Supplementary Box 1 Problems and solutions for study design

Powering the study & **Read depth requirements.** The number of samples and sequencing depth required to be able to detect significant differences will depend on factors such as consistency of microbiome composition between different samples, the inherent microbial diversity of the samples, and effect size of the phenomenon being studied. **Solution**: decisions can be guided by results from previous studies in the same type of environment. In cases where this information is lacking it may be prudent to carry out preliminary marker gene-based studies to gauge the relative impact of each of the factors listed opposite.

Confounding variables and control groups. It is often very difficult to select a control group to compare against the samples of interest that is free from other confounding variables. An example of this is rodent microbiome research, where cage and animal batch effects can result in dramatic differences in microbiome composition, independent of the variable being studied [1]. Another example is the cross-sectional study of the microbiome associated with a disease for cases in which the patients cannot be sampled in the absence of active treatment. **Solution:** Current best practice is to collect as much metadata about each of the study groups as possible and factor these into the subsequent analyses when comparing groups. For clinical samples this typically includes features such as gender, age, antibiotic/medication use, location, dietary habits, and Bristol stool chart scores. For environmental samples this commonly includes associated parameters such as geographic location, season, pH, temperature etc. Further extensive advice for planning rodent microbiome studies is available [1]. Longitudinal sampling from the same patient/location can also act as an additional control, especially when longitudinal changes can be correlated with associated metadata.

Sample collection/preservation. It may be difficult to process and store all samples in exactly the same way (for example when samples are provided from a number of locations by different research groups). With longitudinal studies, samples collected at the final time point may spend less time in frozen storage prior to DNA extraction than samples collected at other time points. Such changes in sampling and preservation procedures may introduce systematic biases. **Solution:** Where possible, collection and preservation methodologies should be standardized throughout for all samples within a given study. All procedures used should also be recorded and included as pertinent metadata when carrying out subsequent data analyses. This should ideally include factors such as time between collection and DNA extraction, length of time in frozen storage, and number of freeze-thaw cycles. For mammalian gut samples there is some evidence that storage in glycerol may result in more representative compositional results following long term frozen storage [2]. Similarly, freeze drying prior to long-term frozen storage may be a prudent approach [3].

Biomass/Contamination. Modern sequence based technologies are highly sensitive, meaning very small amounts of DNA are sufficient for sequencing. However, common laboratory kits and reagents are not sterile, meaning that any contamination that is present in these can potentially overwhelm the “real” signal in samples containing only a very low microbial biomass [4]. **Solution.** It is prudent to gauge the level of biomass present in samples before sequencing using a quantitative approach such as qPCR. Samples containing fewer than 10^5 microbial cells appear to be most impacted by background contamination [4]. Table 1 offers some approaches that may be tried in order to enrich cell numbers/DNA yields from samples prior to sequencing. Negative control samples, that have been processed using the same kits/reagents as the actual samples, should be sequenced in order to determine the types of contaminating microbes present. Sequence data derived from these contaminants might then be removed bioinformatically from the final sequence datasets. Note that the sensitivity of these negative controls can be enhanced by the use of carrier DNA [5].

Choice of DNA extraction methodology. This step can hugely impact the results of a metagenomics study. If the approach selected is not stringent enough to extract DNA from some cell types they will not be represented

accurately in the subsequent sequence data. Fundamentally, the optimal type of DNA extraction approach will depend on the underlying composition of the cell types that are present within a given sample. Unfortunately this can vary greatly, even within the same type of sample (e.g. the faeces of some humans are dominated by Gram negative species with cell walls that are relatively easy to disrupt, while those of others are dominated by relatively recalcitrant Gram positive species). As a result, no one DNA extraction approach will work optimally for all sample types. **Solution**: The use of defined mock community controls [6] consisting of cultures derived from a mixture of the types of species that are common within a given environment can be a useful starting point to test the efficiency and accuracy of different DNA extraction methods. Mock communities can be optimized by including a phylogenetically diverse collection of species that are known to be commonly abundant in the sample type being studied. However, it is difficult to mimic the complexity of real microbial communities using simplified mocks, and impossible to test for the efficiency of the extraction step for unknown/uncultured organisms. Much evidence suggests that incorporating a bead-beating step into the DNA extraction process improves yield and representativeness of resulting species profiles compared to chemical-only lysis [7,8](ref#133 C.Q.,N.J.L.). However, this type of approach does typically result in more sheared DNA, potentially limiting the power of burgeoning long read sequencing technologies. DNA extraction methodology should also be included as crucial metadata when uploading sequence data to public repositories. This allows variance in methodology choices to be factored into subsequent meta-analyses that incorporate metagenomic datasets from different laboratories.

1 Laukens, D., Brinkman, B. M., Raes, J., De Vos, M. & Vandenabeele, P. Heterogeneity of the gut microbiome in mice: guidelines for optimizing experimental design. *FEMS microbiology reviews*, fuv036 (2015).

2 McKain, N., Genc, B., Snelling, T. J. & Wallace, R. J. Differential recovery of bacterial and archaeal 16S rRNA genes from ruminal digesta in response to glycerol as cryoprotectant. *Journal of microbiological methods* **95**, 381-383, doi:10.1016/j.mimet.2013.10.009 (2013).

3 Kia, E. *et al.* Integrity of the Human Faecal Microbiota following Long-Term Sample Storage. *PloS one* **11**, e0163666, doi:10.1371/journal.pone.0163666 (2016).

4 Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology* **12**, 87, doi:10.1186/s12915-014-0087-z (2014).

5 Xu, Z. *et al.* Improving the sensitivity of negative controls in ancient DNA extractions. *Electrophoresis* **30**, 1282-1285, doi:10.1002/elps.200800473 (2009).

6 The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214 (2012).

7 Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z. & Forney, L. J. Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One* **7**, e33865 (2012).

8 Gerasimidis, K. *et al.* The effect of DNA extraction methodology on gut microbiota research applications. *BMC Research Notes* **9**, 365, doi:10.1186/s13104-016-2171-7 (2016).

130.