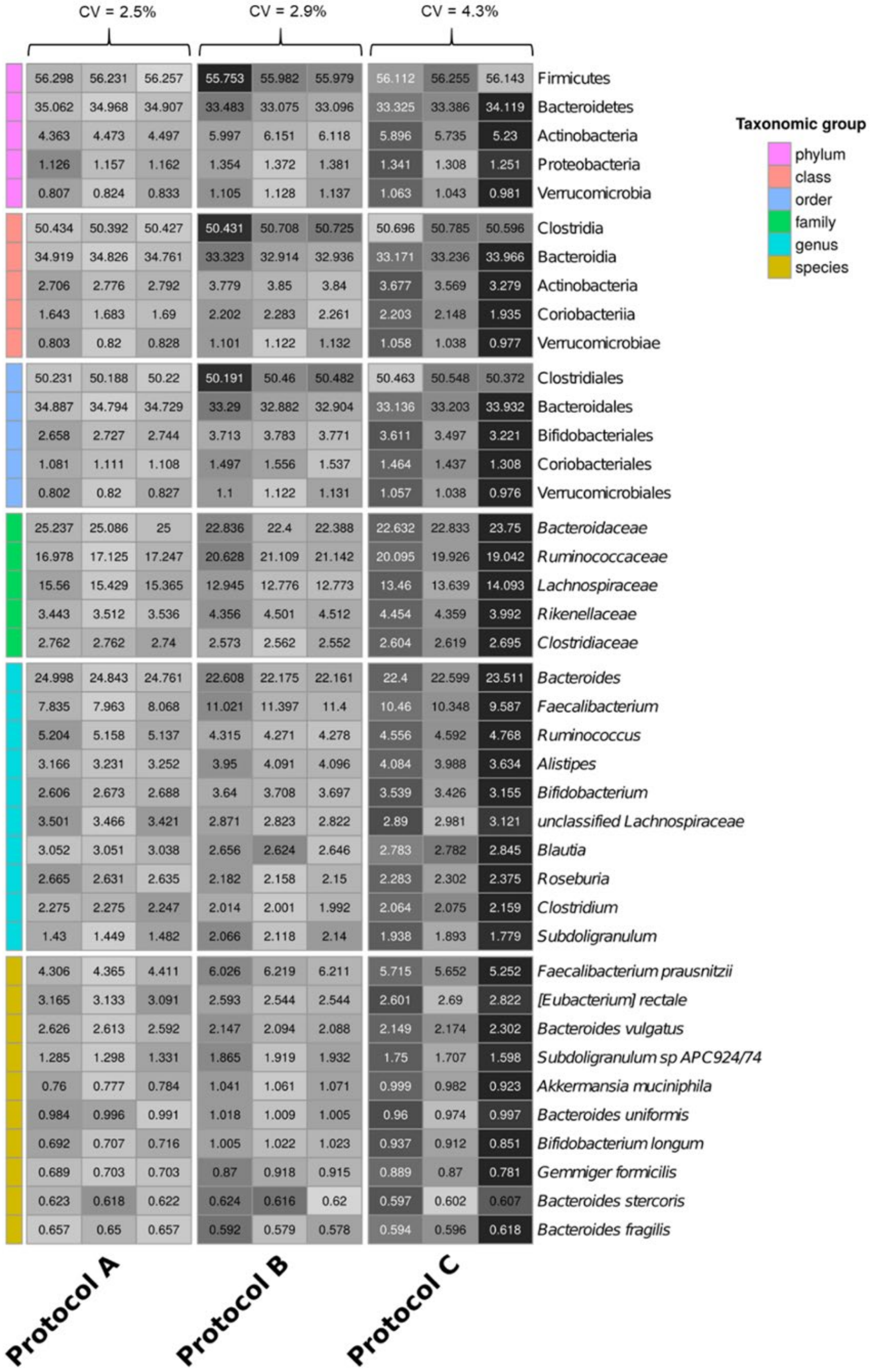


Figure 2: *Comparison of three protocols using a complex microbial DNA pool from stool samples. This graph shows the taxonomic composition of all detected microbial taxa with a relative sequence abundance >0.1%. Each node represents a taxonomic unit (e.g. species at the tips), and the size of the circle indicates the mean relative sequence abundance of the respective taxonomic unit across all samples and protocols. Circles highlighted with small letters represent the most abundant genera and species (annotated on the right). Outer rings represent relative sequence abundances for each sample separately. Darker colors indicate higher relative sequence abundance.*

To estimate how well the three protocols were able to detect the diversity within a sample (alpha diversity), we calculated the species richness and evenness as well as the Shannon Diversity Index. All of these parameters were slightly higher with protocol A (table 2). Furthermore, Protocol A also showed slightly higher repeatability. In comparison with protocol B and C, protocol A had the lowest variation between replicates among the most abundant taxonomic units and also showed the lowest coefficient of variation (CV), when considering all microbial species with a relative sequence abundance >0.01% (figure 3).

	Richness	Evenness	Shannon Diversity Index
Protocol A	267 ± 1	0.804 ± 0.001	4.49 ± 0.00
Protocol B	258 ± 1	0.781 ± 0.001	4.34 ± 0.01
Protocol C	259 ± 3	0.791 ± 0.003	4.40 ± 0.02

Table 2: *Alpha diversity values determined from datasets generated with the three different protocols (Protocol A, B, C). Values represent means ± standard deviations of three library preparations (n=3).*



Protocol A			Protocol B			Protocol C			
CV = 2.5%			CV = 2.9%			CV = 4.3%			
56.298	56.231	56.257	55.753	55.982	55.979	56.112	56.255	56.143	Firmicutes
35.062	34.968	34.907	33.483	33.075	33.096	33.325	33.386	34.119	Bacteroidetes
4.363	4.473	4.497	5.997	6.151	6.118	5.896	5.735	5.23	Actinobacteria
1.126	1.157	1.162	1.354	1.372	1.381	1.341	1.308	1.251	Proteobacteria
0.807	0.824	0.833	1.105	1.128	1.137	1.063	1.043	0.981	Verrucomicrobia
50.434	50.392	50.427	50.431	50.708	50.725	50.696	50.785	50.596	Clostridia
34.919	34.826	34.761	33.323	32.914	32.936	33.171	33.236	33.966	Bacteroidia
2.706	2.776	2.792	3.779	3.85	3.84	3.677	3.569	3.279	Actinobacteria
1.643	1.683	1.69	2.202	2.283	2.261	2.203	2.148	1.935	Coriobacteriia
0.803	0.82	0.828	1.101	1.122	1.132	1.058	1.038	0.977	Verrucomicrobiae
50.231	50.188	50.22	50.191	50.46	50.482	50.463	50.548	50.372	Clostridiales
34.887	34.794	34.729	33.29	32.882	32.904	33.136	33.203	33.932	Bacteroidales
2.658	2.727	2.744	3.713	3.783	3.771	3.611	3.497	3.221	Bifidobacteriales
1.081	1.111	1.108	1.497	1.556	1.537	1.464	1.437	1.308	Coriobacteriales
0.802	0.82	0.827	1.1	1.122	1.131	1.057	1.038	0.976	Verrucomicrobiales
25.237	25.086	25	22.836	22.4	22.388	22.632	22.833	23.75	*Bacteroidaceae*
16.978	17.125	17.247	20.628	21.109	21.142	20.095	19.926	19.042	*Ruminococcaceae*
15.56	15.429	15.365	12.945	12.776	12.773	13.46	13.639	14.093	*Lachnospiraceae*
3.443	3.512	3.536	4.356	4.501	4.512	4.454	4.359	3.992	*Rikenellaceae*
2.762	2.762	2.74	2.573	2.562	2.552	2.604	2.619	2.695	*Clostridiaceae*
24.998	24.843	24.761	22.608	22.175	22.161	22.4	22.599	23.511	*Bacteroides*
7.835	7.963	8.068	11.021	11.397	11.4	10.46	10.348	9.587	*Faecalibacterium*
5.204	5.158	5.137	4.315	4.271	4.278	4.556	4.592	4.768	*Ruminococcus*
3.166	3.231	3.252	3.95	4.091	4.096	4.084	3.988	3.634	*Alistipes*
2.606	2.673	2.688	3.64	3.708	3.697	3.539	3.426	3.155	*Bifidobacterium*
3.501	3.466	3.421	2.871	2.823	2.822	2.89	2.981	3.121	*unclassified Lachnospiraceae*
3.052	3.051	3.038	2.656	2.624	2.646	2.783	2.782	2.845	*Blautia*
2.665	2.631	2.635	2.182	2.158	2.15	2.283	2.302	2.375	*Roseburia*
2.275	2.275	2.247	2.014	2.001	1.992	2.064	2.075	2.159	*Clostridium*
1.43	1.449	1.482	2.066	2.118	2.14	1.938	1.893	1.779	*Subdoligranulum*
4.306	4.365	4.411	6.026	6.219	6.211	5.715	5.652	5.252	*Faecalibacterium prausnitzii*
3.165	3.133	3.091	2.593	2.544	2.544	2.601	2.69	2.822	*[Eubacterium] rectale*
2.626	2.613	2.592	2.147	2.094	2.088	2.149	2.174	2.302	*Bacteroides vulgatus*
1.285	1.298	1.331	1.865	1.919	1.932	1.75	1.707	1.598	*Subdoligranulum sp APC924/74*
0.76	0.777	0.784	1.041	1.061	1.071	0.999	0.982	0.923	*Akkermansia muciniphila*
0.984	0.996	0.991	1.018	1.009	1.005	0.96	0.974	0.997	*Bacteroides uniformis*
0.692	0.707	0.716	1.005	1.022	1.023	0.937	0.912	0.851	*Bifidobacterium longum*
0.689	0.703	0.703	0.87	0.918	0.915	0.889	0.87	0.781	*Gemmiger formicilis*
0.623	0.618	0.622	0.624	0.616	0.62	0.597	0.602	0.607	*Bacteroides stercoris*
0.657	0.65	0.657	0.592	0.579	0.578	0.594	0.596	0.618	*Bacteroides fragilis*

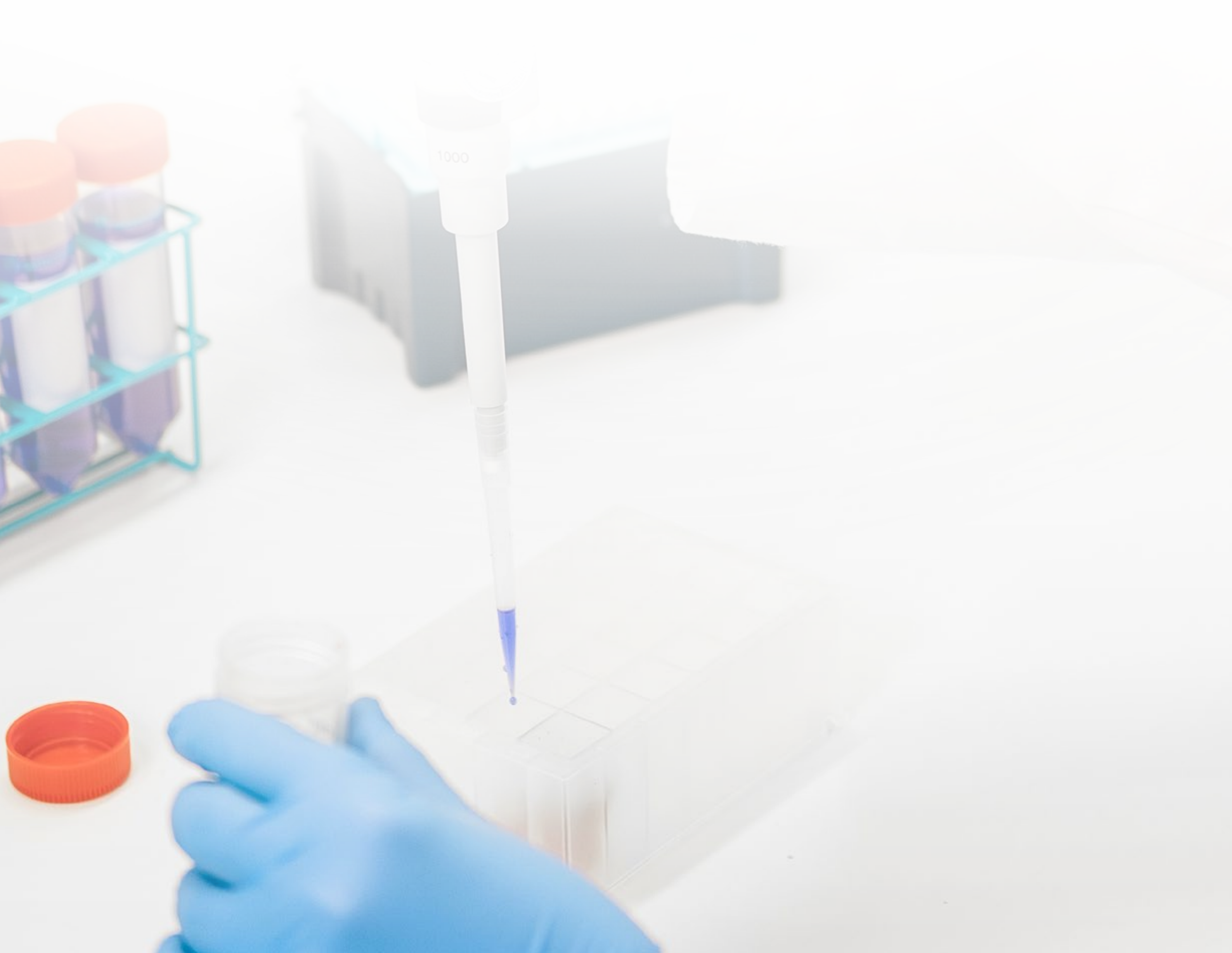
Figure 3: *Comparison of three protocols using a complex microbial DNA pool from stool. This heatmap shows the variation between replicates within each protocol for the most abundant taxonomic units at the phylum, class, order, family, genus and species level. Colors are scaled row-wise to allow an evaluation of each taxonomic unit separately. The darker the color of a field, the higher is the deviation of this value from the mean of all three samples of the respective protocol. Numbers in the fields represent the relative sequence abundance. Coefficients of variation (CV) are mean values of all microbial species with a relative sequence abundance >0.01%.*

Conclusion

Our detailed comparison of three different library preparation protocols for shotgun metagenomics revealed that protocol A showed the best results for both the standard and the complex stool samples. This protocol reflected the theoretical composition of the standard samples very well and had the highest repeatability. Based on these results, protocol A represents an accurate standard protocol and is highly suitable for metagenomic profiling. Furthermore, protocol C, which showed only a slightly lower accuracy than protocol A, seems to be a good alternative for extremely low concentrated DNA samples due to its very low input requirements.

References

Hillmann, Benjamin; Al-Ghalith, Gabriel A.; Shields-Cutler, Robin R.; Zhu, Qiyun; Gohl, Daryl M.; Beckman, Kenneth B. et al. (2018): Evaluating the Information Content of Shallow Shotgun Metagenomics. In: mSystems 3 (6), e00069-18. DOI: 10.1128/mSystems.00069-18.



CeGaT

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 bioinformaticians 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 sequencing conditions on our Illumina platforms.

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

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
Web:	www.cegat.com

DAkkS
Deutsche Akkreditierungsstelle
D-PL-13206-01-00

Accredited by DAkkS according to
DIN EN ISO/IEC 17025:2018

CAP ACCREDITED
COLLEGE of AMERICAN PATHOLOGISTS

CLIA CERTIFIED ID: 99D2130225

2022/11