



HAL
open science

Towards standards for human fecal sample processing in metagenomic studies

Paul I. Costea, Georg Zeller, Shinichi Sunagawa, Eric Pelletier, Adriana A. Alberti, Florence Levenez, Melanie Tramontano, Marja Driessen, Rajna Hercog, Ferris-Elias Jung, et al.

► To cite this version:

Paul I. Costea, Georg Zeller, Shinichi Sunagawa, Eric Pelletier, Adriana A. Alberti, et al.. Towards standards for human fecal sample processing in metagenomic studies. *Nature Biotechnology*, 2017, 35 (11), pp.1069-1076. 10.1038/nbt.3960 . cea-02073086

HAL Id: cea-02073086

<https://cea.hal.science/cea-02073086v1>

Submitted on 15 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**Repository of the Max Delbrück Center for Molecular Medicine (MDC)
in the Helmholtz Association**

<https://edoc.mdc-berlin.de/16793>

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.

This is the final version of the accepted manuscript. The original article has been published in final edited form in:

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

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

Publisher: [Nature America](#) (Springer Nature)

Copyright © 2017 Nature America Inc., part of Springer Nature. All rights reserved.

1 Towards standards for human fecal sample processing in metagenomic 2 studies

3 Paul I. Costea¹, Georg Zeller¹, Shinichi Sunagawa^{1,2}, Eric Pelletier^{3,4,5}, Adriana Alberti³, Florence
4 Levenez⁶, Melanie Tramontano¹, Marja Driessen¹, Rajna Hercog¹, Ferris-Elias Jung¹, Jens Roat
5 Kultima¹, Matthew R. Hayward¹, Luis Pedro Coelho¹, Emma Allen-Vercoe⁷, Laurie Bertrand³, Michael
6 Blaut⁸, Jillian Brown⁹, Thomas Carton¹⁰, Stéphanie Cools-Portier¹¹, Michelle Daigneault⁷, Muriel
7 Derrien¹¹, Anne Druesne¹¹, Willem M. de Vos^{12,13}, B. Brett Finlay¹⁴, Harry J. Flint¹⁵, Francisco
8 Guarner¹⁶, Masahira Hattori^{17,18}, Hans Heilig¹², Ruth Ann Luna¹⁹, Johan van Hylckama Vlieg¹¹, Jana
9 Junick⁸, Ingeborg Klymiuk²⁰, Philippe Langella⁶, Emmanuelle Le Chatelier⁶, Volker Mai²¹, Chaysavanh
10 Manichanh¹⁶, Jennifer C. Martin¹⁵, Clémentine Mery¹⁰, Hidetoshi Morita²², Paul O'Toole⁹, Céline
11 Orvain³, Kiran Raosaheb Patil¹, John Penders²³, Søren Persson²⁴, Nicolas Pons⁶, Milena Popova¹⁰,
12 Anne Salonen¹³, Delphine Saulnier⁸, Karen P. Scott¹⁵, Bhagirath Singh²⁵, Kathleen Slezak⁸, Patrick
13 Veiga¹¹, James Versalovic¹⁹, Liping Zhao²⁶, Erwin G. Zoetendal¹², S. Dusko Ehrlich^{6,27,*}, Joel Dore^{6,*},
14 Peer Bork^{1,28,29,30,*}

15

16 * - 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

17

18 **Abstract**

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

34 Over 3000 publications in the past five years have used DNA- or RNA- based profiling methods to
35 interrogate microbial communities in locations ranging from ice columns in the remote arctic to the
36 human body, resulting in more than 160,000 published metagenomes (both shotgun and 16S rRNA
37 gene)¹. To date, one of the most studied ecosystems is the human gastrointestinal tract. The gut
38 microbiome is of particular interest due to its large volume, high diversity and potential relevance to
39 human health and disease. Numerous studies have found specific microbial fingerprints that may be
40 useful in distinguishing disease states, for example diabetes²⁻⁴, inflammatory bowel disease^{5,6} or
41 colorectal cancer⁷. Others have linked the human gut microbial composition to various factors, such
42 as mode of birth, age, diet and medication⁸⁻¹¹. Such studies have almost exclusively used their own
43 specific, demographically distinct cohort and methodology. Given the many reports of batch effects¹²
44 and known differences when analyzing data generated using different protocols¹³⁻¹⁸, comparisons or
45 meta-analyses are limited in their interpretability. For example, healthy Americans from the HMP
46 study showed lower taxonomic diversity in their stool than patients with inflammatory bowel disease
47 (IBD) from a European study¹⁹, although it is established that IBD patients worldwide have reduced
48 taxonomic diversity²⁰. It is thus currently very difficult to disentangle biological from technical
49 variation when comparing across multiple studies²¹.

50 In metagenomic studies, the calculation of compositional profiles and ecological indices is preceded
51 by a complex data generation process, consisting of multiple steps (Figure 1), each of which is subject
52 to technical variability²². Usually, a small sample is collected by an individual shortly after passing
53 stool and stored in a domestic freezer, prior to shipment to a laboratory. The location within the
54 specimen that the sample is taken from has been shown to impact the measured composition²³,
55 which is why in some studies²⁴ larger quantities were homogenized prior to storage in order to
56 generate multiple, identical aliquots. Furthermore, different fixation methods can be used to
57 preserve the sample for shipping and long-term storage. Freezing at below -20°C is the standard,
58 though more practical alternatives exist²³⁻²⁵. Eventually, the sample is subjected to DNA extraction,
59 library preparation, sequencing and downstream bioinformatics analysis (Figure 1).

60 Here we examined the extent to which DNA extraction influences the quantification of microbial
61 composition, and compared it to other sources of technical and biological variation. The majority of
62 the protocol comparison studies to date have used a 16S rRNA gene amplification approach, which
63 suffers from additional issues. Specifically, the choice of primer, PCR bias and even the choice of
64 polymerase can affect the results²⁶, which may lead to different conclusions when performing the
65 same DNA extraction comparison in a different setup – issues that are minimized using metagenomic
66 sequencing. We compared a wide range of extraction methods, using metagenomic shotgun
67 sequencing, in respect to both taxonomic and functional variability, while keeping all other steps
68 standardized. We investigated the most commonly used extraction kits with varying modifications
69 and additional protocols which do not make use of commercially available kits (see Supplementary
70 Table 1 and Supplementary Information). While other studies have previously investigated the
71 differences between extraction methods in a given setting^{12,15,16,27}, we here systematically tested for
72 reproducibility within and across laboratories on three continents, by applying strict and consistent
73 quality criteria. We further assessed the accuracy of the best performing extraction methods by using
74 a mock community of ten bacterial species whose exact relative abundance was known. This
75 community included both gram-positive and gram-negative bacteria and their relative abundance
76 spanned three orders of magnitude. Based on these analyses we recommend a standardized
77 protocol for DNA extraction from human stool samples, which, if accepted by the research
78 community, will greatly enhance comparability among metagenomic studies.

79 **Results**

80 **Study design**

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

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

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

108 **Quality control for DNA yield and fragmentation**

109 Maximizing DNA concentration while also minimizing fragmentation are key aspects to consider
110 when selecting an extraction protocol. This is both because good quality libraries are required for
111 shotgun sequencing and because protocols that consistently recover low yield or highly fragmented
112 DNA are likely to skew the measured composition. We found considerable variation in the quantity of
113 extracted DNA, in line with previous observations²⁸ (Figure 2). For example, protocol 18 recovered
114 100 times more DNA than protocols 3 and 12, and 10 times more than protocols 8, 19 and 20 (Figure
115 2). Furthermore, there was considerable variation in the fragmentation of the recovered DNA, as
116 measured by the percentage of total DNA in fragments below 1.8 kb in length; for example protocols
117 4, 10 and 12 consistently yielded highly fragmented DNA while for protocol 1 no fragmentation was
118 observed. For subsequent analysis, samples that yielded below 500ng of DNA or were very
119 fragmented (median sample fragmentation above 25%), were not subjected to sequencing. In total,
120 143 libraries, extracted using 21 different protocols passed the quality requirements imposed above,
121 though as an example only four of 18 samples extracted with protocol 16 (one sample A and three
122 sample B replicates) met the requirements (Supplementary Table 3). For other protocols, a small
123 number of samples were discarded for lack of compliance with quality/quantity criteria.

124 **Quality control for variability in taxonomic and functional composition**

125 All metagenomes were compared with respect to taxonomic and functional compositions to quantify
126 the relative abundances of microbial taxa and their respective gene-encoded functions (Methods).
127 Briefly, based on the extracted DNA, shotgun sequencing libraries were prepared and subjected to
128 sequencing on the Illumina HiSeq2000 platform, yielding a mean of 3.8 Gb (+/- 0.7 Gb) per sample.
129 Raw sequencing data were then processed using the MOCAT²⁹ pipeline and relative taxonomic and
130 gene functional abundances were computed by mapping high-quality reads to a database of single
131 copy taxonomic marker genes (mOTUs)¹⁹ and annotated human gut microbial reference genes³⁰,
132 respectively (Methods).

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

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

146 individual variation, as measured between different sampling time points for the same individuals.
147 This effect was much smaller than the between individual variation, resulting in individual-specific
148 microbial composition preservation over time as noted before^{19,23,32}. The smallest contributor
149 observed, quantified on a small number of samples (n=7), was within specimen variation, resulting
150 from sampling different parts of the stool itself²³. In terms of technical sources of variation we have
151 considered measurement errors (assessed through technical replication), library preparation, and
152 effects introduced by the two most widely used preservation^{23,24} methods (fresh freezing and
153 RNAlater). It is important to note that these effects have not all been measured independently of
154 each other, resulting in some of the quantified variations being a convolute of multiple effects
155 (Figure3 – checkboxes).

156 Different distance measures can be used to assess the magnitude of these effects. We focused here
157 on two, which are complementary in terms of the features of the data they consider and thus the
158 dimensions, which become relevant. These distance measures were computed on both
159 metagenomics operational taxonomic units (mOTUs²¹) and clusters of orthologous groups (COGs³³)
160 abundance data, to derive species and functional variation (see Methods). Firstly, we used a
161 Spearman correlation to assess how well species abundance rankings are preserved and found that
162 the variation between most extraction protocols is smaller than the technical within-specimen
163 variation (summarized by the median, Figure 3a). This suggests that, with the exception of protocols
164 8 and 12, all others recover comparable species rankings. Consequently, if only the ranks are of
165 interest, most of the available protocols would provide highly comparable results. However, for
166 many applications the abundances of the taxonomic units are important and need to be
167 commensurable. Using a Euclidean distance (which cumulates abundance deviations) we found that
168 many protocols were not comparable and actually introduce large batch effects at the species level,
169 with the median between-protocol distance being higher than the within-specimen variation (Figure
170 3a), hampering the comparability of samples generated with different extraction methods. To assess
171 similarity between extraction protocol effects, we used principle coordinate analysis (PCoA, see
172 Methods) to visualize these distance spaces (Supplementary Figure 2). These indicated that protocol
173 12, and to a lesser extent also protocols 3, 8, 11, 16 and 18, had abundance profiles that were
174 different from most of the other protocols.

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

181 Among the sources of technical variation, the within-protocol variation (i.e. measurement error) was
182 consistently smallest, with the magnitude of the library preparation effect being comparable (Figure
183 3a,b). The variation introduced by storage method (RNA Later vs. frozen) was larger than within-
184 protocol variation, and, as previously shown, smaller than within-specimen variation in taxonomic
185 space^{23,24}.

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

189 **Quality control for species-specific abundance variation**

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

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

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

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

220 **Factors influencing DNA extraction outcome**

221 Using diversity as an optimality criterion, we determined protocol parameters that are significantly
222 associated with this indicator (Figure 5). For this purpose we focused on protocols that use Qiagen
223 kits¹⁵, namely numbers 5, 6, 8, 9, 11, 13, 15 and 20, which reduces the number of variables that can
224 influence the outcome. We find that “mechanical lysis”, “zirconia beads” and “shaking” are positively
225 associated with diversity. We note that there is no association with DNA fragmentation, as all of the
226 samples extracted with these protocols had a low number of fragments below 1.8 kb (Figure 2). This
227 was consistent with the notion that mechanical lysis and bead beating are necessary to efficiently
228 extract the DNA of gram-positive bacteria that have cell walls that are harder to break³⁵ and also in
229 line with our postulation that effective gram-positive recovery will increase the observed diversity.
230 The only significant negative association was with the InhibitEX tablet, which was included in the kit
231 and which the manufacturer recommends for “absorb[ing] substances that can degrade DNA and

232 inhibit downstream enzymatic reactions so that they can easily be removed by a quick centrifugation
233 step³⁶, though our assessment suggests an adverse effect on DNA extraction quality. This analysis
234 suggests specific modifications with which – also currently suboptimal – extraction methods could be
235 improved, independent of all other variables. For example, introducing a bead beating step is likely
236 to improve the extraction, independent of the specific commercial kit used; adding such a step to the
237 only protocol using Mobio’s PowerSoil kit (protocol 3) would be expected to improve its
238 performance. Our results may therefore generally inform the future development of better DNA
239 extraction protocols.

240 **Protocol reproducibility and transferability across laboratories**

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

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

249 Each extraction method was established and performed in three other laboratories, which had no
250 experience with the respective protocol, in order to assess the wider applicability of each as a
251 standard extraction protocol. All three methods were reproducible across locations, though only
252 protocol H had an effect below that of the smallest biological variation (i.e. within-sample). Protocols
253 W and Q introduced a cross-lab effect comparable to within-sample variation (Supplementary Figure
254 5).

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

262 **Protocol extraction accuracy**

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

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

284 Discussion

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

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

300 We anticipate that procedures for DNA extraction will likely further improve in the future, but put
301 forward protocol Q as a potential benchmark for these new methods. While we have only tested this
302 methodology on stool, we believe it to be applicable to other kinds of samples. However, we caution
303 that additional considerations may apply, such as that of kit contamination³⁹, which may differ
304 between the protocols investigated here and would, for example, have a high impact on samples
305 with low biomass.

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

311 **Online methods**

312 **Library preparation and sequencing**

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

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

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

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

330 **Determining taxonomic and functional profiles**

331 For determining the taxonomic composition of each sample, shotgun sequencing reads were mapped
332 to a database of selected single copy phylogenetic marker genes¹⁹ and summarized into species-level
333 (mOTU) relative abundances. Functional profiles of clusters of orthologous groups (COGs) were
334 computed using MOCAT²⁹ by mapping shotgun sequencing reads to an annotated reference gene
335 catalogue as described in Voigt et al.²³. COG category abundances were calculated by summing the
336 abundance of the respective COGs belonging to each category per sample, excluding NOGs.

337 **Comparison to other technical and biological variation**

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

342 For assessing the variation induced by different preservation methods (namely freezing and RNA-
343 later) we use the data from Franzosa et al.²⁴ and compared the same sample, preserved with the two
344 different methods. For within specimen variation we used data from Voigt et al.²³, where they have
345 sampled the same stool multiple times at different locations along the specimen. As this study also
346 used different storage methods for some samples, we are able to quantify the effect of both within-
347 specimen variation and storage together. For the between time point and individual effect
348 assessment we used the data from the time series data from Voigt et al.²³ as well as a subset of stool
349 samples from the Human Microbiome Project³¹. To ensure comparability across such different
350 studies we have computed distances between all samples on the same subset of relatively abundant

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

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

356 **Determining significantly different species**

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

360 **Mock community cultivation**

361 Bacteria were cultivated at 37°C under anaerobic conditions in a Vinyl Anaerobic Chamber (COY)
362 inflated with a gas mix of approximately 15% carbon dioxide, 83% nitrogen and 2% hydrogen. For
363 long-term storage, cryovials containing freshly prepared bacterial cultures plus 7% DMSO were
364 tightly sealed and frozen at -80°C. Prior to the experiment, bacteria were pre-cultivated twice using
365 modified Gifu Anaerobic Medium broth (mGAM, 05433, HyServe). Bacteria were mixed based on
366 their OD, pelleted by centrifugation and re-suspended in 0.05 Vol RNAlater® Stabilization Solution
367 (AM7020, Thermo Fisher Scientific). 50 µL of this suspension were distributed to 2 mL safe-lock tubes
368 (30120094, Eppendorf) and frozen at -80°C for later DNA extraction and sequencing.

369 When assessing the relative abundances obtained from sequencing the mock community alone, we
370 note the presence of ~6% *Escherichia coli* across all extractions, likely a contamination of the mock
371 community itself and not a result of the DNA extraction. As we did not quantify the input of *E. coli* it
372 was not considered in subsequent evaluation. Apart from this and after rarefying to comparable
373 numbers of reads across the three tested protocols we find no evidence of extraction specific
374 contaminants. However, this may be due to the large quantity of input material which would mask
375 the kit contaminants that are likely in low abundance¹⁶.

376 **Flow Cytometry**

377 Bacterial cells were fixed in 70% Ethanol and stored at 4°C for later analysis at the cytometer. Cells
378 were pelleted and rehydrated in PBS with 1mM EDTA aiming at a dilution of 0.6 OD₆₀₀. We used
379 propidium iodide (PI, Sigma-Aldrich, stock concentration 1 mg/mL resuspended in milliQ H₂O) at a
380 final concentration of 20 µg/mL as fluorescent probe to label bacterial DNA. The cell suspension was
381 sonicated five times for 10 seconds (0.5 seconds ON, 0.5 seconds OFF, 10% amplitude, Branson
382 Sonifier W-250 D, Heinemann) interrupted by 4 min of cooling.

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

390 **Principal coordinate analysis**

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

393

394

395 **References**

- 396 1. Meyer, F. *et al.* The metagenomics RAST server - a public resource for the automatic
397 phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).
- 398 2. Larsen, N. *et al.* Gut microbiota in human adults with type 2 diabetes differs from non-
399 diabetic adults. *PLoS One* **5**, e9085 (2010).
- 400 3. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes.
401 *Nature* **490**, 55–60 (2012).
- 402 4. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the
403 human gut microbiota. *Nature* **528**, 262–266 (2015).
- 404 5. Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn’s disease revealed by a
405 metagenomic approach. *Gut* **55**, 205–11 (2006).
- 406 6. Carroll, I. M. *et al.* Molecular analysis of the luminal- and mucosal-associated intestinal
407 microbiota in diarrhea-predominant irritable bowel syndrome. *Am. J. Physiol. Gastrointest.*
408 *Liver Physiol.* **301**, G799-807 (2011).
- 409 7. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer.
410 *Mol. Syst. Biol.* **10**, 766 (2014).
- 411 8. Dethlefsen, L., McFall-Ngai, M. & Relman, D. a. An ecological and evolutionary perspective on
412 human-microbe mutualism and disease. *Nature* **449**, 811–8 (2007).
- 413 9. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial
414 microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U. S. A.* **107**,
415 11971–5 (2010).
- 416 10. Yatsunencko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**,
417 222–7 (2012).
- 418 11. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers.
419 *Nature* **500**, 541–546 (2013).
- 420 12. Wesolowska-Andersen, A. *et al.* Choice of bacterial DNA extraction method from fecal
421 material influences community structure as evaluated by metagenomic analysis. *Microbiome*
422 **2**, 19 (2014).
- 423 13. McOrist, A. L., Jackson, M. & Bird, A. R. A comparison of five methods for extraction of
424 bacterial DNA from human faecal samples. *J. Microbiol. Methods* **50**, 131–139 (2002).
- 425 14. Smith, B., Li, N., Andersen, A. S., Slotved, H. C. & Kroghfelt, K. A. Optimising bacterial DNA
426 extraction from faecal samples: comparison of three methods. *Open Microbiol. J.* **5**, 14–7
427 (2011).
- 428 15. Maukonen, J., Simões, C. & Saarela, M. The currently used commercial DNA-extraction
429 methods give different results of clostridial and actinobacterial populations derived from
430 human fecal samples. *FEMS Microbiol. Ecol.* **79**, 697–708 (2012).
- 431 16. Kennedy, N. A. *et al.* The impact of different DNA extraction kits and laboratories upon the
432 assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLoS One* **9**,
433 e88982 (2014).
- 434 17. Salonen, A. *et al.* Comparative analysis of fecal DNA extraction methods with phylogenetic
435 microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J.*
436 *Microbiol. Methods* **81**, 127–34 (2010).
- 437 18. Ariefdjohan, M. W., Savaiano, D. A. & Nakatsu, C. H. Comparison of DNA extraction kits for

- 438 PCR-DGGE analysis of human intestinal microbial communities from fecal specimens. *Nutr. J.*
439 **9**, 23 (2010).
- 440 19. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes.
441 *Nat. Methods* **10**, 1196–9 (2013).
- 442 20. Manichanh, C., Borruel, N., Casellas, F. & Guarner, F. The gut microbiota in IBD. *Nat. Rev.*
443 *Gastroenterol. Hepatol.* **9**, 599–608 (2012).
- 444 21. Lozupone, C. A. *et al.* Meta-analyses of studies of the human microbiota. *Genome Res.* **23**,
445 1704–14 (2013).
- 446 22. Raes, J. & Bork, P. Molecular eco-systems biology: towards an understanding of community
447 function. *Nat. Rev. Microbiol.* **6**, 693–9 (2008).
- 448 23. Voigt, A. Y. *et al.* Temporal and technical variability of human gut metagenomes. *Genome Biol.*
449 **16**, 73 (2015).
- 450 24. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut.
451 *Proc. Natl. Acad. Sci. U. S. A.* **111**, E2329–38 (2014).
- 452 25. Song, S. J. *et al.* Preservation Methods Differ in Fecal Microbiome Stability, Affecting
453 Suitability for Field Studies. *mSystems* **1**, (2016).
- 454 26. Gohl, D. M. *et al.* Systematic improvement of amplicon marker gene methods for increased
455 accuracy in microbiome studies. *Nat. Biotechnol.* **34**, 942–949 (2016).
- 456 27. Claassen, S. *et al.* A comparison of the efficiency of five different commercial DNA extraction
457 kits for extraction of DNA from faecal samples. *J. Microbiol. Methods* **94**, 103–10 (2013).
- 458 28. Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z. & Forney, L. J. Evaluation of methods for the
459 extraction and purification of DNA from the human microbiome. *PLoS One* **7**, e33865 (2012).
- 460 29. Kultima, J. R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One*
461 **7**, e47656 (2012).
- 462 30. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing.
463 *Nature* **464**, 59–65 (2010).
- 464 31. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome.
465 *Nature* **486**, 207–214 (2012).
- 466 32. Franzosa, E. A. *et al.* Identifying personal microbiomes using metagenomic codes. *Proc. Natl.*
467 *Acad. Sci.* **112**, 201423854 (2015).
- 468 33. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different
469 taxonomic ranges. *Nucleic Acids Res.* **40**, D284–9 (2012).
- 470 34. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability
471 and resilience of the human gut microbiota. *Nature* **489**, 220–30 (2012).
- 472 35. Santiago, A. *et al.* Processing faecal samples: a step forward for standards in microbial
473 community analysis. *BMC Microbiol.* **14**, 112 (2014).
- 474 36. InhibitEx Tablets - QIAGEN Online Shop. Available at: <https://www.qiagen.com/fr/shop/lab-basics/buffers-and-reagents/inhibitex-tablets/>.
- 476 37. Henderson, G. *et al.* Effect of DNA extraction methods and sampling techniques on the
477 apparent structure of cow and sheep rumen microbial communities. *PLoS One* **8**, e74787
478 (2013).
- 479 38. Jones, M. B. *et al.* Library preparation methodology can influence genomic and functional

480 predictions in human microbiome research. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 14024–9 (2015).
481 39. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based
482 microbiome analyses. *BMC Biol.* **12**, 87 (2014).
483
484

485 **Acknowledgements**

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

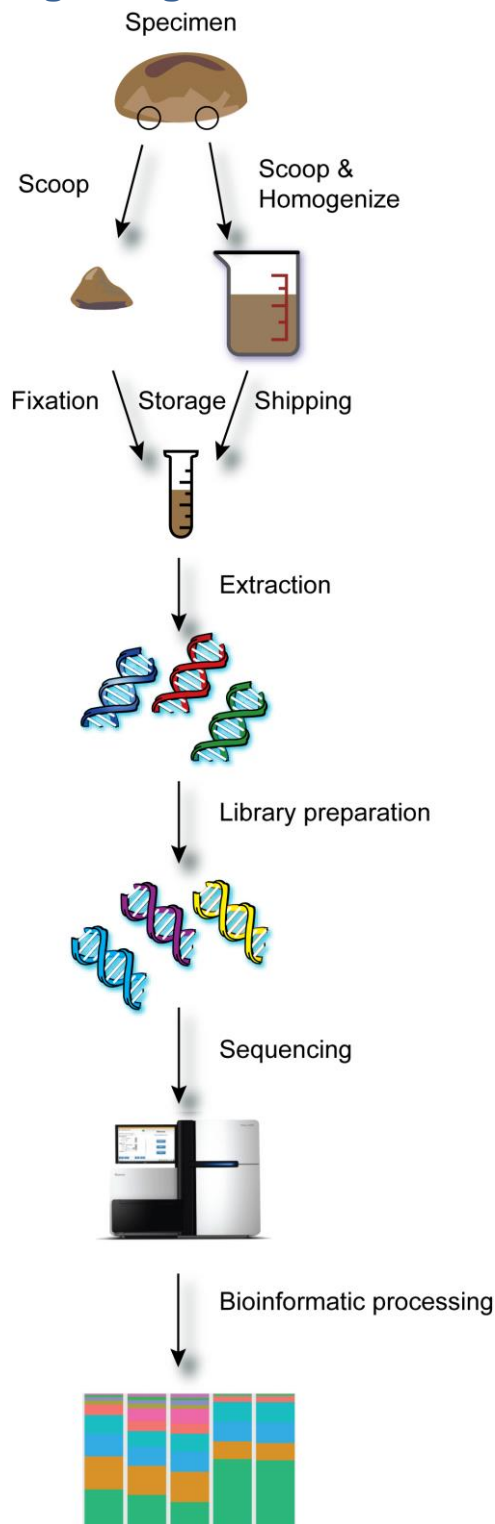
492 **Author contributions**

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

499 **Competing interest statement**

500 The authors declare no competing financial interests.

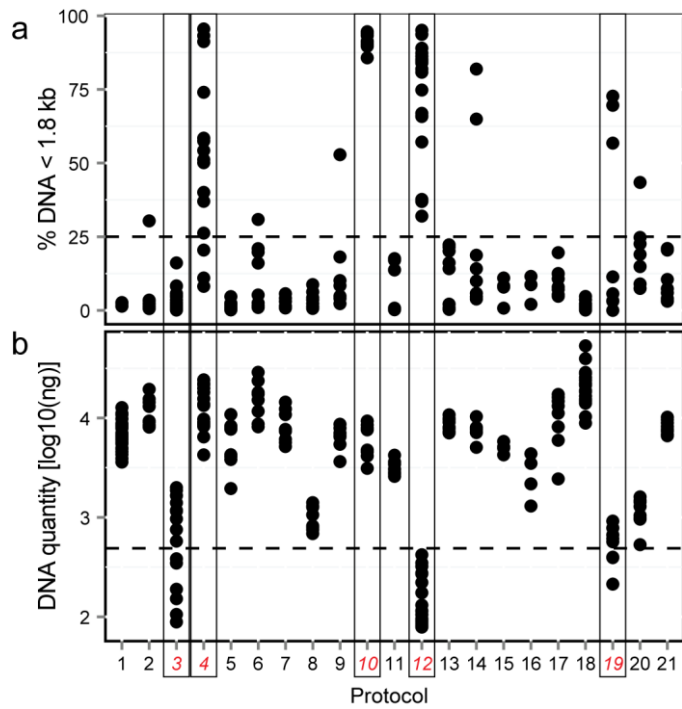
501 **Figure legends**



502

503 **Figure 1: Schematic workflow of human fecal samples processing.**

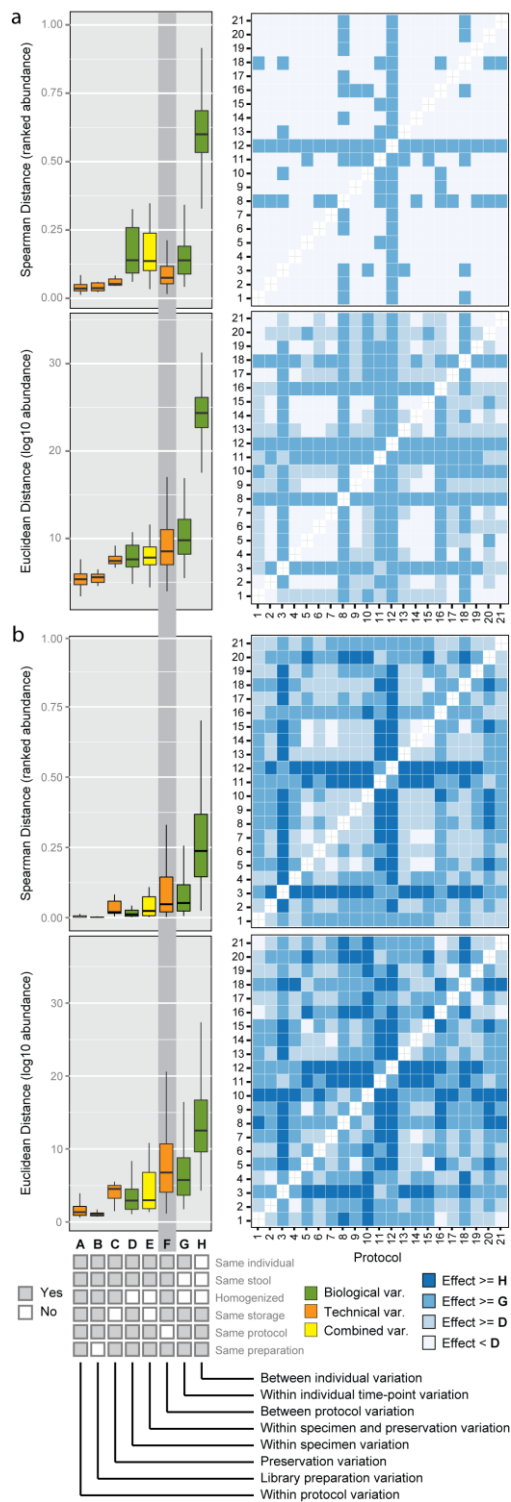
504 Illustration of the main steps involved in extracting and analyzing DNA sequences from human fecal
505 samples, from collection to bioinformatics analysis. Importantly, none of the outlined steps are
506 standardized, which may introduce strong effects between different studies, making their results
507 hard to compare. For example, differences between freezing and RNA-later fixation have been
508 previously described²³ to bias the measured sample composition.



509

510 **Figure 2: Quality control of extracted DNA**

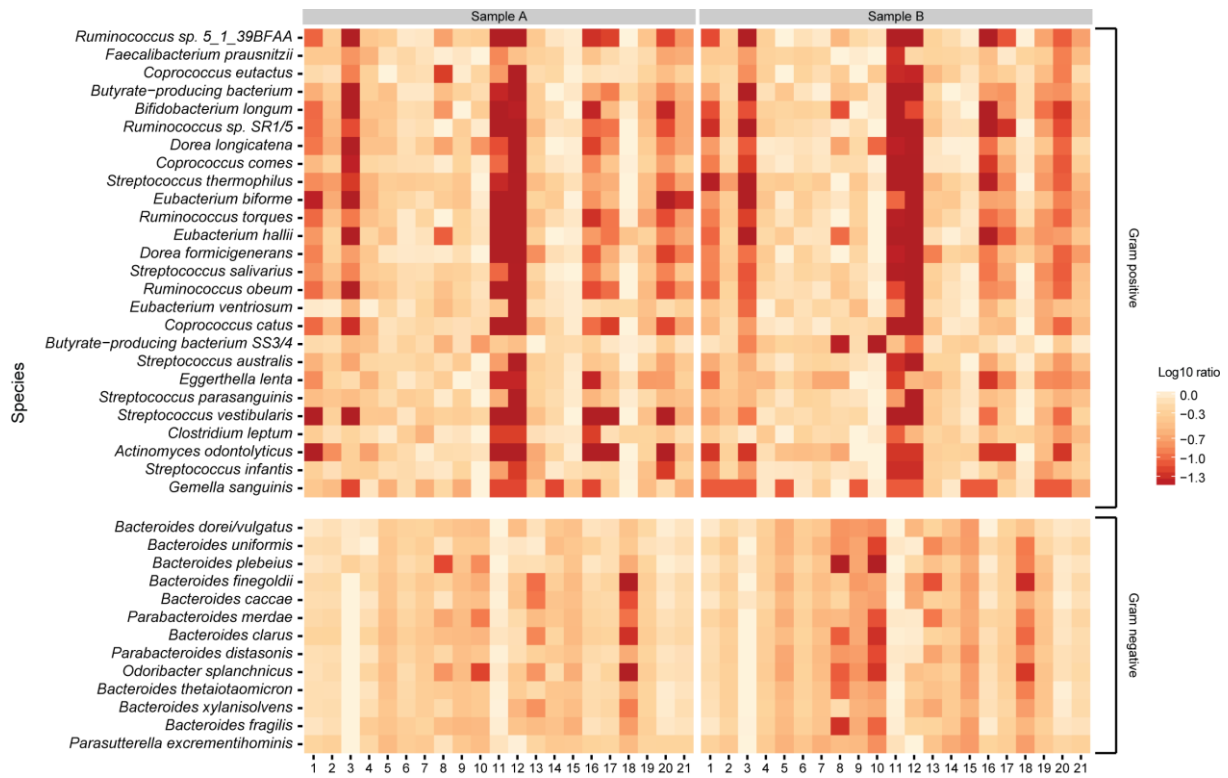
511 Quality (a) and quantity (b) of extracted DNA from 21 different protocols. a) Percentage of DNA
 512 molecules shorter than 1.8 kb, b) quantity of extracted DNA. Protocols failing quality cut-offs
 513 (indicated by dashed lines) for either measurement are highlighted in red and boxed.



514

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

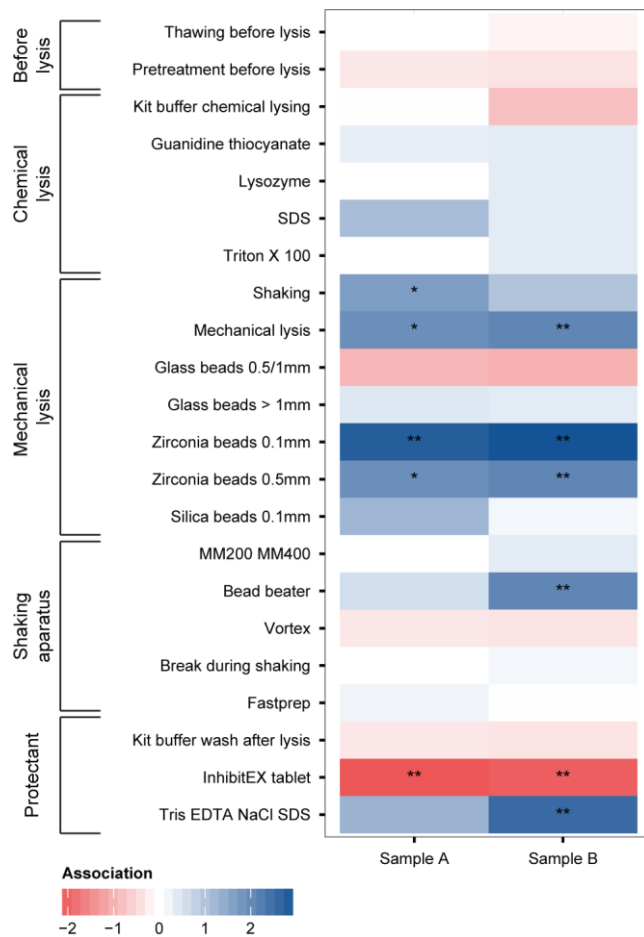
516 Using both a Euclidean and an Spearman distance measure (see Methods) on species abundances
 517 (using mOTU¹⁹) (a) as well functional abundances (using COGs³³) (b), shows the relative effect size of
 518 different sources of variation. These have been assessed on independent samples from different
 519 studies and thus also capture additional differences. The library preparation and the within-protocol
 520 variation are the smallest effects, while the between protocol variation may be greater than some
 521 biological effects^{23,24}. Heat maps on the right show all pairwise distances between protocols,
 522 highlighting which protocol may be considered comparable and which not under different measures
 523 of similarity as encoded by letters D,H and G on the bottom-right.



524

525 **Figure 4: Species specific abundance variation**

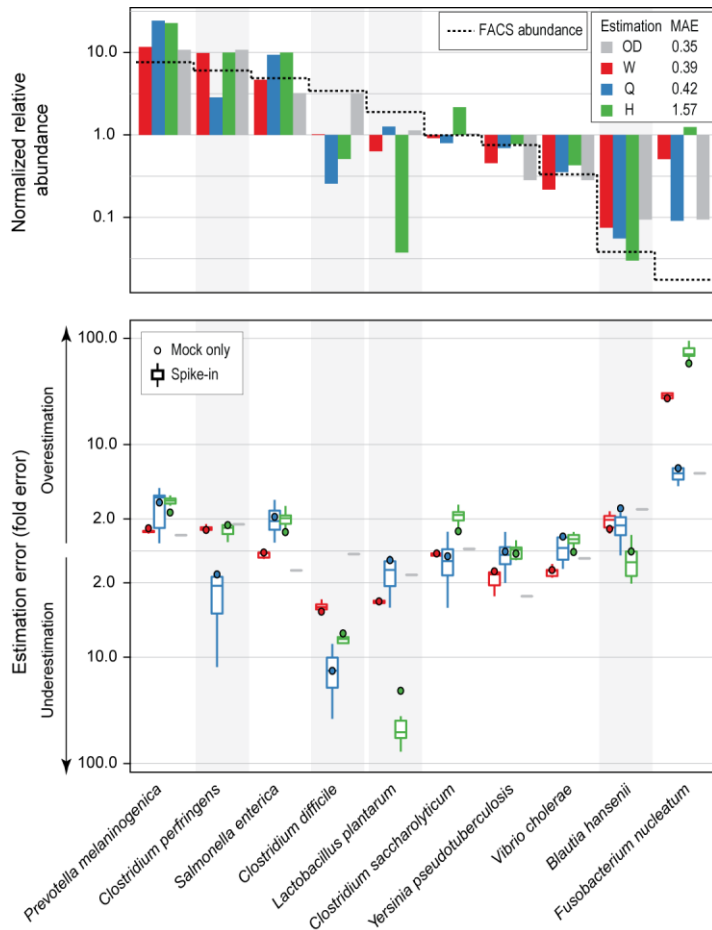
526 Assessing variation of species abundances shows that biases are consistent across the two samples.
 527 Considering species for which the abundances are significantly different between extraction
 528 protocols (Kruskal-Wallis test, FRD corrected p-value < 0.05) we show that gram-positive bacteria are
 529 heavily under-estimated compared to the mean across the five highest recovered ratios, while gram-
 530 negative bacteria are only slightly, though significantly skewed. Abundances are calculated using
 531 mOTUs¹⁹, with only those having a species level annotation being shown.



532

533 **Figure 5: Effects of protocol manipulations on sample composition**

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



538

539 **Figure 6: Mock community extraction quality**

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

550