Authors

Bill Mohn, PhD
Dr. William (Bill) Mohn is a Professor in the Department of Microbiology & Immunology in the Life Sciences Institute at the University of British Columbia, and also serves as co-founder and co-CSO of Microbiome Insights. Bill has done landmark research on bacterial metabolism of chlorinated compounds, terpenoids, lignocellulose, and steroids. He is a leader in microbial ecology and the development of methods to investigate highly complex microbial communities and was elected to both The American Academy of Microbiology and the Board of Directors of the International Society for Microbial Ecology.

Pedro Dimitriu, PhD
Pedro Dimitriu is Microbiome Insights’ Senior Scientist and Senior Director of Bioinformatics. He is in charge of integrating sequencing data, ecological statistics, and machine learning methods to study host–microbe interactions. Before joining the company, Pedro was a postdoctoral fellow at UBC’s Department of Microbiology and Immunology.

Erick Cardenas, PhD
Erick Cardenas is Microbiome Insights’ Senior Director of Bioinformatics and was project lead for an IBD disease monitoring program based on microbiome profiling and machine learning. Erick has conducted research on the role of microbial communities in bioremediation, plant-growth promotion in biofuel crops, and nutrient cycling in forest soils. He brings to Microbiome Insights expertise in the analysis of metagenomes, and in the development of statistical and bioinformatics tools.
Ben Tantika, MSc
Ben is Director of Lab Operations at Microbiome Insights, bringing experience in advanced techniques in both wet and dry lab settings. During his B.Sc. and M.Sc. thesis study, he conducted research on wine microbes with the goal of understanding how microorganisms affect wine’s sensorial and chemical profiles. Ben previously worked as a research technician and Next Generation Sequencing (NGS) technician specializing in all things related to Sanger and Illumina NGS DNA Sequencing technology.

Kristina Campbell, MSc
Kristina Campbell is a science and medical writer who specializes in creating resources on emerging translational aspects of the gut microbiome for scientists and healthcare professionals. She reports on microbiome science for online and print media throughout Europe and North America, and is the co-author of a gut microbiota textbook for dietitians as well as two books for the general public.

Reviewers:
Brett Finlay, PhD, Claire Arrieta, PhD and Greg Hillebrand, PhD
Planning a microbiome study: The comprehensive guide

Who is this guide for?

This guide was created for anyone interested in studying microbial communities. We designed the guide so scientists who are new to microbiome research can get up to speed quickly on the main considerations and options, and also so scientists with previous experience in microbiome research can dig deeper into the nuances of their study design and microbiome analysis tools. After reading this guide, PhD students and postdoctoral fellows, for example, will be able to plan a study or to work more effectively with a core facility or external service provider. Industry scientists will be in a better position to evaluate various options for microbiome studies and assess the suitability of in-house capabilities. Overall, this resource can be used to help plan and conduct useful microbiome studies.

What does this guide cover?

This document encompasses the combined multi-decade experience of our team in microbiome research and analysis. It covers all the key steps for microbiome sequencing studies, from study design through to analysis and interpretation, including useful resources and supporting documents. After reading this guide, you will better understand the important considerations for formulating your research question and planning your study, collecting and managing your samples, sequencing, and interpreting the data. The guide does not cover conventional culture-based methods for microbiome research, although in many cases, data from these culture-based methods can be a useful complement to metagenomics data.
# Table of Contents

## 1 Planning Your Study
- Getting Started  
  - 7
- Study Process  
  - 7
- Main considerations for a study  
  - 8

## 2 Sample Types & Collection Methods
- Sample Types  
  - 11
  - I. Fecal samples  
  - 11
  - II. Skin samples  
  - 11
  - III. Oral samples  
  - 12
  - IV. Urine samples  
  - 13
  - V. Vaginal samples  
  - 13
  - VI. Soil samples  
  - 13
  - VII. Water samples  
  - 13
- Sample considerations  
  - 14

## 3 DNA Extraction
- Library preparation considerations  
  - 16
- Contamination  
  - 17
  - Best practices for controlling contamination  
  - 17
- Other sources of variation (PCR bias)  
  - 17

## 4 Total Bacterial Quantification via qPCR  
  - 18
## Metagenomic Sequencing Options

I. Sequencing of 16S, 18S, or ITS2
II. Shotgun Metagenome Sequencing (SMS)
III. Shallow Shotgun Metagenome Sequencing (SSMS)

## Metatranscriptomics

## Short-Chain Fatty Acid Analysis

## Bioinformatics, Data Visualization, and Reporting

- Quality control plots
- Data exploration
- Hypothesis tests
- Diversity
- Example Reports

## Glossary
Getting Started

Microbiome research has come a long way since the data generated for the two largest initial human-focused projects in the field: the Human Microbiome Project (HMP) 1 and 2 in North America, and the Metagenomics of the Human Intestinal Tract (MetaHIT) project in Europe and elsewhere. The number of microbiome scientific publications continues to grow year by year, with study designs addressing more complex questions and analyses becoming more sophisticated. The microbiome has become an essential variable in a wide range of studies – particularly in human research related to diet and drug products.

In the current landscape of microbiome research, funders, journal editors, and other stakeholders are no longer satisfied with including microbiome data arbitrarily. To meet the higher expectations of these stakeholders, researchers must think strategically about why they are including microbiome data in their studies and what they hope to learn from it. Gaining meaningful insights from microbiome data in your study takes careful planning, and many factors can significantly impact the overall quality of the resulting data. Microbiome Insights is here to help you achieve your study goals and can assist at any point in the process to provide the scientific expertise, technical capabilities, and professional, friendly support to set you up for success.

Study process

A microbiome study includes these basic steps:

- Sampling, storage, and shipment (if applicable)
- DNA extraction
- DNA amplification (if applicable)
- Profiling, including DNA sequencing and/or metabolite analysis
- Bioinformatics, data visualization, and reporting

As you plan your study, it is important to think through each of these steps and your local resources for handling them. For example, will all of your sampling devices be available (i.e. delivered to where you need them) by the time you want to collect your samples? Immediately upon collection, how will you store the samples at the appropriate temperature? What needs to be done to prepare your samples for shipping, if applicable? If you encounter difficulties with any of these steps, you can consider asking a contract research organization (CRO) such as Microbiome Insights to help you.

At Microbiome Insights, we have customers who outsource the whole study to us and others who only require some steps, such as sequencing or bioinformatics analysis.
Main considerations for a study

Whether you are planning a study in your own lab or decide to partner with a reputable CRO, here are the main considerations for planning a study and writing a study protocol:

- **Hypothesis or question**: Traditionally, scientific studies tend to have a hypothesis that the researchers would like to support or reject. However, in the microbiome field many strong studies are exploratory and do not need a specific hypothesis. Determining the central question for your microbiome study will enable you to do a better analysis. This helps you plan the study, setting the right expectations and allowing you to determine what is in scope and out of scope.

- **Objectives**: Now that you have envisioned the central question for your study, what are your objectives? These are the specific, measurable ways you will address your question. These ‘subthemes’ of your question help you focus on the type of data you will need. It is important to have objectives that reflect the biological question you want to answer, which will then help guide the choice of technology (including microbiome sequencing) that you use to answer that question.

- **Context**: Before designing your study, thoroughly review the previously published work (and ongoing research, if possible) on the subject you are studying. Be sure to read the methods sections of previous studies and evaluate how well the researchers managed to answer the central question. Make note of any questions or gaps and use these to inform your own study.

- **Study design**: Another important consideration is the type of study that will give you the most compelling data to answer your research question. For example, an observational study collects data on a group of subjects without attempting to change their behaviours. In this type of study, accounting for confounding factors (e.g. age, sex, diet, and medications) is very important. Some of these factors may even help stratify your subject cohort. By contrast, interventional studies evaluate the effect of a particular intervention, such as a medication, probiotic supplement, or dietary intervention. These studies are strengthened when a placebo intervention is used and when the placebo and intervention groups are matched on as many factors as possible. If you are using an intervention with animals, make sure you minimize confounders such as genetics, maternal influences, and cage effects (e.g. by housing animals that receive the intervention with those that do not receive it). Also think through the possible confounders in a human interventional study – for example, a medication taken by a disease cohort and not taken by a healthy cohort. Appropriate meta-data is also important to consider when designing your study. For example, data on bowel habits may be important to gather when obtaining fecal samples, or images of skin when obtaining skin microbiome samples.

- **Sample type**: When you address a microbiome hypothesis, a broad range of sample types can be collected to help you answer your question, depending on what is most relevant: fecal, skin, oral, urine, vaginal, soil, plant, water, and others. You may need to delve into the published literature to identify the best sample type or types for your study, since the ideal sample type often depends on...
the disease or condition of interest. For example, fecal samples may be ideal for studies on inflammatory bowel disease given the known role of microorganisms in disease processes in the colon; but samples from the oral cavity may be relevant for studies on rheumatoid arthritis since previous work shows oral microbiome characteristics correlate with systemic autoimmune features. The sample type may also depend on how many you need per subject: fecal samples are a tractable sample type for longitudinal studies, as they can be collected at multiple time points from the same individuals with minimal invasiveness, in contrast with more invasive sample types, such as tissue obtained through colonoscopy. Also remember that the type of sample you collect may determine your options for sequencing type and data depth. Learn more about sample types and their applications below in this guide.

- **Sampling frequency:** The overall question and objectives, as well as practical considerations, will determine the appropriate sampling frequency for your microbiome study. Is it better to recruit more subjects and sample all of them at a single point in time (e.g. for a cross-sectional design), or to ask the same subjects to provide multiple samples over time (e.g. for a repeated measures design)? Normally, given the same number of samples, your analysis will be less powerful with a cross-sectional design than with a repeated measures design. Furthermore, a repeated measures design leads to more accurate conclusions about your question. Depending on the reproducibility of the sampling method, you may also want to consider taking more than one sample at each time point—for example, samples in triplicate to reduce sampling error.

- **Study size:** The size of your study is likely limited by your budget, but it is worth considering whether you want to collect more data (or more complex data) from fewer subjects, or less data from a greater number of subjects. Where treatments or subject groups are compared, the number of samples is critical because you must have enough statistical power to detect differences. Related previous studies will often suggest an appropriate study size based on a statistical power calculation. However, in some cases, you cannot know a priori how many samples will be necessary, and the decision must be based on your best estimate and availability of resources.

- **Key success metrics / criteria:** At the outset, determine the specific markers of success for your study. Remember that study integrity, objectivity, and accuracy are more important than supporting a hypothesis per se. You can ask what metrics will show you’ve learned something or advanced the field; if results do not end up as significant, it may simply mean that a more powerful study is needed, or a slightly altered question is required.

- **Ethics:** Research must be conducted ethically, adhering to scientific best practices. If you are conducting a study with humans, you’ll need to obtain informed consent, define how any personal data will be stored, and specify how you will protect patient privacy. Institutional Review Board (IRB) approvals are commonly required, and most journals will not accept human research that has not been approved by an IRB. For studies involving animals, approval from other associations such as the American Association for Laboratory Animal Science (IACUC) may be required. If you are working with soil or plant samples, field experiments will often have applicable regulations, too. Human studies should be registered at clinicaltrials.gov. Many journals now prefer to have clinical studies registered in advance of study execution for manuscripts to be considered for publication.
• **Envisioning your final report:** Even as you design your study, you should look ahead to what you will need your final report to look like. Think about questions such as: Who will I be presenting the study results to? What kinds of things will my target audience be looking for to inform future decisions? How can I help them better understand the data (or the study in general) to make these decisions? What are the publication standards in the relevant field? First, this helps you make sure you’re collecting the right kind of data for the analyses you want. And second, particularly if you are partnering with a CRO, setting these expectations early in the process can help determine the scope of work and save you time. The service provider may offer a basic study template and modify it according to your needs.

After you have carefully planned your study and taken into account the considerations above, it’s time to start collecting samples. This guide will go through the next critical steps.
2 Sample Types & Collection Methods

This section covers some of the sample types commonly used in microbiome research, along with some key considerations for sample collection, storage, and shipment. For detailed collection instructions for each of the sample types you can follow the links to the dedicated sections on our website.

Sample Types

I. Fecal samples

Human fecal samples can be obtained non-invasively, making them the most commonly used way to obtain a ‘snapshot’ of the gut microbiome. Fresh fecal samples contain rich microbial communities but proper conditions are required to maintain the integrity of this microbiome until it reaches the lab for analysis. For all samples, they should be stored as cold as possible (at least -20°C) as quickly as possible after collection. For long-term storage -80°C is recommended, and freeze-thaw cycles should be avoided or minimized. A fecal swab is the easiest method for human stool sampling, but the sample volume may be insufficient for some analyses. Collecting whole stool yields the larger volumes that may be required for certain analyses (e.g. metabolomics) or for multiple analyses. Whole fecal samples may be collected using a sterile sealed container that can withstand -80°C, or using a kit with a stabilization buffer that enables short-term storage and transport at ambient temperature.

Instructions for collection: Fecal Samples

II. Skin samples

Skin is the human body's largest organ and harbors an important layer of microbes: bacteria, archaea, fungi, and viruses. Compared with the gut, however, the skin does not support a high-biomass microbiome. Thus, microbiome sampling and sequencing for skin require special knowledge and handling.

Our internal validations have shown that skin microbiome recovery is optimized using swabs pre-moistened with sterile water with a mild surfactant (available in vials), as compared to dry swabs or skin tapes. This optimization results in overall higher microbial DNA content while minimizing host DNA content. We recommend using the Becton-Dickinson, BBL Cultureswab EZ II (or similar), which includes a double swab encased in a non-breathable transport tube. Because many skin microorganisms are firmly attached, the most representative samples are obtained when the
swab is vigorously swiped back and forth over the skin. When samples are collected in this way, skin cells are inevitably also recovered, yielding substantial host (human or animal) DNA in the analysis. For amplicon-based analyses, host DNA may impact amplification efficiency but is not generally a concern. For shotgun metagenomic analyses, on the other hand, host DNA may interfere with results. For human samples your analysis provider may suggest obtaining host DNA beforehand, or deleting the host reads bioinformatically by mapping reads to a human reference. Sequencing more deeply (i.e. more reads per sample) helps mitigate problems with host DNA.

Skin samples that represent the microbiome in deeper layers of the skin can be obtained with skin biopsies; however, this method is invasive and thus more complicated and expensive to execute compared to swabs for surface organisms. It may be an option for some animal studies.

Skin washing and use of some skin products can transiently but drastically impact the composition of the skin microbiome. It is recommended that the study protocol limits washing and use of products on the area(s) to be sampled for a defined period prior to sampling (e.g. 12 hours).

Learn more about planning a skin microbiome study:

How to Design a Skin Microbiome Study, Part I: Sampling

How to Design a Skin Microbiome Study, Part II: Amplicon Sequencing

Instructions for collection: Skin Samples

III. Oral samples

The oral microbiome is dense, with a highly diverse community of bacteria, as well as other microbes such as viruses, fungi, and archaea. The oral microbiome is shaped by host factors such as anti-microbial compounds in saliva, while also containing transient microbes from the surrounding environment. To accurately capture the oral microbiome it is imperative to stabilize the sample immediately upon collection (e.g. using a stabilization buffer and / or transporting on dry ice).

The mouth has distinct microbiomes in various locations (saliva, dental hard tissues, hard and soft palate, tongue, inner cheek, etc.). Therefore, the sampling location must be planned to target the microbiome that will best address the research hypothesis. It is also worth noting that oral samples may have a high amount of host DNA.

Instructions for collection: Oral Samples
IV. Urine samples

In the past, many assumed urine was free of bacteria and other microbes in individuals without infections. However, the healthy urinary tract does have a complement of microbes and certain conditions can result in higher levels of detectable microbes in the urinary system. Urine samples are considered low biomass and are also typically low in diversity. In common with other high-volume, low microbial biomass samples such as air particulates or seawater, microbes within a urine sample are typically concentrated through filtration and/or centrifugation prior to DNA extraction.

Instructions for collection: Urine Samples

V. Vaginal samples

The vaginal microbiome is a low-diversity microbial community, normally dominated by *Lactobacillus*. These microbiomes cluster into ‘types’ depending on the dominant bacterial species. When investigating the vaginal microbiome in relation to health issues such as bacterial vaginosis, sexually transmitted infections (STIs), or urinary tract infections (UTIs), the patient giving the sample may have sensitivities in the vagina and surrounding area so sample collection must be done by a professional in a clinic, or at home using a validated self-collection kit with clear instructions.

Instructions for collection: Vaginal Samples

VI. Soil samples

Soil microbial communities are both numerous ($10^{10}$ individual microorganisms/gram soil) and diverse (104 bacterial species/gram soil). Soils can also present high spatial variability in the form of microenvironments, so replication is very important when designing soil microbiome studies. The most studied agricultural microbiome is the plant’s rhizosphere: the microorganisms (including mycorrhizal fungi) living at the interface of the plant’s root system and the surrounding soil. Soil microbes are an important component of many biogeochemical cycles including Carbon, Nitrogen, Oxygen, and Sulfur cycles.

Instructions for collection: Soil Samples

VII. Water samples

Water samples from the environment often contain a high diversity of microbes, including bacteria and archaea. Check out this link to a protocol prepared by Dr. Steven Hallam’s lab, which outlines small volume filtration of water samples. An important step is to obtain at least 100 mL...
Instructions for collection: Water Samples

External contamination can occur when microorganisms from other sources are present in your sample. Many potential sources of external contamination exist: cells from the person taking the sample, microorganisms from the air or nearby surfaces, or the ‘kitome’ from various components of commercial reagents and kits. Low-biomass samples (such as those with increased host DNA) are especially susceptible to external contamination. You may decide to provide an external contamination control, for example, by taking a fresh swab and swirling it in the air in the sampling environment. A service provider can further discuss appropriate controls for your study.

Sample considerations

Find below the important considerations for sample collection, storage, and shipment to the lab for analysis.

- **Sample quantity:** You should obtain a quantity of sample that is adequate for at least 2 processing rounds. The amount needed for analysis varies by sample type. For instance, stool samples require very little quantity, but stool consistency can significantly impact downstream microbiome analyses and can be considered a confounding factor. The Bristol Stool Scale is often used to record stool consistency at the time of collection. Learn more about the [Bristol Stool Scale here](#). Also, consider a higher sample quantity if the body site is low in microbial biomass.

- **Sampling device/strategy:** The practicalities of the sampling device and protocol are determined by factors such as the facility used for sample collection and personnel that will be collecting them. For example, will the samples be collected by research staff, or by the study participants themselves? Where possible, samples should be collected in 2 sets so you have a backup sample in case of unforeseen problems such as shipping delays or loss. Multiple collection device types exist: swabs, tubes, hats, and others. In choosing your device, also consider what needs to be analyzed in addition to DNA – for example, RNA, short-chain fatty acids, or additional metabolites.

- **Sample storage:** Consider the storage conditions required for your samples from the moment they are collected to when they are received in the lab. These conditions are important for preserving the samples until further processing. The gold standard for fecal and other samples is to ‘snap freeze’ the sample (using dry ice) at -80°C as quickly as possible, or to immediately extract the DNA. Soil samples, tissue samples, and rodent fecal samples should all be frozen immediately if possible.

Some sample collection kits enable the samples to be stored at ambient temperature for a period of time. Generally, storage kits provide better support in case of unfortunate events such as thaw during sample shipment. If you are shipping your samples to a lab, check with the lab about whether it has an appropriate freezer, and adequate space in the freezer, to store your samples. Whatever sample collection device you are using, you can look for storage requirements validated by the studies in the scientific literature. Kit providers will have both instructions and...
validation study papers posted on their websites or available upon request. Note that if you are planning for multiple downstream applications, it’s best to consider multiple samples with appropriate storage for each one. For example, if you are interested in both SCFAs and V4 region of the 16S rRNA gene and you cannot store the samples at -80°C, your 16S samples can be preserved via a collection kit storage device at ambient temperature for several days, while your SCFA samples need to be preserved on dry ice or in a -20°C freezer for the short term. Some companies now offer storage devices for fecal metabolomic analysis, although the accuracy can be reduced due to the inconsistency of the buffer composition.

Some available devices for sampling and storage are produced by the following companies:

- DNA Genotek
- Norgen Biotek
- RNAlater
- Zymo Research (also used for mock community controls)

Learn more on choosing the right collection device here:

Considerations when choosing a collection device for your fecal samples

- **Sample collection timing:** Special timing considerations are needed if you intend to use metatranscriptomic profiling. Since mRNA is unstable and short-lived it is important to select a sampling period and location in which you expect your activity of interest to be occurring. For example, if you are interested in cold resistance genes, you need to sample after a transition to cold conditions. Additionally, special care is needed to stabilize the sample using very low temperatures (e.g. snap freezing to -80 °C using liquid nitrogen) or stabilization solutions (e.g., RNAlater) to avoid mRNA degradation.

- **Shipping:** Samples being stored frozen should be shipped on dry ice with enough dry ice to last for the duration of transport. Samples collected using a stabilization buffer for storage at ambient temperature need not be kept cold unless you will be storing your samples longer than the kit’s stability timeframe as outlined by the collection kit manufacturer. Before you collect your samples, you should understand applicable shipping regulations to your country of destination, to avoid the loss of valuable samples. Permits are required to import certain sample types, such as soil. Generally, a waybill and commercial invoice are required for the samples to cross an international border. When preparing your samples for shipping to an external lab, you are usually required to submit the order request form or sample manifest before shipping your samples. It is important to ensure all samples are labeled according to the identification numbers provided on the sample manifest. Samples should also be packaged in the same order as listed in the sample manifest.
DNA Extraction

A variety of DNA extraction kits are available. Extraction involves (1) mechanical and chemical lysis of cells, and (2) removal of inhibitors (humic acid and phenol, primarily). It’s best to use both approaches to account for different cell walls and cell membranes that may be present in the samples. It is important to note that DNA extraction can introduce bias (see below), and as such it influences the analysis results. A whole cell reference reagent can be used to evaluate DNA extraction choices, as well as downstream sequencing and bioinformatic pipelines.

In choosing a DNA extraction protocol, you will need to consider both the sample type and the required yield – purity as well as volume and concentration. Further considerations are:

- **Host DNA:** For low-biomass samples, consider whether host DNA will interfere with the type of analysis you desire. Various host DNA depletion methods (e.g. kits) are available, but their use must be undertaken cautiously because these methods may simultaneously reduce your ability to detect certain microorganisms in the sample.

- **Automation:** Some labs have steps that are automated (handled by a robot) as opposed to manual handling. Automation helps increase consistency in sample processing, because there is less handling and contact, but automated pipelines may be more error-prone under certain circumstances (e.g. when there is a low microbial load). Time and cost are the main considerations for deciding whether to use automated tools.

- **Quality control:** A quality-control step after DNA extraction should be considered, especially in pilot projects. Qubit is good for quantification / absolute DNA concentration, while Nanodrop is good for purity assessment (as it tends to over-estimate DNA concentration). Typically, a lab might Qubit quantify a subset of samples from a project to determine what range of DNA quantity was extracted. Technicians also might Qubit quantify positive and negative controls to determine if the DNA extraction process performed as expected. After this initial DNA extraction quality control step, post-PCR each sample might be run into gel electrophoresis to determine successful amplification.

**Library preparation considerations**

Before DNA samples can be sequenced by next-generation sequencing, they must be fragmented, end-repaired, and collected into adapter-ligated libraries. Library preparation protocols can influence the results generated by your sequencing experiments. These practices can help avoid problems with library preparations:

- Collect all samples, randomize, extract, and sequence all at once if possible.
• Be aware of “repeat offenders” – contaminants such as *Staphylococcus aureus* and *Escherichia coli* – that are commonly seen.

**Contamination**

Contamination is when microbial sequences inferred to belong to one sample do not in fact originate from that sample. This becomes increasingly problematic with decreasing biomass in the sample. Here are common sources of contamination:

• **Well-to-well:** This contamination occurs between wells in a well plate, usually between neighboring ones. The greatest risk of well-to-well contamination exists in DNA extraction, and to a lesser extent in library preparation.

• **Index hopping:** Also called index swapping or barcode mis-assignment, this phenomenon occurs when multiplexed samples are sequenced on next-generation sequencing platforms and some sequencing reads are assigned to an incorrect sample in a pool.

**Best practices for controlling contamination**

Microbiome Insights has its own set of negative and positive controls, which guard against any external contamination that may possibly come from the lab. The following practices can help control contamination:

• Sequencing a positive control not expected in real samples
• Including at least 2 types of negative controls (for instance, a DNA template-free control and an extraction control)
• Using mock communities for assessing bias in DNA extraction
• Using dual barcodes to help control index hopping
• Isolating pre- and post- PCR work
• Avoiding a liquid handling robot for DNA extraction

**Other sources of variation (PCR bias)**

• Choice of primers
• Chimeras
• Misincorporation of nucleotides
• PCR drift (accumulation of random amplification)
• Choice of high-fidelity polymerase and number of PCR cycles

[Resources on PCR bias in Nature Biotechnology](#)

[Resources on PCR bias in mSphere](#)
Quantitative polymerase chain reaction (qPCR), also known as Real-Time PCR, is a method that measures the number of copies of a DNA region defined with a particular PCR primer or primers. Using this method, technicians specifically amplify the 16S rRNA gene amplicon (V4 region) and quantify the rRNA gene copy number in each sample to estimate the total bacterial load. This technique can also be used to determine the feasibility of your microbiome study by providing an exact measure of bacterial abundance in your samples. Using this technique is highly recommended prior to sequencing samples suspected to have low biomass, to determine whether these samples will result in positive amplification.

Species-specific qPCR may also be an option. Microbiome Insights offers S. aureus and Cutibacterium acnes species-specific qPCR quantification methods, which are validated and well-referenced. These are either SYBR-based qPCR or Taqman Probe-based qPCR. Not all assays target the 16S rRNA gene.
A range of sequencing options is available, each method having its benefits and limitations. Typically, the kinds of analyses you can expect from a lab specializing in microbiome studies are as follows:

I. Sequencing of 16S, 18S, or ITS2

Amplicon sequencing, or rRNA gene sequencing, is performed to determine the relative abundance of taxa in a bacterial community – in other words, “what is there” – and to compare between groups of interest. This level of analysis can help determine changes in the overall microbial profile over time, or between treatment groups. Sequencing of short-read amplicons will give you reliable identification to the genus level. Depending on the sample type, the following amplicons are commonly used:

- Prokaryotic communities (16S gene V4 region)
- Eukaryotic communities (18S gene)
- Fungal communities (ITS2)

Amplicon sequencing is cost-effective and good for providing an overview of bacteria within a microbial community when functional profiling is not required. This method has a medium to high risk of bias, with the analyzed taxonomic composition depending on the selected primers and targeted variable region.
For amplicon sequencing, the sequencing reads are clustered into operational taxonomic units (OTUs). Whereas OTUs are clustered based on 97% similarity, an alternative approach using amplicon sequence variants (ASVs) clusters based on 100% sequence similarity. While in theory the ASV approach yields better taxonomic resolution, the achievable resolution is inherently limited by the use of short sequences. In some cases, the ASV approach may result in a single bacterial genome being erroneously split into separate clusters. Because ASVs are “exact”, they can be reused across studies, a key advantage over OTUs inferred de novo. But recent algorithmic advances now enable clustering of new sequences into existing OTUs.

II. Shotgun Metagenome Sequencing (SMS)

Shotgun metagenome sequencing is performed for high-resolution taxonomic profiling (diversity and abundance), as well as functional analysis, or “what the community members are capable of doing”. This technique allows for sequencing of DNA from all organisms within the community, with the ability to identify species with greater accuracy than short-read amplicon sequencing. The data generation allows for more advanced reporting and genome assemblies. SMS is defined operationally as anything higher in depth than 2M reads per sample.

SMS provides more information than 16S rRNA gene sequencing, but the increased cost per sample can be a limitation. Cost generally depends on the sequencing depth required. For certain
sample types (skin, for example; see above), this method is highly sensitive to host DNA contamination because it sequences all DNA, including that of the host.

Sometimes SMS is paired with transcriptomic analysis to give insights into both functional capacity and activities of the microbes.

### III. Shallow Shotgun Metagenome Sequencing (SSMS)

This technique involves sequencing samples at a shallower depth than is applied in full shotgun metagenome sequencing. SSMS yields high-resolution output for taxa and functional profiles, similar to SMS. By combining many more samples into a single sequencing run and using a modified protocol that uses a lower volume of reagents for sequencing library preparation, SSMS is an economical way to provide compositional and functional sequencing data similar to deep shotgun metagenome sequencing. For host-associated samples, however, host contamination can pose a problem. SSMS is best suited for samples that are rich in microbes and contain low levels of host DNA contamination such as fecal samples. Skin swabs and biopsies can have 30-90% host contamination, for example, which can lead to insufficient usable data with a SSMS approach.
When SSMS is technically feasible, it offers a relatively economical way to obtain data, particularly in exploratory and high-biomass-sample studies.

16S rRNA gene sequencing vs. shotgun metagenomic sequencing

Shallow shotgun sequencing: A primer
Metatranscriptomics involves profiling gene expression in your samples, via analysis of the actively transcribed ribosomal and messenger RNA from a microbial community, to discover “what the community members are actively doing”. The power of this technique is in elucidating how microorganisms respond to rapid changes in their environments – i.e. the cellular response to various conditions. High-throughput RNA sequencing can provide insights into gene expression patterns as well as metabolic activities and functional roles of different microbial species in an ecosystem.

RNA extraction is the most critical step for obtaining high-quality RNA for cDNA synthesis and sequencing. After extraction, roughly 80% of the total RNA is ribosomal; therefore, rRNA depletion is sometimes used to reduce the rRNA load and increase the messenger RNA (mRNA) content in the mixture. This procedure varies in efficiency and is not recommended in low biomass samples. Another option is RNA/DNA coextraction, which enables paired metagenomics and metatranscriptomics sequencing.

The two options for RNA library preparation are rRNA depletion and poly(A) mRNA selection. Option 1 is the preferred method, in particular, for prokaryotic/bacterial organisms.

Following total RNA extraction and purification, including the removal of residual DNA by DNase treatment, rRNA must be depleted to obtain adequate sequencing coverage. Quality control checks must be carried out after these processes. Prior to sequencing, RNA is generally fragmented, size-selected, and converted to double-stranded cDNA using single-strand cDNA conversion and second-strand preparation. This process eliminates strandedness; analysis depending on strand selection (e.g. Anti-sense Adaptors and barcodes) are then ligated to the cDNA fragments during library preparation.
Metagenomics Pipeline

Sample

↓

RNA extraction + rRNA depletion

↓

Library preparation

↓

Illumina reads

↓

Quality filter

↓

Genome search

Gene prediction

↓

Functional annotation (KEGG, COG, pfam)

↓

Read mapping

Genome dereplication
7 Short-Chain Fatty Acid Analysis

Short-chain fatty acids (SCFAs) are the fermentation products of indigestible fibers from diet (e.g. cellulose, resistant starch) by the bacteria in the gut, and include acetic acid, propionic acid, butyric acid, isobutyric acid, valeric acid, isovaleric acid, hexanoic acid, and heptanoic acid. These fatty acids have been shown to play an important role in regulating metabolism via the gut and are closely associated with certain gastrointestinal conditions and other diseases such as type 2 diabetes. SCFAs in stool are seen as a proxy for gut health and are related to inflammation. Learn more about how to interpret your fecal short-chain fatty acid data below.

Interpreting your fecal short-chain fatty acid data

The volatility of SCFAs in stool samples, however, poses a challenge when it comes to research protocols. If you are interested in incorporating SCFA analysis into your study, it is important to consider your sample collection and stabilization options, including temperature control. Frozen samples provide the best results, and ideally, samples for SCFA analysis do not contain any buffer, since it may affect accuracy and the amount of buffer must be known. In response to questions from our clients, our team put together a short internal investigation on how storage time, temperature, and stabilization reagent affect the quantities of SCFAs detected in a sample. Our white paper details the results of this investigation and is available below.

Factors that affect short-chain fatty acid preservation in human stool samples
Bioinformatics is an interdisciplinary field of science that combines biology, computer science, information engineering, mathematics, and statistics to analyze and interpret biological data. Once sequencing has been completed, you should be left with the raw sequencing files in FASTQ format – the standard for storing the output of high-throughput sequencing instruments such as the Illumina sequencers. From here, several subsequent steps are required to process the data before you are ready for analysis and report generation.

For amplicon-based studies, processing involves removing low quality sequences, primer regions, and non-specific amplicons, then comparing the sequences against reference databases to identify the organisms present. The most popular amplicon processing workflows are Mothur (the default at Microbiome Insights), QIIME, and DADA2. For shotgun metagenomics, processing involves removing low-quality sequences, artifacts, and sequences matching against the host. High-quality sequences are then compared against taxonomic and functional databases. In addition, sequences can be assembled into contigs or partial genomes, depending on the objectives of the study, and more importantly, sequencing depth.

**Quality control plots**

The purpose of this step is to evaluate the sequencing quality and look for samples that could skew the analysis. Sequence base call quality is plotted across the sequence length using software such as FastQC. Quality is expected to decrease across the sequence, particularly for the reverse (R2) reads. Q values of 30 are usually good (99.9% accuracy). Sequencing depth and summaries of the QC process are also useful in evaluating the quality of the sequencing data. Our team typically removes samples with poor sequencing depth compared with the rest of the project samples. For SMS samples, we additionally evaluate the host contamination to look for outliers. Variable host contamination could lead to highly variable sequencing depth, which makes the data noisier.

**Data exploration**

Barplot, heatmaps, and ordinations allow researchers to compare the composition of the samples across sample groups. Barplots are usually done for taxonomic profiles at different levels of taxonomy and for different taxa. For barplots, our team uses relative abundances (adding up to 100%) unless some
quantitative information is also available, such as qPCR data. Heatmaps are frequently used for functional profiles to highlight changes in specific taxa or functions. Ordinations are data reduction techniques to visualize the similarity of taxonomic or functional profiles of the project samples. Samples that are closer in the ordination have more similar profiles.

Hypothesis tests

Depending on the study design, different hypothesis tests can be used to look for changes across groups or gradients or for individual taxa/functions associated with a condition. Commonly used techniques are PERMANOVA, which involves looking for the multidimensional centroid of a group of samples and comparing it against the centroid of other groups to identify differences in the profiles due to one or more specific categorical factors (e.g. taxonomic profiles changed after treatment A). Differential abundance testing involves looking for specific groups that have a specific pattern across sample groups (e.g. taxon C was more abundant in treatment B samples).

When multiple datasets are available (for example, two or more -omics types) it is possible to integrate them using methods such as variation partitioning (which explores the contribution of different data matrices to the microbiome data), and tools such as Mixomics.

Diversity

One way of summarizing data is to calculate diversity. For calculation of this parameter, you have the choice of which taxonomic level to use (genus, species, or strain) and which estimate of diversity parameter.

- **Alpha-diversity** is a summary of the diversity of one community. Species richness, Simpson’s index, inverse Simpson, Shannon’s diversity, and Shannon’s richness are different ways of estimating diversity. For example, species richness is the total number of microbial taxa present in a sample, while Shannon’s index accounts for both taxonomic abundance and evenness (how evenly distributed the OTUs are).

- **Beta-diversity** compares multiple communities; it captures how the diversity of one community relates to another. All profiles are compared in a pair-wise fashion to determine a distance score and store it in a distance matrix. Similarity is a term often used in beta-diversity comparisons, so two identical samples have a distance of 0 and a similarity of 1. Some measurements of distance use only the presence/absence information while others, such as the Bray-Curtis dissimilarity, also include the abundance (Bray and Curtis, 1957).

Although diversity is often calculated, it is important to remember its limitations. Real abundances or number of species remain unknown. Moreover, a change in diversity will not tell you which taxa have changed.
Example Reports

Reports are powerful tools for telling the ‘story’ of the data, provided you carefully consider the context as well as the specific design. If you use a microbiome analysis service provider, reports can be customized to suit your scientific question.

Available report formats

- For amplicon studies, standard bioinformatic reporting includes plots showing taxonomic composition, alpha-diversity, beta-diversity (ordination), differential abundance testing, and multivariate analysis (permutational analysis of variance). This is accompanied by raw bioinformatics outputs (OTU table, taxonomic classification of each OTU), differential abundance results, and relative abundances of taxa at various taxonomic levels.

Download Amplicon Sequencing Sample Report

- For shotgun metagenome sequencing, high-resolution taxonomic and functional profiles are analyzed, observing the same principles that guide the analysis of amplicons. Our team’s main metagenomics pipeline is based on tools developed to characterize environmental microbiomes – specifically Kraken2 and Superfocus – but we can also accommodate other approaches such as Metaphlan4 and Humann3.

Download Shotgun Sequencing Sample Report

You need to settle on your statistical approaches and models: for example, factorial, nested, or repeated measures design. Compared with amplicon sequencing, metagenomics has different aims for its output, involving much more thorough functional and gene mapping as a necessary element of reporting.

For metatranscriptomic sequencing, functional annotation and reference alignment are key steps, followed by differential expression analysis.

- Advanced bioinformatic reporting may include:
  - Microbial epidemiology
  - Multiple regression
  - Partitioning sources of variation
  - Multi-omics integration analysis
  - Clustering algorithms
  - Machine-learning-based modeling
  - Biomarker discovery
We conclude this guide with a glossary of some technical terms used in the microbiome literature.

**Differential abundance (DA) testing**
The goal of DA testing is to identify microbial features whose abundance differs across conditions (e.g. treatment A vs. control). There are multiple avenues for estimating the statistical significance of DA tests. Our team typically models feature counts using the negative binomial distribution, but DA testing is an area of active research. See likelihood-ratio test below.

**Dissimilarity matrix**
A matrix of sample-to-sample dissimilarity values as calculated by the metric of choice. The Bray-Curtis dissimilarity matrix is the most conventionally used method to assess beta-diversity, for example.

**Diversity**
The amount of variation of microbial community structure. Measures of diversity are derived from tables of relative abundance and/or prevalence.

**Likelihood-ratio test (LRT)**
The LRT examines two models for the counts: a full model with a certain number of terms, and a reduced model, in which some of the terms of the full model are removed. The test determines if the increased likelihood of the data using the extra terms in the full model is more than expected if those extra terms are truly zero. The LRT is therefore useful for testing multiple terms at once – for example, testing 3 or more levels of a factor at once, or all interactions between two variables.

**Amplicon Sequence Variants (ASVs)**
ASVs are a way of clustering 16S rRNA gene sequences into taxonomic groups, as an alternative to OTUs (below). The ASV approach clusters sequences based on 100% similarity. This approach has gained popularity in recent years, and often (but not always) results in better taxonomic resolution. ASVs, when used alongside good de-noising procedures, are superior at reconstructing both low and high diversity mock communities.

**Operational taxonomic units (OTUs)**
OTUs are mathematical units used to cluster groups of closely related microbial DNA sequences into taxonomic groups. OTUs defined at 97% sequence similarity are loosely estimated as a species. This approach ‘blurs’ highly similar sequences in order to minimize the influence of any sequencing errors that may have occurred. The choice of algorithms strongly influences OTU calculations.
**Negative binomial distribution**
This distribution is fairly common to model overdispersed count-based data (such as microbiome), where the number of biological replicates may be low, and the data are not normally distributed. Statistical tests that utilize assumptions of normality cannot be used to accurately assess means and variances of microbiome data.

**Ordinations**
Ordinations are dimensional-reduction techniques that are used to visualize complex relationships between communities. Using a dissimilarity matrix (see above in beta-diversity), ordinations attempt to capture sample-to-sample relationships in a 2-dimensional plot. Common methods include Principal Coordinates Analysis (PCoA) and Non-Metric Multidimensional Scaling (NMDS) analysis.

**PERMANOVA**
Permutational Analysis of Variance (PERMANOVA) tests for significant differences in the whole microbiome among discrete categorical or continuous variables. The samples are randomly reassigned to the various sample categories (Monte-Carlo Permutations), and the between-category differences are compared to the true between-category differences. PERMANOVA utilizes the sample-to-sample distance matrix (e.g., Bray-Curtis) directly to perform the calculation.

**Relative abundance**
Relative abundance refers to how common or rare an OTU is relative to other OTUs in the same community. Relative abundance is useful for monitoring changes in a taxon’s abundance across samples. The abundance scores are the counts of reads assigned to each OTU.

**Richness**
Richness is the sum of unique OTUs found in each sample. Shannon's diversity utilizes the richness of a sample along with the evenness (how evenly the OTUs are distributed) to calculate a diversity index.
Have a question for us?
Would you like to talk with one of our scientists? We’re here to help.

Are you planning a study and require a sequencing provider?

Request a quote

Contact Us:
microbiomeinsights.com/contact-us

Visit our website:
microbiomeinsights.com