Towards standards for human fecal sample processing in metagenomic studies

Costea, P.I., Zeller, G., Sunagawa, S., Pelletier, E., Alberti, A., Levenez, F., Tramontano, M., Driessen, M., Hercog, R., Jung, F.E., Kultima, J.R., Hayward, M.R., Coelho, L.P., Allen-Vercoe, E., Bertrand, L., Blaut, M., Brown, J.R.M., Carton, T., Cools-Portier, S., Daigneault, M., Derrien, M., Druesne, A., de Vos, W.M., Finlay, B.B., Flint, H.J., Guarner, F., Hattori, M., Heilig, H., Luna, R.A., van Hylckama Vlieg, J., Junick, J., Klymiuk, I., Langella, P., Le Chatelier, Z., Mai, V., Manichanh, C., Martin, J.C., Mery, C., Morita, H., O'Toole, P.W., Orvain, C., Patil, K.R., Penders, J., Persson, S., Pons, N., Popova, M., Salonen, A., Saulnier, D., Scott, K.P., Singh, B., Slezak, K., Veiga, P., Versalovic, J., Zhao, L., Zoetendal, E.G., Ehrlich, S.D., Dore, J., Bork, P.

Nature Biotechnology
2017 NOV ; 35(11): 1069-1076
2017 OCT 02 (first published online)
doi: 10.1038/nbt.3960

URL: https://www.nature.com/articles/nbt.3960

Publisher: Nature America (Springer Nature)

Towards standards for human fecal sample processing in metagenomic studies

Paul I. Costea[1], Georg Zeller[1], Shinichi Sunagawa[1,2], Eric Pelletier[3,4,5], Adriana Alberti[3], Florence Levenez[6], Melanie Tramontano[1], Marja Driessen[1], Rajna Hercog[1], Ferris-Elias Jung[1], Jens Roat Kultima[1], Matthew R. Hayward[1], Luis Pedro Coelho[1], Emma Allen-Vercoe[7], Laurie Bertrand[3], Michael Blaut[8], Jillian Brown[9], Thomas Carton[10], Stéphanie Cools-Portier[11], Michelle Daigneault[7], Muriel Derrien[11], Anne Druesne[11], Willem M. de Vos[12,13], B. Brett Finlay[14], Harry J. Flint[15], Francisco Guarner[16], Masahira Hattori[17,18], Hans Heilig[12], Ruth Ann Luna[19], Johan van Hylckama Vlieg[11], Jana Junick[8], Ingeborg Klymiuk[20], Philippe Langella[6], Emmanuelle Le Chatelier[6], Volker Mai[21], Chaysavanh Manichanh[16], Jennifer C. Martin[15], Clémentine Mery[10], Hidetoshi Morita[22], Paul O'Toole[9], Céline Orvain[3], Kiran Raosaheb Patil[1], John Penders[23], Søren Persson[24], Nicolas Pons[6], Milena Popova[10], Anne Salonen[13], Delphine Saulnier[8], Karen P. Scott[15], Bhagirath Singh[25], Kathleen Slezak[8], Patrick Veiga[11], James Versalovic[19], Liping Zhao[26], Erwin G. Zoetendal[12], S. Dusko Ehrlich[6,27,*], Joel Dore[6,*], Peer Bork[1,28,29,30,*]

* - Corresponding authors

[1] European Molecular Biology Laboratory, Germany
[2] Department of Biology, Institute of Microbiology, ETH Zurich, CH-8092 Zurich, Switzerland
[3] CEA-Institut de Génomique, Genoscope, Centre National de Séquençage, Evry, France
[4] CNRS UMR8030, Evry France
[5] Université Evry Val d'Essonne, Evry, France
[6] Metagenopolis, Institut National de la Recherche Agronomique, Jouy en Josas, France
[7] The University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada
[8] Department of Gastrointestinal Microbiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Arthur-Scheunert-Allee 114-116, 14558 Nuthetal, Germany
[9] School of Microbiology & APC Microbiome Institute, University College Cork, T12 Y337 Cork, Ireland
[10] Biofortis, Mérieux NutriSciences, France
[11] Danone Nutricia Research, Palaiseau, France
[12] Laboratory of Microbiology, Wageningen University, Stippeneng 4, 6708 WE, Wageningen, The Netherlands
[13] Department of Bacteriology and Immunology, Immunobiology Research Program, Haartmaninkatu 3 (PO Box 21), FIN-00014 University of Helsinki, Finland
[14] Michael Smith Laboratories, University of British Columbia, Vancouver, B.C., Canada
[15] Rowett Institute of Nutrition and Health, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, UK
[16] Digestive System Research Unit, Vall d'Hebron Research Institute, CIBEREHD, Barcelona, Spain
[17] Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8561, Japan
[18] Graduate School of Advanced Science and Engineering, Waseda University, Tokyo 169-8555, Japan
[19] Texas Children's Hospital, 1102 Bates Avenue, Feigin Center, Houston, TX 77030, United States
[20] Center for Medical Research, Medical University of Graz, Graz, Austria
[21] Department of Epidemiology, College of Public Health and Health Professions and College of Medicine, Emerging Pathogens Institute, University of Florida, 2055 Mowry Rd., Gainesville, FL 32610-0009, United States
[22] Graduate School of Environmental and Life Science, Okayama University, Okayama 700-8530, Japan
[23] School of Nutrition and Translational Research in Metabolism (NUTRIM) and Care and Public Health Research Institute (Caphri), Department of Medical Microbiology, Maastricht University Medical Center, Maastricht, The Netherlands
[24] Unit of Foodborne Infections, Department of Bacteria, Parasites & Fungi, Statens Serum Institut, Artillerivej 5, 2300 Copenhagen, Denmark
[25] Centre for Human Immunology, Department of Microbiology & Immunology and Robarts Research Institute, University of Western Ontario, London, Ontario N6A 5C1 Canada
[26] Ministry of Education Key Laboratory for Systems Biomedicine, Shanghai Centre for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, PR China
[27] King's College London, Centre for Host-Microbiome Interactions, Dental Institute Central Office, Guy's Hospital, UK
[28] Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany
[29] Molecular Medicine Partnership Unit, 69120 Heidelberg, Germany
[30] Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany

Abstract

Metagenomic analysis of fecal samples suffers from challenges in comparability and reproducibility that need to be addressed in order to better establish microbiota contributions to human health. To test and improve current protocols, we quantified the effect of DNA extraction on the observed microbial composition, by comparing 21 representative protocols. Furthermore, we estimated the effect of sequencing, sample storage and biological variability on observed composition, and show that the DNA extraction process is the strongest technical factor to impact the results. We characterized the biases of different methods, introduced a quality scoring scheme and quantified transferability of the best methods across labs. Finally, we propose a standardized DNA extraction methodology for human fecal samples, and confirm its accuracy using a mock community in which the relative abundances are known. Use of this methodology will greatly improve the comparability and consistency of different human gut microbiome studies and facilitate future meta-analyses.

Over 3000 publications in the past five years have used DNA- or RNA- based profiling methods to interrogate microbial communities in locations ranging from ice columns in the remote arctic to the human body, resulting in more than 160,000 published metagenomes (both shotgun and 16S rRNA gene)[1]. To date, one of the most studied ecosystems is the human gastrointestinal tract. The gut microbiome is of particular interest due to its large volume, high diversity and potential relevance to human health and disease. Numerous studies have found specific microbial fingerprints that may be useful in distinguishing disease states, for example diabetes[2–4], inflammatory bowel disease[5,6] or colorectal cancer[7]. Others have linked the human gut microbial composition to various factors, such as mode of birth, age, diet and medication[8–11]. Such studies have almost exclusively used their own specific, demographically distinct cohort and methodology. Given the many reports of batch effects[12] and known differences when analyzing data generated using different protocols[13–18], comparisons or meta-analyses are limited in their interpretability. For example, healthy Americans from the HMP study showed lower taxonomic diversity in their stool than patients with inflammatory bowel disease (IBD) from a European study[19], although it is established that IBD patients worldwide have reduced taxonomic diversity[20]. It is thus currently very difficult to disentangle biological from technical variation when comparing across multiple studies[21].

In metagenomic studies, the calculation of compositional profiles and ecological indices is preceded by a complex data generation process, consisting of multiple steps (Figure 1), each of which is subject to technical variability[22]. Usually, a small sample is collected by an individual shortly after passing stool and stored in a domestic freezer, prior to shipment to a laboratory. The location within the specimen that the sample is taken from has been shown to impact the measured composition[23], which is why in some studies[24] larger quantities were homogenized prior to storage in order to generate multiple, identical aliquots. Furthermore, different fixation methods can be used to preserve the sample for shipping and long-term storage. Freezing at below -20°C is the standard, though more practical alternatives exist[23–25]. Eventually, the sample is subjected to DNA extraction, library preparation, sequencing and downstream bioinformatics analysis (Figure 1).

Here we examined the extent to which DNA extraction influences the quantification of microbial composition, and compared it to other sources of technical and biological variation. The majority of the protocol comparison studies to date have used a 16S rRNA gene amplification approach, which suffers from additional issues. Specifically, the choice of primer, PCR bias and even the choice of polymerase can affect the results[26], which may lead to different conclusions when performing the same DNA extraction comparison in a different setup – issues that are minimized using metagenomic sequencing. We compared a wide range of extraction methods, using metagenomic shotgun sequencing, in respect to both taxonomic and functional variability, while keeping all other steps standardized. We investigated the most commonly used extraction kits with varying modifications and additional protocols which do not make use of commercially available kits (see Supplementary Table 1 and Supplementary Information). While other studies have previously investigated the differences between extraction methods in a given setting[12,15,16,27], we here systematically tested for reproducibility within and across laboratories on three continents, by applying strict and consistent quality criteria. We further assessed the accuracy of the best performing extraction methods by using a mock community of ten bacterial species whose exact relative abundance was known. This community included both gram-positive and gram-negative bacteria and their relative abundance spanned three orders of magnitude. Based on these analyses we recommend a standardized protocol for DNA extraction from human stool samples, which, if accepted by the research community, will greatly enhance comparability among metagenomic studies.

Results

Study design

This study consisted of two phases. In the first phase, in order to assess the variability introduced by different extraction methods, we produced multiple aliquots of two stools samples (obtained from two individuals, referred to as sample A and B). Within two hours of emission, the samples were homogenized in an anaerobic cabinet to ensure that the different aliquots have identical microbial compositions, and subsequently aliquoted in 200mg amounts, frozen at -80°C within four hours and shipped frozen on dry-ice to 21 collaborating laboratories, spanning 11 countries over three continents. These laboratories employed extraction methodologies ranging from the seven most commonly used extraction kits (Invitek's PSPStool, Mobio's PowerSoil, Omega Bio Tek's EZNAstool, Promega Maxwell, Qiagen's QIAampStoolMinikit, Bio101's G'Nome, MP-Biomedicals's FastDNAspinSoil and Roche's MagNAPureIII) to non-kit-based protocols (Supplementary Table 1 and Supplementary Information). Once extracted, the DNA was shipped to a single sequencing center (GENOSCOPE, France), which tested two different library preparation methods (see Methods), before performing identical sequencing and analytical methods in an attempt to minimize other possible sources of variation.

In a second phase, after applying a panel of quality criteria, including quantity and integrity of extracted DNA, recovered diversity and ratio of recovered gram-positive bacteria, we selected five protocols (1, 6, 7, 9, and 15). Extractions were then performed in the original laboratory applying the protocol and in three other laboratories, which had not used the method before, in order to assess reproducibility of these protocols and their transferability between laboratories. For the same samples A and B, three replicates/aliquots were provided per sample per laboratory, as detailed above. To quantify the absolute extraction error of the selected protocols, a mock community consisting of 10 bacterial species that are generally absent in the stool of healthy individuals

(Supplementary Table 2) was prepared, such that the cell density of all species in the mock community was determined. DNA was extracted from the mock alone as well as from eight additional samples, consisting of stool spiked with the mock in order to emulate a realistic setting. All extractions were done at a lab that had not previously used any of the three extraction methods, further testing the reproducibility of the methods.

Quality control for DNA yield and fragmentation

Maximizing DNA concentration while also minimizing fragmentation are key aspects to consider when selecting an extraction protocol. This is both because good quality libraries are required for shotgun sequencing and because protocols that consistently recover low yield or highly fragmented DNA are likely to skew the measured composition. We found considerable variation in the quantity of extracted DNA, in line with previous observations[28] (Figure 2). For example, protocol 18 recovered 100 times more DNA than protocols 3 and 12, and 10 times more than protocols 8, 19 and 20 (Figure 2). Furthermore, there was considerable variation in the fragmentation of the recovered DNA, as measured by the percentage of total DNA in fragments below 1.8 kb in length; for example protocols 4, 10 and 12 consistently yielded highly fragmented DNA while for protocol 1 no fragmentation was observed. For subsequent analysis, samples that yielded below 500ng of DNA or were very fragmented (median sample fragmentation above 25%), were not subjected to sequencing. In total, 143 libraries, extracted using 21 different protocols passed the quality requirements imposed above, though as an example only four of 18 samples extracted with protocol 16 (one sample A and three sample B replicates) met the requirements (Supplementary Table 3). For other protocols, a small number of samples were discarded for lack of compliance with quality/quantity criteria.

Quality control for variability in taxonomic and functional composition

All metagenomes were compared with respect to taxonomic and functional compositions to quantify the relative abundances of microbial taxa and their respective gene-encoded functions (Methods). Briefly, based on the extracted DNA, shotgun sequencing libraries were prepared and subjected to sequencing on the Illumina HiSeq2000 platform, yielding a mean of 3.8 Gb (+/- 0.7 Gb) per sample. Raw sequencing data were then processed using the MOCAT[29] pipeline and relative taxonomic and gene functional abundances were computed by mapping high-quality reads to a database of single copy taxonomic marker genes (mOTUs)[19] and annotated human gut microbial reference genes[30], respectively (Methods).

There are, as outlined above (Figure 1), many steps in which sample handling can differ and batch effects can be introduced. The resulting variation in taxonomic and gene functional composition estimates should be considered in terms of both effect size and consistency: if protocol differences lead to an effect larger than the biological variation of interest (e.g. in an intervention study), it will mask that signal. Consistent “batch effects” will introduce bias that can distort any meta-analysis even if their absolute size is comparatively small. It is thus important to minimize these biases in order to facilitate cross-study comparisons.

To contextualize the magnitude of the extraction effect, we compared the technical variation quantified here (caused by extraction protocol) to other technical and biological effects (Figure 3), assessed on available data from multiple other studies[23,24,31] (Methods). The greatest difference was observed between individuals, though we note incongruences in the size of this effect between cohorts, due to the extraction method used; protocols that generally underestimate diversity will cause samples to look more similar to each other (Supplementary Figure 1). Next was the within

individual variation, as measured between different sampling time points for the same individuals. This effect was much smaller than the between individual variation, resulting in individual-specific microbial composition preservation over time as noted before[19,23,32]. The smallest contributor observed, quantified on a small number of samples (n=7), was within specimen variation, resulting from sampling different parts of the stool itself[23]. In terms of technical sources of variation we have considered measurement errors (assessed through technical replication), library preparation, and effects introduced by the two most widely used preservation[23,24] methods (fresh freezing and RNAlater). It is important to note that these effects have not all been measured independently of each other, resulting in some of the quantified variations being a convolute of multiple effects (Figure3 – checkboxes).

Different distance measures can be used to assess the magnitude of these effects. We focused here on two, which are complementary in terms of the features of the data they consider and thus the dimensions, which become relevant. These distance measures were computed on both metagenomics operational taxonomic units (mOTUs[21]) and clusters of orthologous groups (COGs[33]) abundance data, to derive species and functional variation (see Methods). Firstly, we used a Spearman correlation to assess how well species abundance rankings are preserved and found that the variation between most extraction protocols is smaller than the technical within-specimen variation (summarized by the median, Figure 3a). This suggests that, with the exception of protocols 8 and 12, all others recover comparable species rankings. Consequently, if only the ranks are of interest, most of the available protocols would provide highly comparable results. However, for many applications the abundances of the taxonomic units are important and need to be commensurable. Using a Euclidean distance (which cumulates abundance deviations) we found that many protocols were not comparable and actually introduce large batch effects at the species level, with the median between-protocol distance being higher than the within-specimen variation (Figure 3a), hampering the comparability of samples generated with different extraction methods. To assess similarity between extraction protocol effects, we used principle coordinate analysis (PCoA, see Methods) to visualize these distance spaces (Supplementary Figure 2). These indicated that protocol 12, and to a lesser extent also protocols 3, 8, 11, 16 and 18, had abundance profiles that were different from most of the other protocols.

Analysis of functional microbiome composition, based on COGs (see Methods, Figure 3b), shows that the majority of extraction protocol effects were greater than biological variation within specimen and across time points within the same individual (Figure 3b), with some of them being greater even than between-subject variability. This may in part be due to the known relatively low variation between individuals in this space[31,34] and would dramatically influence conclusions taken from comparative studies.

Among the sources of technical variation, the within-protocol variation (i.e. measurement error) was consistently smallest, with the magnitude of the library preparation effect being comparable (Figure 3a,b). The variation introduced by storage method (RNA Later vs. frozen) was larger than within-protocol variation, and, as previously shown, smaller than within-specimen variation in taxonomic space[23,24].

Taken together, our analysis demonstrates that usage of different DNA extraction protocols resulted in large technical variation, both in taxonomic and in functional space, highlighting that this is a crucial parameter to consider when designing microbiome studies.

Quality control for species-specific abundance variation

Having quantified and contextualized the different biological and technical sources of variation, we next assessed the quality of different DNA extraction protocols[18,28,35] by investigating species-specific effects and measured diversity. We argue that this provides a good proxy for the estimation accuracy and is in principle applicable to any metagenomic sample without additional sequencing and cultivation efforts.

We investigated species-specific abundance variation to assess which were most influenced by the extraction protocols. For this, we compared the estimated abundance of a given species in all replicates of a given protocol to the abundances of that species in all replicates of all other protocols, by performing a Kruskal-Wallis test (see Methods). We then applied a false discovery rate (FDR) correction to the obtained p-values. Of the 366 tested species, we found 90 that were significantly affected by extraction protocol (q-value< 0.05). The majority of these were gram-positive, accounting for 37% (+/- 7%) of the sample abundance on average (Figure 4).

These results are in line with previous observations that gram-positive bacteria are more likely to be affected by extraction method[13,35] and are also to be expected based on our extensive knowledge of gram-positive cell walls and their considerably higher mechanical strength. These differences do not reflect the overall performance of any of the protocols, but highlight upper limits of the effect size that may be observed for these species. For a fair comparison, we contrasted the recovered abundance of some of the significantly affected species, to the mean of the top five highest estimates. This clearly showed that most protocols estimated considerably lower gram-positive bacteria fractions, while the variation in gram-negative abundance estimations is comparatively small (Figure 4).

As the observed biases hint at protocol-dependent incomplete lysis of gram-positive bacteria, we hypothesized that this would correspond to decreased diversity. We thus evaluated whether diversity is a good general indicator of DNA extraction performance. Using the Shannon diversity, which accounts for both richness and evenness, we saw that the recovered relative abundance of gram-positive bacteria correlates with the observed diversity, with a higher fraction of gram-positives resulting in higher diversity (Supplementary Figure 3). Furthermore, we found dramatically reduced diversity in protocols already determined to perform poorly from a DNA quality perspective (i.e. protocols 3, 11 and 12) (Supplementary Figure 4). We conclude that a diversity measure is a good proxy for overall protocol performance and accuracy of the recovered abundance profile.

Factors influencing DNA extraction outcome

Using diversity as an optimality criterion, we determined protocol parameters that are significantly associated with this indicator (Figure 5). For this purpose we focused on protocols that use Qiagen kits[15], namely numbers 5, 6, 8, 9, 11, 13, 15 and 20, which reduces the number of variables that can influence the outcome. We find that “mechanical lysis”, “zirconia beads” and “shaking” are positively associated with diversity. We note that there is no association with DNA fragmentation, as all of the samples extracted with these protocols had a low number of fragments below 1.8 kb (Figure 2). This was consistent with the notion that mechanical lysis and bead beating are necessary to efficiently extract the DNA of gram-positive bacteria that have cell walls that are harder to break[35] and also in line with our postulation that effective gram-positive recovery will increase the observed diversity. The only significant negative association was with the InhibitEX tablet, which was included in the kit and which the manufacturer recommends for “absorb[ing] substances that can degrade DNA and

inhibit downstream enzymatic reactions so that they can easily be removed by a quick centrifugation step"[36], though our assessment suggests an adverse effect on DNA extraction quality. This analysis suggests specific modifications with which – also currently suboptimal – extraction methods could be improved, independent of all other variables. For example, introducing a bead beating step is likely to improve the extraction, independent of the specific commercial kit used; adding such a step to the only protocol using Mobio's PowerSoil kit (protocol 3) would be expected to improve its performance. Our results may therefore generally inform the future development of better DNA extraction protocols.

Protocol reproducibility and transferability across laboratories

Based on the quality of the extracted DNA, species diversity as well as species-specific biases, we selected the five best performing protocols: 15, 7, 6, 9, and 1 (in this order), to be tested for reproducibility across laboratories (phase II). Protocols 15, 6 and 9 use the same Qiagen-based lysis and extraction kit and were combined into a slightly modified protocol, "Q" (Supplementary Information). Protocols 1 and 7 were coded as H and W, respectively.

Laboratories that originally delivered DNA based on the protocol implementations Q, W and H, replicated those extractions in phase II, ensuring that the variability was comparable to that observed in the first set of extractions (Supplementary Figure 5).

Each extraction method was established and performed in three other laboratories, which had no experience with the respective protocol, in order to assess the wider applicability of each as a standard extraction protocol. All three methods were reproducible across locations, though only protocol H had an effect below that of the smallest biological variation (i.e. within-sample). Protocols W and Q introduced a cross-lab effect comparable to within-sample variation (Supplementary Figure 5).

Although protocol H seemed to be more reproducible across facilities, it underestimated gram-positive bacteria compared to the other two protocols (Supplementary Figure 5, and protocol 1 in Figure 4) and so yielded less diverse estimates of microbial composition. Protocol W, while also more reproducible (Supplementary Figure 5 and protocol 7 in Figure 4), is impractical and hard to automate as it involves the use of phenol-chloroform. Protocol Q recovers a highly diverse estimate of the microbial composition which it appears to achieve through lysis of gram-positive bacteria and does so in a way that is easy to implement and use across facilities.

Protocol extraction accuracy

In order to estimate the accuracy of the proposed extraction methods, we designed a mock community, with known bacterial species and respective abundances, to use as a baseline quantification. While this provides a standard to compare to, the culturing, mixing and accurate abundance estimation of such a community are complex. Historically, multiple attempts have met with problems in recovering the expected abundance profiles with either metagenomic or 16S rRNA gene amplicon sequencing[18,28,35]. Thus, we have designed our mock community with a focus on the recovery of gram positive and gram negative bacteria, highlighted here and in previous studies as an important source of variation between extraction methods[16,37]. As such, the mock community consists of 10 bacterial strains that are generally absent from the healthy gut microbiome. We accurately quantified cell numbers for each of the cultured species using optical density and cell counting by fluorescence activated cell sorting (FACS), before mixing them in such a way that their

abundances in the mock community span three orders of magnitude to allow assessing the quantification accuracy over a large dynamic range (see Methods and Supplementary Table 2). We then added the mock community into stool samples from eight additional individuals and extracted DNA using the three best performing protocols. Using the mock spike-in as a baseline, we estimated extraction biases in the background of inter-individual microbiome variation. We found all three protocols to perform well (Figure 6) with protocol W performing best (median absolute error [MAE] of 0.39x) as expected from the previous analysis, closely followed by protocol Q (MAE = 0.42x). While the estimated abundances deviated less than 0.5 fold in most cases, the estimation of *Clostridia* abundances showed considerable variance (between 0.5 and 10 fold) even under the best performing protocols, highlighting directions for future improvements.

Discussion

We have shown that of all the factors quantified herein, variations in DNA extraction protocol have the largest effects on the observed microbial composition. The outcome of extraction protocols can be influenced by many variables and implementation details, creating a parameter space which is challenging to test exhaustively. This led us to consider methodologies already established across the field and thus compare between extraction protocols already in use in different laboratories. In this context we recognize the limits of our recommendations regarding which protocol steps are most crucial to prevent distortions, though we also note a good agreement between the ones identified here to results of previous, more focused comparisons[13,14,35,37,38].

Protocols were compared in their extraction quality and validated for transferability, ensuring reproducible use. Although for particular applications some of the tests are more important than others (e.g. in a multisite consortium reproducibility across labs is more important than in an in-depth study in one location), overall protocol Q seems a compromise that should suit most applications. We further tested the quantification accuracy of the best performing protocols by using a mock community, and showed that protocol Q has a median absolute quantification error of less than 0.5x.

We anticipate that procedures for DNA extraction will likely further improve in the future, but put forward protocol Q as a potential benchmark for these new methods. While we have only tested this methodology on stool, we believe it to be applicable to other kinds of samples. However, we caution that additional considerations may apply, such as that of kit contamination[39], which may differ between the protocols investigated here and would, for example, have a high impact on samples with low biomass.

The proposed protocol, together with standard practices for sample collection and the library preparation used can be found on the IHMS website (http://www.microbiome-standards.org/). Taken together, our recommendations, if implemented across laboratories, will greatly improve cross-study comparability and with this our ability to make stronger inferences about the properties of the microbiome.

Online methods

Library preparation and sequencing

Library preparation started with fragmentation of 250 ng genomic DNA to a 150-700 bp range using the Covaris E210 instrument (Covaris, Inc., USA). The SPRIWorks Library Preparation System and SPRI TE instrument (Beckmann Coulter Genomics) were used to perform end repair, A tailing and Illumina compatible adaptors (BiooScientific) ligation. We also performed a 300-600 bp size selection in order to recover most of the fragments. DNA concentration measurements were all performed at Genoscope, using Qubit (fluorimetric dosage) and DNA quality was assessed by 0,7 % gel migration.

DNA fragments were then amplified by 12 cycles PCR using Platinum Pfx Taq Polymerase Kit (Life Technologies) and Illumina adapter-specific primers. Libraries were purified with 0.8x AMPure XP beads (Beckmann Coulter). After library profile analysis by Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and qPCR quantification, the libraries were sequenced using 100 base-length read chemistry in paired-end flow cell on the Illumina HiSeq2000 (Illumina, San Diego, USA).

In the second library preparation protocol, the three enzymatic reactions were performed by a high throughput liquid handler, the Biomek® FX Laboratory Automation Workstation (Beckmann Coulter Genomics) especially conceived for library preparation of 96 samples simultaneously. The size selection was skipped. DNA amplification and sequencing were then performed as in the case of the first approach.

Raw reads for all sequences samples have been deposited to ENA under BioProjectID ERP016524.

Determining taxonomic and functional profiles

For determining the taxonomic composition of each sample, shotgun sequencing reads were mapped to a database of selected single copy phylogenetic marker genes[19] and summarized into species-level (mOTU) relative abundances. Functional profiles of clusters of orthologous groups (COGs) were computed using MOCAT[29] by mapping shotgun sequencing reads to an annotated reference gene catalogue as described in Voigt et al.[23]. COG category abundances were calculated by summing the abundance of the respective COGs belonging to each category per sample, excluding NOGs.

Comparison to other technical and biological variation

To contextualize the size of the effect introduced by different extraction methods, we have assessed different effects caused by either technical or biological factors. These are due to: within protocol variation, library preparation, sample preservation, within specimen variation, between time-points samples from the same individual and between individuals.

For assessing the variation induced by different preservation methods (namely freezing and RNA-later) we use the data from Franzosa et al.[24] and compared the same sample, preserved with the two different methods. For within specimen variation we used data from Voigt et al.[23], where they have sampled the same stool multiple times at different locations along the specimen. As this study also used different storage methods for some samples, we are able to quantify the effect of both within-specimen variation and storage together. For the between time point and individual effect assessment we used the data from the time series data from Voigt et al.[23] as well as a subset of stool samples from the Human Microbiome Project[31]. To ensure comparability across such different studies we have computed distances between all samples on the same subset of relatively abundant

microbes, by removing mOTUs whose summed abundance over all samples was below 0.01% of the total microbial abundance.

For assessing library preparation induced variation, we used the same extracted DNA and subjected it to two library preparation methods (Supplementary Information). The first method was the one routinely used for all library preparations presented in the study.

Determining significantly different species

A Kruskal-Wallis test was applied for each species with non-zero abundance in at least two protocols, across both samples. To account for multiple testing, we applied a Bonferroni correction to the test p-values and rejected the null for any corrected values below 0.05.

Mock community cultivation

Bacteria were cultivated at 37°C under anaerobic conditions in a Vinyl Anaerobic Chamber (COY) inflated with a gas mix of approximately 15% carbon dioxide, 83% nitrogen and 2% hydrogen. For long-term storage, cryovials containing freshly prepared bacterial cultures plus 7% DMSO were tightly sealed and frozen at -80°C. Prior to the experiment, bacteria were pre-cultivated twice using modified Gifu Anaerobic Medium broth (mGAM, 05433, HyServe). Bacteria were mixed based on their OD, pelleted by centrifugation and re-suspended in 0.05 Vol RNAlater® Stabilization Solution (AM7020, Thermo Fisher Scientific). 50 µL of this suspension were distributed to 2 mL safe-lock tubes (30120094, Eppendorf) and frozen at -80°C for later DNA extraction and sequencing.

When assessing the relative abundances obtained from sequencing the mock community alone, we note the presence of ~6% *Escherichia coli* across all extractions, likely a contamination of the mock community itself and not a result of the DNA extraction. As we did not quantify the input of *E. coli* it was not considered in subsequent evaluation. Apart from this and after rarefying to comparable numbers of reads across the three tested protocols we find no evidence of extraction specific contaminants. However, this may be due to the large quantity of input material which would mask the kit contaminants that are likely in low abundance[16].

Flow Cytometry

Bacterial cells were fixed in 70% Ethanol and stored at 4°C for later analysis at the cytometer. Cells were pelleted and rehydrated in PBS with 1mM EDTA aiming at a dilution of 0.6 OD_{600}. We used propidium iodide (PI, Sigma-Aldrich, stock concentration 1 mg/mL resuspended in milliQ H_2O) at a final concentration of 20 µg/mL. as fluorescent probe to label bacterial DNA. The cell suspension was sonicated five times for 10 seconds (0.5 seconds ON, 0.5 seconds OFF, 10% amplitude, Branson Sonifier W-250 D, Heinemann) interrupted by 4 min of cooling.

Samples were analyzed using a BD Accuri™ C6 Cytometer (BD Biosciences) equipped with a 488nm laser. PI fluorescence signal was collected using a 585/40 bandpass filter. Absolute bacterial cell numbers were determined by addition of 50 µL of CountBright™ absolute counting beads (C36950, Thermo Fisher Scientific) with known concentration. At least 2000 beads were acquired for each sample and bacterial numbers were calculated following the manufacturer´s indications. Post-acquisition analysis was done with FlowJo software 10.0.8 (Tree Star, Inc.). Sampling and FACS analysis was performed in duplicate and.

Principal coordinate analysis

Principal coordinate analysis was performed with the R ade4 package (version 1.6.2), using the *dudi.pco* function.

References

1. Meyer, F. *et al.* The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9,** 386 (2008).

2. Larsen, N. *et al.* Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* **5,** e9085 (2010).

3. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).

4. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528,** 262–266 (2015).

5. Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55,** 205–11 (2006).

6. Carroll, I. M. *et al.* Molecular analysis of the luminal- and mucosal-associated intestinal microbiota in diarrhea-predominant irritable bowel syndrome. *Am. J. Physiol. Gastrointest. Liver Physiol.* **301,** G799-807 (2011).

7. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10,** 766 (2014).

8. Dethlefsen, L., McFall-Ngai, M. & Relman, D. a. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* **449,** 811–8 (2007).

9. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 11971–5 (2010).

10. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–7 (2012).

11. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–546 (2013).

12. Wesolowska-Andersen, A. *et al.* Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* **2,** 19 (2014).

13. McOrist, A. L., Jackson, M. & Bird, A. R. A comparison of five methods for extraction of bacterial DNA from human faecal samples. *J. Microbiol. Methods* **50,** 131–139 (2002).

14. Smith, B., Li, N., Andersen, A. S., Slotved, H. C. & Krogfelt, K. A. Optimising bacterial DNA extraction from faecal samples: comparison of three methods. *Open Microbiol. J.* **5,** 14–7 (2011).

15. Maukonen, J., Simões, C. & Saarela, M. The currently used commercial DNA-extraction methods give different results of clostridial and actinobacterial populations derived from human fecal samples. *FEMS Microbiol. Ecol.* **79,** 697–708 (2012).

16. Kennedy, N. A. *et al.* The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLoS One* **9,** e88982 (2014).

17. Salonen, A. *et al.* Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J. Microbiol. Methods* **81,** 127–34 (2010).

18. Ariefdjohan, M. W., Savaiano, D. A. & Nakatsu, C. H. Comparison of DNA extraction kits for

PCR-DGGE analysis of human intestinal microbial communities from fecal specimens. *Nutr. J.* **9,** 23 (2010).

19. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10,** 1196–9 (2013).

20. Manichanh, C., Borruel, N., Casellas, F. & Guarner, F. The gut microbiota in IBD. *Nat. Rev. Gastroenterol. Hepatol.* **9,** 599–608 (2012).

21. Lozupone, C. A. *et al.* Meta-analyses of studies of the human microbiota. *Genome Res.* **23,** 1704–14 (2013).

22. Raes, J. & Bork, P. Molecular eco-systems biology: towards an understanding of community function. *Nat. Rev. Microbiol.* **6,** 693–9 (2008).

23. Voigt, A. Y. *et al.* Temporal and technical variability of human gut metagenomes. *Genome Biol.* **16,** 73 (2015).

24. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E2329-38 (2014).

25. Song, S. J. *et al.* Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. *mSystems* **1,** (2016).

26. Gohl, D. M. *et al.* Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* **34,** 942–949 (2016).

27. Claassen, S. *et al.* A comparison of the efficiency of five different commercial DNA extraction kits for extraction of DNA from faecal samples. *J. Microbiol. Methods* **94,** 103–10 (2013).

28. 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).

29. Kultima, J. R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One* **7,** e47656 (2012).

30. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).

31. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).

32. Franzosa, E. A. *et al.* Identifying personal microbiomes using metagenomic codes. *Proc. Natl. Acad. Sci.* **112,** 201423854 (2015).

33. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40,** D284-9 (2012).

34. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489,** 220–30 (2012).

35. Santiago, A. *et al.* Processing faecal samples: a step forward for standards in microbial community analysis. *BMC Microbiol.* **14,** 112 (2014).

36. InhibitEx Tablets - QIAGEN Online Shop. Available at: https://www.qiagen.com/fr/shop/lab-basics/buffers-and-reagents/inhibitex-tablets/.

37. Henderson, G. *et al.* Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. *PLoS One* **8,** e74787 (2013).

38. Jones, M. B. *et al.* Library preparation methodology can influence genomic and functional

predictions in human microbiome research. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 14024–9 (2015).

39. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12,** 87 (2014).

Acknowledgements

We would like to acknowledge the help of Sebastian Burz and Kevin Weizer for the editing and web-posting of the SOPs. We thank D. Ordonez and N.P. Gabrielli Lopez for advice on flow cytometry, which was provided by the Flow Cytometry Core Facility, EMBL. This study was funded by the European Community's Seventh Framework Programme via International Human Microbiome Standards (HEALTH-F4-2010-261376) grant. We also received support from Scottish Government Rural and Environmental Science and Analytical Services.

Author contributions

PIC, SS, GZ analyzed data, drafter and finalized the manuscript. EP and AA analyzed data, sequenced samples and wrote the manuscript. FL, JRK, MRH, LPC and EAV analyzed data and wrote the manuscript. MT, MD, RH, FJ and KRP created and quantified the mock community. MB, JB, LB, TC, SCP, MD, AD, WMV, BBF, HJF, FG, MH, HH, JHV, JJ, IK, PL, ELC, VM, CM, JCM, CM, HM, CO, POT, JP, SP, NP, MP, AS, DS, KPS, BS, KS, PV, JV, LZ, EGZ extracted samples and wrote the manuscript. SDE, JD and PB designed the study and wrote the manuscript.

Competing interest statement

The authors declare no competing financial interests.

Figure legends

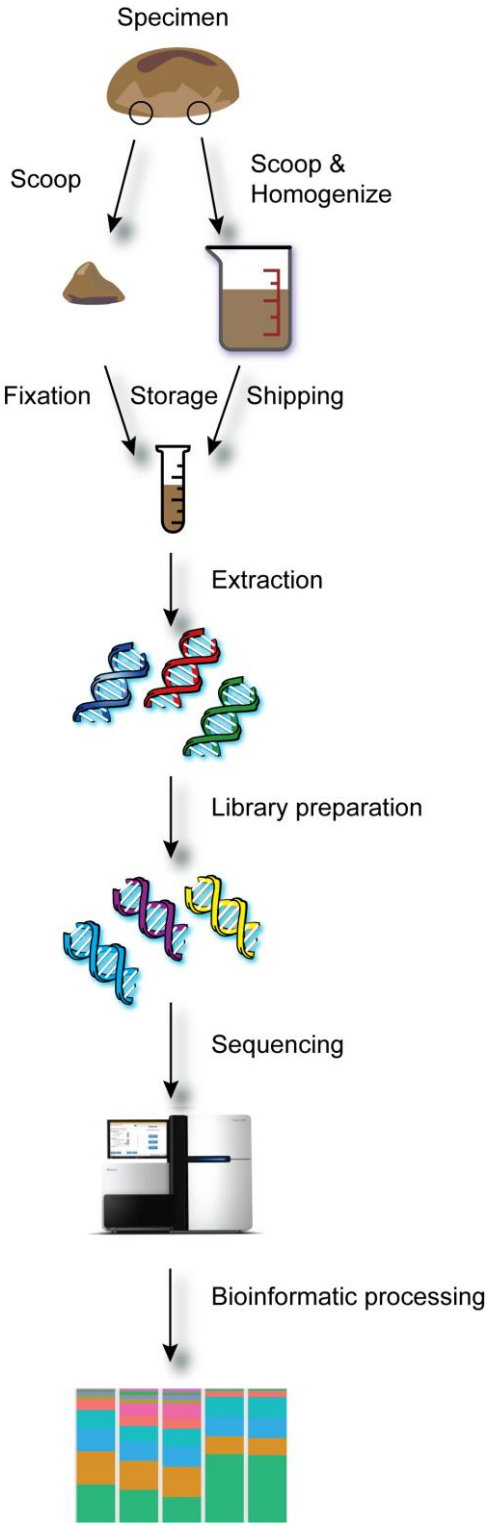

Figure 1: Schematic workflow of human fecal samples processing.

Illustration of the main steps involved in extracting and analyzing DNA sequences from human fecal samples, from collection to bioinformatics analysis. Importantly, none of the outlined steps are standardized, which may introduce strong effects between different studies, making their results hard to compare. For example, differences between freezing and RNA-later fixation have been previously described[23] to bias the measured sample composition.

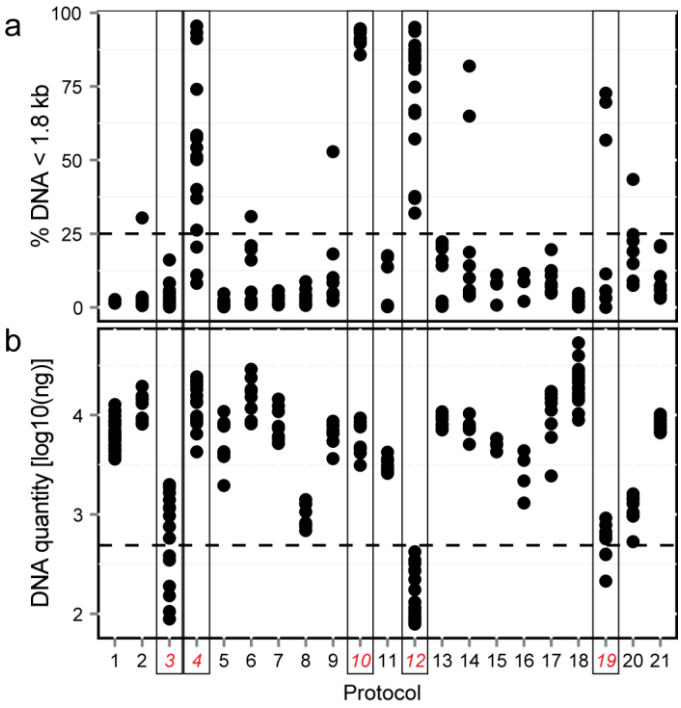

Figure 2: Quality control of extracted DNA

Quality (a) and quantity (b) of extracted DNA from 21 different protocols. a) Percentage of DNA molecules shorter than 1.8 kb, b) quantity of extracted DNA. Protocols failing quality cut-offs (indicated by dashed lines) for either measurement are highlighted in red and boxed.

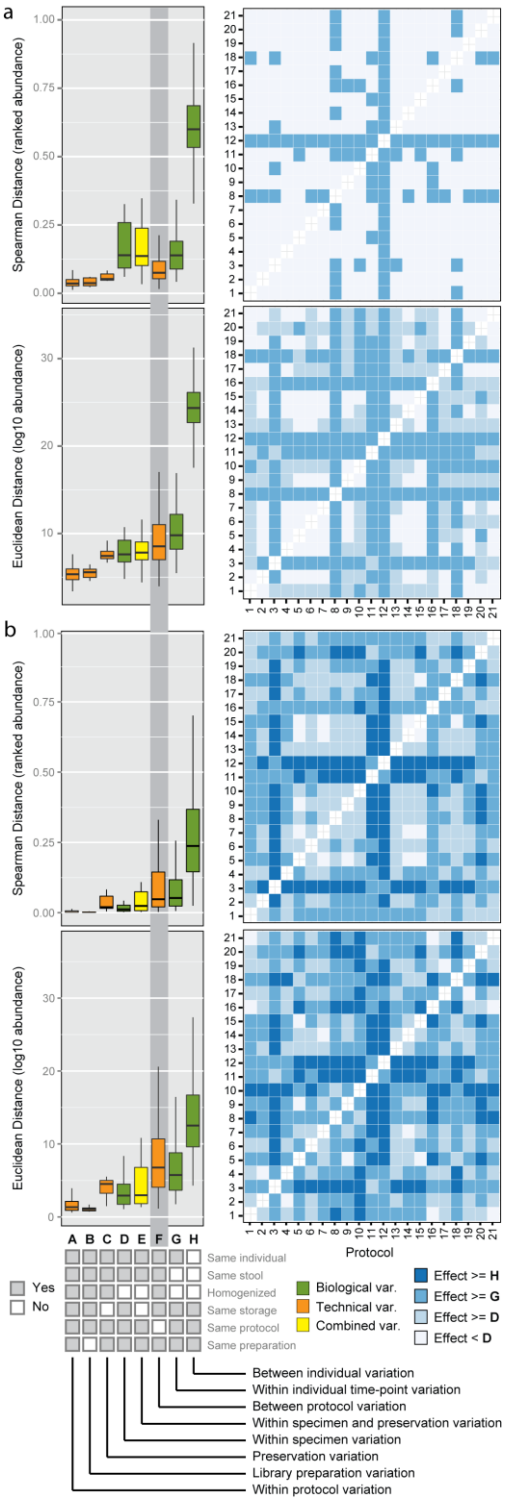

Figure 3: Effect of DNA extraction protocol and library preparation on sample composition

Using both a Euclidean and an Spearman distance measure (see Methods) on species abundances (using mOTU[19]) (a) as well functional abundances (using COGs[33]) (b), shows the relative effect size of different sources of variation. These have been assessed on independent samples from different studies and thus also capture additional differences. The library preparation and the within-protocol variation are the smallest effects, while the between protocol variation may be greater than some biological effects[23,24]. Heat maps on the right show all pairwise distances between protocols, highlighting which protocol may be considered comparable and which not under different measures of similarity as encoded by letters D,H and G on the bottom-right.

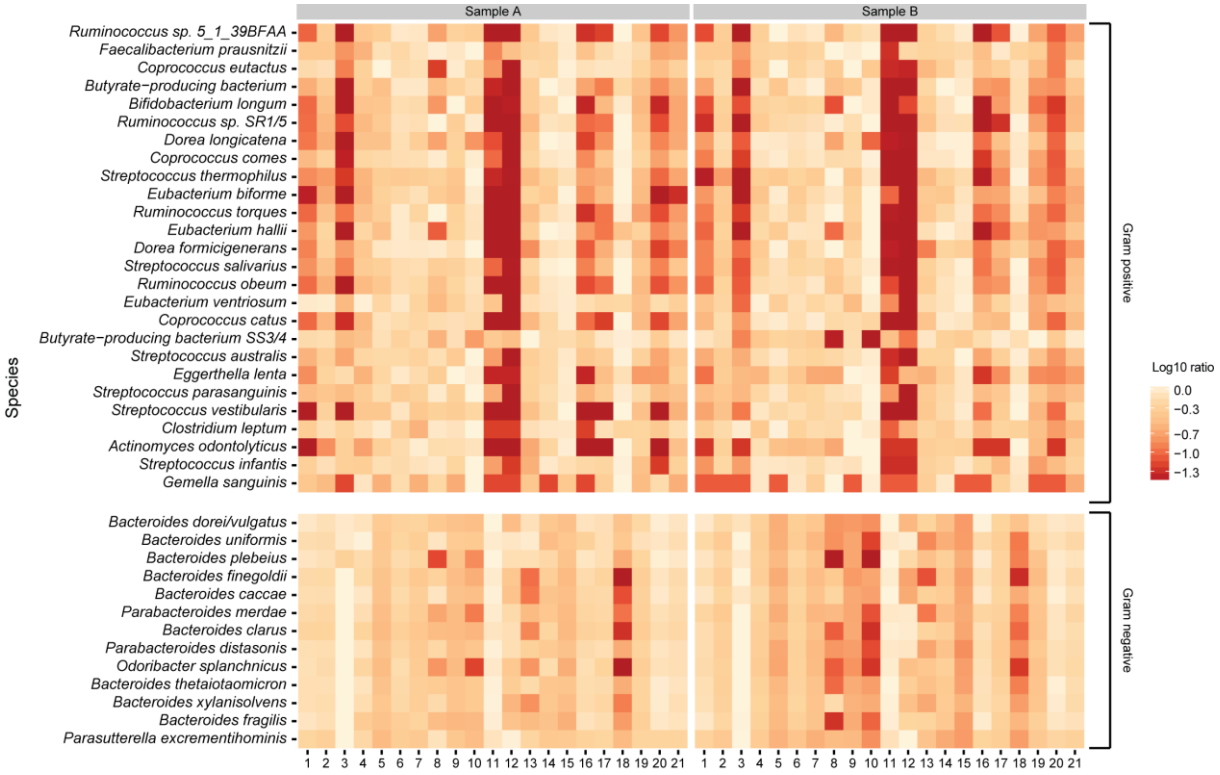

Figure 4: Species specific abundance variation

Assessing variation of species abundances shows that biases are consistent across the two samples. Considering species for which the abundances are significantly different between extraction protocols (Kruskal-Wallis test, FRD corrected p-value < 0.05) we show that gram-positive bacteria are heavily under-estimated compared to the mean across the five highest recovered ratios, while gram-negative bacteria are only slightly, though significantly skewed. Abundances are calculated using mOTUs[19] , with only those having a species level annotation being shown.

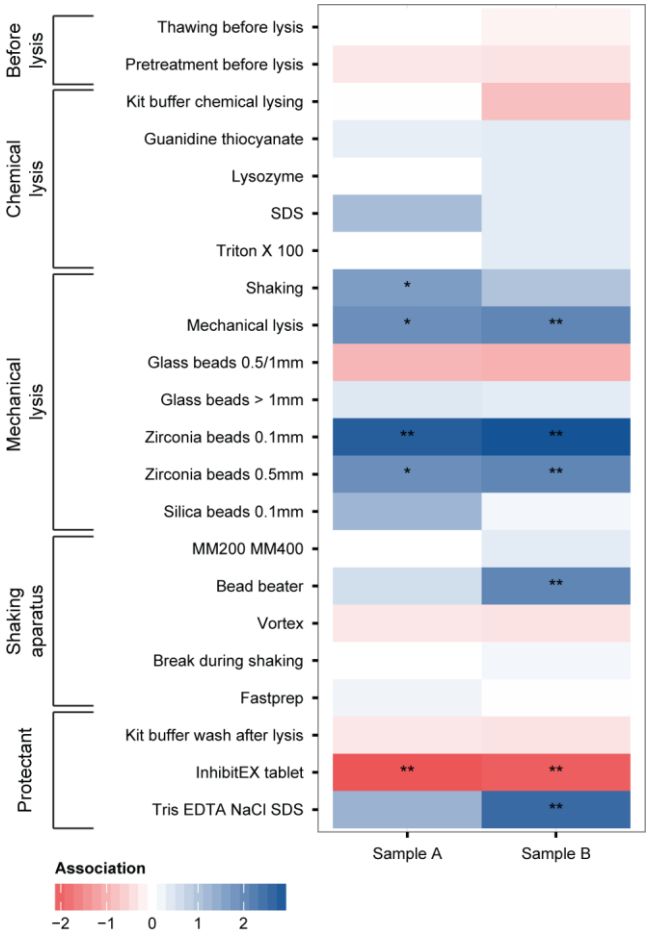

Figure 5: Effects of protocol manipulations on sample composition

Out of 22 protocol descriptors that vary between the Qiagen based methods, 7 are significantly associated with diversity outcomes. Associations are coded as negative (red) and positive (blue), with significance highlighted by * < 0.05 and ** < 0.01. P-values have been FDR corrected for multiple testing.

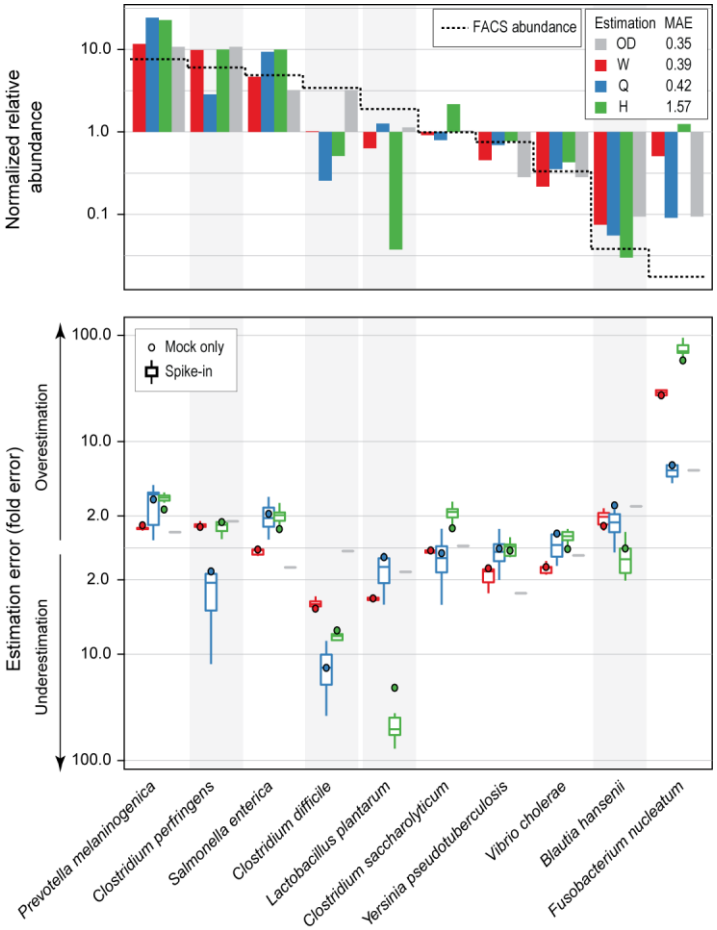

Figure 6: Mock community extraction quality

Using 10 bacterial species, mixed at known relative abundances, as a baseline, we show that the estimation obtained from the different extraction methods are generally correct, using a median absolute error measure. To account for compositional effects, we report log-ratio transformed values, relative to the geometric mean. The top panel shows the median estimated abundance across ten extractions, with the ground truth value indicated by a dashed line for each species. With gray bars we show the estimated abundance from optical density measurements of the mock community. In the bottom panel we show the full distribution of the estimated abundances and highlight that obtained by extracting DNA from the mock community itself, as opposed to extracting DNA from a sample to which the mock community has been added before extraction. Gram positive bacterial are highlighted by a gray background the two panels.