3

Collecting samples for metagenomics

3.1 Nonmicrobial components in the sample that could influence DNA extraction and sequencing amount

The Bristol's stool score (BSS), whereby fecal samples are graded according to shape (Fig. 3.1A), is a useful approximation for water content, gut transit time, as well the number of bacterial cells per gram of feces [1–3]. With illustrations in questionnaires, volunteers can fill out the BSS as they collect their own feces. Such self-reported BSS showed correlations with enzymes detected in the fecal metagenome that metabolize secondary bile acids [3], which makes mechanistic sense.

A mountable toilet system can automatically record values such as fecal BSS, volume, and flow rate of urine, but the published version does not collect samples yet [4].

The number of microbial cells per gram of feces could be affected by the presence of food debris, which becomes substantial when the diet is as fibrous as that of a panda (Fig. 3.1A). A study of fecal samples from healthy adults on a British diet (1980), with vigorous agitation and detergent treatment of the feces, reported the fecal mass to be 55% bacteria, 17% fiber, and 24% soluble material. Proportion of the dry mass can actually be different with different microbes, but we are not too worried about that.

Nylon swabs with detachable ends can be used for fecal and mucosal samples (Fig. 3.1B). Swabs for skin and nasal samples are typically wetted with physiological saline or buffer solutions before sampling, but it is not clear how such transient exposure to minerals might change the microbial community. Larger volumes may be necessary for some low-biomass samples. Make sure to screw the lids tightly, to avoid liquid leaks or drying during transport. If the samples are going to be collected by volunteers themselves, clear illustrations or videos

Investigating Human Diseases with the Microbiome: Metagenomics Bench to Bedside.
https://doi.org/10.1016/B978-0-323-91369-0.00001-7

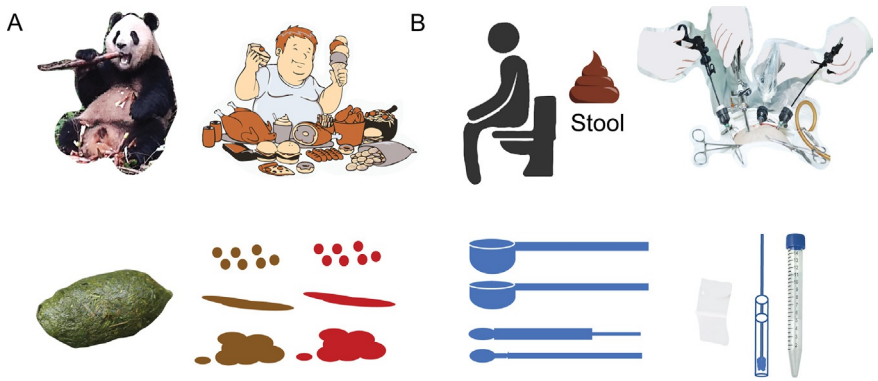

Fig. 3.1 Getting enough microbes when collecting samples. (A) Metagenomic samples may contain materials other than microbes. The human feces are illustrated on a simplified Bristol's Stool Scale (BSS), the hardest has a BSS of 1, a BSS of 4 in the middle, and the watery form has a BSS of 7. (B) Scoops or swabs for solid samples or surfaces, and tubes for liquids. Laparoscopes or other new technologies that make a small cut are presumably less prone to contamination for the collection of microbiome samples in the operation room. Skin and other surfaces are decontaminated. Swabs or brushes can be in protective tubing before reaching a sampling site. Credit: Huijue Jia, Xin Tong of BGI-Shenzhen.

for the procedure would be a good idea, as well as photo records for the sample. Compliance rates could vary between cohorts. Quality checks should be performed early.

Fecal samples can become bloody with inflammatory bowel diseases, hemorrhoids, etc., making the proportion of human reads in a metagenomic sample to go much higher than 1% (Table 3.1). Human microbiome samples other than feces and supragingival plaques all have a higher percentage of the human genomic sequence, going over 99% in some tissue samples (Table 3.1). With informed consent, the low-depth human genome sequence can be analyzed together with the microbiome, while beware of tissue differences in comparison to blood.

Removal of host cells by low-speed centrifugation has been shown to lead to loss of bacterial species in bronchoalveolar lavage samples (BAL) [17]. Experimental removal of host sequences using molecular biology or chemical ways also affects the microbiome composition [15,18] but may be optimized in the future. So we currently recommend bioinformatic removal after unbiased sequencing.

The combination of technologies used depends on the questions we would like to address (e.g., bacteria in pancreatic cancer, Chapter 1, Fig. 1.8). For a good sample, it would be a pity not to fully investigate its potential, just because the sequencing amount appears intimidating. After all, a lot of the microbial evolution and interactions would be local.

Table 3.1 Varying percentage of human sequences in shotgun metagenomic data of samples from different body sites.

Body region	Body site	% human sequences	Reference
Gut	Feces	1%	[5,6]
Gut	Feces (Crohn's disease)	20% or more in some samples	[7]
Oral	Buccal mucosa	82%–90%	[5,8,9]
Oral	Supragingival plaque	40%; 5.55%	[5,10]
Oral	Subgingival plaque	79%	[5]
Oral	Tongue dorsum	30%	[9,11]
Oral	Saliva	77%–91%	[5,11]
Skin	Dry (e.g., volar forearm)	36%	[12]
Skin	Moist (e.g., antecubital fossa)	44%	[12]
Skin	Sebaceous (e.g., retroauricular crease)	59%–73%	[5,12]
Urogenital	Vagina (including the posterior fornix)	90%–98%	[5,13]
Urogenital	Cervical orifice	98%	[14]
Urogenital	Peritoneal fluid	99.8%	[15]
Urogenital	Placenta	> 99%	[16]
Airway	Anterior nares	96%	[5]

Approximate average values are summarized from multiple studies, and if different, separated by ";". This is not meant to be an exhaustive list. It should serve the purpose for a general understanding, when researchers and clinicians decide on the sequencing amount for metagenomic shotgun sequencing (e.g., 10 Gb of paired-end 100 bp reads), or opt for amplicon sequencing, in situ hybridization, etc. Credit: Huijue Jia.

3.2 Beware of contamination in each step, from stools to low-biomass samples

As briefly noted in Chapter 1, claims for the presence of a microbiome in samples with a low microbial biomass, e.g., placenta samples (Chapter 1, Fig. 1.7), are greeted with skepticism, but were more readily accepted in tumor samples (e.g., Chapter 1, Fig. 1.8) [19–24]. Studying these low biomass samples helps us reach a clear understanding of things to take care of for all kinds of metagenomic samples (Fig. 3.2). On the other hand, it may turn out to be more difficult to prove the absence of microbes at a site, than the presence. What kind of extreme environment can be free of any bacteria, archaea, and fungi?

For fecal, oral (Chapter 2, Fig. 2.6), and vaginal samples, there are typically over 10^{10} microbial cells (cfu, colony forming units on a culture plate) in a swab sample. This high number of microbial genomes relieves the caution for contamination during sampling, DNA extraction, library construction, and sequencing. For low biomass samples (Figs. 3.2 and 3.3), the reagents used for sampling and

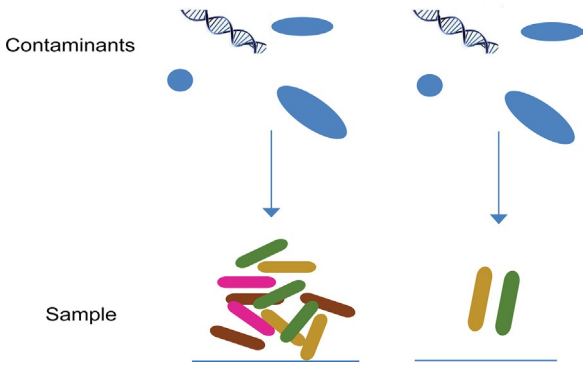

Fig. 3.2 Contaminating DNA or microbes can be introduced at any step from sampling to sequencing. It is just that samples with low biomass are more easily overwhelmed by the presence of contaminants. Credit: Huijue Jia, Xin Tong of BGI-Shenzhen.

subsequent steps, and the facility should all be routinely checked for the presence of live or dead microbes [25], e.g., by using amplicon sequencing (Fig. 3.4). The most cited study on contamination performed serial dilution of *Salmonella bongori*, and other taxa dominated the 16S rRNA gene amplicon results at the 5th serial dilution, corresponding to about 10^3 *Salmonella bongori* cells [26]. While the focus of the study was on reagents contamination [26], the author could not find information regarding the pipettes for dilution (the PCR amplification was performed in a hood using autoclaved microcentrifuge tubes and filtered pipette tips) and the protection from human contamination, as is well known in the ancient DNA field. More caution against batch effects should also be taken with amplicon sequencing (e.g., [27]).

Hospital rooms, or even space station rooms, do accumulate microbes after being used [28–32], with potential microbiome contribution from both patients and staff. Air and ventilation systems in a hospital room may show up positive with bacteria or fungi (Fig. 3.3). As long as they are different from the samples of interest (e.g., Worked sample 3.3), one can still tentatively believe that there is a microbiome in the samples and look for more evidence. The negative controls also apply to live culture on plates, and quantitative polymerase chain reactions (qPCR).

In samples dominated by human sequences, however, amplicon sequencing may pick up lots of human sequences that were nonspecifically amplified (e.g., in kidney stones [33]). Besides optimizing the PCR condition, such samples may need to be purified according to fragment length, or subject to targeted sequencing, while controlling for contamination.

Having multiple samples from the same individual, including samples from different body sites, could also lend some support that the microbes detected are not random contamination, or systematic contamination in reagents, but show patterns that are specific for each individual [15,34,35]. When studying the female reproductive tract in nonpregnant women without inflammation, our collaborating

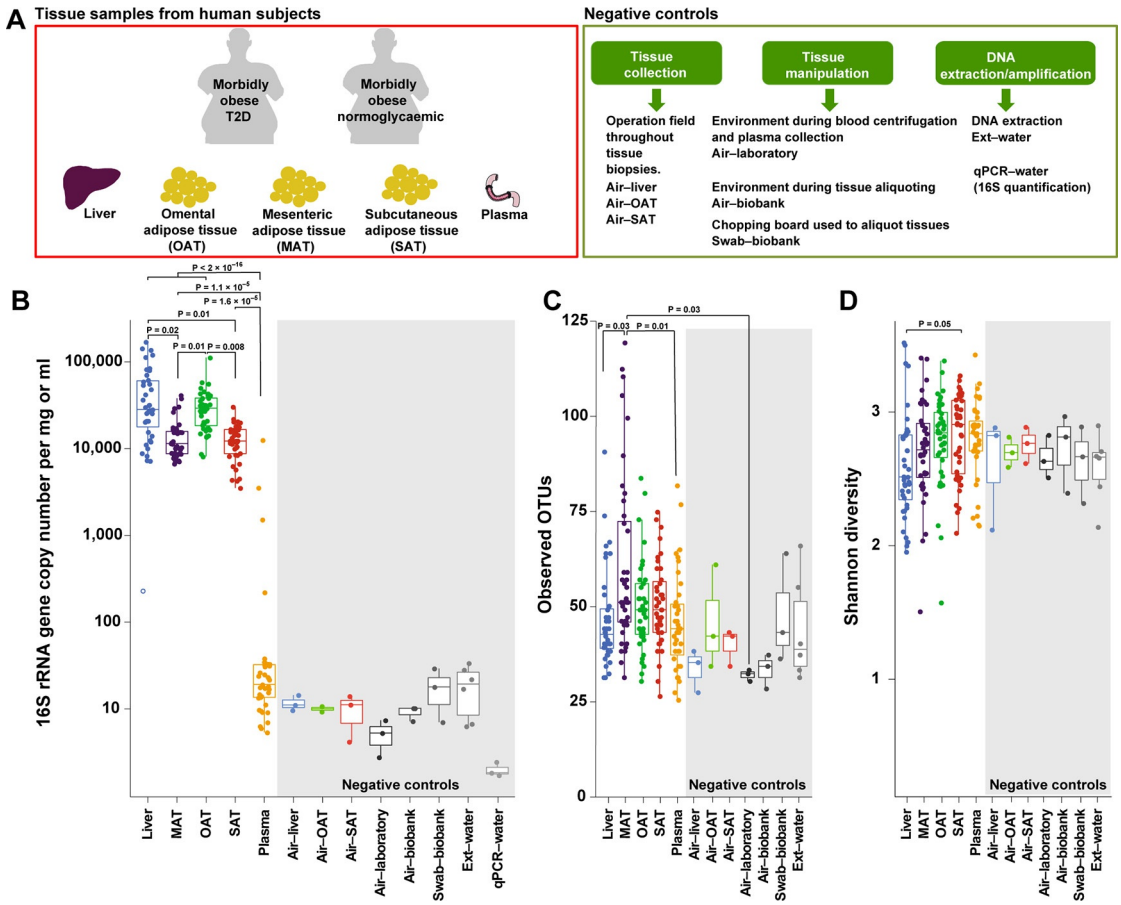

Fig. 3.3 An example of study design for low-biomass samples, including negative controls as well as different physiological conditions for association study. (A) Liver, three different adipose tissue depots (OAT, MAT, and SAT), and plasma samples were collected from individuals with morbid obesity who had T2D ($n=20$) and from those who had normoglycaemia ($n=20$). DNA extraction and amplification procedures were carried out using optimized conditions for bacterial DNA detection in blood plasma and tissues. A comprehensive set of negative controls was tested to control for environmental sample contamination at major steps in the analysis: tissue collection, tissue manipulation, and DNA extraction and amplification. During tissue collection, tubes were kept open next to the operation field throughout the entire procedure (air-liver, air-OAT, and air-SAT). Contamination coming from tissue manipulation was controlled by another set of tubes that were kept open next to the operator throughout blood centrifugation and plasma collection (air-laboratory) as well as during tissue aliquoting (air-biobank). The chopping board used to aliquot tissues was sampled prior to tissue manipulation (swab-biobank). Water samples were used to control for labware, reagent, and/or environmental contamination during DNA extraction (ext-water) and amplification steps for tissue 16S rRNA quantification by quantitative PCR (qPCR-water). After thorough validation of negative controls on a case-by-case basis, 16S quantification and sequencing data were used in the discovery of tissue-specific bacterial signatures linked to T2D. (B–D) Number of bacteria across body sites. (B) 16S rRNA gene counts. (C) Observed OTUs. (D) Shannon index in the liver, three different adipose tissue depots (OAT, MAT, and SAT), and plasma of participants with obesity. Negative controls were tested to control for environmental sample contamination at major steps in the analysis: tissue collection (air-liver, air-OAT, air-SAT), tissue manipulation (air-laboratory, air-biobank, and swab-biobank), and DNA extraction or amplification (ext-water, qPCR-water). In panels (B–D), groups were compared using a Kruskal-Wallis one-way ANOVA followed by Dunn's test for pairwise comparison. Box plots depict the first and the third quartile with the median represented by a vertical line within the box; the whiskers extend from the first and third quartiles to the highest and lowest observation, respectively, not exceeding 1.5×IQR. Credit: Cropped from Fig. 1, Fig. 2 of Anhê FF, Jensen BAH, Varin TV, Servant F, Van Blerk S, Richard D, et al. Type 2 diabetes influences bacterial tissue compartmentalisation in human obesity. Nat Metab 2020;2:233–42. https://doi.org/10.1038/s42255-020-0178-9.

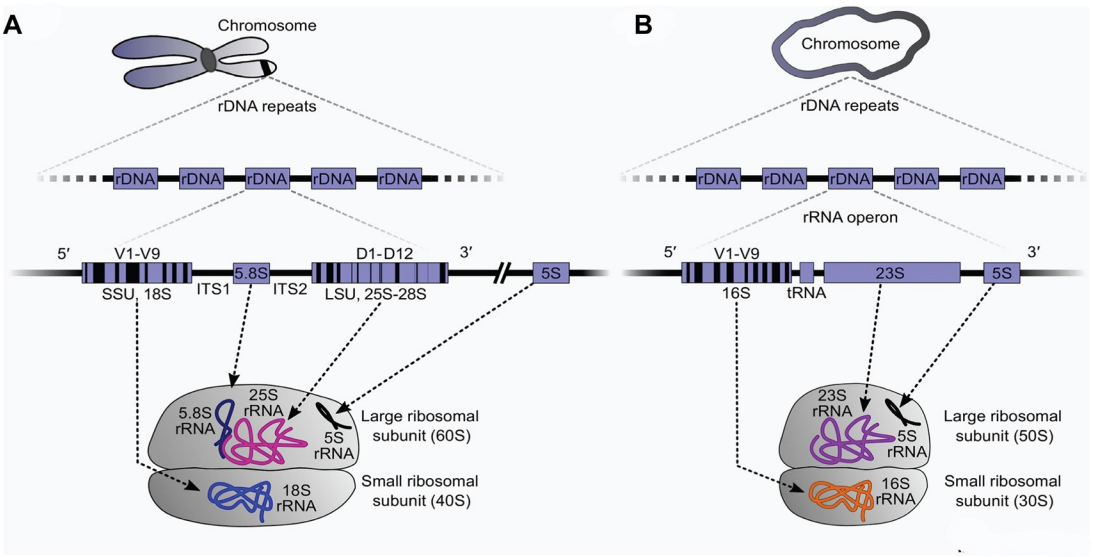

Fig. 3.4 Schematic representation of the ribosomal RNA (rRNA) gene cluster (or rDNA). The variable regions of (A) eukaryotic and (B) prokaryotic rRNA loci are commonly used to characterize microbial taxa and resolve their phylogenetic relationships by amplicon sequencing and analyses. In most fungi, the rRNA gene cluster includes the small ribosomal subunit (SSU, 18S), with internal transcribed spacer regions (ITS1 and ITS2) flanking the 5.8S, and large ribosomal subunit (LSU, 25–28S) regions. In bacteria, the rRNA operon comprises the SSU (16S), LSU (23S), and 5S loci. *Black vertical lines* in serial order illustrate the variable regions in SSU (V1–V9) and LSU (D1–D12), best suited for biodiversity assessments through microbial communities profiling. Credit: From Fig. 1 of Lavrinienko A, Jernfors T, Koskimäki JJ, Pirttilä AM, Watts PC. Does intraspecific variation in rDNA copy number affect analysis of microbial communities? Trends Microbiol 2021;29:19–27. https://doi.org/10.1016/j.tim.2020.05.019.

gynecologist Dr. Ruifang Wu made sure that after the subcentimeter cut on the skin, her team started with the peritoneal fluid from the pouch of Douglas, before the laparoscope moved on to the fallopian tubes (e.g., for those with an obstruction there), and then the endometrium (fibroids, endometriosis, or adenomyosis) that we suspected to have a denser microbial community. The vaginal and cervical samples were taken on the day of the initial visit to the clinic, with protective tubing for the cervical mucus sampling before it reached into the cervix. Despite the overdominance of Lactobacilli, the vaginal and cervical samples nicely matched upper reproductive tract samples that were collected a few days later in the operation room for each volunteer [34,35]. The peritoneal fluid, with a more neutral pH, might also be a source community (Chapter 2) of low biomass but diverse microbes [34,35].

The placenta microbiome has remained controversial (Chapter 1, Fig. 1.7), as it is pushing the detection limit of existing methods. According to fluorescent in situ hybridization (FISH) against conserved regions in the 16S rRNA, Dr. Kjersti M. Aagaard reported that small clusters of bacteria were mostly localized to the villous parenchyma or syncytiotrophoblast, and less commonly in the chorion and the ma-

ternal intervillous space [36]. A recent study collected samples from the placental terminal villi, required bacteria to be detected by both 16S rRNA gene amplicon sequencing (V1–V2 region, Fig. 3.4) and metagenomic shotgun sequencing, and discredited results identified in controls including vaginal samples [25]. Only the neonatal pathogen *Streptococcus agalactiae* (Group B *Streptococcus*) remained as a placental microbe [25], which we could detect in the cervix [14].

Contamination from blood has always been a concern (sampled in the adipose tissue microbiome study, Fig. 3.3). The sequencing amount in the placenta study was so low (26.5 million reads on average, > 99% human genome (Table 3.1); e.g., if 99.9% human, there would be 26.5 thousand nonhuman reads per sample on average, less than 1 × coverage for the genome of a single bacterium) that broad functional capacity seen from KEGG level 2 pathways fluctuated among samples [16]. For cervical orifice samples, we observed a correlation between the dominant bacteria species and the percentage of human sequences [9,14]. However, some of the taxa identified in the original study may still be genuinely present in the placenta. The placenta microbiome in this metagenomic study showed a high relative abundance of *Escherichia coli* [16], which is known in the meconium (newborn feces) [37–41], along with potentially gut or oral bacteria such as *Bacteroides*, *Streptococcus parasanguinis*, *Prevotella melaninogenica*, reproductive tract bacteria such as *Cutibacterium acnes* (renamed from *Propionibacterium acnes*), *Lactobacillus iners*, *L. crispatus*, which were more or less recapitulated in the abovementioned recent study with plenty of controls [25]. Bacteria such as *Cutibacterium acnes* and *Streptococcus parasanguinis* are also part of the infant oral and gut microbiome [16,38,40,42]. And if the species assignment for *Neisseria lactamica* holds true despite the low sequencing coverage [16] (more on taxonomy in Chapter 5), it will provide a potential explanation for carriage of the bacteria in the nasopharynx of young children. More taxa have been detected by 16S rRNA gene amplicon sequencing [36], which was not overloaded with the human sequences.

According to interesting studies on bacteria in adipose tissues and their dysbiosis in T2D patients, apparent bacterial diversity in plasma and in negative controls was not lower than in adipose tissue samples, but the copy numbers of the total 16S rRNA gene were orders of magnitude higher than the negative controls (Fig. 3.3) [43,44]. The per mg (milligram) tissue unit [44] looks better than the per μg DNA unit [43], because the latter includes the human genome (2 copies of 3.2 GB). The diameter of white adipocytes varies from less than 30 μm to over 300 μm [45]. Try some calculations in light of Chapter 1. If these are all very small adipocytes with a diameter of 20 μm, the bacteria cell (~ 16S rRNA gene copy number in Fig. 3.3, per unit tissue weight or volume [44]) to human cell ratio would be roughly 1:10. With a more mediocre adipocyte diameter of 100 μm, however, the bacteria cell to human cell ratio would be flip to about 10:1.

As we move on to study the brain, the lungs (Figs. 3.5 and 3.6), or other tissues, are we collecting all the related samples, and in a good order? For bronchoalveolar lavage (BAL) samples, the good news is that there appeared to be no difference between samples taken via the mouth versus samples taken via the nose (with protective tubing) [46],

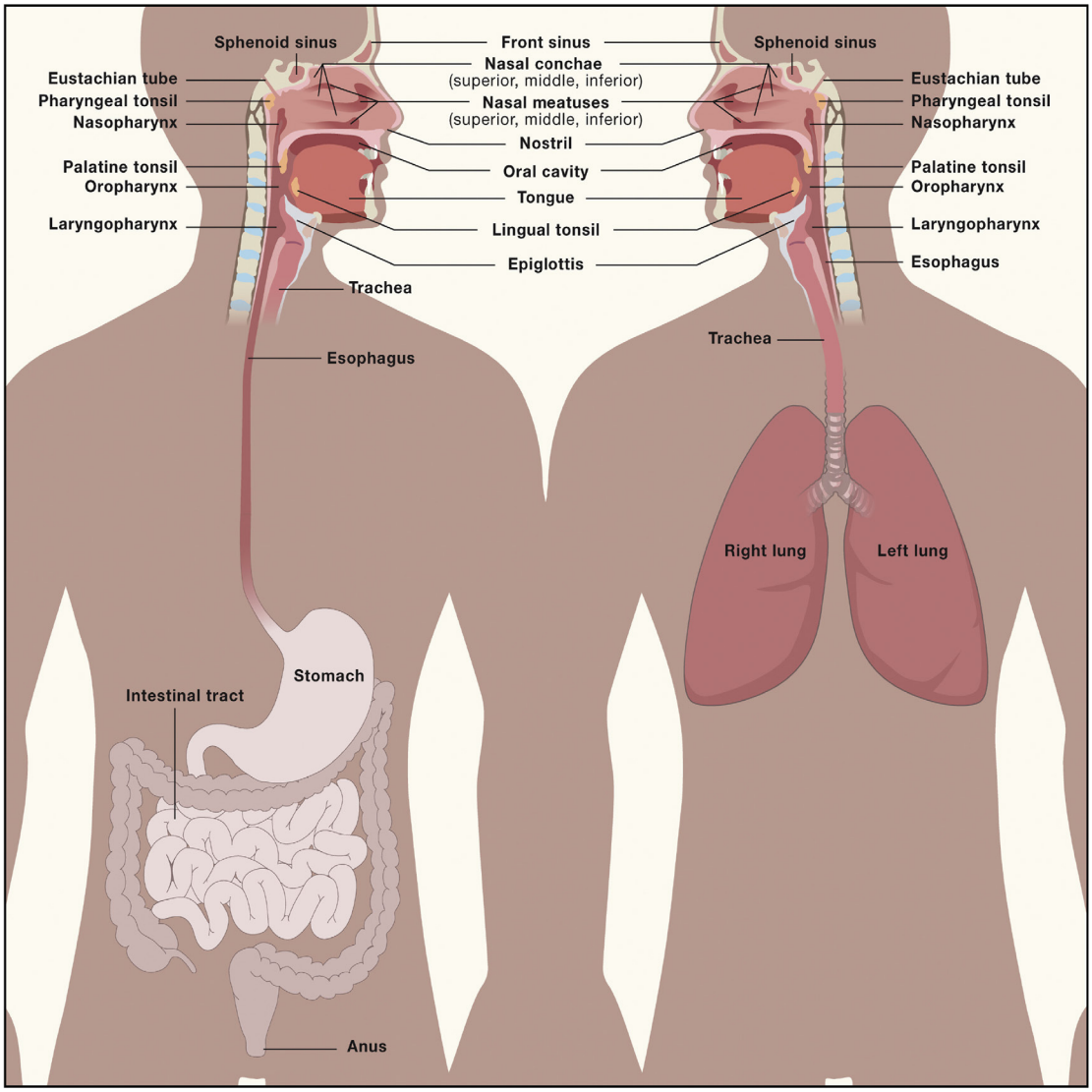

Fig. 3.5 Understanding microbial dispersal across the human body. A schematic illustrating some of the distal body sites to which microbes may disperse from the nose, mouth, or throat, as well as some interconnections among sites. For example, the pharyngeal tonsils, also known as the adenoids (a major site of lymphoid tissue in the oral/nasal pharyngeal), may serve as a reservoir for middle ear infections as a result of dispersal via the eustachian tube. Credit: Fig. 3 of Proctor DM, Relman DA. The landscape ecology and microbiota of the human nose, mouth, and throat. Cell Host Microbe 2017;21:421–32. https://doi.org/10.1016/j.chom.2017.03.011.

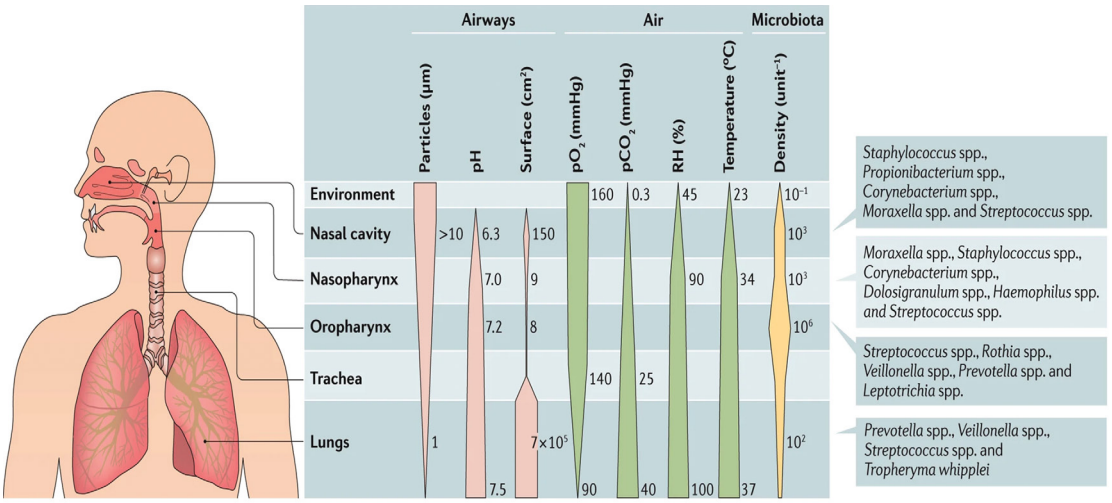

Fig. 3.6 Physiological and microbial gradients along the respiratory tract. Physiological and microbial gradients exist along the nasal cavity, nasopharynx, oropharynx, trachea, and the lungs. The pH gradually increases along the respiratory tract, whereas most of the increases in relative humidity (RH) and temperature occur in the nasal cavity Furthermore, the partial pressures of oxygen (pO_2) and carbon dioxide (pCO_2) have opposing gradients that are determined by environmental air conditions and gas exchange at the surface of the lungs. Inhalation results in the deposition of particles from the environment into the respiratory tract; inhaled particles that are more than 10 µm in diameter are deposited in the upper respiratory tract, whereas particles less than 1 µm in diameter can reach the lungs. These particles include bacteria-containing and virus-containing particles, which are typically larger than 0.4 µm in diameter. These physiological parameters determine the niche-specific selective growth conditions that ultimately shape the microbial communities along the respiratory tract. The unit by which bacterial density is measured varies per niche; the density in the environment is depicted as bacteria per cm^3 (indoor) air, density measures in the nasal cavity and nasopharynx are shown as an estimated number of bacteria per nasal swab, and the densities in the oropharynx and the lungs represent the estimated number of bacteria per ml of oral wash or bronchoalveolar lavage (BAL), respectively. Credit: Fig. 1 of Man WH, de Steenhuijsen Piters WAA, Bogaert D. The microbiota of the respiratory tract: gatekeeper to respiratory health. Nat Rev Microbiol 2017;15:259–70. https://doi.org/10.1038/nrmicro.2017.14.

although the related samples should be compared pairwise for each individual instead of shown in a crude PCA (Principle Component Analysis). How many microbial cells (Chapter 1) are we going to have in the samples? For the viruses and fungi, would we need other information to better understand the local habitat and morphology?

3.3 Reagents that prevent microbial growth after sampling

In the early days of metagenomic studies, fecal samples with a relative abundance of the facultative anaerobe *E. coli* that exceeded ~ 30% were often suspected of prolonged exposure to room temperature, but they can reflect genuine conditions such as colorectal cancer, Crohn's disease, IgA deficiency, Type 2 diabetes [7,47–50].

Nowadays, researchers no longer have to ask volunteers to temporarily store feces in their household fridge, or have dry ice at the clinic every day. The freezing procedure can also affect the composition of the metagenomic sample, e.g., due to pH and other concentration changes during crystallization of water, and components in the metagenomic sample can affect the freezing efficiency. Commercial reagents are available that allow room-temperature preservation of microbiome samples for 2 or 4 weeks (e.g., from DNA Genotek Inc., Mawi DNA Technologies LLC., MGIEasy from MGI Tech Co. Ltd.), much longer than delivery time for courier mails in many places. Filter paper has also been used for fecal and cervical samples, which are air-dried after sampling and then sealed, without much consensus for how to minimize contamination. One has to make sure if the amount of DNA from filter paper is sufficient for shotgun sequencing, not just 16S rRNA gene amplicon sequencing.

There are not enough studies published for metatranscriptomics, which in addition to high-quality preservation of RNA, also require (however incomplete) removal of ribosomal RNAs before sequencing [51,52]. Metaproteomics is also on the rise, and we have only tried fresh or frozen samples.

The stabilizing reagents only stop bacterial growth and decay, and do not kill bacteria. So at least some of the microbes could still be cultured on a plate. Similarly, analyzing the metabolome of the same samples using mass spectrometry typically requires the swabs to be in a reagent (e.g., 50:50 ethanol:water for skin swabs [53]) different from the one used for microbiome storage, but there are commercial products that try to work for both purposes.

Worked sample 3.1

Please think of a kind of microbiome sample you are interested in.

What is the current estimate for the number of microbial cells and the number of species at this body site? Do you expect the numbers to change for the disease in question?

How much DNA would you need for constructing a metagenomic library (e.g., 0.5 μg)? Would you use a swab or some other plasticware?

Would you be able to process or freeze the samples immediately (e.g., save some for other omics in the future)? If a commercial reagent would be used to preserve the samples for a few weeks at room temperature and possibly through courier mail, do you think it might affect some microbes more than the others in this particular microbial community?

Would you have a standard mix and real samples to compare the fresh, frozen, and reagent-preserved community?

3.4 DNA extraction for metagenomic samples

Even in the absence of too many plant fibers or protists (Fig. 3.1), DNA extraction for metagenomic samples is more complex than DNA extraction for mammalian cells or single microbial species, because of the myriad of cell walls that we all try to break (Fig. 3.7) [54–56]. Physical and chemical steps can all be added, and control samples could be used in parallel or as a quantitative spike-in, if sufficiently different in sequence. Nanopore sequencing that requires the DNA fragments to be more than 20 kb long would be a different story.

For bead beating, the harder tetragonal Zirconium polycrystals are superior to glass beads. Size of the beads also needs to be considered for the bacteria and fungi in the sample. Smaller beads make more contact points within a given time but may not have enough momentum to break some fungi.

Unlike amplicon sequencing, metagenomic shotgun sequencing workflows that do not involve any PCR steps can better tolerate impurities in DNA. For example, DNA from the fecal sample may still come out a little yellowish.

Automated platforms (e.g., 96 samples per plate) can achieve a more uniform quality of sample processing, and reduce the chance of random contamination, in addition to saving time. Such automation could also better protect the staff from whatever is in the clinical samples.

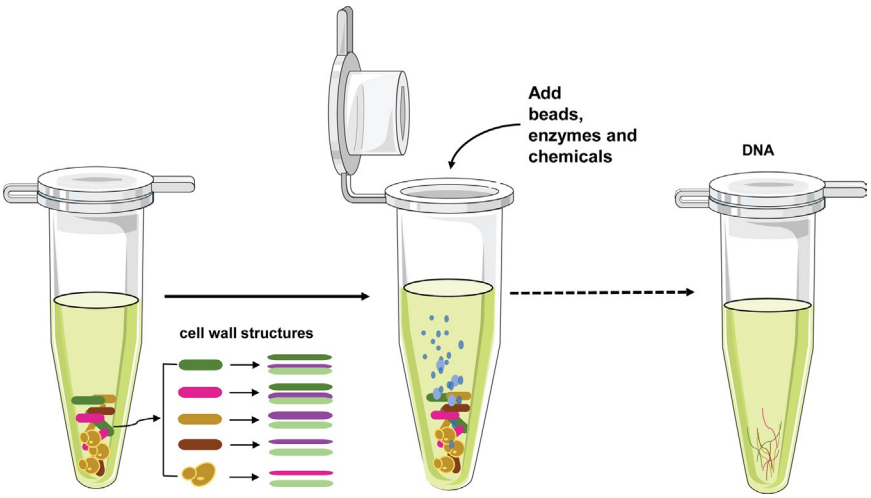

Fig. 3.7 DNA extraction for a complex community of microorganisms. The thick lines represent different cell wall structures for gram-negative, gram-positive bacteria, as well as fungi. Credit: Xin Tong of BGI-Shenzhen.

For constructing sequencing libraries, fragmentation of the extracted DNA used to be performed by sonication. The throughput can be much higher using enzymatic reactions (e.g., Tn5 transposase), which is compatible with automation.

3.5 Sequencing amount

In theory, the unique alignment of a single sequencing read is sufficient to detect a microbe (more on taxonomy in Chapter 5). Metagenomic shotgun sequencing that has no PCR amplification step has a negligible error rate [57]. For a metagenomic sample sequenced with 100 million reads (e.g., paired-end 100 bp reads, making $100 \times 10^6 \times 100 = 10$ Gb data), the lowest possible relative abundance that is directly detected for a taxa or a gene is 10^{-8}. For a sample containing 10^{11} microbial cells (< 1 g from the total biomass of ~ 200 g [58,59], Chapter 1), the detection limit should be single-cell, which could be verified with serially diluted spike-ins [22,60]. For low-abundance taxa, there is still a large gap in sequencing amount. It may be wise to look for samples in the same individual or in other people where this taxon is more abundant or try to culture it first.

Due to the physics of polymer (DNA) bending, bridge-PCR-based sequencing platforms tend to oversequence high-GC regions (e.g., Bifidobacterium genomes have a GC content of ~ 60%), and numerical adjustment may be needed for the abundances [57,61,62].

16S rRNA gene amplicon sequencing for bacteria and ITS (internal transcribed spacer regions) amplicon sequencing for fungi (Fig. 3.4) could detect microbes even when the concentration of DNA in a sample is below detection using ultraviolet light (e.g., in some urine samples [35]). The different hypervariable regions (e.g., V4–V5, V1–V2) have different taxonomic resolutions. Full-length amplicon sequencing spanning the entire region of 16S for bacteria, and 18S-ITS for fungi, is a promising way of reliably classifying microbial species using amplicon sequencing (e.g., [63]).

Worked sample 3.2

Shown below are DNA extraction results from 3 samples:

Sample	Concentration (ng/μL)	Volume (μL)	Total DNA (μg)
A001	8.56	80	0.68
A002	1.32	80	0.11
A003	24.4	80	1.95

If the average genome size in the above fecal samples is 5 Mb, about how many bacterial cells do we have in each sample?
If these are vaginal samples with an average bacterial genome size of 2.3 Mb, with 96% human sequences in the metagenome, about how many bacterial cells do we have in each sample?
If these vaginal samples are dominated by *L. iners*, whose genome size is only 1.3 Mb, about how many bacterial cells do we have in each sample?
If these vaginal samples are dominated by the fungus *Candida albicans*, how many copies of the fungal genome do we have in each sample?

3.6 Taxonomic and functional profiles, absolute abundance

Taxonomic and functional profiles can then be derived from the sequencing data, typically according to genes (More in Chapter 5). Some of the retroviral sequences in the human genome are understudied and may show up as RNA viruses in the taxonomic assignment.

When the total relative abundance is normalized to 1, please beware of the unclassified portion, which can be high in some samples.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a classic database for functional annotation [64], with curated new add-ons such as the gut-brain modules [65,66]. Many KOs (KEGG Orthology groups) are present in multiple modules, and a cutoff such as 60% completeness is often imposed when reporting the presence of a module in a sample. The exact function of an enzyme may lie in a single amino acid difference from the database output (e.g., production of imidazole propionate [67]). There is a lot of basic research to catch on here.

If one takes pains to quantitatively perform each of the above-mentioned steps from sampling to sequencing, in comparison to standards, one can estimate the absolute number of particular microbes in a sample. A recent fecal microbiome study on preterm infants used spike-in cells and reported dynamics in absolute abundance, e.g., the fungi *Candida albicans* appeared to inhibit many bacteria [68], or rather it boomed when not effectively outcompeted by bacteria. The bacterium *Salinibacter ruber* DSM 13855, archaea *Haloarcula hispanica* ATCC 33960 and fungus *Trichoderma reesei* ATCC 13631 chosen as spike-ins were all absent from the samples, and were only detected in the samples because they were added to weighed aliquots of feces.

Worked sample 3.3

Find the taxonomic profile (16S rRNA gene amplicon sequencing) from the following study, and try to plot the differences between the breast milk samples (some pumped, some not) and the negative controls. Besides a table, what do you think would be a good plot, PCoA (Principal Coordinate Analysis) using Bray-Curtis distance? Box plot or violin plot for each taxon?

Core milk microbiota[a] among 393 mothers in the CHILD (Canadian Healthy Infant Longitudinal Development) cohort in comparison to the negative controls (Table S3 of [69])

		Samples ($n=393$)			Negative controls ($n=15$)
Lineage	**Genus**	**Mean ± SD**	**Maximum**	**Prevalence**	**Prevalence**
Proteobacteria—Burkholderiales	Unclassified	5.86 ± 3.43	12.56	100%	13%
Firmicutes—*Staphylococcaceae*	*Staphylococcus*	4.86 ± 11.5	87.5	100%	20%
Proteobacteria—*Oxalobacteraceae*	*Ralstonia*	4.79 ± 2.76	9.41	100%	7%
Proteobacteria—*Comamonadaceae*	Unclassified	4.42 ± 2.58	9.75	100%	7%
Proteobacteria—*Comamonadaceae*	*Acidovorax*	3.95 ± 2.34	13.33	100%	20%
Proteobacteria—*Oxalobacteraceae*	*Massilia*	2.37 ± 1.40	6.47	100%	13%
Proteobacteria—Uncl. Alteromonadales	*Rheinheimera*	1.89 ± 1.15	4.74	100%	0
Proteobacteria—*Rhizobiaceae*	*Agrobacterium*	1.85 ± 1.08	4.51	100%	7%
Proteobacteria—*Rhodospirillaceae*	Unclassified	1.61 ± 1.07	5.24	100%	7%
Proteobacteria—*Neisseriaceae*	*Vogesella*	1.23 ± 0.74	3.04	100%	0
Actinobacteria—*Nocardioidaceae*	*Nocardioides*	1.09 ± 0.65	2.61	100%	13%
Proteobacteria—Burkholderiales	Unclassified	1.07 ± 0.64	2.63	100%	0

[a] Defined as amplicon sequence variants (ASVs) present in at least 95% of samples with a mean relative abundance of more than 1% after removing potential reagent contaminants. Uncl, unclassified. The unit for the "Mean ± SD" and the "Maximum" columns is probably % relative abundance.

3.7 Sample size for metagenome-wide association studies

Microbiome data are notoriously heterogeneous, fat-tailed (high probability of extreme values, in contrast to normal distributions), and zero-inflated (a taxon can have zero value in many samples). Therefore, all the commonly used statistical tests for metagenomic data are nonparametric, and do not depend on a distribution. For example, Wilcoxon rank-sum test, also known as Mann-Whitney *U* test, is commonly used to look for differences between two groups, and significant results do not need to have a large effect.

For sample size estimation and power analyses, we have a modified version of [70] which removed the limitation on the initial statistical sampling space. Ideally, one needs to start with about 40 samples to get the pseudo-R^2 to then accurately calculate the required sample size for a power of 0.8 or higher. But that would have already exceeded the sample size needed for comparing groups of samples with a strong and relatively homogenous difference, such as dental plaques from rheumatoid arthritis patients and healthy controls, fecal samples from IgA deficient people and their normal spouse. Including more samples could be adding heterogeneity (Fig. 3.8), and would not necessarily follow the same statistical distribution, e.g., different people might get the same disease in a different season. So another dataset that differs in some way should rather be an independent validation, and could also allow the discovery of some different biomarkers.

Current knowledge of a disease can help us decide on the most important body sites and the number of samples to collect for metagenome-wide association studies (MWAS), along with other metadata that needs to be recorded or controlled for. For example, gingival samples for rheumatoid arthritis, fecal samples for Crohn's disease or IgA-deficient individuals show a strong dysbiosis with only a few samples [7,10,48,71–73]. Fecal microbiome markers for colorectal cancer converged nicely with a few dozens of samples from each country (more in Chapter 7), while for metabolic diseases such as obesity and type 2 diabetes (T2D) (more in Chapter 6), one either goes for extreme phenotypes or larger sample sizes [74,75]. Treatment-naïve samples come from population-wide disease screening efforts (Chapter 7), or from remote regions without access to modern treatments. For most doctors, checking on previous patients before a relapse would be a more realistic and clinically meaningful study design. Patients who relapsed after being without medication for at least 3 months could show the same disease markers, despite other differences in their microbiome with longer disease duration and previous medication [10,66]. Otherwise, medication information is recorded and statistically adjusted for in analyses.

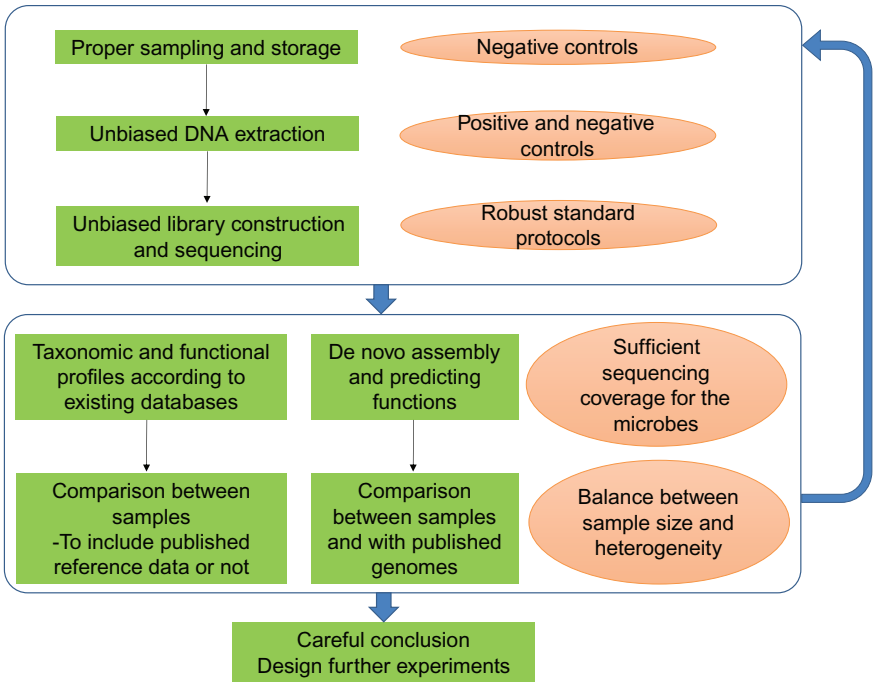

Fig. 3.8 A simplified step-by-step guide for an ideal metagenomics study from samples to data analyses and conclusions. After about two dozen samples, it would be a good idea to check the phenotype distributions and see whether the recruitment of volunteers is working well to address the questions one intends to study. Credit: Huijue Jia.

Permutational Analyses of Variances (PERMANOVA) [76,77] is a nonparametric statistical test that is suitable for analyzing potential influence from phenotypes and questionnaire entries on metagenomic data. Although it is a multivariant test, a typical start is to analyze each phenotype by itself. Questionnaires and other omics data, if well designed for a particular body site, e.g., pregnancy history and hormones for the cervical microbiome, oral hygiene and immune features for the oral microbiome, can explain notable portions of the microbiome variances [14,78], more than the single-digit percentiles typically seen for fecal studies of adult cohorts [3,79–81].

For body sites that are known to have clearly different community types, e.g., fecal microbiome, vaginal microbiome, the samples may need to be stratified according to the community types during analyses, at the cost of apparent sample size [9,82,83]. Stratifying the Human Microbiome Project (HMP) fecal data according to enterotypes was reported to increase statistical power [84]. Sex differences in microbiome composition and immune responses also mean that stratifying a

microbiome study according to sex could lead to interesting discoveries that differ from the overall analyses [11,83].

Researchers need to know the important factors in a cohort, so that whenever we see some differences between groups, we can then make sure that the difference is due to the disease (or other condition) we are comparing, instead of a difference in the other factors between groups (e.g. genetic and lifestyle differences behind ethnic or geographical groups [50,85,86]). Brute force statistical adjustments would always lead to loss of true signals, before we can rationally model the microbiome. In a study of colorectal cancer, levels of the iron-binding protein ferritin and weekly intake of red meat significantly differed among the control, adenoma, and carcinoma groups, which are relevant for colorectal cancer and needs no adjustment [47] (more on causality in Chapter 6). Days since the last menstrual cycle (in numbers instead of divided into two or three categories of menstrual phase) associated with many bacteria in the upper and the lower reproductive tract (from the peritoneal fluid, uterus, cervix, to mid-vagina), so we simply cannot be sure about some of the bacteria's associations with hysteromyoma (uterine fibroids) and adenomyosis [34]. These are not to be discarded because of statistics, but to be further studied in more conclusive ways.

In addition to healthy controls, other conditions that share some feature of the disease should also be considered for sampling. For example, hyposalivation due to other reasons are included, to better understand the oral microbiome in Sjögren's syndrome [87]. A study of the respiratory microbiome in people with obstructive sleep apnea probably needs to consider differences in body fat and cardiovascular health in the volunteers. Again, if published data are included as the related conditions for comparison, considerable heterogeneity may be introduced [52,88]. One way to circumvent the metadata and sample handling differences is to always compare the disease samples with control samples from the same study [6,89] before better standards are worked out for specific fields of study [90].

3.8 Summary

Low biomass samples, i.e., where the microbial populations are small (Chapter 1) and not much above current detection limits, are more sensitive to contamination during sample collection, DNA extraction, etc. Multiple methods in addition to metagenomic sequencing would be needed to estimate the number of microbes in low biomass samples and to link the microbes to potential functions. It will always be a good idea to collect related samples from the same

Worked sample 3.4

Seeing the PERMANOVA results below from the fecal microbiome of monozygotic and dizygotic twins in the TwinsUK cohort (Table S2 of [50]; smaller sample size than the amplicon studies on the cohort [91,92]), which are the phenotypes you would be cautious about when making conclusions?

Try to have a visual sense of the distribution of phenotypes, by making plots. BMI, for example, is mostly lean in some East Asian cohorts (e.g., Ref. [3]), and would not show up as strongly as for the old ladies here; Waist-to-hip ratio, muscle mass, and fat mass estimated from conductance or scans, could better reflect the belly fat than BMI.

In what ways do you think "Year of birth" might be different from "Age of metagenomics sample?"

According to current knowledge and technologies, how do you think the potential influence of physical exercises on the gut microbiome could be better investigated? (also for Chapter 8)

Phenotypes	Number of twins	Groups	Sample size	Degree of freedom	Sums of squares	Mean square	F model	Pseudo-R^2	P(> F)	P adjusted (BH)
Twin pair number	246	NA	246	122	43.28	0.355	1.196	0.543	0.000	0.002
BMI	249	NA	249	1	0.511	0.511	1.571	0.006	0.002	0.017
Drugs (diabetic tablets)	230	Y	5	1	0.463	0.463	1.428	0.006	0.015	0.083
		N	225							
Has a doctor ever diagnosed or treated you for any of the following conditions?/diabetes	222	Y	10	1	0.429	0.429	1.319	0.006	0.028	0.091
		N	212							
Year of birth	250	NA	250	1	0.424	0.424	1.303	0.005	0.035	0.091
Current location (Geo-clusters)	247	Cluster 1	10	3	1.133	0.378	1.161	0.014	0.037	0.091
		Cluster 2	134							
		Cluster 3	52							
		Cluster 4	51							
Vegetarian or vegan	198	N	179	1	0.420	0.420	1.291	0.007	0.041	0.091
		Y	19							
Age at metagenomic sample	250	NA	250	1	0.417	0.417	1.280	0.005	0.043	0.091
Number of units of alcohol drunk per week	241	1–5 units	88	1	0.389	0.389	1.193	0.005	0.100	0.173
		6–10 units	21							
		11–15 units	56							
		16–20 units	7							
		21–40 units	34							
		40 + units	4							
		None	31							

Phenotypes	Number of twins	Groups	Sample size	Degree of freedom	Sums of squares	Mean square	F model	Pseudo-R^2	P(> F)	P adjusted (BH)
Menopausal status	240	Postmenopausal Premenopausal Going through menopause	174 38 28	2	0.734	0.367	1.129	0.009	0.102	0.173
Smoking status	249	Smoker Never smoked	92 157	1	0.367	0.367	1.126	0.005	0.160	0.247
Currently, how many minutes per week do you spend walking briskly/ gardening vigorously?	196	0 ≥ 1	22 174	1	0.354	0.354	1.096	0.006	0.214	0.303
Currently, how many minutes per week do you spend in nonweight bearing activity? e.g., swimming, cycling, yoga, aqua aerobics etc.	188	0 ≥ 1	102 86	1	0.323	0.323	0.998	0.005	0.447	0.559
Drugs (Insulin)	230	Y N	4 226	1	0.321	0.321	0.988	0.004	0.460	0.559
Currently, how many minutes per week do you spend on weight-bearing activity? E.g., aerobics, running, dance, football, basketball, racquet sports, etc. (do not include walking or gardening)	191	0 ≥ 1	103 88	1	0.309	0.309	0.953	0.005	0.589	0.667
Outdoor sports	108	Y N	56 52	1	0.29	0.29	0.885	0.008	0.798	0.848

PERMANOVA for the influence of each phenotype on the gut microbial gene profile (11.4 million genes of the fecal microbiome [50]). 9999 permutations, Bray-Curtis distance. As one PERMANOVA test was performed for each phenotype, multiple testing was controlled using the Benjamini-Hochberg procedure. The phenotypes are not analyzed in combination. Y, yes; N, no.

individual. As complex communities, caution needs to be taken in storage, DNA extraction, and sequencing to avoid underrepresentation of some of the taxa in a sample. We will see in Chapter 5 that metagenomic assembly would require a higher sequencing coverage than metagenomic detection. Questionnaires and other information need to be optimized for each body site, in order not to miss important information during statistical analyses, and to consistently identify biomarkers from MWAS on multiple cohorts.

References

[1] Vandeputte D, Falony G, Vieira-Silva S, Tito RY, Joossens M, Raes J. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. Gut 2016;65:57–62. https://doi.org/10.1136/gutjnl-2015-309618.

[2] Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, et al. Quantitative microbiome profiling links gut community variation to microbial load. Nature 2017;551:507–11. https://doi.org/10.1038/nature24460.

[3] Jie Z, Liang S, Ding Q, Li F, Tang S, Wang D, et al. A transomic cohort as a reference point for promoting a healthy gut microbiome. Med Microecol 2021;8:100039. https://doi.org/10.1016/j.medmic.2021.100039.

[4] Park S, Won DD, Lee BJ, Escobedo D, Esteva A, Aalipour A, et al. A mountable toilet system for personalized health monitoring via the analysis of excreta. Nat Biomed Eng 2020;4:624–35. https://doi.org/10.1038/s41551-020-0534-9.

[5] Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, et al. A framework for human microbiome research. Nature 2012;486:215–21. https://doi.org/10.1038/nature11209.

[6] Jie Z, Liang S, Ding Q, Li F, Tang S, Sun X, et al. Disease trends in a young Chinese cohort according to fecal metagenome and plasma metabolites. Med Microecol 2021. https://doi.org/10.1016/j.medmic.2021.100037.

[7] He Q, Gao Y, Jie Z, Yu X, Laursen JM, Xiao L, et al. Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. Gigascience 2017;6:1–11. https://doi.org/10.1093/gigascience/gix050.

[8] Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded human microbiome project. Nature 2017. https://doi.org/10.1038/nature23889.

[9] Chen C, Hao L, Zhang Z, Tian L, Song L, Zhang X, et al. Dynamics in the vaginocervical microbiome after oral probiotics. J Genet Genomics 2021. https://doi.org/10.1101/2020.06.16.155929.

[10] Zhang X, Zhang D, Jia H, Feng Q, Wang D, Di Liang D, et al. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med 2015;21:895–905. https://doi.org/10.1038/nm.3914.

[11] Zhu J, Tian L, Chen P, Han M, Song L, Tong X, et al. Over 50,000 metagenomically assembled draft genomes for the human oral microbiome reveal new taxa. Genomics Proteomics Bioinformatics 2021. https://doi.org/10.1016/j.gpb.2021.05.001.

[12] Oh J, Byrd AL, Park M, Kong HH, Segre JA. Temporal stability of the human skin microbiome. Cell 2016;165:854–66. https://doi.org/10.1016/j.cell.2016.04.008.

[13] Fettweis JM, Serrano MG, Brooks JP, Edwards DJ, Girerd PH, Parikh HI, et al. The vaginal microbiome and preterm birth. Nat Med 2019;25:1012–21. https://doi.org/10.1038/s41591-019-0450-2.

[14] Jie Z, Chen C, Hao L, Li F, Song L, Zhang X, et al. Life history recorded in the vagino-cervical microbiome along with multi-omics. Genomics Proteomics Bioinformatics 2021. https://doi.org/10.1016/j.gpb.2021.01.005.
[15] Li F, Chen C, Wei W, Wang Z, Dai J, Hao L, et al. The metagenome of the female upper reproductive tract. Gigascience 2018;7. https://doi.org/10.1093/gigascience/giy107.
[16] Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. The placenta harbors a unique microbiome. Sci Transl Med 2014;6:237ra65. https://doi.org/10.1126/scitranslmed.3008599.
[17] Dickson RP, Erb-Downward JR, Prescott HC, Martinez FJ, Curtis JL, Lama VN, et al. Cell-associated bacteria in the human lung microbiome. Microbiome 2014;2:28. https://doi.org/10.1186/2049-2618-2-28.
[18] 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. https://doi.org/10.1186/s40168-018-0426-3.
[19] Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al. *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. Genome Res 2012;22:299–306. https://doi.org/10.1101/gr.126516.111.
[20] Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res 2012;22:292–8. https://doi.org/10.1101/gr.126573.111.
[21] Riley DR, Sieber KB, Robinson KM, White JR, Ganesan A, Nourbakhsh S, et al. Bacteria-human somatic cell lateral gene transfer is enriched in cancer samples. PLoS Comput Biol 2013;9. https://doi.org/10.1371/journal.pcbi.1003107, e1003107.
[22] Geller LT, Barzily-Rokni M, Danino T, Jonas OH, Shental N, Nejman D, et al. Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. Science 2017;357:1156–60. https://doi.org/10.1126/science.aah5043.
[23] Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 2020;579:567–74. https://doi.org/10.1038/s41586-020-2095-1.
[24] Nejman D, Livyatan I, Fuks G, Gavert N, Zwang Y, Geller LT, et al. The human tumor microbiome is composed of tumor type–specific intracellular bacteria. Science 2020;368:973–80. https://doi.org/10.1126/science.aay9189.
[25] de Goffau MC, Lager S, Sovio U, Gaccioli F, Cook E, Peacock SJ, et al. Human placenta has no microbiome but can contain potential pathogens. Nature 2019;572:329–34. https://doi.org/10.1038/s41586-019-1451-5.
[26] Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 2014;12:87. https://doi.org/10.1186/s12915-014-0087-z.
[27] Sun X, Hu Y-H, Wang J, Fang C, Li J, Han M, et al. Efficient and stable metabarcoding sequencing data using a DNBSEQ-G400 sequencer validated by comprehensive community analyses. GigaByte 2021. https://doi.org/10.46471/gigabyte.16.
[28] Lax S, Sangwan N, Smith D, Larsen P, Handley KM, Richardson M, et al. Bacterial colonization and succession in a newly opened hospital. Sci Transl Med 2017;9:1–11.
[29] Pidot SJ, Gao W, Buultjens AH, Monk IR, Guerillot R, Carter GP, et al. Increasing tolerance of hospital *Enterococcus faecium* to handwash alcohols. Sci Transl Med 2018;10:eaar6115. https://doi.org/10.1126/scitranslmed.aar6115.
[30] Mora M, Wink L, Kögler I, Mahnert A, Rettberg P, Schwendner P, et al. Space station conditions are selective but do not alter microbial characteristics relevant to human health. Nat Commun 2019;10:3990. https://doi.org/10.1038/s41467-019-11682-z.

[31] Checinska A, Probst AJ, Vaishampayan P, White JR, Kumar D, Stepanov VG, et al. Microbiomes of the dust particles collected from the international space station and spacecraft assembly facilities. Microbiome 2015;3:50. https://doi.org/10.1186/s40168-015-0116-3.
[32] Lee MD, O'Rourke A, Lorenzi H, Bebout BM, Dupont CL, Everroad RC. Reference-guided metagenomics reveals genome-level evidence of potential microbial transmission from the ISS environment to an astronaut's microbiome. IScience 2021;24:102114. https://doi.org/10.1016/j.isci.2021.102114.
[33] Saw JJ, Sivaguru M, Wilson EM, Dong Y, Sanford RA, Fields CJ, et al. In vivo entombment of bacteria and fungi during calcium oxalate, brushite, and struvite urolithiasis. Kidney360 2021;2:298–311. https://doi.org/10.34067/kid.0006942020.
[34] Chen C, Song X, Wei W, Zhong H, Dai J, Lan Z, et al. The microbiota continuum along the female reproductive tract and its relation to uterine-related diseases. Nat Commun 2017;8:875. https://doi.org/10.1038/s41467-017-00901-0.
[35] Chen C, Hao L, Wei W, Li F, Song L, Zhang X, et al. The female urinary microbiota in relation to the reproductive tract microbiota. Gigabyte 2020;2020:1–9. https://doi.org/10.46471/gigabyte.9.
[36] Seferovic MD, Pace RM, Carroll M, Belfort B, Major AM, Chu DM, et al. Visualization of microbes by 16S in situ hybridization in term and preterm placentas without intraamniotic infection. Am J Obstet Gynecol 2019;221:146.e1–146.e23. https://doi.org/10.1016/j.ajog.2019.04.036.
[37] Gosalbes MJ, Llop S, Vallès Y, Moya A, Ballester F, Francino MP. Meconium microbiota types dominated by lactic acid or enteric bacteria are differentially associated with maternal eczema and respiratory problems in infants. Clin Exp Allergy 2013;43:198–211. https://doi.org/10.1111/cea.12063.
[38] Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe 2015;17:690–703. https://doi.org/10.1016/j.chom.2015.04.004.
[39] Collado MC, Rautava S, Aakko J, Isolauri E, Salminen S. Human gut colonisation may be initiated in utero by distinct microbial communities in the placenta and amniotic fluid. Sci Rep 2016;6:23129. https://doi.org/10.1038/srep23129.
[40] Wang J, Zheng J, Shi W, Du N, Xu X, Zhang Y, et al. Dysbiosis of maternal and neonatal microbiota associated with gestational diabetes mellitus. Gut 2018. https://doi.org/10.1136/gutjnl-2018-315988. gutjnl-2018-315988.
[41] He Q, Kwok L-Y, Xi X, Zhong Z, Ma T, Xu H, et al. The meconium microbiota shares more features with the amniotic fluid microbiota than the maternal fecal and vaginal microbiota. Gut Microbes 2020;12:1794266. https://doi.org/10.1080/19490976.2020.1794266.
[42] Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. Cell Host Microbe 2018;24:133–145.e5. https://doi.org/10.1016/j.chom.2018.06.005.
[43] Massier L, Chakaroun R, Tabei S, Crane A, Didt KD, Fallmann J, et al. Adipose tissue derived bacteria are associated with inflammation in obesity and type 2 diabetes. Gut 2020;69(10):1796–806. https://doi.org/10.1136/gutjnl-2019-320118.
[44] Anhê FF, Jensen BAH, Varin TV, Servant F, Van Blerk S, Richard D, et al. Type 2 diabetes influences bacterial tissue compartmentalisation in human obesity. Nat Metab 2020;2(3):233–42. https://doi.org/10.1038/s42255-020-0178-9.
[45] Stenkula KG, Erlanson-Albertsson C. Adipose cell size: importance in health and disease. Am J Physiol Integr Comp Physiol 2018;315:R284–95. https://doi.org/10.1152/ajpregu.00257.2017.
[46] Dickson RP, Erb-Downward JR, Martinez FJ, Huffnagle GB. The microbiome and the respiratory tract. Annu Rev Physiol 2016;78:481–504. https://doi.org/10.1146/annurev-physiol-021115-105238.

[47] Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nat Commun 2015;6:6528. https://doi.org/10.1038/ncomms7528.
[48] Moll JM, Myers PN, Zhang C, Eriksen C, Wolf J, Appelberg KS, et al. Gut microbiota perturbation in IgA deficiency is influenced by IgA-autoantibody status. Gastroenterology 2021. https://doi.org/10.1053/j.gastro.2021.02.053.
[49] Zhong H, Ren H, Lu Y, Fang C, Hou G, Yang Z, et al. Distinct gut metagenomics and metaproteomics signatures in prediabetics and treatment-naïve type 2 diabetics. EBioMedicine 2019. https://doi.org/10.1016/j.ebiom.2019.08.048.
[50] Xie H, Guo R, Zhong H, Feng Q, Lan Z, Qin B, et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. Cell Syst 2016;3:572–584.e3. https://doi.org/10.1016/j.cels.2016.10.004.
[51] David LA, CFC M, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature 2013;505:559–63. https://doi.org/10.1038/nature12820.
[52] Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol 2014;32:834–41. https://doi.org/10.1038/nbt.2942.
[53] Bouslimani A, Porto C, Rath CM, Wang M, Guo Y, Gonzalez A, et al. Molecular cartography of the human skin surface in 3D. Proc Natl Acad Sci U S A 2015;112:E2120–9. https://doi.org/10.1073/pnas.1424409112.
[54] Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, et al. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol 2017. https://doi.org/10.1038/nbt.3960.
[55] Tourlousse DM, Narita K, Miura T, Sakamoto M, Ohashi A, Shiina K, et al. Validation and standardization of DNA extraction and library construction methods for metagenomics-based human fecal microbiome measurements. Microbiome 2021;9:95. https://doi.org/10.1186/s40168-021-01048-3.
[56] Yang F, Sun J, Luo H, Ren H, Zhou H, Lin Y, et al. Assessment of fecal DNA extraction protocols for metagenomic studies. Gigascience 2020;9(7). https://doi.org/10.1093/gigascience/giaa071.
[57] Fang C, Zhong H, Lin Y, Chen B, Han M, Ren H, et al. Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. Gigascience 2018;7:1–8. https://doi.org/10.1093/gigascience/gix133.
[58] Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol 2016;14. https://doi.org/10.1371/journal.pbio.1002533, e1002533.
[59] Stephen AM, Cummings JH. The microbial contribution to human faecal mass. J Med Microbiol 1980;13:45–56. https://doi.org/10.1099/00222615-13-1-45.
[60] Lager S, de Goffau MC, Sovio U, Peacock SJ, Parkhill J, Charnock-Jones DS, et al. Detecting eukaryotic microbiota with single-cell sensitivity in human tissue. Microbiome 2018;6:151. https://doi.org/10.1186/s40168-018-0529-x.
[61] Patterson J, Carpenter EJ, Zhu Z, An D, Liang X, Geng C, et al. Impact of sequencing depth and technology on de novo RNA-Seq assembly. BMC Genomics 2019;20:604. https://doi.org/10.1186/s12864-019-5965-x.
[62] Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, et al. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. Gigascience 2020;9. https://doi.org/10.1093/gigascience/giaa008.
[63] Fang C, Sun X, Fan F, Zhang X, Wang O, Zheng H, et al. High-resolution single-molecule long-fragment rRNA gene amplicon sequencing for uncultured bacterial and fungal communities. bioRxiv 2021. https://doi.org/10.1101/2021.03.29.437457.
[64] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012;40:D109–14. https://doi.org/10.1093/nar/gkr988.

[65] Valles-Colomer M, Falony G, Darzi Y, Tigchelaar EF, Wang J, Tito RY, et al. The neuroactive potential of the human gut microbiota in quality of life and depression. Nat Microbiol 2019. https://doi.org/10.1038/s41564-018-0337-x.

[66] Zhu F, Ju Y, Wang W, Wang Q, Guo R, Ma Q, et al. Metagenome-wide association of gut microbiome features for schizophrenia. Nat Commun 2020;11:1612. https://doi.org/10.1038/s41467-020-15457-9.

[67] Koh A, Molinaro A, Ståhlman M, Khan MT, Schmidt C, Mannerås-Holm L, et al. Microbially produced imidazole propionate impairs insulin signaling through mTORC1. Cell 2018;175:947–961.e17. https://doi.org/10.1016/j.cell.2018.09.055.

[68] Rao C, Coyte KZ, Bainter W, Geha RS, Martin CR, Rakoff-Nahoum S. Multi-kingdom ecological drivers of microbiota assembly in preterm infants. Nature 2021. https://doi.org/10.1038/s41586-021-03241-8.

[69] Moossavi S, Sepehri S, Robertson B, Bode L, Goruk S, Field CJ, et al. Composition and variation of the human milk microbiota are influenced by maternal and early-life factors. Cell Host Microbe 2019;25:324–335.e4. https://doi.org/10.1016/j.chom.2019.01.011.

[70] Kelly BJ, Gross R, Bittinger K, Sherrill-Mix S, Lewis JD, Collman RG, et al. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. Bioinformatics 2015;31:2461–8. https://doi.org/10.1093/bioinformatics/btv183.

[71] Qin J, Li R, Raes J, Arumugam M, Burgdorf KSS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464:59–65. https://doi.org/10.1038/nature08821.

[72] Zou M, Jie Z, Cui B, Wang H, Feng Q, Zou Y, et al. Fecal microbiota transplantation results in bacterial strain displacement in patients with inflammatory bowel diseases. FEBS Open Bio 2019. https://doi.org/10.1002/2211-5463.12744.

[73] Fadlallah J, El Kafsi H, Sterlin D, Juste C, Parizot C, Dorgham K, et al. Microbial ecology perturbation in human IgA deficiency. Sci Transl Med 2018;10. https://doi.org/10.1126/scitranslmed.aan1217, eaan1217.

[74] Wang J, Jia H. Metagenome-wide association studies: fine-mining the microbiome. Nat Rev Microbiol 2016;14:508–22. https://doi.org/10.1038/nrmicro.2016.83.

[75] Liu R, Hong J, Xu X, Feng Q, Zhang D, Gu Y, et al. Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. Nat Med 2017;23(7):859–68. https://doi.org/10.1038/nm.4358.

[76] Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol 2001. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x.

[77] Anderson MJ, Walsh Daniel CI. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? Ecol Monogr 2013. https://doi.org/10.1890/12-2010.1.

[78] Liu X, Tong X, Zhu J, Liu T, Jie Z, Zou Y, et al. Metagenome-genome-wide association studies reveal human genetic impact on the oral microbiome. Biorxiv 2021. https://doi.org/10.1101/2021.05.06.443017.

[79] Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, et al. Population-level analysis of gut microbiome variation. Science 2016;352:560–4. https://doi.org/10.1126/science.aad3503.

[80] Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 2016;352:565–9. https://doi.org/10.1126/science.aad3369.

[81] Wang J, Thingholm LB, Skiecevičienė J, Rausch P, Kummen M, Hov JR, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nat Genet 2016;48:1396–406. https://doi.org/10.1038/ng.3695.

[82] Gu Y, Wang X, Li J, Zhang Y, Zhong H, Liu R, et al. Analyses of gut microbiota and plasma bile acids enable stratification of patients for antidiabetic treatment. Nat Commun 2017;8:1785. https://doi.org/10.1038/s41467-017-01682-2.

[83] Liu X, Tang S, Zhong H, Tong X, Jie Z, Ding Q, et al. A genome-wide association study for gut metagenome in Chinese adults illuminates complex diseases. Cell Discov 2021;7:9. https://doi.org/10.1038/s41421-020-00239-w.

[84] Mattiello F, Verbist B, Faust K, Raes J, Shannon WD, Bijnens L, et al. A web application for sample size and power calculation in case-control microbiome studies. Bioinformatics 2016;32:2038–40. https://doi.org/10.1093/bioinformatics/btw099.

[85] He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med 2018;24:1532–5. https://doi.org/10.1038/s41591-018-0164-x.

[86] Deschasaux M, Bouter KE, Prodan A, Levin E, Groen AK, Herrema H, et al. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. Nat Med 2018;24:1526–31. https://doi.org/10.1038/s41591-018-0160-1.

[87] AlmståhI A, Wikström M, Stenberg I, Jakobsson A, Fagerberg-Mohlin B. Oral microbiota associated with hyposalivation of different origins. Oral Microbiol Immunol 2003;18:1–8.

[88] Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature 2013;498:99–103. https://doi.org/10.1038/nature12198.

[89] Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, et al. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun 2017;8:845. https://doi.org/10.1038/s41467-017-00900-1.

[90] Kachroo N, Lange D, Penniston KL, Stern J, Tasian G, Bajic P, et al. Standardization of microbiome studies for urolithiasis: an international consensus agreement. Nat Rev Urol 2021;18:303–11. https://doi.org/10.1038/s41585-021-00450-8.

[91] Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al. Human genetics shape the gut microbiome. Cell 2014;159(4):789–99. https://doi.org/10.1016/j.cell.2014.09.053.

[92] Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, et al. Genetic determinants of the gut microbiome in UK twins. Cell Host Microbe 2016;19(5):731–43. https://doi.org/10.1016/j.chom.2016.04.017.