Amplicon Metagenome Analyses of Microbial Communities – Doing It the Right Way

16S metagenomics powers human, animal, and environmental microbiome studies of any scale.

Microbiota and NGS - a Dream Team

Dynamic host-microbe interactions have shaped the evolution of life. Virtually all plants and animals are colonized by microbiota, which are ecological communities of commensal, symbiotic, and pathogenic microorganisms. And it is increasingly recognized that the biological processes of hosts and their associated microbial communities function in tandem, often as biological partners forming a collective entity known as a metaorganism [1]. Microorganisms form very diverse communities and a characteristic of these communities is that a few taxa dominate them, while a very large number of taxa occur with lower frequency [2]. Furthermore, taxa that cannot be cultivated may also occur and therefore such taxa cannot be detected by classical methods.

The rapidly growing interest in microbiome research has been reinforced by the ability to profile different microbial communities using Next Generation Sequencing (NGS). This culture-free, high-throughput technology enables the identification and comparison of entire microbial communities, which is known as metagenomics [3]. Metagenomics typically involves two different sequencing strategies: the first sequencing strategy is amplicon sequencing, which is usually of the 16S rRNA gene as a phylogenetic marker, while the second sequencing strategy is shotgun metagenome sequencing, and this is a whole genome sequencing approach [3]. Therefore, metagenomics provides comprehensive answers to a range of important questions, including the influence of the human intestinal flora on health.

The use of the 16S ribosomal RNA gene in prokaryotes and the ITS sequence (internal transcribed spacers of rDNA) in fungi as phylogenetic markers has proven to be an efficient and cost-effective strategy for microbiome analysis. In fact, experiments revolving around 16S rRNA allow even the imputation of functional contents based on taxon frequencies [4] [5].

On the other hand, shotgun metagenomics enables researchers to measure the functional relationships between hosts and bacteria by directly determining the functional content of samples. In addition, shotgun metagenomics has a theoretically unbiased coverage of all taxonomies found in a DNA sample. However, contamination with host DNA and the occurrence of low frequency taxa requires very deep sequencing if one is to achieve the same taxonomic resolution as 16S rRNA sequencing. This means a manifold increase in both the costs and the data load. In this White Paper we will limit ourselves to 16S amplicon sequencing and the factors that need to be taken into account when conducting it.

Typical Barcoding Loci for Bacteria and Fungi

16S rDNA as barcoding locus in prokaryotes

The 16S rRNA gene is the DNA sequence corresponding to ribosomal RNA and it is essential for the synthesis of all prokaryotic proteins. The 16S rRNA gene occurs in all bacteria and it is highly conserved and highly specific. The internal structure of the 16S rRNA gene is composed of variable regions (see ***Figure 1***). Having varying

Microsynth AG, Switzerland
Schützenstrasse 15 · P.O. Box · CH-9436 Balgach · Phone +41 71 722 83 33 · Fax +41 71 722 87 58 · info@microsynth.ch · www.microsynth.ch

degrees of difference among the different bacteria makes it possible to identify the taxonomic identity of microbes, at the genus level and often at the species level. Since it is not practical to sequence the complete 16S rDNA using NGS, only a part of the variable regions is sequenced.

Microsynth currently offers different standard primer sets that have been developed based on the recommendations of the Human Microbiome Project Consortium [6]. Specifically, the V4 and V34 locus primers are suitable for most bacterial metacommunity projects. It has been demonstrated in several studies that the aforementioned primers facilitate the detection of a broad range of taxa (see ***Figure 1A***). Depending on the project requirements, customer-specific primer sets can also be applied or even developed and validated.

ITS regions as barcoding locus for fungi

The ITS region has become the gold standard for the classification of fungi [7]. With few exceptions [8], this locus is suitable for distinguishing fungi up to species level. In addition to the ITS regions, other loci have been used for barcoding fungi, although the data basis is much smaller [9].

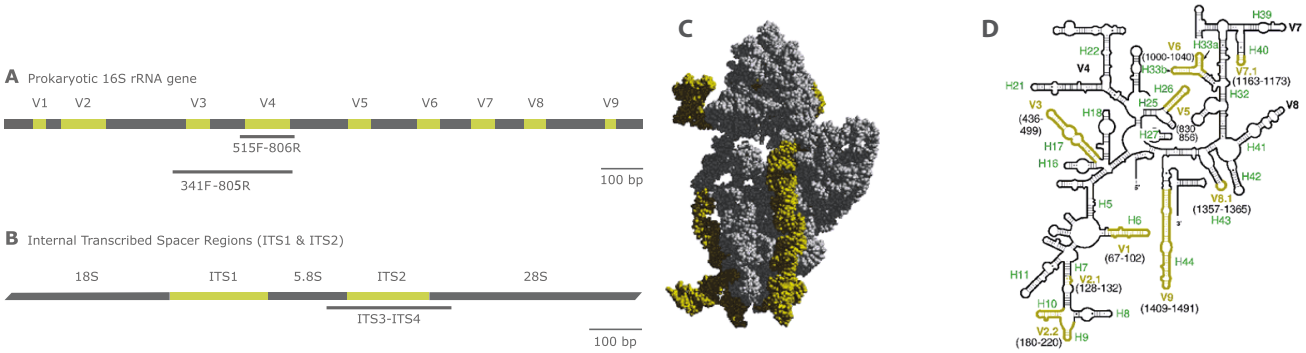


Figure 1: *Overview of ribosomal gene loci commonly used for the taxonomic analysis of microbial communities. Hypervariable regions are marked in green, while conserved regions are marked in gray. A. Structure of the prokaryotic 16S rRNA gene showing the nine hypervariable regions (V1-V9) and the regions targeted by the commonly used primer systems. B. Organization of the fungal rRNA gene operon showing two internal transcribed spacer regions (ITS). ITS2 is used most often for profiling fungal communities. C. 3D structural model of the 16S rRNA of Escherichia coli according to Tung et al [10]. The variable regions in the 16S rRNA are shown in green. D. Secondary structure of the 16S rRNA with variable regions in green.*

Choosing the Right Primer Set is the Key to Success!

Amplicon metagenomics is based on NGS sequencing of the microbial 16S rRNA gene. Since NGS single read lengths are limited to 300 base pairs (600 bp for paired end reads) when using Illumina high-throughput platforms, only parts of the 16S rRNA gene can be amplified and sequenced. In prokaryotes, the analysis targets hypervariable regions (V1-9) on the 16S rRNA gene. Meanwhile, in fungi the internal transcribed spacer regions (ITS) are used for taxonomic profiling (see ***Figure 1***).

For 16S/ITS amplicon metagenomics, it is important to give high priority to the choice of primers. An ideal primer system should be sufficiently universal to cover a broad range of taxonomic groups, while the resulting amplicon must provide sufficient taxonomic information for a reliable taxonomic classification. Based on our experience and the validation of our 16S/ITS analysis pipeline, we recommend the V34 primer system for a broad and accurate bacterial classification. However, if Archaea are also expected, the V4 primer system should be used in order to obtain a good taxonomic classification. It should be emphasized that our service is not limited to the presented marker genes and primer systems; other phylogenetic marker genes (e.g. cytochrome c oxidase I) and primer systems can also be used. Furthermore, a pilot study can be very helpful in terms of finding the best primer system for your specific research question. Finally, metagenome analyses should only compare communities generated with identical primer sets. This is because of the biases of the different primer systems.

Besides the choice of the locus specific PCR primers, the isolation of high-quality DNA from environmental samples has a major impact on the outcome of the study.

What Does a Typical Project Schedule Look Like?

The first step in an amplicon metagenome project is deciding which primer set is the most promising. At Microsynth, the question of whether to work with a standard primer set or to instead develop a customer-specific primer set is discussed in detail with the customer. In addition, it must be defined whether DNA is to be isolated internally or whether it should be outsourced (see ***Figure 2***). The DNA is then amplified by PCR and Microsynth uses a "two-step" PCR protocol for the amplification. In a first step, the locus is amplified with short template specific primers as well as an Illumina tail adapter. From there, in a second step, the so-called NGS fusion primers, including an index for multiplexing as well as an Illumina adapter, are used. Previous studies have shown that this protocol generates high quality multiplex amplicon libraries and ensures high reproducibility [11].

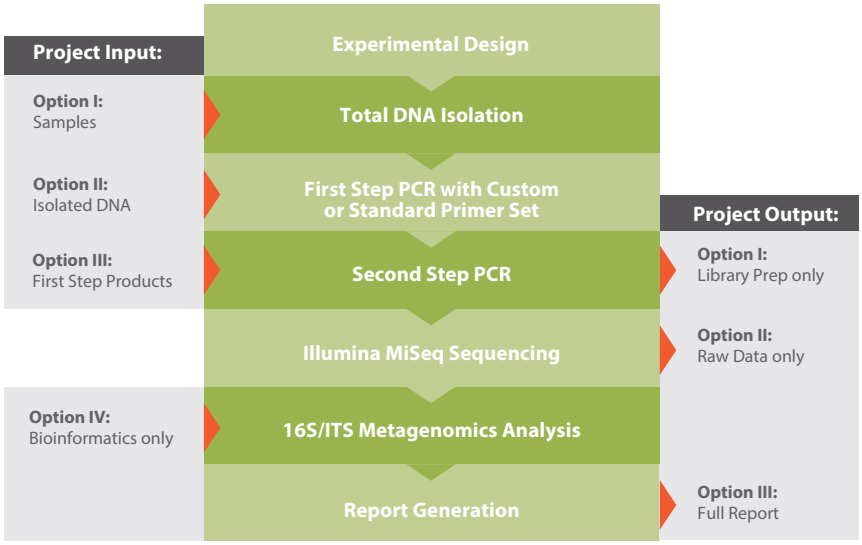


Figure 2: *Schematic representation of a project procedure for metagenome analysis of microbial communities. Depending on the initial situation and the problem, the first step is to determine the suitable primer set. However, should the customer request it, a new primer set can be developed. Furthermore, either the entire process - including DNA isolation - can be outsourced to Microsynth, or the DNA can be isolated by the customer. At the end of the process the customer will receive a clearly structured report that can be used for further analysis.*

Instead of only using a single forward and reverse primer for first step amplification, some protocols use several forward primers that differ in length by adding various numbers of degenerate bases (wobbles, Ns) at the 5' end of the locus specific primer. An alternative but similar approach is to use a fixed number of degenerate bases (see ***Figure 3***). Both concepts aim at adding sequence diversity as this improves the quality and quantity of reads generated on an Illumina MiSeq platform (see ***Figure 4***). The fixed length degenerate bases are less effective in terms of introducing sequence diversity compared to the staggered degenerate base approach. However, they represent a good trade-off in practice and are also easier to handle in downstream analysis. These protocols are especially useful for high-throughput projects where sequencing throughput is particularly critical and many samples are pooled.

For projects involving very low amounts of starting material we recommend a three-step PCR protocol including two subsequent locus-specific PCRs to increase the yield of sequenceable amplicons. After equimolar pooling of the PCR products, NGS sequencing is performed on the Illumina MiSeq. Meanwhile, the final and perhaps most important step is the bioinformatic analysis. For this step, we have established a designated pipeline that comprises state-of-the-art algorithms and software designed to extract as much valuable information as possible from the generated data and to visualize the data in a captivating form. In the following section we will show in more detail how a bioinformatic analysis and its output in metagenome projects can look.

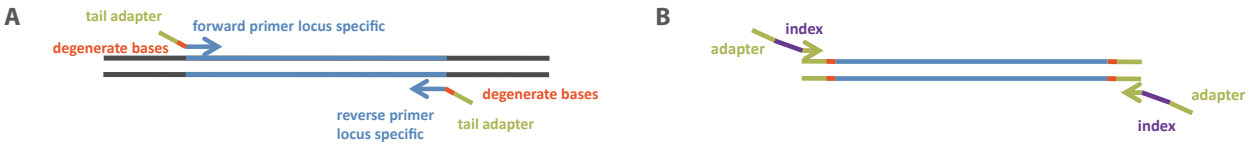


Figure 3: *Principle of the degenerate base approach for primers. A: In the first step PCR, the target is amplified and the tail adapter is appended. 5 degenerate bases are appended between the tail adapter and the locus specific primer. B: In the second step PCR the index and the adapter are appended. The second PCR product binds to the Illumina flow cell and it is sequenced starting at the degenerate bases, which ensures sequence diversity.*

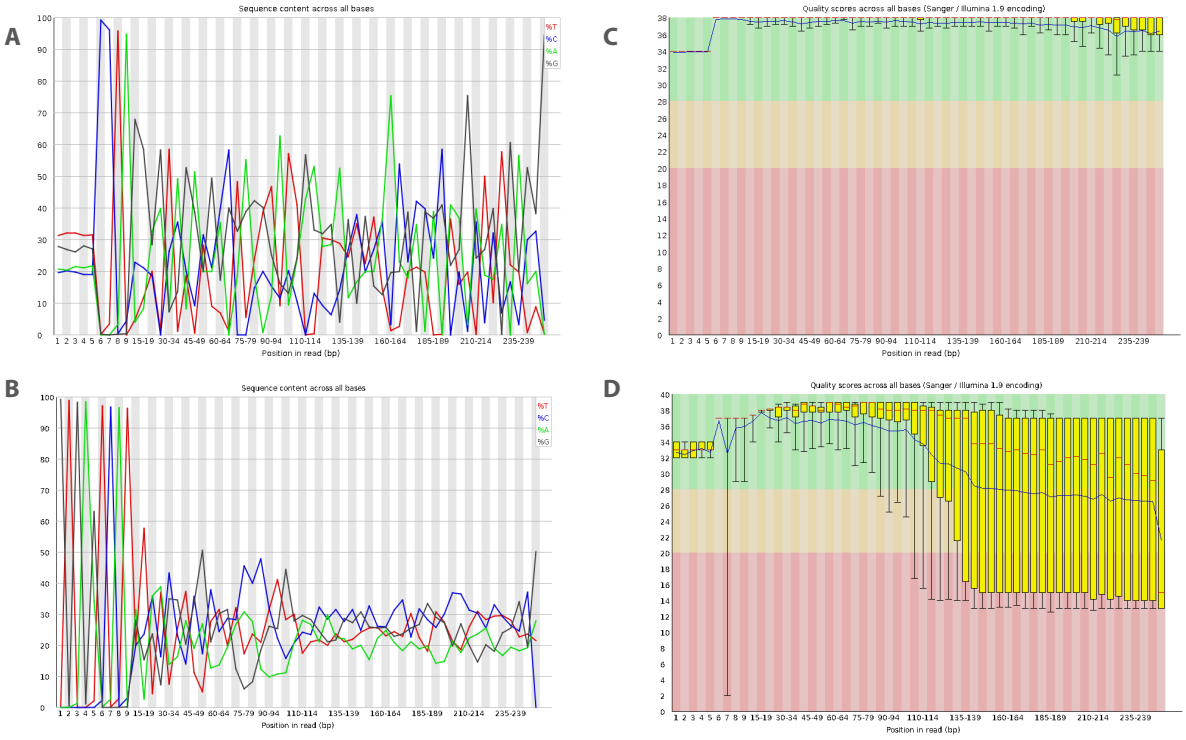


Figure 4: *These charts show the effect of degenerate bases on the quality of the sequencing. A: In the chart of the sequence content across all bases, you can clearly see the 5 degenerate bases at the beginning of the sequence due to their typical distribution (approximately 25%). B: When no degenerate bases were included, the beginning of the sequencing was always the same. C: The chart of the per base quality shows high quality scores for all bases when degenerated bases were used. D: When no degenerate bases were used in the primers, the quality score soon decreased for longer reads.*

State of the Art - the Bioinformatics Analysis Pipeline from Microsynth

For the bioinformatic analysis of the sequencing data, the sequenced paired-end reads are first subjected to de-multiplexing and trimming of Illumina adaptor residuals. In a second preparation step, paired-end reads are filtered for their quality and their locus specific primers are trimmed as well. From there, the remaining paired-end reads are de-noised [12] to form operational taxonomic units (OTUs), while in the process discarding singletons and chimeras. The resulting OTU abundances are then filtered for possible barcode bleed-in contaminations [13] to reduce noise. Following this, the OTU sequences are compared to a reference sequence database, such as RDP [14], in order to predict their taxonomies and corresponding confidence scores. The resulting metagenome is visualized by an interactive Krona chart [15] (see ***Figure 5***) that provides a quick and easy overview of the data. This enables the scientist to intuitively explore the intricacies of the analyzed bacterial community. The diversity of the metagenomic community is expressed in simple and comparable terms as alpha and beta diversity scores. The alpha diversity describes the intra-diversity of each sample, while the beta diversity describes the inter-diversity of all samples together [16]. Different and widely used alpha diversity scoring metrics are displayed in ***Figure 6B***. On the X axis the analyzed samples are annotated, and on the Y axis their individual scores for each metric are displayed. Rarefaction curves are calculated in order to estimate if a microbiome has been sufficiently characterized. If the curves end in a plateau, this signifies that the microbiome was sufficiently covered (see ***Figure 6A***). The complex interaction of multiple bacterial communities in a given environment is described by beta diversity based on a distance metric such as Unifrac [17]. Principal component analysis (PCA) helps in terms of simplifying, understanding, and visualizing such interactions (see ***Figure 7A***). In

Figure 7A, the PCA was able to explain 90% of the variance of the original data with just two components (75% on the first principal component and 15% on the second principal component). If an experimental design exists, dividing samples into different categories such as treated and control samples, the differential OTU analysis reveals detailed changes in the microbiome of the analyzed groups [18] (see ***Figure 7B***). Finally, functional profiles are predicted [19] (see ***Table 1***) using various publicly available databases. The pathways and their abundance within the different samples are shown in ***Table 1***.

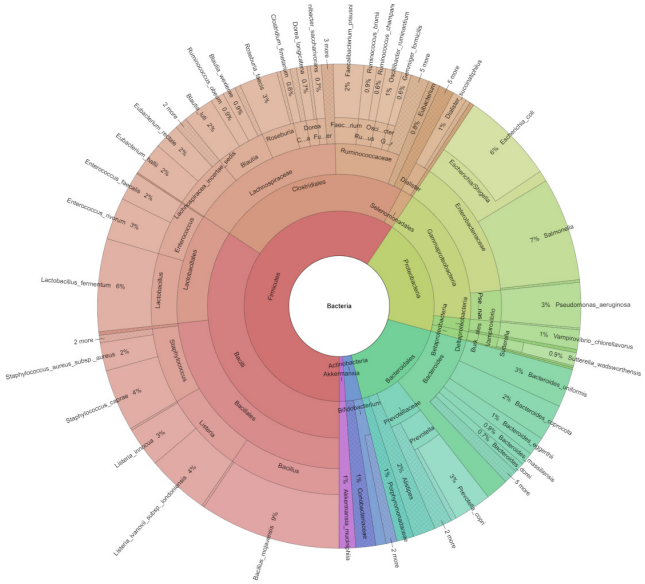


Figure 5. *Interactive Krona chart of the bacteria represented by 16S rRNA gene amplicon-based bacterial diversity in a feces sample. Each circle represents the phylum, class, order, family, genus, and species from the inside to the outside of the circle, respectively. In addition, the relative abundance of each taxa is annotated on the chart.*

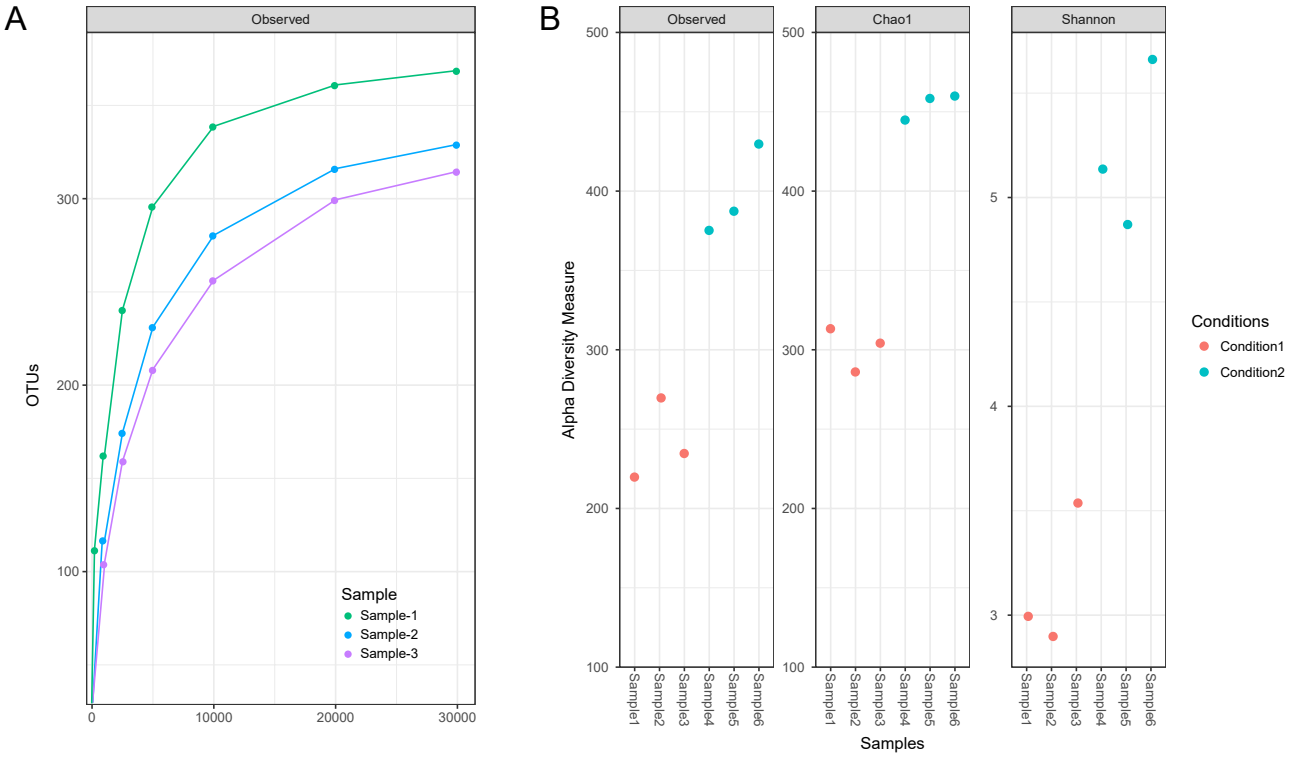
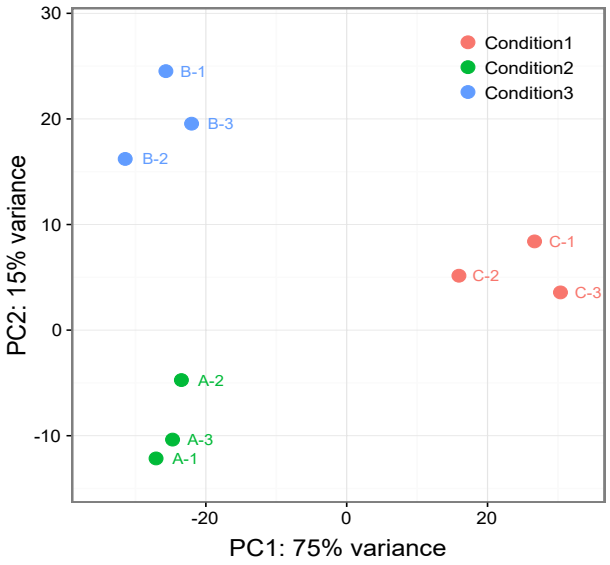


Figure 6. *Examples of alpha diversity results. 6A. Rarefaction curves indicating whether sampling and sequencing covered the sample richness (the x axis displays the sampled number of reads, while the y axis displays the number of detected OTUs). 6B. Alpha diversity measures for the analyzed community including observed richness; the Chao 1 index and the Shannon diversity index.*

A



B

Condition-1 vs Condition-2 OTU table

10 records per page

OTU_ID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
OTU1	5960	-0.170	0.0308	-5.54	3.09e-08	4.11e-08
OTU19	5410	0.765	0.0668	11.40	2.45e-30	1.96e-29
OTU2	5080	0.165	0.0428	3.85	1.19e-04	1.19e-04
OTU3	4510	-0.176	0.0258	-6.80	1.06e-11	1.69e-11
OTU4	4260	0.263	0.0269	9.79	1.28e-22	3.41e-22
OTU5	3500	-0.228	0.0281	-8.11	5.25e-16	1.05e-15
OTU6	2540	-0.457	0.0435	-10.50	9.27e-26	3.71e-25
OTU7	2090	0.204	0.0432	4.72	2.34e-06	2.67e-06

Showing 1 to 8 of 8 entries

Figure 7. *Examples of beta diversity results. 7A. Principal component analysis plot to visualize sample clustering. 7B. This excerpt from a results table shows differential abundance of OTUs between two conditions, including statistical measures for differential abundance (log fold change) and significance (adjusted p-value).*

Table 1. *Functional profiles predicted according to OTUs, their predicted taxonomies, and their abundance in each of the samples.*

pathway	description	Sample_V34_1a	Sample_V34_1b	Sample_V34_1c
NONOXIPENT-PWY	pentose phosphate pathway (non-oxidative branch)	45500	46300	46500
CALVIN-PWY	Calvin-Benson-Bassham cycle	39700	40200	40400
PWY-7220	adenosine deoxyribonucleotides de novo biosynthesis II	39700	39800	40100
PWY-7222	guanosine deoxyribonucleotides de novo biosynthesis II	39700	39800	40100
PWY-7663	gondoate biosynthesis (anaerobic)	38200	38600	38800
PWY-6737	starch degradation V	37400	37500	37800
PWY-5101	L-isoleucine biosynthesis II	37100	37800	37800
GLYCOCAT-PWY	glycogen degradation I (bacterial)	36800	37600	37300
PWY-7229	superpathway of adenosine nucleotides de novo biosynthesis I	36500	37000	37200
ANAGLYCOLYSIS-PWY	glycolysis III (from glucose)	36100	36600	36700

Showing 1 to 10 of 336 entries

Conclusion

The complexity of metagenome analysis has increased in line with technological progress. In order to perform meaningful metagenome analyses, our experience has taught us that the following success criteria are important:

- It must be considered which variable DNA region (e.g. on the 16S rDNA or ITS regions) is best suited for the intended study. Based on these considerations, the optimal primer set is determined (and, if possible, it is pre-tested in a pilot study).
- To be able to amplify the different taxa representatively and to sequence them afterwards, a robust and functional DNA isolation procedure must be available. On the other hand, we recommend a "two-step" protocol if possible, to enable the use of high-quality amplicon libraries for the subsequent sequencing.
- Data analysis should be performed using a few but meaningful bioinformatics tools.
- If you are lacking experience in one or more of the above criteria, we recommend outsourcing the study to an experienced service provider. Microsynth has more than 10 years of experience.

Literature

[1] Rausch, P., Rühlemann, M., Hermes, B.M. et al. Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. Microbiome 7, 133 (2019). https://doi.org/10.1186/s40168-019-0743-1

[2] McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., Dornelas, M., Enquist, B.J., Green, J.L., He, F.L., Hurlbert, A.H., Magurran, A.E., Marquet, P.A., Maurer, B.A., Ostling, A., Soykan, C.U., Ugland, K.I. & White, E.P. (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecology Letters, 10: 995–1015.

[3] Hugenholtz, P. (2002) Exploring prokaryotic diversity in the genomic era. Genome Biology 3: reviews0003.1–reviews0003.8.

[4] Morgan XC, Huttenhower C. Chapter 12: human microbiome analysis. PLoS Comput Biol. 2012;8(12):e1002808.

[5] Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31(9):814–21.

[6] The Human Microbiome Project Consortium (2012) A framework for human microbiome research. Nature, 486: 215-221.

[7] Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Bergeron MJ, Hamelin RC, Vialle A, and Fungal Barcoding Consortium. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Science 109: 6241-6246. (doi: 10.1073/pnas.1117018109)

[8] Grünig, C.R.; Brunner, P. C.; Duo, A. & Sieber, T. N. (2007) Suitability of methods for species recognition in the Phialocephala fortinii - Acephala applanata species complex using DNA analysis. Fungal Genetics and Biology, 44: 773-788.

[9] Ratnasingham S, Hebert PDN (2013) A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. PLoS ONE 8(8): e66213. DOI:10.1371/journal.pone.0066213

[10] Tung, C.-S., Joseph, S. & Sanbonmatsu, K.Y. (2003) All-atom homology model of the Escherichia coli 30S ribosomal subunit. Nature Structural Biology, 9: 750-755

[11] Berry, D., Mahfoudh, K.B., Wagner, M. & Loy, A. (2011) Barcoded primers used in multiplex amplicon pyrosequencing bias amplification, Applied and Environmental Microbiology, 77: 7846- 7849.

[12] R.C. Edgar (2016), UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing, https://doi.org/10.1101/081257

[13] R.C. Edgar (2018), UNCROSS2: identification of cross-talk in 16S rRNA OTU tables, https://doi.org/10.1101/400762

[14] Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis Nucl. Acids Res. 42(Database issue):D633-D642; doi: 10.1093/nar/gkt1244 [PMID: 24288368]

[15] Ondov BD, Bergman NH, and Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011 Sep 30; 12(1):385.

[16] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061217

[17] Lozupone C, Knight R . (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71: 8228–8235.

[18] https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8 (DESeq2)

[19] Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Langille, M. G.I.*; Zaneveld, J.*; Caporaso, J. G.; McDonald, D.; Knights, D.; a Reyes, J.; Clemente, J. C.; Burkepile, D. E.; Vega Thurber, R. L.; Knight, R.; Beiko, R. G.; and Huttenhower, C. Nature Biotechnology, 1-10. 8 2013.

Contact

Microsynth AG
Schützenstrasse 15
CH-9436 Balgach
Switzerland
phone: +41-71-722 83 33
web: www.microsynth.ch
email: genome@microsynth.ch