nature biotechnology

Towards standards for human fecal sample processing in metagenomic studies

Paul I Costea[1], Georg Zeller[1], Shinichi Sunagawa[1,2], Eric Pelletier[3–5], 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 R M Brown[9], Thomas Carton[10], Stéphanie Cools-Portier[11], Michelle Daigneault[6], 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 W 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[1,28–30]

Technical variation in metagenomic analysis must be minimized to confidently assess the contributions of microbiota to human health. Here we tested 21 representative DNA extraction protocols on the same fecal samples and quantified differences in observed microbial community composition. We compared them with differences due to library preparation and sample storage, which we contrasted with observed biological variation within the same specimen or within an individual over time. We found that DNA extraction had the largest effect on the outcome of metagenomic analysis. To rank DNA extraction protocols, we considered resulting DNA quantity and quality, and we ascertained biases in estimates of community diversity and the ratio between Gram-positive and Gram-negative bacteria. We recommend a standardized DNA extraction method for human fecal samples, for which transferability across labs was established and which was further benchmarked using a mock community of known composition. Its adoption will improve comparability of human gut microbiome studies and facilitate meta-analyses.

More than 3,000 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 (including shotgun and 16S rRNA gene sequences)[1]. The human gastrointestinal tract is one of the most studied of these ecosystems. The gut microbiome is of particular interest due to its large volume, high diversity and 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 (IBD)[5,6] or colorectal cancer[7]. Others have linked human gut microbial composition to factors such as mode of birth, age, diet or medication[8–11]. Such studies have almost exclusively used their own methodology and a demographically distinct cohort. Numerous reports of batch effects[12] and known differences when analyzing data generated using different protocols[13–18] mean that comparisons or meta-analyses are limited in their interpretability. For example, healthy Americans from the Human Microbiome Project study showed lower taxonomic diversity in their stool than patients with IBD from a European study[19], although it is established that IBD patients worldwide have reduced taxonomic diversity[20]. This illustrates the difficulties in disentangling 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 (**Fig. 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 before being shipped to a laboratory. The location within the specimen from which the sample is taken has been shown to impact the measured composition[23], which is why in some studies[24] larger quantities are homogenized before storage in order to generate multiple identical aliquots. Furthermore, different fixation methods are used to preserve samples for shipping and long-term storage. Freezing at below −20 °C is the norm, though more practical alternatives exist[23–25]. Eventually, DNA is extracted from the sample, followed by library preparation, sequencing and downstream bioinformatics analysis (**Fig. 1**).

We examined the extent to which DNA extraction influences the quantification of microbial composition, and compared this variable with other sources of technical and biological variation. Most protocol comparison studies to date have used a 16S rRNA gene amplification

A full list of affiliations appears at the end of the paper.

Received 8 July 2016; accepted 11 August 2017; published online 2 October 2017; doi:10.1038/nbt.3960

approach, which suffers from additional problems. Specifically, the choice of primer, PCR biases and even the choice of polymerase can affect the results of 16S rDNA studies[26], which may result in different outcomes when carrying out the same DNA extraction comparison in a different laboratory. Fortunately, these problems do not affect shotgun metagenomic sequencing.

We compared a wide range of DNA extraction methods, using shotgun metagenomic sequencing as the readout, and assessed taxonomic and functional variability while keeping all of the other steps standardized. We investigated the most commonly used extraction kits with varying modifications, as well as additional protocols that do not make use of commercially available kits (**Supplementary Data 1** and Online Methods). Although other studies have investigated differences between DNA extraction methods in one setting[12,15,16,27], we systematically tested reproducibility within and across laboratories on three continents, by applying strict and consistent quality criteria.

Finally, we checked the accuracy of the best-performing DNA extraction methods using a mock community of ten bacterial species whose exact relative abundance is known. This community includes both Gram-positive and Gram-negative bacteria, and their relative abundances span three orders of magnitude. Based on these results we recommend a standardized protocol for DNA extraction from human stool samples. If accepted by the research community, we believe that this protocol will greatly enhance comparability among metagenomic studies.

RESULTS

Study design

Our study had two phases. In the first phase, to assess the variability introduced by different DNA extraction methods, we produced multiple aliquots of two stool samples (obtained from two individuals and named sample A and B). Within 2 h of emission, the samples were homogenized in an anaerobic cabinet (to ensure that the different aliquots had identical microbial compositions) and aliquots of 200 mg were frozen at −80 °C within 4 h. We shipped four aliquots of each sample, frozen on dry ice, to 21 collaborating laboratories in 11 countries on three continents. Participating laboratories used 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 Data 1** and Online Methods). Each lab performed at least four extractions, after which the DNA was shipped to a single sequencing center (GENOSCOPE, France), which tested two different library preparation methods on a subset of samples (Online Methods) before performing identical sequencing and analytical methods in an attempt to minimize other possible sources of variation.

In a second phase, after considering the quantity and integrity of extracted DNA, the recovered diversity and the observed ratio of Gram-positive bacteria, we selected five protocols (numbers 1, 6, 7, 9 and 15; see **Supplementary Methods** for full descriptions of all 21 protocols). As three of these were very similar (6, 9 and 15), they were combined into one, resulting in three protocols being compared in this phase. For each protocol, extractions were then performed in the original laboratory that used that protocol and in three additional laboratories, none of which had used the method before, in order to assess reproducibility of this subset of protocols among laboratories. Three aliquots of samples A and B were provided to each laboratory, as detailed above.

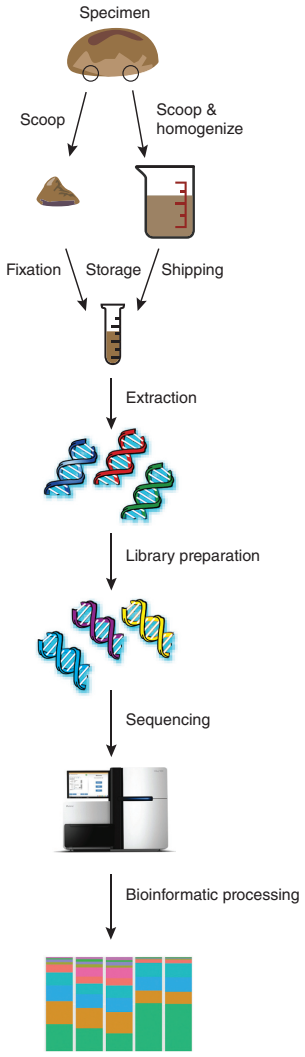

Figure 1 Schematic workflow of human fecal samples processing. Illustration of the main steps involved in extracting and analyzing DNA from human fecal samples, from collection to bioinformatics analysis. Notably, none of the outlined steps are standardized, which may introduce variability among different studies, making their results difficult to compare. For example, differences between freezing and RNAlater fixation have been reported[23] to bias the measured sample composition.

To quantify the absolute extraction error of the three protocols, we prepared a mock community consisting of 10 bacterial species that are generally absent from the stool of healthy individuals (**Supplementary Data 2**), such that the cell density of all species in the mock community was predetermined. DNA was extracted from the mock sample and from eight additional samples, each consisting of stool from a unique individual spiked with similar proportions of the mock community (in order to emulate a realistic setting). All of these extractions were carried out 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 high 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 community composition. We found considerable variation in the quantity of extracted DNA, in line with previous observations[28] (**Fig. 2**). For example, protocol 18 recovered 100 times more DNA than protocols 3 and 12 and recovered 10 times more than protocols 8, 19 and 20 (**Fig. 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 no fragmentation was observed for protocol 1. For subsequent analysis, samples that yielded less than 500 ng 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 4 of 18 samples extracted with protocol 16 (one sample A and three sample B replicates) met the requirements (**Supplementary Data 3**). For other protocols, small numbers of samples were discarded for lack of compliance with quality and/or quantity criteria.

Quality control for variability in microbial composition

All metagenomes were compared with respect to taxonomic and functional composition to quantify the relative abundances of microbial taxa and their respective gene-encoded functions (Online 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 abundances and gene functional abundances were computed by mapping high-quality reads to a database of single-copy taxonomic marker genes[19] and to annotated human gut microbial reference genes[30], respectively (Online Methods).

There are, as outlined above (**Fig. 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 result in an effect larger than the biological variation of interest (for example, in an intervention study), it will mask that signal. Consistent 'batch effects' will introduce bias that can distort a meta-analysis even if their absolute size is comparatively small. It is 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 (**Fig. 3**) assessed on available data from multiple other studies[23,24,31] (Online 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 Fig. 1**). The next-largest difference 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 composite of multiple effects (**Fig. 3**).

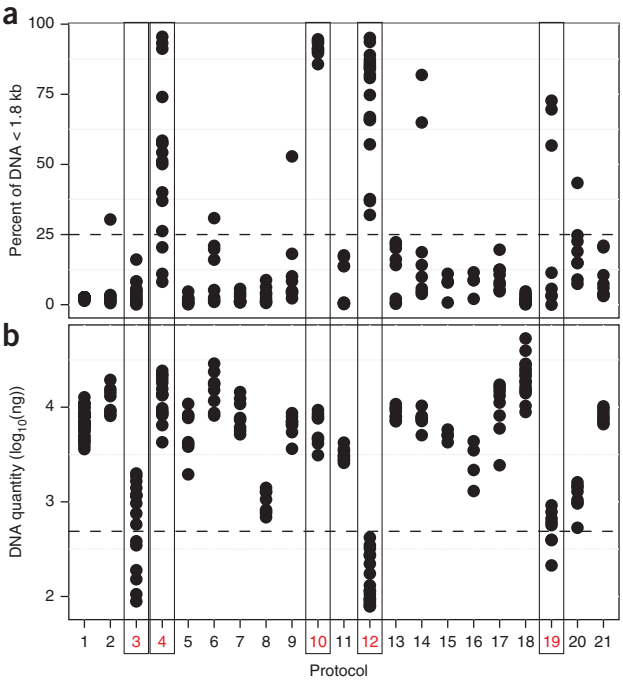

Figure 2 Quality control of extracted DNA. Quality (**a**) and quantity (**b**) of extracted DNA from 21 different protocols. (**a**) Percentage of DNA molecules shorter than 1.8 kb. (**b**) Quantity of extracted DNA. Protocols failing to meet quality cut-offs (indicated by dashed lines) for either measurement are highlighted in red and boxed.

Various 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 abundance data from both single-copy taxonomic marker genes[21] and clusters of orthologous groups[33] to derive species and functional variation (Online Methods). First, we used a Spearman correlation to assess how well species abundance rankings are preserved and found that the variation between most extraction protocols was smaller than the technical within-specimen variation (**Fig. 3a**). This suggests that, with the exception of protocols 8 and 12, all others recovered 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 (**Fig. 3a**), hampering the comparability of samples generated with different extraction methods. To assess similarity between extraction protocol effects, we used principle coordinate analysis (Online Methods) to visualize these distance spaces (**Supplementary Fig. 2**). These indicated that protocol 12, and to a lesser extent protocols 3, 8, 11, 16 and 18 as well, had an abundance profile that differed from most of the other protocols.

Analysis of functional microbiome composition, based on clusters of orthologous groups (**Fig. 3b** and Online Methods), showed

that the majority of extraction protocol effects were greater than biological variation within specimens and across time points within the same individual (**Fig. 3b**), and some of them were 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, comparable to in magnitude to the library preparation effect being (**Fig. 3a**,**b**). The variation introduced by storage method (RNAlater versus frozen) was larger than within-protocol variation and, as previously shown, smaller than within-specimen variation in taxonomic space[23,24]. Taken together, these results show that different DNA extraction protocols result in substantial technical variation, both in taxonomic and in functional space, highlighting that this is a crucial step in any microbiome study.

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 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 (Online Methods). We then applied a false discovery rate correction to the obtained *P* values. Of the 366 tested species, we found 90 that were significantly affected by extraction protocol ($q < 0.05$). The majority of these were Gram-positive, accounting for 37% (±7%) of the sample abundance, on average (**Fig. 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 with the mean abundance 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 was comparatively small (**Fig. 4**).

As the observed biases suggested protocol-dependent incomplete lysis of Gram-positive bacteria, we hypothesized that this would result in 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 correlated with the observed diversity, with a higher fraction of Gram-positives resulting in higher diversity (**Supplementary Fig. 3**). Furthermore, we found dramatically reduced diversity in protocols already determined to perform poorly with regards to DNA quality (i.e., protocols 3, 11 and 12; **Supplementary Fig. 4**). We conclude that a diversity measure is a good proxy for overall protocol performance and accuracy of the recovered abundance profile.

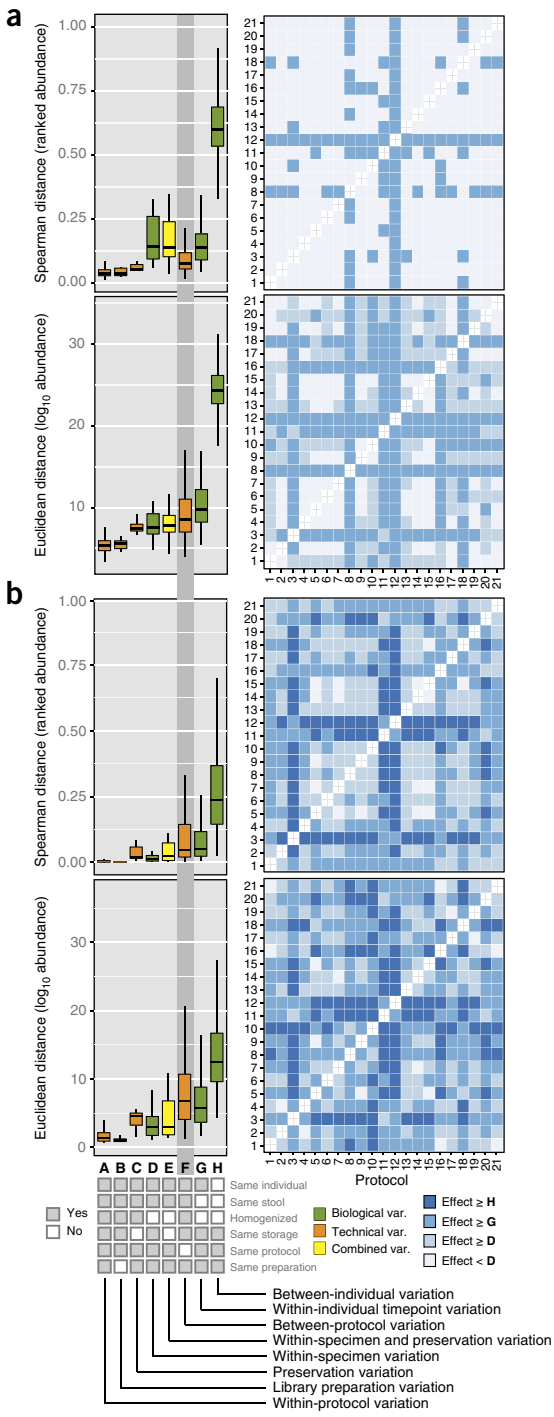

Figure 3 Effect of DNA extraction protocol and library preparation on sample composition. Using both Euclidean and Spearman distance measurements (Online Methods) on (**a**) species abundances (using single-copy taxonomic marker genes (mOTUs)[19]), as well as on (**b**) functional abundances (using clusters of orthologous groups (COGs)[33]), 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. Library preparation and within-protocol variation (var.) have the smallest effects, while between-protocol variations may be greater than some biological effects[23,24]. Right: heat maps show all pairwise distances between protocols, highlighting which protocols may be considered comparable and which not under different measures of similarity, as encoded by letters D, H and G on the bottom right.

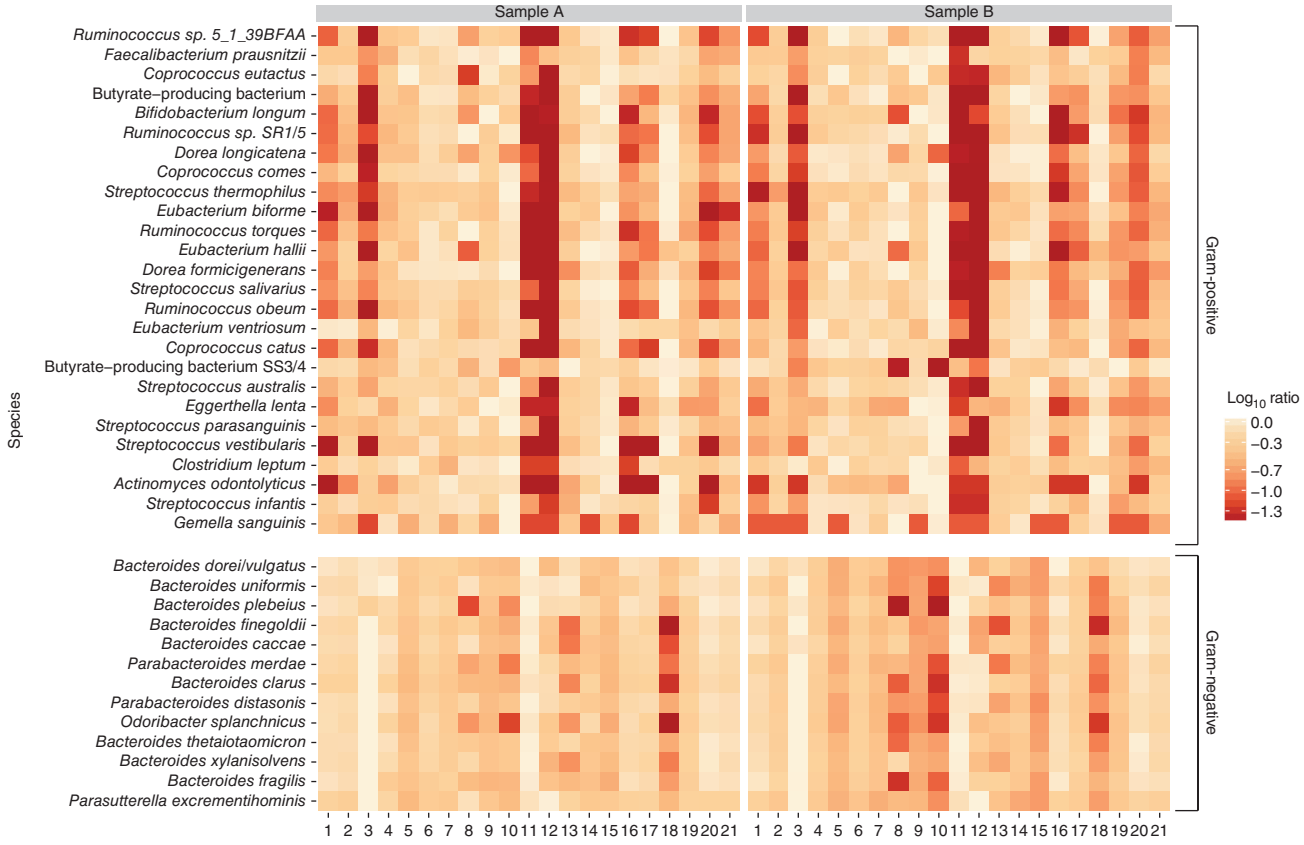

Figure 4 Species-specific abundances variation shows that biases are consistent across the two samples. Considering species for which the abundances are significantly different between extraction protocols (Kruskal-Wallis test, false discovery rate (FDR) corrected $P < 0.05$), we show that Gram-positive bacteria are heavily underestimated 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.

Factors influencing DNA extraction outcome

Using diversity as an optimality criterion, we determined protocol parameters that were significantly associated with this indicator ($P < 0.05$; **Fig. 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 reduced the number of variables that can influence the outcome. We found that mechanical lysis, zirconia beads and shaking were positively associated with diversity. We note that there was no association with DNA fragmentation, as all of the samples extracted with these protocols had few fragments below 1.8 kb (**Fig. 2**). This was consistent with the notion that mechanical lysis and bead beating are necessary to efficiently extract the DNA of Gram-positive bacteria, which 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 (**Fig. 5**) 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] However, our assessment suggests that this tablet has an adverse effect on DNA extraction quality. This analysis suggests specific modifications with which extraction methods—also currently suboptimal—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 among laboratories

Based on the quality of the extracted DNA, species diversity and species-specific biases, we selected the five best performing protocols, numbers 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, which we coded Q (**Supplementary Methods**). 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 Fig. 5**).

Each extraction method was established and performed in triplicate in three other laboratories, which had no experience with the respective protocol they were assigned, 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 Fig. 5**).

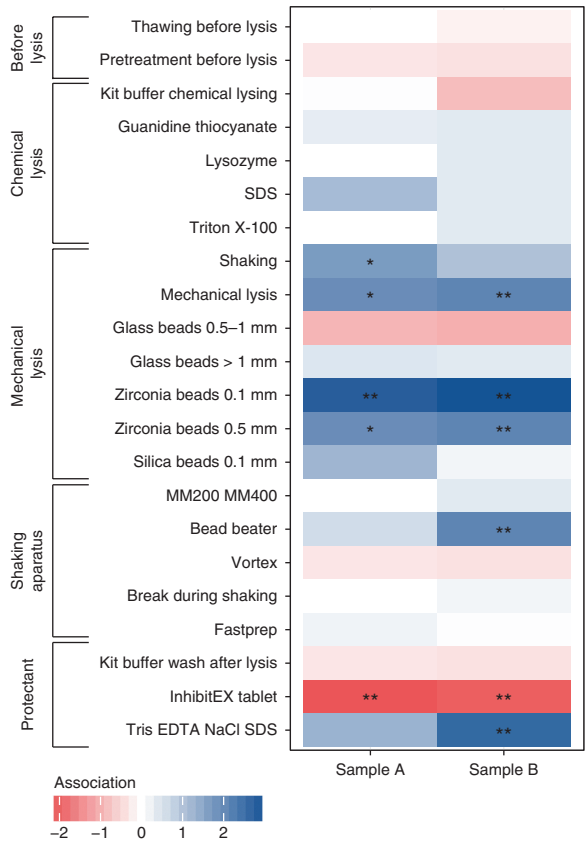

Figure 5 Effects of protocol manipulations on sample composition. Of 22 protocol descriptors that vary between the Qiagen based methods, seven were significantly associated with diversity outcomes. Associations are coded as negative (red) or positive (blue), with significance highlighted by $*P < 0.05$ and $**P < 0.01$. *P* values have been FDR-corrected for multiple testing.

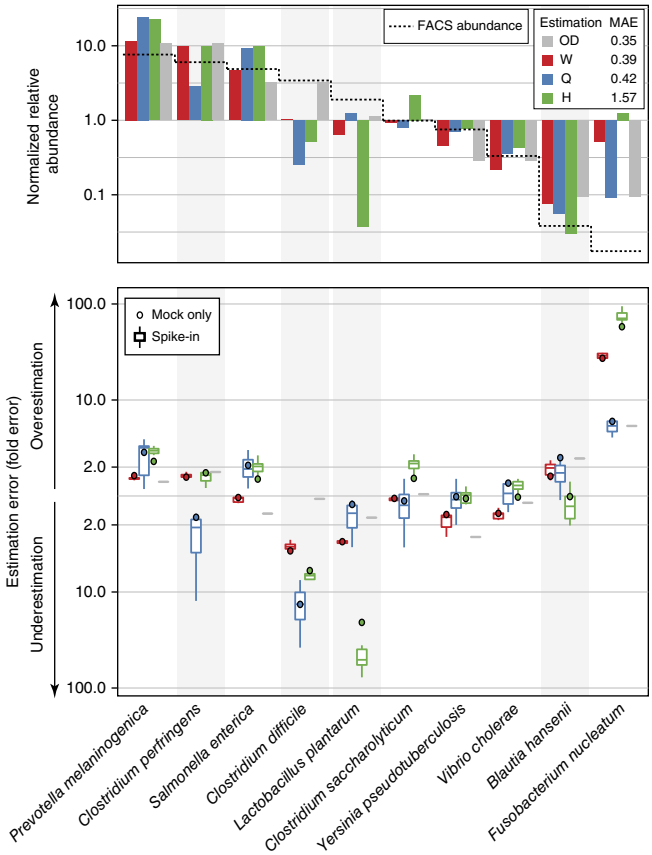

Figure 6 Mock community extraction quality. Using 10 bacterial species, mixed at known relative abundances, as a baseline, we show that the estimation obtained from the various extraction methods are generally correct, using a median absolute error measure. To account for compositional effects, we report log-ratio transformed values relative to the geometric mean. Top: the median estimated abundance across ten extractions, with the ground truth value indicated by a dashed line for each species. Gray bars show the estimated abundance from optical density measurements of the mock community. Bottom: the full distribution of the estimated abundances, highlighting 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 bacteria are highlighted by a tan background the two panels.

Although protocol H seemed to be more reproducible across laboratories, it underestimated Gram-positive bacteria compared to the other two protocols (**Supplementary Fig. 5** and protocol 1 in **Fig. 4**) and so yielded less diverse estimates of microbial composition. Protocol W, while also more reproducible (**Supplementary Fig. 5** and protocol 7 in **Fig. 4**), is difficult to automate because it involves the use of phenol-chloroform. Protocol Q recovered a highly diverse estimate of the microbial composition, which it seems to achieve through efficient 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. While this provided 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]. We 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]. It consisted of 10 bacterial strains that are generally absent from the healthy gut microbiome, so we could assess their abundance when mixed into real samples. We accurately quantified cell numbers for each of the cultured species using optical density and cell counting by fluorescence activated cell sorting before mixing them in such a way that their abundances in the mock community spanned three orders of magnitude, to allow us to assess the quantification accuracy over a large dynamic range (Online Methods and **Supplementary Data 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 interindividual microbiome variation. We found that all three protocols performed well (**Fig. 6**), with protocol W performing best (median absolute error of 0.39×), as expected from the previous analysis, closely followed by protocol Q (median absolute error = 0.42×). 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 the potential for further improvement.

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, creating a parameter space that is challenging to test exhaustively. Instead we chose to evaluate methodologies already in use in the microbiome field and thus compare between already established extraction protocols. In this context, we recognize that our study has the limitation of identifying only those protocol steps that are in common use, though we also note good agreement between the variables identified to be important with results of previous, more focused comparisons[13,14,35,37,38].

Protocols were compared for extraction quality and validated for transferability, ensuring reproducible use. Although for particular applications some of the tests are more important than others (for example, in a multisite consortium, reproducibility across labs is more important than in an in-depth study in one location), protocol Q seems to be the best overall and 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 had a median absolute quantification error of less than 0.5×.

We anticipate that procedures for DNA extraction will likely further improve in the future and propose that protocol Q can serve as a potential benchmark for new methods. While we have only tested this protocol on stool, it might also work well with other 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.

Protocol Q, together with standard methods for sample collection and library preparation, can be found on the IHMS website (http://www.microbiome-standards.org/) and in the Online Methods. Taken together, our recommendations, if implemented across laboratories, may improve cross-study comparability and with this, our ability to make meaningful inferences about the properties of the microbiome.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Burz and K. Weizer for editing and web-posting 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 as well as from EMBL.

AUTHOR CONTRIBUTIONS

P.I.C., S.S. and G.Z. analyzed data and drafted and finalized the manuscript. E.P. and A.A. analyzed data, sequenced samples and wrote the manuscript. F.L., J.R.K., M.R.H., L.P.C. and E.A.-V. analyzed data and wrote the manuscript. M.T., M. Driessen, R.H., F.-E.J. and K.R.P. created and quantified the mock community. M.B., J.R.M.B., L.B., T.C., S.C.-P., M. Derrien, A.D., M. Daigneault, R.A.L., W.M.d.V., B.B.F., H.J.F., F.G., M.H., H.H., J.v.H.V., J.J., I.K., P.L., E.L.C., V.M., C. Manichanh, J.C.M., C. Mery, H.M., C.O., P.W.O., J.P., S.P., N.P., M.P., A.S., D.S., K.P.S., B.S., K.S., P.V., J.V., L.Z. and E.G.Z. extracted samples and wrote the manuscript. S.D.E., J.D. and P.B. designed the study and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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**, 205–211 (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–G807 (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**, 811–818 (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. USA* **107**, 11971–11975 (2010).
10. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (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**, 14–17 (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**, 127–134 (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**, 1196–1199 (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**, 1704–1714 (2013).
22. Raes, J. & Bork, P. Molecular eco-systems biology: towards an understanding of community function. *Nat. Rev. Microbiol.* **6**, 693–699 (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. USA* **111**, E2329–E2338 (2014).
25. Song, S.J. *et al.* Preservation methods differ in fecal microbiome stability, affecting suitability for field studies. *mSystems* https://dx.doi.org/10.1128/mSystems.00021-16 (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**, 103–110 (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.* Human Microbiome Project Consortium. 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. USA* **112**, E2930–E2938 (2015).
33. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40**, D284–D289 (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**, 220–230 (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. USA* **112**, 14024–14029 (2015).
39. Salter, S.J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).

[1]Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany. [2]Department of Biology, Institute of Microbiology, ETH Zurich, Zurich, Switzerland. [3]CEA - Institut François Jacob - Genoscope, Evry, France. [4]CNRS UMR-8030, Evry, France. [5]Université Evry Val d'Essonne, Evry, France. [6]Metagenopolis, Institut National de la Recherche Agronomique, Jouy en Josas, France. [7]Department of Molecular and Cellular Biology, The University of Guelph, Guelph, Ontario, Canada. [8]Department of Gastrointestinal Microbiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany. [9]School of Microbiology & APC Microbiome Institute, University College Cork, Cork, Ireland. [10]Biofortis, Mérieux NutriSciences, Nantes, France. [11]Danone Nutricia Research, Palaiseau, France. [12]Laboratory of Microbiology, Wageningen University & Research, Wageningen, the Netherlands. [13]Immunobiology Research Program, Department of Bacteriology and Immunology, University of Helsinki, Helsinki, Finland. [14]Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada. [15]Rowett Institute of Nutrition and Health, University of Aberdeen, Aberdeen, UK. [16]Digestive System Research Unit, Vall d'Hebron Research Institute, Barcelona, Spain. [17]Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan. [18]Graduate School of Advanced Science and Engineering, Waseda University, Tokyo, Japan. [19]Texas Children's Hospital, Feigin Center, Houston, Texas, USA. [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, Gainesville, Florida, USA. [22]Graduate School of Environmental and Life Science, Okayama University, Okayama, 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, Copenhagen, Denmark. [25]Centre for Human Immunology, Department of Microbiology & Immunology and Robarts Research Institute, University of Western Ontario, London, Ontario, 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, London, UK. [28]Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany. [29]Molecular Medicine Partnership Unit, Heidelberg, Germany. [30]Max-Delbrück-Centre for Molecular Medicine, Berlin, Germany. Correspondence should be addressed to: S.D.E. (dusko.ehrlich@inra.fr), J.D. (joel.dore@inra.fr) or P.B. (bork@embl.de).

ONLINE METHODS

Sample collection. Samples were donated by nine male adults and one female adult with normal gastrointestinal health. Informed consent was obtained from all ten donors. For the first two, from whom the phase I samples were obtained, approval for the Institut National de la Recherche Agronomique to manage human-derived biological samples was granted by the Ministry of Research and Education under number DC-2012-1728. The collection of the remaining eight samples was approved by the EMBL Bioethics Internal Advisory Board and is in agreement with the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report, under number BIAC 2015-009.

Data availability. Raw reads for all sequenced samples have been deposited to ENA under BioProject ID ERP016524.

DNA extraction methods. The 21 extraction protocols compared in the present study were implemented and used in the participating laboratories. They employ a wide range of commercial kits, as summarized in **Supplementary Data 1**. Detailed step by step descriptions are available in **Supplementary Methods**, numbered accordingly.

Library preparation and sequencing. Standard 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 adaptor (Bioo Scientific) ligation. We also performed a 300–600-bp size selection to recover most of the fragments. DNA concentration measurements were all performed at Genoscope, using Qubit (fluorometric dosage), and DNA quality was assessed by 0.7% gel migration.

DNA fragments were then amplified by 12 cycles of PCR, using Platinum Pfx Taq Polymerase Kit (Life Technologies) and Illumina adaptor-specific primers. Libraries were purified with 0.8× 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 cells 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. We term this the high-throughput prep, and it was applied in parallel to a subset of 12 samples, allowing us to compare the differences introduced by this step alone. Size selection was skipped. DNA amplification and sequencing were then performed as in the first approach.

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 COGs were computed using MOCAT[29] by mapping shotgun sequencing reads to an annotated reference gene catalog 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 assessed different effects caused by either technical or biological factors. These are due to variations within protocols, in library preparation, in sample preservation, within specimens, between timepoints sampled from the same individual and between individuals.

To assess the variation induced by different preservation methods (namely freezing and RNAlater) we used 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], in which they sampled the same stool multiple times at different locations along the specimen. As this study also used different storage methods for some samples, we were able to quantify the effect of both within-specimen variation and storage together. For the between-timepoint and individual-effect assessments, we used time-series data from Voigt *et al.*[23] as well as data from a subset of stool samples from the Human Microbiome Project[31]. To ensure comparability across such different studies, we 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 variation induced by library preparation, we used the same extracted DNA from 12 samples and subjected it to two library preparation methods (the standard and high-throughput approaches). The standard method was the one routinely used for library preparations presented in this study.

Statistical analysis. To determine which species significantly differed in abundance between extraction methods, a Kruskal-Wallis test was applied for each species with nonzero abundance in at least two protocols, across both samples. To account for multiple testing, we applied a Bonferroni correction to the resulting *P* values and rejected the null hypothesis for any corrected value below 0.05.

Principal coordinate analysis was performed with the R ade4 package (version 1.6.2), using the dudi.pco function. A **Life Sciences Reporting Summary** is available.

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 precultivated twice using modified Gifu anaerobic medium broth (mGAM, 05433, HyServe). Bacteria were mixed based on their OD, pelleted by centrifugation and resuspended in 0.05 vol RNAlater Stabilization Solution (AM7020, Thermo Fisher Scientific). We distributed 50 µL of this suspension to 2-mL safe-lock tubes (30120094, Eppendorf) and froze them 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 found 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 1 mM 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 a fluorescent probe to label bacterial DNA. The cell suspension was sonicated five times for 10 s (0.5 s ON, 0.5 s 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 488-nm 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 2,000 beads were acquired for each sample, and bacterial numbers were calculated following the manufacturer's indications. Postacquisition analysis was done with FlowJo software 10.0.8 (Tree Star, Inc.). All sampling and FACS analysis was performed in duplicate.

nature research

Corresponding author(s): Peer Bork

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined. — Does not apply

2. Data exclusions

   Describe any data exclusions. — Does not apply

3. Replication

   Describe whether the experimental findings were reliably reproduced. — Does not apply

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups. — Does not apply

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis. — Does not apply

   Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Nature Biotechnology: doi:10.1038/nbt.3960

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	All software is publicly available and details of usage are included in Online Methods

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	Does not apply

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	Does not apply

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	Does not apply
b. Describe the method of cell line authentication used.	Does not apply
c. Report whether the cell lines were tested for mycoplasma contamination.	Does not apply
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	Does not apply

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	Does not apply

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Does not apply

Nature Biotechnology: doi:10.1038/nbt.3960