Microb Health Dis 2021; 3: e474

European Helicobacter & Microbiota Study Group

GUT MICROBIOME PIPELINE – THE ROAD SO FAR

J.W. Debelius, M. Pereira, L. Engstrand

Centre for Translational Microbiome Research, CTMR, Department of Microbiology, Tumour and Cell Biology (MTC), Karolinska Institutet, Science for Life Laboratory, Solna, Sweden

Corresponding Author: Lars Engstrand, MD; e-mail: lars.engstrand@ki.se

Abstract: Quantitive bioinformatic studies are a main stay of modern microbiome research. Despite almost 15 years since the first studies were published, a wide variety of techniques exist for analysis with new techniques emerging every year. In this review, we explore the latest developments in sample preservation, extraction, and sequencing. We describe new bioinformatic developments in both marker-gene and metagenomic sequencing. While these represent novel analytical approaches, in 2020 and beyond, we look forward to better benchmarks of the existing tools.

Keywords: Sample extraction, Bioinformatic analysis, Bias.

INTRODUCTION

The use of non-cultured based approaches to study the microbiome provided us with unprecedented ability to understand the complex and vital relationship between humans and their commensal organisms. In recent years, the decreasing cost of next generation sequencing has exceeded Moore's law and computational advancements, amplicon-based and whole metagenome shotgun sequencing (WGSS) technologies have made large scale microbiome studies affordable. However, the interdisciplinary nature of the field has led to the development of a series of ad-hoc protocols with no clear gold standard, despite the attempts of several initiatives, such as the "Microbiome Quality Control project (MQCP)" and "International Human Microbiome Standards (IHMS)"[1-3]. Due to variability among protocols, bias can be added at each step of the pipeline, making the data hard to reproduce and interpret[4]. In the past few years, efforts have been made to extend the coverage of commensal organisms (mycobiome and virome) and to improve DNA extraction pipeline and the fidelity in sequencing technologies while decreasing costs per sample. In this review, we aim to show the most relevant advances in the field and an overview of each step of the pipeline can be found in Figure 1.

SAMPLE COLLECTION

One long-discussed issue in study design is the sample preservation technique, which need to be relatively easy to use and provide good conservation to the sample. At the moment, RNAlater, OMNIgeneGUT, 95% ethanol, card-devices (such as FOBT) and FIT tubes seem to be the best preservation options when immediately freezing is not possible. Although they also bring their limitations: OMNIgeneGUT may cause a small shift in bacterial composition; eth-

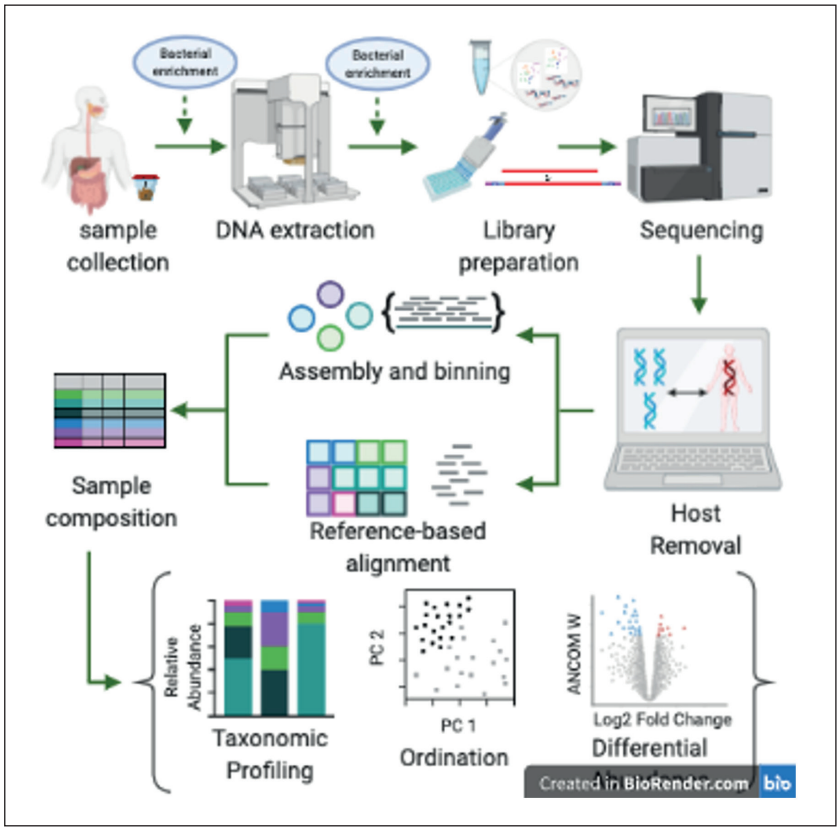

Figure 1. Microbiome analysis pipeline.

anol requires special handling while shipping and seems to have poor performance in WGSS; RNAlater is not suitable for long storage at room temperature; and FOBT and FIT provide little biological material (for more detailed evaluation[4-8]). Therefore, each researcher should carefully evaluate the pros and cons of each method, having in mind the particularities of their study.

Another source of bias is the homogenization technique and the issue of repetitive freeze and thaw of biological material[4,7,9]. As mentioned by Wu et al[7], a bad homogenization can fail to fully capture the microbiome profile since different portions of the stool can vary in composition. Furthermore, repetitive freeze and thaw may allow bacterial growth and will degrade DNA promoting profile shift, especially when aiming to study long amplicons and cDNA libraries. To circumvent such an issue, a promising but not yet well-validated option is the use of CryoXtractR (a device that can aliquot frozen samples with the help of a high-speed rotation needle[7]).

DNA EXTRACTION

Although studies showed that the inter-individual variation is more prominent than kit variation, less effective extraction kits will mask group variability and increase the risk for false-negative results. It is currently believed that bead-beating accompanied by chemical lysis-based protocols are the best choices since they can provide higher yield, better-quality DNA, and a good recovery ratio between gram-positive and gram-negative[2-4,10]. Notably, kit variability issues are mostly related to the recovery of gram-positive bacteria[3,4,7] and this is still an ongoing issue that requires further development.

The most recent benchmarking papers concluded that Qiagen's QIAamp Stool Mini kit (with or without modifications) was the best protocol for stool extraction among the popular choices[2,11]. However, due to the complexity of dysbiosis conditions, the need to combine microbiome, virome, and mycobiome analysis has recently been raised. And, when evaluating

capacity to extract virus and yeast, the literature is quite controversial regarding the results obtained with the leading commercial kits[12-14]. Moreover, not only stool samples are studied from the gastrointestinal tract, biopsies and fluids such as pancreatic juice are also tested, raising the need for protocol adaptations for lower microbial biomass. Low biomass samples are especially affected by the effect of contaminants and high amounts of host DNA[15,16]. The primary source of contamination with foreign DNA – which may mask the real microbiome composition – are extraction reagents and laboratory environments. However, reports showed that even PCR reagents and ultrapure water could be a source of contamination[10].

Attempting to overcome some of the limitations mentioned before, the use of suitable negative and positive controls is mandatory. The negative control should evaluate the effect of extraction reagents and sample buffer separately, and arguments have been made for the use of a single organism control rather than a true blank, since a certain number of cells are required for detection[17]. The use of commercially available mock communities is the current standard for positive controls in most environments[3,18]. However, despite a few criticisms, the use of a pool of several samples or a chemostat community from in vitro microbial model system is a promising alternative, since it would mimic better the reality of the samples extracted[3].

Finally, high amount of host DNA is also an issue when aiming for WGSS. To overcome the issue, the researcher may increase sequencing depth[3,19-21] or circumvent the issue using bacterial enrichment techniques. The currently available methods for bacterial enrichment are NebNext, QIAamp, HostZERO, and MoLysis kit or lyPMA and Benzonase treatment[19-21]. These kits seem to be highly sensitive to the level of DNA shearing, presence of lysed cells, and conservation buffer used, requiring optimization before use. Nonetheless, they are promising options for host DNA depletion and should be further studied and optimized.

Altogether, we are still far from a gold standard for sample conservation, homogenization, extraction, and even control choice. We hope to see new multicenter benchmark studies for a revaluation of best practices.

SEQUENCE STRATEGIES

Amplicon based sequencing (16S rRNA sequencing)

The 16s rRNA gene is a standard target for bacterial identification due to its relative universality, low cost, specificity, and computational tractability and has been the primary method for microbiome characterization for the last decade. While the internal transcribed spacer (ITS) linker is the primary target for evaluating mycobiome composition[22-25], although the appropriate ITS region is still under discussion.

However, amplicon-based methods have some well-known limitations. Amplicon sequencing is PCR dependent and relies on the universality of primers targeting a hypervariable region to be able fully cover all species in an environment[4,26-28].

Historically, species and strain level resolution have been a limitation in marker gene sequencing. The publication of denoising algorithms such as Dada2, Unoise, and Deblur, in 2016-2017 provided single nucleotide resolution amplicon sequence variants (ASVs) for Illumina and 454 sequencing[29-32]. This resolution is often sufficient to find strain-specific differences in biogeography or behavior in closely related organisms[33]. However, it brings the second challenge in marker gene sequencing: taxonomic annotation. Taxonomic annotation in a messy proposition in and of itself, due to differences between morphology-based taxonomic classifications and molecular phylogeny[34], annotation-based differences between commonly used databases[35], rapidly changes in organism naming (including a 2020 updated to genus Lactobacillus)[36].

The 2018 release of the Genome Taxonomy Database (GTDB) proposed a new naming scheme for organisms based on phylogenetic relationships to improve concordance between name and clade[34]. It provides a more standard naming scheme for newly introduced organisms and a toolkit for classification[37]. These annotations were incorporated into the last two releases of the popular Silva database[38]. A 2020 preprint expands the GTDB to provide improved Archeal taxonomy[39]. Although these improvements provide a framework going for-

ward, extensive work is required to improve annotation on currently uncultured organisms, likely through the development of a culture-based database using newer anaerobic methods and improved curation.

A second proposed solution to improve taxonomic resolution is through the use of bespoke classifiers. In general, taxonomic prediction algorithms assume that all sequences in a taxonomic database are equally likely in a given environment. The clawback algorithm, released in 2019, uses a large collection of publicly available studies to improve classification by weighting taxonomy based on the environment[40]. This classifier improved species level accuracy by 10% over traditional classifiers.

Full length amplicon sequencing has been proposed to solve resolution-associated issues. Aiming to overcome incomplete coverage and absence of strain resolution, new options were developed to cover the full length of the 16S rRNA[41-43]. Oxford Nanopore and PacBio sequencers are capable of covering the whole 16S rRNA (~1500bp). The main drawback of these technologies is a higher error rate compared to shorter reads[42], but efforts are taking place to improve accuracy[43]. Another proposed approach for increased marker gene resolution is to reconstruct the full-length sequence from multiple short regions. These leverages existing short read technology by either sheering full length 16s rRNA sequences into short, barcoded reads and re-assembling the full length reads[44] or amplifying with multiple primers and then scaffolding against a database[45]. Both techniques have the potential build on existing denoising techniques and higher accuracy short read technology. Full length, and synthetic full-length sequencing claim they allows for complete coverage of community and claims to decrease analysis bias, but only a few studies were performed in the platform (mostly using mock communities). Therefore, further studies[41,42] showing performance in biological samples are necessary.

In 2020 and beyond, amplicon-based methods will need to address a few fundamental issues. First, the field would benefit from environment-specific community wide standards. Second, databases need to be expanded and improved to both cover the broad diversity of life and provide specificity for organisms, and finally, methodological improvements are needed to improve the feasibility of high resolution, full length amplicon sequencing.

Shotgun metagenomic sequencing (WGSS)

WGSS overcomes several of the limitations faced by 16S sequencing, although the potential pitfalls have yet to be fully explored. Among the advantages of WGSS are that it is PCR free, allows functional analysis of organisms, allows sub-species and strain-level annotation, and may expand community profiling to DNA viruses, fungi and protists. However, due to the untargeted nature, WGSS is more sensitive to environmental and host DNA than amplicon-based methods which makes it difficult in many environments[19-21]. WGSS is also associated with both high sample-preparation and computational costs, making it prohibitive for many groups.

Among the current technologies Oxford nanopore, PacBio, Illumina and MGI are the most cited[46-49]. Until now, no independent benchmarking comparison study was performed with all of them to compare technologies efficiency.

Several solutions have been proposed to address high costs. For certain sample types, shallow shotgun sequencing may provide meaningful enough data for a comprehensive analysis[50]. A second recent proposal[51] was miniaturization of enzyme-based protocols with acoustic liquid handling robots in combination with co-assembly techniques, which significantly reduced the average per-sample reagent costs. Finally, improvements in sequencing technologies have brought down the cost per read.

The primary annotation strategies for metagenomic data can be split into either reference-based method, which involves the alignment of sequences against an existing reference database, and de novo assembly-based approaches, where genetic content is inferred by assembling reads into longer contiguous reads (contigs) and then binning these contigs into genomes[52].

Recent advances in reference-based methods have focused in two key areas. First, improvements have been made in memory consumption and runtime for alignment algorithms including an update to the popular k-mer based alignment algorithm, Kraken2[53]. However,

several alternatives for kmer based approaches have been proposed, including Ganon, which uses a novel indexing scheme; and Metalign which performs a two-step alignment procedure. Both methods claim improved accuracy over Kraken2, although only Ganon promises improved speed and memory consumption.

De novo approaches have also seen new developments in the last year. One popular assembly-based proposal has been translation from DNA to proteins. Protein-based methods have shown improved precision and accuracy over nucleotide-based methods because they limit some of the major challenges in nucleotide-based assembly. In 2019, updated version Guided Reference-based Assembly of Short Peptides (GRASP)[54] and the novel Protein Level ASSebler, (Plass) algorithm[55] were published.

Several new binning algorithms were also released in 2019 and 2020. MetaBat2 promises automated tuning parameters, and claims to outperform other binning approaches in completeness, runtime, and memory consumption[56]. MetaBMF claims species level resolution, and outperformed several popular binning algorithms, including MetaBat1, in terms of accuracy, precision and runtime[57]. A third algorithm draws on image signal processing and machine learning to provide binning for a single sample[58]. However, the method underperforms on dense data, which may suggest a pre-clustering step is needed.

Methods for long read technology have also been evolving. Minerva[59] was published in 2019, which performs long read deconvolution for synthetic long reads. Advances in the annotation and binning of true long read technology included reference-based approaches to address read-associated errors and improved alignment in the past year[60-62]. Additional work has also been invested in assembly with long reads alone[63,64] or in combination with short reads[51,65] have proven promising.

In the next year, we hope to see improved understanding of how wet lab preparation affect results and improved protocols for handling contamination[16,18]. While there has been extensive development in metagenomic assembly techniques, independent benchmarks are severely needed to address the suitability of newly developed algorithms and provide a community standard. We also anticipate additional results from the Critical Assessment of Metagenome Interpretation (CAMI) challenge[66], which will provide more insight into the best strategies in a new and rapidly growing field.

BIOSTATISTICAL ANALYSES AND DIFFERENTIAL ABUNDANCE

One of the fundamental statistical challenges in microbiome analysis is the compositional nature of the data[67]. Most culture-free microbiome techniques decouple true microbial biomass from observed read abundance, meaning that observed sequences represent a relative abundance which sums to one. This can be addressed through the use of an internal control, biomass quantification, or compositionally aware methods. In 2019 and 2020, several methods were proposed, developed, and evaluated to address the challenges of differential abundance in the microbiome.

A somewhat controversial 2020 preprint claimed[68] that using the same volume of sample (rather than the same DNA concentration) would maintain the concentration-based information in combination with a Bayesian model to estimate differences in orders of magnitude. We look forward to the peer-reviewed version of this study because, if effective, this technique might prove a useful solution for many groups.

Several bioinformatic solutions have also been proposed. Morton et al[69] presented Songbird, a rank-based approach, where relative differences are ranked in comparison to other taxa in the community using a multivariate-regression model. Their computational method was benchmarked against measures of absolute abundance, and they found their relative taxonomic ranks replicated the results for absolute abundance better than two popular compositional methods. Alternative proposals have included pre-prints which suggest models relying on reference taxa to address compositionality[70,71]. However, the methods potentially suffer from a major issue in most common methods for handling microbial differential abundance: zero-substitution to allow the use of log-ratios. Several recent methods have proposed the use of zero-inflated methods to address this issue, including the use of a zero-inflated Dirichlet multinomial[72] and a Bayesian modeling with a hierarchical negative binomial and

accounts for sampling depth in the model[73]. Finally, the Adaptive multivariate two-sample test for Microbial Differential Abundance (AMDA) also debuted in 2019[74]. This test is restricted to comparisons between two groups and uses a two-step protocol where the features are first selected using a permutation-based approach and then a maximum mean discrepancy (MMD) test is applied to the remaining features. They benchmarked against other MMD tests but have not compared against non-MMD tests.

There have been several other advances between 2019 and 2020. These included the publication of DECOIDE, a compositionally aware method for dimensionality reduction which retains species relationships[75]; a new tool to model metabolic flux and infer metabolic interactions in metagenomic data[76] and a technique for integrating microbiome/metabolome interactions with machine learning[77]. Additionally, QIIME 2, a follow-up to the highly cited amplicon centric QIIME analysis platform, was published in 2019[78]. The updated version integrates and wraps several popular existing tools and includes integrated provenance tracking, a community-based plugin development system, and multiple interfaces depending on user experience and needs.

In the next years, we hope to see a new, independent benchmark of the flood of differential abundance methods. Most methods only compare against a subset of others, few leverage an internal control, and the last major independent comparison of microbiome differential abundance techniques in 2019 failed to account for several popular compositional methods [28968702]. We also hope to see improvements in method interpretability, as this remains a major challenge in most microbiome methods.

CONCLUSIONS

The last years have seen several important advances in our ability to profile the microbiome, particularly in the field of metagenomics and differential abundance. However, several issues remain outstanding in the field. First, work needs to be done to address bias through the wet lab pipeline. Community standards and consensus are needed for data generation to ensure high quality, comparable data across labs and sequencing platforms. This includes the identification of appropriate controls. We hope to see more full length 16s rRNA sequencing, particularly in environments where metagenomics cannot be performed due to large amounts of human DNA whether this comes from a combination of short reads or due to improvements in long read technology. Finally, we hope to continue to see analytical improvements through better databases, independent benchmarks of annotation, alignment, binning, and differential abundance analysis, and better approaches to address the complexity of microbiome data.

Conflict of interest

The authors declare that they have no conflict of interest.

REFERENCES

1. Sinha R, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, Amir A, Schwager E, Crabtree j, Ma S, TMQC consortium, Abnet CC, Knight R, White O, Huttenhower C. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nature Biotechnol 2017; 35: 1077-1086. doi.org/10.1038/nbt.3981.
2. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M, Hercog R, Jung FE, Kultima JR, Hayward MR, Coelho LP, Allen-Vercoe E, Bertrand L, Blaut M, Brown JRM, Carton T, Cools-Portier S, Daigneault M, Derrien M, Druesne A, de Vos WM, Finlay BB, Flint HJ, Guarner F, Hattori M, Heilig H, Luna RA, Vlieg JH, Junick J, Klymiuk I, Langella P, Le Chatelier E, Mai V, Manichanh C, Martin JC, Mery C, Morita H, O´Toole PW, Orvain C, Patil KR, Penders J, Persson S, Pons N, Popova M, Salonen A, Saulnier D, Scott KP, Singh B, Slezak K, Veiga P, Versalovic J, Zhao L, Zoetaendal EG, Ehrlich SD, Dore J, Bork P. Towards standards for human fecal sample processing in metagenomic studies. Nature Biotechnol 2017; 35: 1069-1076. doi.org/10.1038/nbt.3960.
3. Greathouse KL, Sinha R, Vogtmann E. DNA extraction for human microbiome studies: the issue of standardization. Genome Biol 2019; 20: 212. doi.org/10.1186/s13059-019-1843-8.

4. Conrads G, Abdelbary MM. Challenges of next-generation sequencing targeting anaerobes. Anaerobe 2019; 58: 47-52. doi.org/10.1016/j.anaerobe.2019.02.006.
5. Vandeputte D, Tito RY, Vanleeuwen R, Falony G, Raes J. Practical considerations for large-scale gut microbiome studies. FEMS Microbiol 2017; 41(Supp1): S154-S167. doi.org/10.1093/femsre/fux027.
6. Jenkins SV, Vang KB, Gies A, Griffin RJ, Jun SR, Nookaew I, Dings RPM. Sample storage conditions induce post-collection biases in microbiome profiles. BMC Microbiol 2018; 18: 227. doi.org/10.1186/s12866-018-1359-5.
7. Wu WK, Chen CC, Panyod S, Chen RA, Wu MS, Sheen LY, Chang SC. Optimization of fecal sample processing for microbiome study--The journey from bathroom to bench. J Formos Med Assoc 2019; 118: 545-555. doi.org/10.1016/j.jfma.2018.02.005.
8. Byrd DA, Sinha R, Hoffman KL, Chen J, Hua X,Shi J, Chia N, Petrosino J, Vogtman E. Comparison of methods to collect fecal samples for microbiome studies using whole-genome shotgun metagenomic sequencing. mSphere 2020; 5: e00827-19. doi: 10.1128/mSphere.00827-19.
9. Antosca K, Hoen AG, Palys T, Hilliard M, Morrison HG, Coker M, Madan J, Karagas MR. Reliability of stool microbiome methods for DNA yields and sequencing among infants and young children. Microbiol Open 2020; e1018. doi.org/10.1002/mbo3.1018.
10. Velásquez-Mejía EP, de la Cuesta-Zuluaga J, Escobar, JS. Impact of DNA extraction, sample dilution, and reagent contamination on 16S rRNA gene sequencing of human feces. Appl Microbiol Biotechnol 2018; 102: 403-411. doi.org/10.1007/s00253-017-8583-z
11. Panek M, ČipčićPaljetak H, BarešiĆA, PeriĆM, MatijašiĆM, LojkiĆI, VranešiĆBD, KrznariĆŽ, Verbanac D.Methodology challenges in studying human gut microbiota--effects of collection, storage, DNA extraction and next generation sequencing technologies. Sci Rep 2018; 8: 5143-5143. doi.org/10.1038/s41598-018-23296-4.
12. Frau A, Kenny JG, Lenzi L, Campbell BJ, Ijaz UZ, Duckworth CA, Burkitt MD, Hall N, Anson J, Darby AC, Probert CSJ. DNA extraction and amplicon production strategies deeply influence the outcome of gut mycobiome studies. Sci Rep 2019; 9: 9328. doi:10.1038/s41598-019-44974-x.
13. Fiedorová K, Radvanský M, Němcová E, Grombiříková H, Bosák J, Černochová M, Lexa M, Šmajs D, Freiberger T. The impact of dna extraction methods on stool bacterial and fungal microbiota community recovery. Front Microbiol 2019; 17; 10:821. doi: 10.3389/fmicb.2019.00821.
14. Deng L, Silins R, Castro-Mejía JL, Kot W, Jessen L, Thorsen J, Shah S, Stokholm J, Bisgaard H, Moineau S, Nielsen DS. A protocol for extraction of infective viromes suitable for metagenomics sequencing from low volume fecal samples. Viruses 2019; 11. doi.org/10.3390/v11070667.
15. Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, Lauder A, Sherrill-Mix S, Chehoud C, Kelsen J, Conrad M, Collman R, Baldassano R, Bushman F, Bittinger K. Optimizing methods and dodging pitfalls in microbiome research. Microbiome 2017; 5: 52. doi.org/10.1186/s40168-017-0267-5.
16. Zinter MS, Mayday MY, Ryckman KK, Jelliffe-Pawlowski LL, Derisi JL. Towards precision quantification of contamination in metagenomic sequencing experiments. Microbiome 2019; 7: 62. doi.org/10.1186/s40168-019-0678-6.
17. Minich JJ, Zhu Q, Janssen S, Hendrickson R, Amir A, Vetter R, Hyde J, Doty MM, Stillwell K, Benardini J, Kim JH, Allen EE, Venkateswaran K, Knight R. KatharoSeq enables high-throughput microbiome analysis from low-biomass samples. mSystems 2018; 3. doi.org/10.1128/msystems.00218-17.
18. Mclaren MR, Willis AD, & Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. eLife 2019; 8. doi.org/10.7554/eLife.46923.
19. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome 2018; 6: 42. doi.org/10.1186/s40168-018-0426-3.
20. Nelson MT, Pope C E, Marsh RL, Wolter DJ, Weiss EJ, Hager KR, Vo AT, Brittnacher MJ, Radey MC, Hayden HS, Eng A, Miller SI, Borenstein E, Hoffman LR. Human and extracellular DNA depletion for metagenomic analysis of complex clinical infection samples yields optimized viable microbiome profiles. Cell Rep 2019; 26: 2227-2240. e5. doi.org/10.1016/j.celrep.2019.01.091.
21. Heravi FS, Zakrzewski M, Vickery K, Hu H. Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples. J Microbiol Methods 2020; 170: 105856. doi.org/10.1016/j.mimet.2020.105856.
22. Auchtung TA, Fofanova TY, Stewart CJ, Nash AK, Wong MC, Gesell JR, Auchtung JM, Ajami NJ, Petrosino, JF. Investigating colonization of the healthy adult gastrointestinal tract by fungi. mSphere 2018; 3. doi.org/10.1128/mSphere.00092-18.
23. Lai GC, Tan TG, Pavelka N. The mammalian mycobiome: a complex system in a dynamic relationship with the host. Wiley Interdiscip Rev Syst Biol Med 2019; 11: e1438. doi.org/10.1002/wsbm.1438.
24. Raimondi S, Amaretti A, Gozzoli C, Simone M, Righini L, Candeliere F, Brun P. Longitudinal survey of fungi in the human gut: its profiling, phenotyping, and colonization. Front Microbiol 2019; 10: 1575. doi.org/10.3389/fmicb.2019.01575.
25. Hooks KB, O'Malley MA. Contrasting strategies: human eukaryotic versus bacterial microbiome research. J Eukaryotic Microbiol 2019; 67: 279-295. doi.org/10.1111/jeu.12766.
26. Mukherjee C, Beall CJ, Griffen AL, Leys EJ. High-resolution ISR amplicon sequencing reveals personalized oral microbiome. Microbiome 2018; 6: 153. doi.org/10.1186/s40168-018-0535-z.
27. Sze MA, Schloss PD, McMahon K. The impact of DNA polymerase and number of rounds of amplification in PCR on 16s rRNA gene sequence data. mSphere 2019; 4. doi.org/10.1128/mSphere.00163-19.
28. Sze MA, Schloss PD. Leveraging existing 16s rRNA gene surveys to identify reproducible biomarkers in individuals with colorectal tumors. mBio 2018; 9. doi.org/10.1128/mBio.00630-18.
29. Nearing JT, Douglas GM, Comeau AM, Langille MGI. Denoising the denoisers: an independent evaluation of microbiome sequence error-correction approaches. Peer J 2018; 6: e5364. doi.org/10.7717/peerj.5364.
30. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 2016; 13: 581-583. doi.org/10.1038/nmeth.3869.

31. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2017; doi.org/10.1128/msystems.00191-16.
32. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv 2016; 081257. doi.org/10.1101/081257.
33. Eren AM, Borisy GG, Huse SM, Mark Welch JL. Oligotyping analysis of the human oral microbiome. Proc Natl Acad Sci USA 2014; 111: E2875-E2884. doi.org/10.1073/pnas.1409644111.
34. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 2018; 36: 996-1004. doi.org/10.1038/nbt.4229.
35. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. Nucl Acids Res 2014; 42: D643-D648. doi.org/10.1093/nar/gkt1209.
36. Zheng J, Wittouck S, Salvetti E, Franz CMAP, Harris HMB, Mattarelli P, O'Toole PW, Pot B, Vandamme P, Walter J, Watanabe K, Wuyts S, Felis GE, Gänzle MG, Lebeer S. A taxonomic note on the genus Lactobacillus: Description of 23 novel genera, emended description of the genus Lactobacillus Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. Int J Syst Evol Microbiol 2020; 70: 2782-2858. doi.org/10.1099/ijsem.0.004107.
37. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 2019; doi.org/10.1093/bioinformatics/btz848.
38. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 2013; 41: D590-596. doi.org/10.1093/nar/gks1219.
39. Rinke C, Chuvochina M, Mussig AJ, Chaumeil PA, Waite DW, Whitman WB, Parks DH, Hugenholtz P. A rank-normalized archaeal taxonomy based on genome phylogeny resolves widespread incomplete and uneven classifications. bioRxiv 2020; 2020.03.01.972265. doi.org/10.1101/2020.03.01.972265.
40. Kaehler BD, Bokulich NA, McDonald D, Knight R, Caporaso JG, Huttley GA. Species abundance information improves sequence taxonomy classification accuracy. Nat Commun 2019; 10: 4643. doi.org/10.1038/s41467-019-12669-6.
41. Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the operon. F1000Res 2018; 7: 1755. doi.org/10.12688/f1000research.16817.1.
42. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nature 2019; 10: 1-11. doi.org/10.1038/s41467-019-13036-1.
43. Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, Knight R, Albertsen M. Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. bioRxiv 2020; 645903. doi: 10.1101/645903.
44. Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. Nat Biotechnol 2018; 36: 190-195. doi: 10.1038/nbt.4045.
45. Fuks G, Elgart M, Amir A, Zeisel A, Turnbaugh PJ, Soen Y, Shental N. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. Microbiome 2018; 6: 17. doi.org/10.1186/s40168-017-0396-x.
46. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nature Biotechnol 2017; 35. doi.org/10.1038/nbt.3935.
47. Malla MA, Dubey A, Kumar A, Yadav S, Hashem A, Abd-Allah EF. Exploring the human microbiome: the potential future role of next-generation sequencing in disease diagnosis and treatment. Front Immunol 2019; 9: 2868. doi.org/10.3389/fimmu.2018.02868.
48. Sevim V, Lee J, Egan R, Clum A, Hundley H, Lee J, R. Everroad C, Detweiler AM, Bebout BM, Pett-Ridge J, Göker M, Murray AE, Lindemann SR, Klenk HP, O'malley R, Zane M, Cheng JF, Copeland A. Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. Sci Data 2019; 6: 1-9. doi.org/10.1038/s41597-019-0287-z.
49. Korostin D, Kulemin N, Naumov V, Belova V, Kwon D, Gorbachev A. Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole genome sequencing. PLoS One 2020; 15: e0230301. doi.org/10.1371/journal.pone.0230301.
50. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, Kenneth BKnights D. Evaluating the information content of shallow shotgun metagenomics. mSystems 2018; 3. doi.org/10.1128/mSystems.00069-18.
51. Sanders JG, Nurk S, Salido RA, Minich J, Xu ZZ, Zhu Q, Martino C, Fedarko M, Arthur TD, Chen F, Boland BS, Humphrey GC, Brennan C, Sanders K, Gaffney J, Jepsen K, Khosroheidari M, Green C, Liyanage M, Dang JW, Phelan VV, Quinn RA, Bankevich A, Chang JT, Rana TM, Conrad DJ, Sandborn WJ, Smarr L, Dorrestein PC, Pevzner PA, Knight R. Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. Genome Biol 2019; 20: 226. doi.org/10.1186/s13059-019-1834-9.
52. Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, Pop M. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief Bioinformatics 2019; 20: 1140-1150. doi.org/10.1093/bib/bbx098.
53. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019; 20: 257. doi.org/10.1186/s13059-019-1891-0.

54. Zhong C, Yang Y, Yooseph S. GRASP2: fast and memory-efficient gene-centric assembly and homolog search for metagenomic sequencing data. BMC Bioinformatics 2019; 20: 276. doi.org/10.1186/s12859-019-2818-1.
55. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat Methods 2019; 16: 603-606. doi.org/10.1038/s41592-019-0437-4.
56. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. Peer J 2019; 7: e7359. doi.org/10.7717/peerj.7359.
57. Ma T, Xiao D, Xing X. MetaBMF: a scalable binning algorithm for large-scale reference-free metagenomic studies. Bioinformatics 2020; 36: 356-363. doi.org/10.1093/bioinformatics/btz577.
58. Kouchaki S, Tapinos A, Robertson DL. A signal processing method for alignment-free metagenomic binning: multi-resolution genomic binary patterns. Sci Rep 2019; 9: 2159. doi.org/10.1038/s41598-018-38197-9.
59. Danko D, Meleshko D, Bezdan D, Mason C, Hajirasouliha I. Minerva: an alignment and reference free approach to deconvolve linked-reads for metagenomics. Genome Res 2018; gr.235499.118. doi.org/10.1101/gr.235499.118.
60. Arumugam K, Bačı C, Bessarab I, Beier S, Buchfink B, Górska A, Qiu G, Huson DH, Williams RBH. Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data. Microbiome 2019; 7: 61. doi.org/10.1186/s40168-019-0665-y.
61. Dilthey AT, Jain C, Koren S, Phillippy AM. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. Nat Commun 2019; 10: 3066. doi.org/10.1038/s41467-019-10934-2.
62. Damme R Van, Hölzer M, Viehweger A, Müller B, Bongcam-Rudloff E, Brandt C. Metagenomics workflow for hybrid assembly, differential coverage binning, transcriptomics and pathway analysis (MUFFIN). bioRxiv 2020; 2020.02.08.939843. doi.org/[10.1101/2020.02.08.939843.
63. Kolmogorov M, Rayko M, Yuan J, Polevikov E, Pevzner P. MetaFlye: scalable long-read metagenome assembly using repeat graphs. bioRxiv 2019; 637637. doi.org/10.1101/2020.03.12.974238.
64. Arumugam K, Bessarab I, Haryono MAS, Liu X, Zuniga-Montanez RE, Roy S, Qiu G, Moses DI, Law Y, Wuertz S, Lauro FM, Huson D, Williams RBH. Analysis procedures for assessing recovery of high quality, complete, closed genomes from Nanopore long read metagenome sequencing. bioRxiv 2020; 2020.03.12.974238. doi.org/10.1101/637637.
65. Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, Dvornicic M, Soldo JP, Koh JY, Tong C, Ng OT, Barkham T, Young B, Marimuthu K, Chng KR, Sikic M, Nagarajan N. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nat Biotechnol 2019; 37: 937-944. doi.org/10.1038/s41587-019-0191-2.
66. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E, Bremges A, Fritz A, Garrido-Oter R, Jørgensen TS, Shapiro N, Blood PD, Gurevich A, Bai Y, Turaev D, DeMaere MZ, Chikhi R, Nagarajan N, Quince C, Meyer F, BalvoČtČM, Hansen LH, Sørensen SJ, Chia BKH, Denis B, Froula JL, Wang Z, Egan R, Don Kang D, Cook JJ, Deltel C, Beckstette M, Lemaitre C, Peterlongo P, Rizk G, Lavenier D, Wu YW, Singer SW, Jain C, Strous M, Klingenberg H, Meinicke P, Barton MD, Lingner T, Lin HH, Liao YC, Silva GGZ, Cuevas DA, Edwards RA, Saha S, Piro VC, Renard BY, Pop M, Klenk HP, Göker M, Kyrpides NC, Woyke T, Vorholt JA, Schulze-Lefert P, Rubin EM, Darling AE, Rattei T, McHardy AC. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat Methods 2017; 14: 1063-1071. doi.org/10.1038/nmeth.4458.
67. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol 2017; 8: 2224. doi.org/10.3389/fmicb.2017.02224.
68. Cruz GNF, Christoff AP, Oliveira LFV de. Equivolumetric protocol generates library sizes proportional to total microbial load in next-generation sequencing. bioRxiv 2020; 2020.02.03.932301. doi.org/10.1101/2020.02.03.93230.
69. Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, Zengler K, Knight R. Establishing microbial composition measurement standards with reference frames. Nat Commun 2019; 10:2719. doi.org/10.1038/s41467-019-10656-5.
70. Li Z, Tian L, O'Malley AJ, Karagas MR, Hoen AG, Christensen BC, Madan JC, Wu Q, Gharaibeh RZ, Jobin C, Li H. IFAA: Robust association identification and inference for absolute abundance in microbiome analyses. ArXiv 2019; arXiv: 190910101.
71. Brill B, Amir A, Heller R. Testing for differential abundance in compositional counts data, with application to microbiome studies. ArXiv 2020; arXiv:190408937.
72. Tang Z-Z, Chen G. Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. Biostatistics 2019; 20: 698-713. doi.org/10.1093/biostatistics/kxy025.
73. Pendegraft AH, Guo B, Yi N. Bayesian hierarchical negative binomial models for multivariable analyses with applications to human microbiome count data. PLoS One 2019; 14: e0220961. doi.org/10.1371/journal.pone.0220961.
74. Banerjee K, Zhao N, Srinivasan A, Xue L, Hicks SD, Middleton FA, Wu R, Zhan X. An adaptive multivariate two-sample test with application to microbiome differential abundance analysis. Front Genet 2019; 10: 350. doi.org/10.3389/fgene.2019.00350.
75. Martino C, Morton JT, Marotz CA, Thompson LR, Tripathi A, Knight R, Zengler K. A Novel sparse compositional technique reveals microbial perturbations. mSystems 2019; 4. doi.org/10.1128/mSystems.00016-19.
76. Diener C, Gibbons SM, Resendis-Antonio O. MICOM: Metagenome-scale modeling to infer metabolic interactions in the gut microbiota. mSystems 2020; 5. doi.org/10.1128/mSystems.00606-19.
77. Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, Badri MH, Swenson TL, Van Goethem MW, Northen TR, Vazquez-Baeza Y, Wang M, Bokulich NA, Watters A, Song SJ, Bonneau R, Dorrestein PC, Knight R. Learning representations of microbe-metabolite interactions. Nat Methods 2019; 16: 1306-1314. doi.org/10.1038/s41592-019-0616-3.

78. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. 8. Nature Biotechnol 2019; 37: 852-857. doi.org/10.1038/s41587-019-0209-9.