

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16

## **Proper experimental design requires randomization/balancing of molecular ecology experiments**

Miklós Bálint<sup>1</sup>, Orsolya Márton<sup>1,2</sup>, Marlene Schatz<sup>3</sup>, Rolf-Alexander Düring<sup>3</sup>,  
Hans-Peter Grossart<sup>4,5</sup>

<sup>1</sup> Senckenberg Biodiversity and Climate Research Centre, Senckenberganlage 25,  
60325 Frankfurt am Main, Germany.

<sup>2</sup> Institute for Soil Sciences and Agricultural Chemistry, Centre for Agricultural  
Research, Hungarian Academy of Sciences, Herman Ottó str. 15, H-1022 Budapest,  
Hungary

<sup>3</sup> Institut für Bodenkunde und Bodenerhaltung, Heinrich-Buff-Ring 26, 35392 Gießen,  
Germany

<sup>4</sup> Leibniz Institute for Freshwater Ecology and Inland Fisheries, Alte Fischerhuetten 2,  
16775 Stechlin, Germany.

<sup>5</sup> Institute of Biochemistry and Biology, Potsdam University, Maulbeerallee 2, 14476  
Potsdam, Germany

Corresponding author: Miklós Bálint, [mbalint@senckenberg.de](mailto:mbalint@senckenberg.de), +49 69 7542 1856

17

## **Abstract**

18

Properly designed (randomized and/or balanced) experiments are standard in  
19 ecological research. Molecular methods are increasingly used in ecology, but studies  
20 generally do not report the detailed design of sample processing in the laboratory. This may  
21 strongly influence the interpretability of results if the laboratory procedures do not account for  
22 the confounding effects of unexpected laboratory events. We demonstrate this with a simple  
23 experiment where unexpected differences in laboratory processing of samples would have  
24 biased results if randomization in DNA extraction and PCR steps do not provide safeguards.  
25 We emphasize the need for proper experimental design and reporting of the laboratory  
26 phase of molecular ecology research to ensure the reliability and interpretability of results.

27

## **Keywords**

28

Laboratory practice, bias, nondemonic intrusions, PCR, DNA extraction, batch effect,  
29 environmental DNA, metabarcoding, sediment, lake community

30

## Introduction

31

Ecological studies regularly ensure that the experimental setup is randomized and/or balanced. This allows to interpret results with respect to the original questions and to minimize the influence of confounding factors. The importance of randomized experimental setups (Fisher 1936) along with balanced designs (Student & Student 1938) is well-known. Consequently, such designs are enforced today in manipulative or observational ecological research (Hurlbert 1984).

37

This is often handled differently with laboratory experiments in molecular biology. By laboratory experiments we mean the laboratory processing (versus obtaining) of samples to generate quantitative molecular genetic data: DNA extractions, polymerase chain reactions, DNA sequencing, etc., in order to obtain haplotype frequencies, taxonomically informative marker gene counts, gene expression measures, SNP tables, etc. Early genome-wide association studies (GWAS) are examples of how basic experimental design may be ignored and what the consequences are: the analyses are expensive, but the obtained data cannot be interpreted (or are misinterpreted) due to confounding effects of laboratory procedures (e.g. Sebastiani *et al.* 2010). The early problems lead to the current recognition of randomized and/or balanced laboratory experimental designs in medical genomics (Yang *et al.* 2008; Leek *et al.* 2010; Lambert & Black 2012).

48

Complex and expensive molecular genetic datasets are increasingly generated in ecology. It is important that these data are generated appropriately since important conclusions and recommendations are drawn from them, often addressing issues of global importance for nature, society and economy. Randomization or balancing in laboratory experiments is essential to avoid batch effects and other nondemonic intrusions (see Hurlbert 1984). This issue has been already raised by Meirmans (2015) in a recent opinion

53

54 paper on population genetics. Meirmans notes that “It is perfectly possible that such  
55 randomization is already practised in genotyping laboratories everywhere and I am simply  
56 unaware of it. [...], if this is the case, this is nowhere evident in the literature”. We have  
57 similar impressions and the screening of one randomly selected 2016 issue of five relevant  
58 journals supports this assumption (Molecular Ecology, The ISME Journal, Ecology and  
59 Evolution, Journal of Biogeography, Soil Biology and Biochemistry, Appendix 1). Only two of  
60 the 59 relevant studies report some form of randomization during the laboratory processing  
61 of samples. This small literature survey is surely not representative of overall molecular  
62 ecology research, but the pattern is worrying since a simple Web of Science search for the  
63 keyword combination “molecul\* AND ecol\*” resulted in over 1740 hits only from 2016.

64         The omission of randomization in the lab may allow chance events to systematically  
65 influence results. Such chance events are common everytime and everywhere: electric  
66 fallouts happen, sudden flaws incapacitates lab personnel, DNA extraction kits are not  
67 delivered in time or have been stored inappropriate, just to mention some. If samples are  
68 processed in batches, the coincidence of these events confounds the results and renders  
69 interpretations unreliable. The potential diversity of such events is so high that nothing can  
70 protect against them except randomization of lab procedures, potentially in combination with  
71 balanced designs.

72         Hurlbert (1984) notes that most of the time chance events have immeasurably small  
73 effects on the results. However, by nature they are also completely unpredictable, both in  
74 frequency and effect size. Since molecular ecology studies mostly work with high  
75 observation numbers (thousands of SNPs over genomes, thousands of operational  
76 taxonomic units - OTUs - in hundreds of samples etc.), even small chance events may result  
77 in statistically significant results (Carver 1993). Here, we demonstrate this with taxonomically  
78 informative marker gene fragments amplified from environmental DNA (eDNA)

79 metabarcoding). The eDNA was preserved in lake sediments and provides a perspective on  
80 lake ecosystem history over several decades. We looked at three aspects of methodological  
81 or biological interest: extracted DNA concentration, PCR efficiency and community  
82 properties (Fig. 1). We evaluated several sources of variation: 1) expected laboratory biases  
83 (DNA extraction kit, Deiner *et al.* 2015; Barlow *et al.* 2016), 2) unexpected laboratory biases  
84 (this case a sudden change in lab personnel) and 3) an ecologically interesting predictor  
85 (either the age of the sediment or the effects of a power plant).

## 86 **Materials and Methods**

### 87 **Sampling**

88 Two sediment cores of the same location from Lake Stechlin were taken on May 14,  
89 2015 with a gravity corer (UWITEC®, Mondsee, Austria) and Perspex tubes (inner diameter  
90 9 cm, lengths 60 cm). Lake Stechlin (latitude 53°10'N, longitude 13°02'E) is a dimictic  
91 meso-oligotrophic lake (maximum depth 69.5 m; area 4.5 km<sup>2</sup>) in the lowlands of Northern  
92 Germany. GDR's first nuclear power plant was built here between 1960-1966 and operated  
93 until 1990, connecting the lake with the nearby mesotrophic Lake Nehmitz and discharging  
94 its cooling water into Lake Stechlin. After coring, the cores were sliced immediately in the  
95 field in approximately 0.5 mm intervals. The first core was designated to eDNA. All sampling  
96 tools were H<sub>2</sub>O<sub>2</sub>-sterilized after cutting each horizon. Sediment for DNA extraction was taken  
97 only from the central part of the core to avoid contamination by contact with the corer's wall.  
98 Samples were immediately stored in 15-ml Falcon tubes (NeoLab Migge GmbH, Heidelberg,  
99 Germany) at -20 °C until DNA extraction. Horizons from the second core were used for  
100 organohalogenic pesticide measurements.

### 101 **Date approximation**

102 Approximate dates were obtained by comparing DDT decomposition compound  
103 concentrations with sedimentation rates inferred with <sup>137</sup>Cs (Casper 1994): 1.2 mm year<sup>-1</sup>  
104 between 1986-1996 and 1.7 mm year<sup>-1</sup> between 1963-1986. We assumed that DDT  
105 deposition started with World War II when a military training camp was operated near the  
106 lake and it effectively stopped in 1990 when agrochemical subventions of the GDR ceased  
107 with the reunification of Germany. The pesticide concentrations, sedimentation rates and  
108 inferred dates can be consulted in the file Stechlin\_organohalogene.csv, deposited in  
109 Figshare (<https://figshare.com/s/32dbca0a906c7f06449b>, DOI:

110 10.6084/m9.figshare.4579681).

111 Halocarbon compound extraction was performed by shaking 300 mg freeze-dried  
112 sediment sample once in acetone and petroleum ether (40-60 °C) and then only in  
113 petroleum ether (40-60 °C), based on ISO 10382:2002. The clear supernatants were unified  
114 and vortexed after centrifugation, then a 10 ml aliquot was transferred to SPME amber screw  
115 top vials and evaporated under a gentle stream of nitrogen until dryness and dissolved again  
116 in 100 µL methanol and mixed with 10 mL of a 0.01 mol CaCl<sub>2</sub> \*2H<sub>2</sub>O / 3.4 mol NaCl salt  
117 solution. As internal standards 13C 2.4 DDT, 13C 4.4 DDT, α-HCH D6, Trifluralin D14, 4.4  
118 DDD D8, 4.4 DDE D8, 13C HCB were used. Finally samples were extracted by Head Space  
119 Solid Phase Microextraction (SPME- HS) with a PDMS 100 fibre and analysed by GC/MS  
120 ion trap in selected ion monitoring mode (SIM). Separation and detection were accomplished  
121 using a Trace Ultra Gas Chromatograph (Thermo Fisher Scientific Inc., Schwerte, Germany)  
122 provided with a RTX-Dioxin 2 fused-silica capillary column with 0.25 µm film thickness, 0.25  
123 mm ID and 60 m length coupled with an ion trap mass spectrometer in SIM (Thermo Fisher  
124 Scientific Inc., Schwerte, Germany).

## 125 **DNA extraction**

126 We selected the youngest 21 horizons (the upper 13.5 cm of the core) for DNA  
127 extractions. Sample order was randomized before DNA extraction to minimize sampling  
128 biases. Four DNA extractions were carried out from each horizon with two commercial kits  
129 (two replicated extractions with both Macherey-Nagel NucleoSpin Soil - Macherey-Nagel,  
130 Düren, Germany, and MoBio PowerSoil - Carlsbad, CA, USA). The protocols of both kits  
131 were modified to specifically target extracellular DNA: instead of lysis, a saturated phosphate  
132 buffer was used to extract sediment-bound DNA (Taberlet *et al.* 2012). All four extraction  
133 replicates of a horizon horizon were performed in a row (see column extract\_order in  
134 sample\_infos.csv deposited in Figshare (<https://figshare.com/s/32dbca0a906c7f06449b>,

135 DOI: 10.6084/m9.figshare.4579681). Four extraction controls (dH<sub>2</sub>O instead of sediment)  
136 were randomly included into the extractions. Extracted DNA concentrations were estimated  
137 on a Qubit 3.0 Fluorimeter (Thermo Fisher Scientific, Waltham, MA, USA).

### 138 **PCR amplifications**

139 DNA templates were re-randomized before PCR setup. Four PCR negative controls  
140 (with dH<sub>2</sub>O instead of DNA template) and two positive controls (containing DNA from  
141 *Hypsiboas punctatus*, *Ponticola kessleri*, *Aspius aspius*, *Coregonus* sp., *Pacifastacus*  
142 *leniusculus*, *Aphanomyces astaci*, a parasitic Chytridiomycota, a saprotrophic  
143 Chytridiomycota, *Yamagishiella* sp., *Fragilaria crotonensis*, *Staurastrum planktonicum*,  
144 *Chaetomium* sp., *Lutra lutra*) were included. We used AmpliTaq MasterMix for the PCRs  
145 (Thermo Fisher Scientific, Waltham, MA, USA). We used general eukaryote primers that  
146 amplify a short fraction of the V7 region of the 18S gene region (Guardiola *et al.* 2015):  
147 forward - TYTGTCTGSTTRATTSCG, reverse - CACAGACCTGTTATTGC. The primers  
148 contained the Illumina sequencing primers  
149 (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG and  
150 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG). The PCRs were run in 15  $\mu$ l reaction  
151 volume (AmpliTaq mastermix: 7.5  $\mu$ l, water: 4  $\mu$ l, each 5  $\mu$ M primer 1  $\mu$ l, DNA template 1.5  $\mu$ l).  
152 The cycling conditions were 95 °C (10 min), 44 cycles of 95 °C (30 sec), 45 °C (30 sec),  
153 72°C (30 sec), final extension at 72 °C (10 min). The PCR products were visualized on a 2%  
154 agarose gel and purified with Agencourt AMPure XP beads (Beckman Coulter GmbH,  
155 Krefeld, Germany).

### 156 **Multiplexing strategy and sequencing**

157 We indexed all samples for multiplexed sequencing in a subsequent short PCR with  
158 primers that contained a fraction of the Illumina sequencing primer (TCGTCGGCAGCGTC  
159 and GTCTCGTGGGCTCGG), an eight-bp nucleotide index, and the Illumina plate adapters



160 (P5: AATGATACGGCGACCACCGAGATCTACAC, P7:  
161 CAAGCAGAAGACGGCATAACGAGAT). The final products are indexed, ready to sequence  
162 Illumina libraries. Index combinations and sequences are provided in the file  
163 multiplexing\_indices.xlsx at Figshare (<https://figshare.com/s/32dbca0a906c7f06449b>, DOI:  
164 10.6084/m9.figshare.4579681). The procedure follows the Illumina 16S metabarcoding  
165 protocol (Illumina 2016). This protocol eliminates index jumps during library preparation  
166 (although a few index jumps are still known to happen on the sequencing plate (Schnell *et al.*  
167 2015). The indexing PCRs were run in 15  $\mu$ l reaction volume (AmpliTaq mastermix: 7.5  $\mu$ l,  
168 each 5  $\mu$ M primer 1  $\mu$ l, PCR product 6.5  $\mu$ l). The cycling conditions were 95 °C (10 min), 8  
169 cycles of 95 °C (30 sec), 52 °C (30 sec), 72°C (30 sec), final extension at 72 °C (10 min). We  
170 checked the efficiency of each PCR run on a 2% agarose gel. The indexed libraries were  
171 purified with Agencourt AMPure XP beads (Beckman Coulter GmbH, Krefeld, Germany).  
172 The indexed libraries were mixed and purified on four QIAamp MinElute columns (Qiagen,  
173 Hilden, Germany). We did not normalize the PCR template concentrations to obtain a rough  
174 estimate of PCR and sequencing efficiency through the read numbers. Our sequencing kit  
175 potentially produces about 1 million paired-end reads with 2 x 150 bp length. Illumina  
176 sequencing was performed at the Berlin Center for Genomics in Biodiversity Research  
177 ([www.begendiv.de](http://www.begendiv.de)) with the MiSeq sequencing kit v2 nano (300 cycles). Unprocessed  
178 sequence data were deposited in the European Nucleotide Archive as PRJEB19403.

## 179 **Sequence processing and data analysis**

180 Raw sequence data were processed with OBITools (Boyer *et al.* 2015). Potential  
181 contamination and false detection biases were controlled for by following the  
182 recommendations of (Giguet-Covex *et al.* 2014; Boyer *et al.* 2015; Pansu *et al.* 2015) in R  
183 3.3.1. (R Core Team 2016). All OBITools and R commands are documented in the file  
184 stechlin\_analyses.pdf at Figshare (<https://figshare.com/s/32dbca0a906c7f06449b>, DOI:  
185 10.6084/m9.figshare.4579681), with the full code accessible through the GitHub repository

186 <https://github.com/MikiBalint/LaboratoryDesign.git>. Commands were run with GNU ‘parallel’  
187 when possible (Tange 2011). The resulting OTU abundance table is provided in the  
188 `stechlin_assigned_190915.tab` file through Figshare  
189 (<https://figshare.com/s/32dbca0a906c7f06449b>, DOI: 10.6084/m9.figshare.4579681).

190 We fitted linear mixed-effect models with `lme4` (Bates *et al.* 2015) on extracted DNA  
191 concentration, PCR efficiency and measures of diversity (the first three integers from Hill’s  
192 diversity series (Hill 1973)) to estimate the effects of potential laboratory biases and  
193 biological factors of interests. The first three Hill numbers correspond to species richness  
194 (H1), the exponent of Shannon diversity (H2), and the inverse of the Simpson diversity (H3).  
195 The identity of the sediment horizon was used as the random effect in these models. We  
196 used multispecies generalized linear models (GLMs) with the ‘`mvabund`’ R package (Wang  
197 *et al.* 2012) to investigate the effects of the predictors on community composition. The  
198 multispecies GLM cannot handle random effects. The community composition effects were  
199 visualized with a latent variable model-based ordination performed with the `boral` R package  
200 (Hui 2016). Both compositional analyses assume a negative binomial distribution of the data,  
201 accounting for the sparse and overdispersed nature of read counts (Bálint *et al.* 2016). The  
202 input data matrices are accessible through Figshare  
203 (<https://figshare.com/s/32dbca0a906c7f06449b>, DOI: 10.6084/m9.figshare.4579681).

204 The models can be written up as

- 205 1) `conc ~ weight + kit + person + age + l(age^2) + (1|depth.nominal)`
- 206 2) `PCR efficiency ~ conc + kit + person + age + (1|depth.nominal)`
- 207 3) `diversities ~ PCR efficiency + person + kit + nuclear + (1|depth.nominal)`
- 208 4) `Community composition ~ reads + kit + person + nuclear,`

209 where *conc* is the extracted DNA concentration, *weight* is the sediment weight used

210 for DNA extraction, *kit* is the DNA extraction kit, *person* is the lab personnel, *depth.nominal*  
211 is the identity of the sediment horizon, *PCR efficiency* is estimated from HTS read numbers,  
212 *nuclear* is the operational period of the nuclear power plant (Fig. 1).

213

## Results

214           The results are summarized in Table 1 and Fig. 2. Regarding DNA concentrations,  
215 the DNA extraction kit (equivalent to the expected laboratory biases) accounted for most  
216 variation, followed by the age of the sediment horizon (biological signal) and the lab  
217 personnel (equivalent to the unexpected laboratory bias). The starting weight (amount) of the  
218 sediment had limited effects on the extraction efficiency and the effect of the lab personnel  
219 was marginally significant. PCR efficiency (evaluated as non-normalized HTS read numbers  
220 from PCR amplification) was mostly explained by the personnel identity (unexpected lab  
221 bias), followed by the DNA extraction kit (expected laboratory biases), the age of the  
222 sediment horizon, and the DNA template concentration used for the PCR. Here, the effect of  
223 the lab personnel was statistically significant. The most important contributors to variation in  
224 the first three Hill numbers consisted of PCR efficiency and the effects of the nuclear power  
225 plant. The DNA extraction kit contributed relatively little to the observed variation in the  
226 diversity indices. The nuclear power plant effects, however, represented the largest  
227 contributors to the explained variation in community composition, followed by the identity of  
228 the lab personnel and PCR efficiency. The DNA extraction kits contributed the least to the  
229 explained compositional variation (Fig. 3). The effects of the lab personnel were statistically  
230 marginally significant. Additional results and effect plots are available in file  
231 `stechlin_analyses.pdf` at Figshare.

## 232                    **Discussion**

233                    Our results demonstrate that nondemonic intrusions (Hurlbert 1984) in the laboratory  
234                    may produce in strong, statistically significant effects that may severely confound results.  
235                    Such effects render equivocal interpretations impossible if they coincide with effects targeted  
236                    by the study. For example, interpretation of power plant effects on community composition  
237                    would be difficult if samples are processed in batches and the sudden change in laboratory  
238                    personnel coincides with a shift between operation periods.

239                    We don't state that biases with comparable extent always appear in unrandomized,  
240                    not balanced laboratory experiments, but they certainly have the potential to do so. This is  
241                    clear in our example: the effects of unexpected laboratory biases exceed the effects of  
242                    known lab biases (DNA extraction kit effects) and biological signal in several models (Fig. 2).  
243                    Such effects potentially influence all molecular ecology studies and threaten the  
244                    interpretability of results. Their importance and extent is widely known in biomedicine (Yang  
245                    *et al.* 2008; Leek *et al.* 2010; Lambert & Black 2012) and needs to be urgently considered in  
246                    molecular ecology.

247                    Generally, randomization of samples before major laboratory steps (extraction, PCR,  
248                    sequencing) is simple and low-cost. The only case where this might be disputable is the  
249                    processing of highly contamination-prone materials where it is almost a lab rule that DNA  
250                    extraction is performed consecutively from the most contamination-prone toward the least  
251                    contamination-prone samples (although to our knowledge the validity of this still needs to be  
252                    tested). Obviously, nondemonic intrusions (including contamination) in the laboratory easily  
253                    become collinear with the processing order and this makes biological signals difficult to  
254                    interpret (Salter *et al.* 2014).

255                    We recommend the followings: first, researchers involved in molecular ecology

256 labwork need to properly design and report laboratory procedures. Guidelines in biomedicine  
257 exist and may be readily adapted (Masca *et al.* 2015). Second, ecologists who rely on  
258 molecular data generated by laboratory personnel or companies must ensure (and should  
259 not take for granted) that principles of experimental design are followed in the laboratory.  
260 This is the easiest when giving samples to a lab since the ecologist can already rearrange  
261 and relabel his/her samples (but controls of PCR, sequencing, orders, etc. may require  
262 further communication). Third, editors and reviewers of manuscripts and grants should  
263 enforce the reporting of laboratory experimental design. This is as much necessary for  
264 reproducible research as the proper presentation of sampling schemes, details of  
265 manipulative experiments and data analysis. We do not intend to provide a list of important  
266 laboratory biases since there are potentially infinite variations. Therefore, molecular  
267 ecologists must ensure randomization or properly balanced designs in every step of  
268 laboratory work and present the details. There is no excuse for avoiding this since more and  
269 more globally important decisions require reliable molecular ecology data in nature and  
270 biodiversity conservation.

271

## Acknowledgements

272

The authors express gratitude to the working community of the Senckenberg

273

Conservation Genetics Group (Gelnhausen, Germany) for hosting and supporting the

274

labwork. Claudia Wittwer, Silke Van den Wyngaert, Martin Jansen, and Berardino

275

Cocchiararo provided tissues for generating the positive controls. We thank Susan Mbedi

276

from BeGenDiv for suggestions related to multiplexed sequencing library preparation and

277

sequencing. MB is supported by DFG grant BA 4843/2-1.

278 **References**

- 279 Bálint, M., Bahram, M., Eren, A.M., Faust, K., Fuhrman, J.A., Lindahl, B., O'Hara, R.B., Öpik,  
280 M., Sogin, M.L., Unterseher, M. & Tedersoo, L. (2016). Millions of reads, thousands of  
281 taxa: microbial community structure and associations analyzed via marker genes. *FEMS*  
282 *microbiology reviews*, **40**, 686–700.
- 283 Barlow, A., Gonzalez Fortes, G.M., Dalen, L., Pinhasi, R., Gasparyan, B., Rabeder, G.,  
284 Frischchauf, C., Paijmans, J.L.A. & Hofreiter, M. (2016). Massive influence of DNA  
285 isolation and library preparation approaches on palaeogenomic sequencing data.
- 286 Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models  
287 Using lme4. *Journal of statistical software*, **67**.
- 288 Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac, E. (2015). obitools:  
289 aunix-inspired software package for DNA metabarcoding. *Molecular ecology resources*,  
290 **16**, 176–182.
- 291 Carver, R.P. (1993). The Case Against Statistical Significance Testing, Revisited. *Journal of*  
292 *experimental education*, **61**, 287–292.
- 293 Casper, P. (1994). Die Cäsium-Datierung von Sedimenten unterschiedlicher mikrobieller  
294 Aktivität, Erweiterte Zusammenfassungen der Jahrestagung, DGL - Deutsche  
295 Gesellschaft für Limnologie, 386-389.
- 296 Deiner, K., Walser, J.-C., Mächler, E. & Altermatt, F. (2015). Choice of capture and extraction  
297 methods affect detection of freshwater biodiversity from environmental DNA. *Biological*  
298 *conservation*, **183**, 53–63.
- 299 Fisher, R.A. (1936). Design of Experiments. *BMJ* , **1**, 554–554.
- 300 Giguet-Covex, C., Pansu, J., Arnaud, F., Rey, P.-J., Griggo, C., Gielly, L., Domaizon, I.,  
301 Coissac, E., David, F., Choler, P., Poulénard, J. & Taberlet, P. (2014). Long livestock  
302 farming history and human landscape shaping revealed by lake sediment DNA. *Nature*  
303 *communications*, **5**, 3211.
- 304 Guardiola, M., Uriz, M.J., Taberlet, P., Coissac, E., Wangenstein, O.S. & Turon, X. (2015).  
305 Deep-Sea, Deep-Sequencing: Metabarcoding Extracellular DNA from Sediments of  
306 Marine Canyons. *PloS one*, **10**, e0139633.
- 307 Hill, M.O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences.  
308 *Ecology*, **54**, 427–432.
- 309 Hui, F.K.C. (2016). boral- Bayesian Ordination and Regression Analysis of Multivariate  
310 Abundance Data inr. *Methods in ecology and evolution / British Ecological Society*, **7**,  
311 744–750.
- 312 Hurlbert, S.H. (1984). Pseudoreplication and the Design of Ecological Field Experiments.  
313 *Ecological monographs*, **54**, 187–211.
- 314 Illumina (2016). 16S Metagenomic Sequencing Library Preparation. Part # 15044223 Rev. A  
315 pp 28. URL:  
316 <http://support.illumina.com/content/dam/illumina-support/documents/documentation/che>



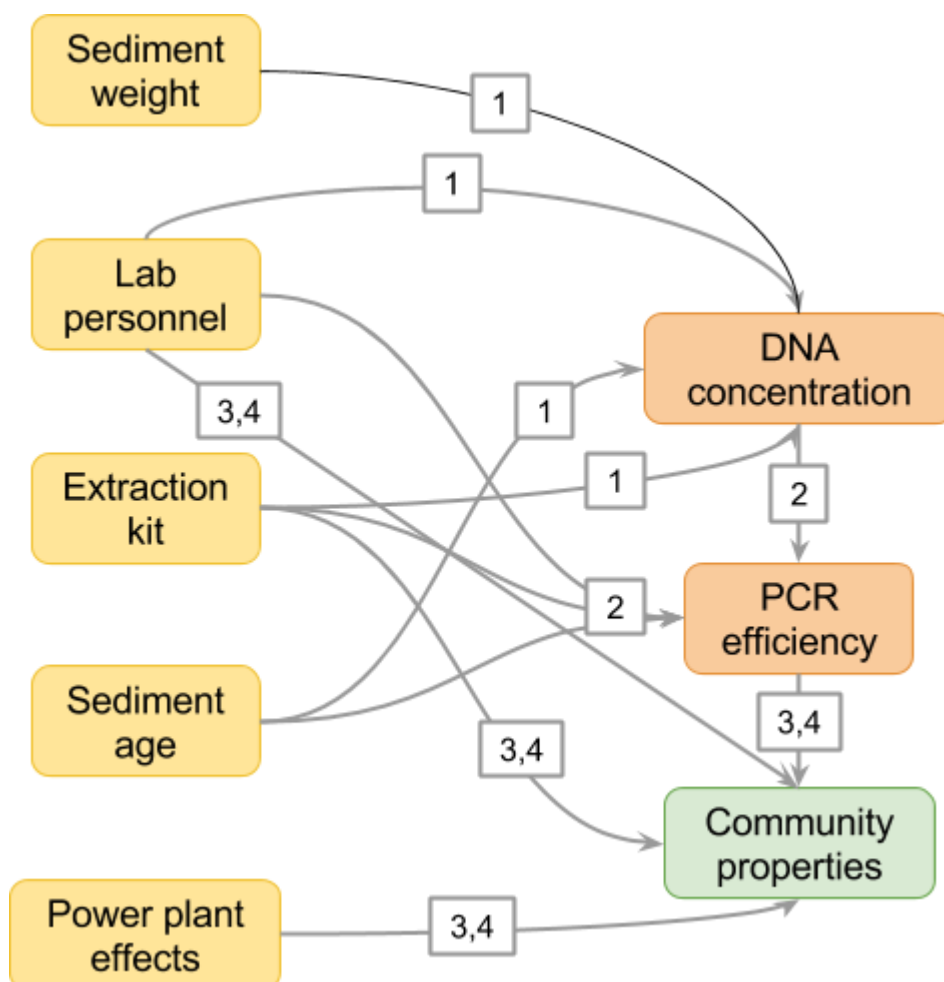
- 317 [mistry\\_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf](#).  
318 Accessed on January 13, 2016.
- 319 Lambert, C.G. & Black, L.J. (2012). Learning from our GWAS mistakes: from experimental  
320 design to scientific method. *Biostatistics*, **13**, 195–203.
- 321 Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman,  
322 D., Baggerly, K. & Irizarry, R.A. (2010). Tackling the widespread and critical impact of  
323 batch effects in high-throughput data. *Nature reviews. Genetics*, **11**, 733–739.
- 324 Masca, N.G., Hensor, E.M., Cornelius, V.R., Buffa, F.M., Marriott, H.M., Eales, J.M.,  
325 Messenger, M.P., Anderson, A.E., Boot, C., Bunce, C., Goldin, R.D., Harris, J.,  
326 Hinchliffe, R.F., Junaid, H., Kingston, S., Martin-Ruiz, C., Nelson, C.P., Peacock, J.,  
327 Seed, P.T., Shinkins, B., Staples, K.J., Toombs, J., Wright, A.K. & Teare, M.D. (2015).  
328 RIPOSTE: a framework for improving the design and analysis of laboratory-based  
329 research. *eLife*, **4**.
- 330 Meirmans, P.G. (2015). Seven common mistakes in population genetics and how to avoid  
331 them. *Molecular ecology*, **24**, 3223–3231.
- 332 Pansu, J., Giguet-Covex, C., Ficetola, G.F., Gielly, L., Boyer, F., Zinger, L., Arnaud, F.,  
333 Poulénard, J., Taberlet, P. & Choler, P. (2015). Reconstructing long-term human impacts  
334 on plant communities: an ecological approach based on lake sediment DNA. *Molecular*  
335 *ecology*, **24**, 1485–1498.
- 336 R: The R Project for Statistical Computing. URL <https://www.R-project.org/> [accessed 23  
337 January 2017]
- 338 Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P.,  
339 Parkhill, J., Loman, N.J. & Walker, A.W. (2014). Reagent and laboratory contamination  
340 can critically impact sequence-based microbiome analyses. *BMC biology*, **12**, 87.
- 341 Schnell, I.B., Bohmann, K. & Gilbert, M.T.P. (2015). Tag jumps illuminated--reducing  
342 sequence-to-sample misidentifications in metabarcoding studies. *Molecular ecology*  
343 *resources*, **15**, 1289–1303.
- 344 Sebastiani, P., Solovieff, N., Puca, A., Hartley, S.W., Melista, E., Andersen, S., Dworkis,  
345 D.A., Wilk, J.B., Myers, R.H., Steinberg, M.H., Montano, M., Baldwin, C.T. & Perls, T.T.  
346 (2010). Genetic signatures of exceptional longevity in humans. *Science*, **2010**.
- 347 Student & Student. (1938). Comparison Between Balanced and Random Arrangements of  
348 Field Plots. *Biometrika*, **29**, 363.
- 349 Taberlet, P., PRUD'HOMME, S.M., Campione, E., Roy, J., Miquel, C., Shehzad, W., Gielly,  
350 L., Rioux, D., Choler, P., Clément, J.-C., Melodelima, C., Pompanon, F. & Coissac, E.  
351 (2012). Soil sampling and isolation of extracellular DNA from large amount of starting  
352 material suitable for metabarcoding studies. *Molecular ecology*, **21**, 1816–1820.
- 353 Tange, O. (2011). GNU Parallel - The Command-Line Power Tool. The USENIX Magazine,  
354 **31**:42-47.
- 355 Wang, Y., Naumann, U., Wright, S.T. & Warton, D.I. (2012). mvabund - an R package for  
356 model-based analysis of multivariate abundance data. *Methods in ecology and evolution*

357 / *British Ecological Society*, **3**, 471–474.

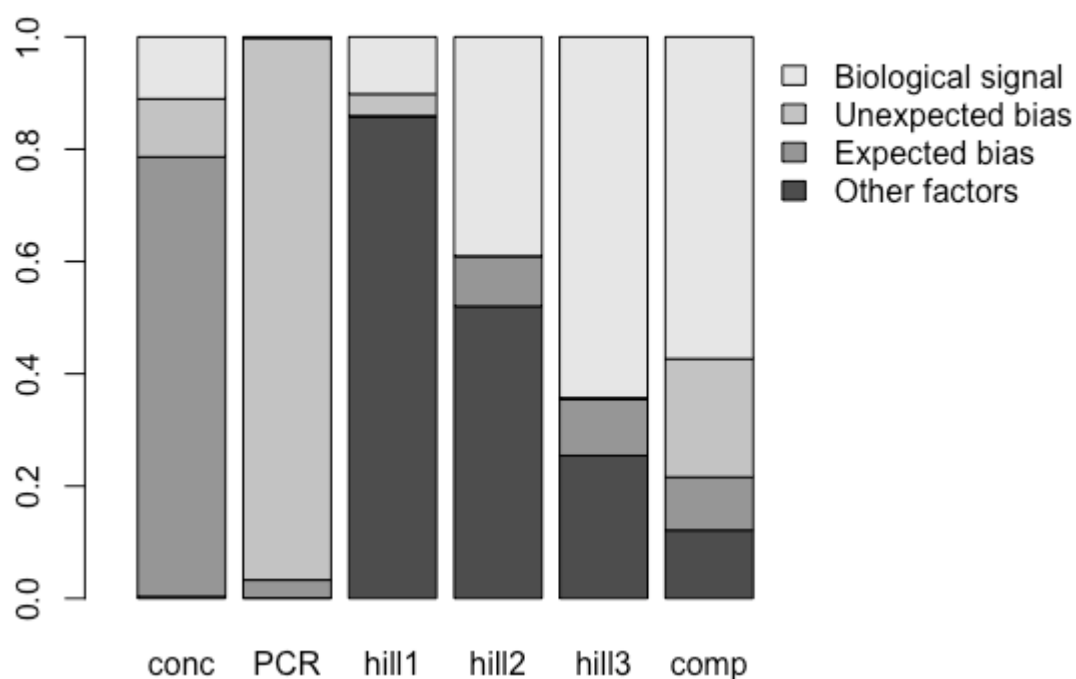
358 Yang, H., Harrington, C.A., Vartanian, K., Coldren, C.D., Hall, R. & Churchill, G.A. (2008).  
359 Randomization in Laboratory Procedure Is Key to Obtaining Reproducible Microarray  
360 Results. *PloS one*, **3**, e3724.

361

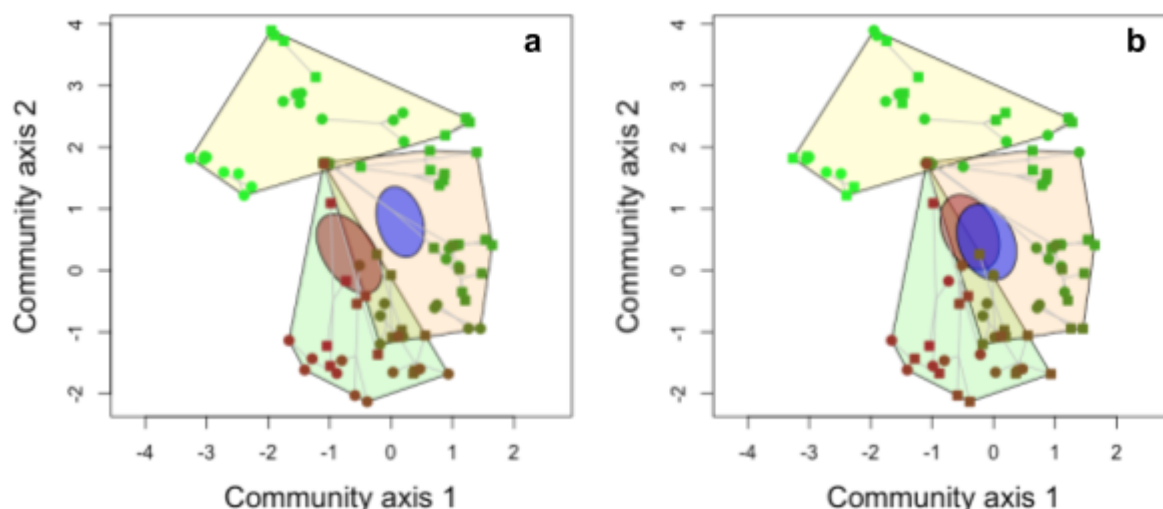
## Figures



362 Fig. 1. Analysis scheme with predictors of variation in high-throughput-sequenced  
363 eDNA data. Numbers on arrows refer to models of DNA concentration, PCR efficiency and  
364 community properties (see Materials and Methods). Yellow color marks variables that were  
365 used only as predictors in models. Orange variables were both predictors and responses in  
366 some of the models. Green marks variables that were only responses in models.



367 Fig. 2. Partitioning of variance explained by expected and unexpected laboratory  
368 biases, and biological signal. The bars represent explained variance in DNA concentration  
369 (conc), PCR efficiency (PCR), diversity indices (hill1-3) and community composition (comp).  
370 Predictors: biological signal: effects of sediment age (conc, PCR) or the power plant  
371 operation periods (hill1-3, comp); unexpected bias: effects of laboratory personnel; expected  
372 bias: effects of DNA extraction kit; other factors: sediment weight (conc), DNA concentration  
373 (PCR); PCR efficiency (hill1-3, comp).



374 Fig. 3. Compositional changes in historic communities explained by expected and  
375 unexpected laboratory biases, and biological signal. Points represent communities  
376 reconstructed from replicated DNA extractions from 21 sediment horizons, representing the  
377 last ~70 years of the lake's history. Symbol color indicates age: dark brown are the oldest,  
378 light green are the youngest communities. Replicated DNA extracts of a horizon are  
379 connected by grey lines. The operational phases of the nuclear power plant are marked with  
380 hulls: green - before building the plant, orange - during power plant operation, yellow - after  
381 operation. a) symbols mark the effects of lab personnel on community composition and the  
382 two ellipses show the 95% confidence interval of the corresponding group centroids. b)  
383 symbols and ellipses mark the effects of the DNA extraction kits.

384

## Tables

385

Table 1. Summary of predictor contributions to variation. \* statistically marginally

386

significant result ( $p < 0.1$ ), \*\* statistically significant result ( $p < 0.05$ ), + statistical significance not

387

tested.

	DNA concentration	PCR efficiency	H1	H2	H3	Community composition
Sediment weight	0.1 <sup>+</sup>		-	-	-	-
Extraction kit	12.9 <sup>+</sup>	2719604 <sup>+</sup>	672.6 <sup>+</sup>	197.9 <sup>+</sup>	67.8 <sup>+</sup>	906
Lab personnel	1.7 <sup>**</sup>	81413118 <sup>**</sup>	37.9	0.9	1.9	2018 <sup>*</sup>
DNA concentration	-	58 222 <sup>+</sup>	-	-	-	-
PCR efficiency	-	-	14834.7 <sup>+</sup>	156.3 <sup>+</sup>	72.3 <sup>+</sup>	1163 <sup>**</sup>
Age/power plant effect	1.8 <sup>+</sup>	198964 <sup>+</sup>	1758.3 <sup>+</sup>	867 <sup>+</sup>	435.5 <sup>+</sup>	5488 <sup>*</sup>

441

## Appendices and Supplementary Materials

442

**Appendix 1.** Molecular ecology studies that report randomization in some part of the

443

work. All relevant articles were screened in a randomly selected issue of four journals for the

444

search term “random”. We deemed articles relevant when they used DNA or RNA methods

445

that are sensitive to laboratory biases (microsatellite genotyping, SNP assays,

446

metabarcoding, metagenomics, (meta)transcriptome comparisons, etc.). We also included

447

also studies that use single genes for molecular identification of species or populations (e.g.

448

barcoding and single-gene biogeographies) since identification may be non-randomly

449

confounded by cross-contamination (a simple example would be cross-contamination when

450

neighboring populations or related species are processed in batches). Randomization in

451

data analysis refers to the use of mixed effect models, the generation of null hypothesis by

452

random data rearrangements, etc.

	Mol. Ecol.	ISME J.	Soil Biol. Biochem.	Ecol. Evol.	J. Biogeogr.
Issue	22	2	1	9	11
Total relevant	14	16	8	9	12
Report randomization in sampling or data analysis	12	4	4	2	6
Report randomization in lab	1	0	0	1	0

486

**Supplementary Files** (accessible through Figshare,

487

<https://figshare.com/s/32dbca0a906c7f06449b>, DOI: 10.6084/m9.figshare.4579681):

488 sample\_infos.csv: the description of samples, negative and positive controls

489 multiplexing\_indices.xlsx: PCR plate setup and nucleotide indices used for sample

490 multiplexing

491 stechlin\_assigned\_190915.tab: OTU abundance table

492 Stechlin\_organohalogene.csv: organohalogene pesticide concentrations in the sediments

493 lab-methods\_OTU\_anova.RData: ANOVA table of the multispecies generalized linear model

494 (100 bootstraps)

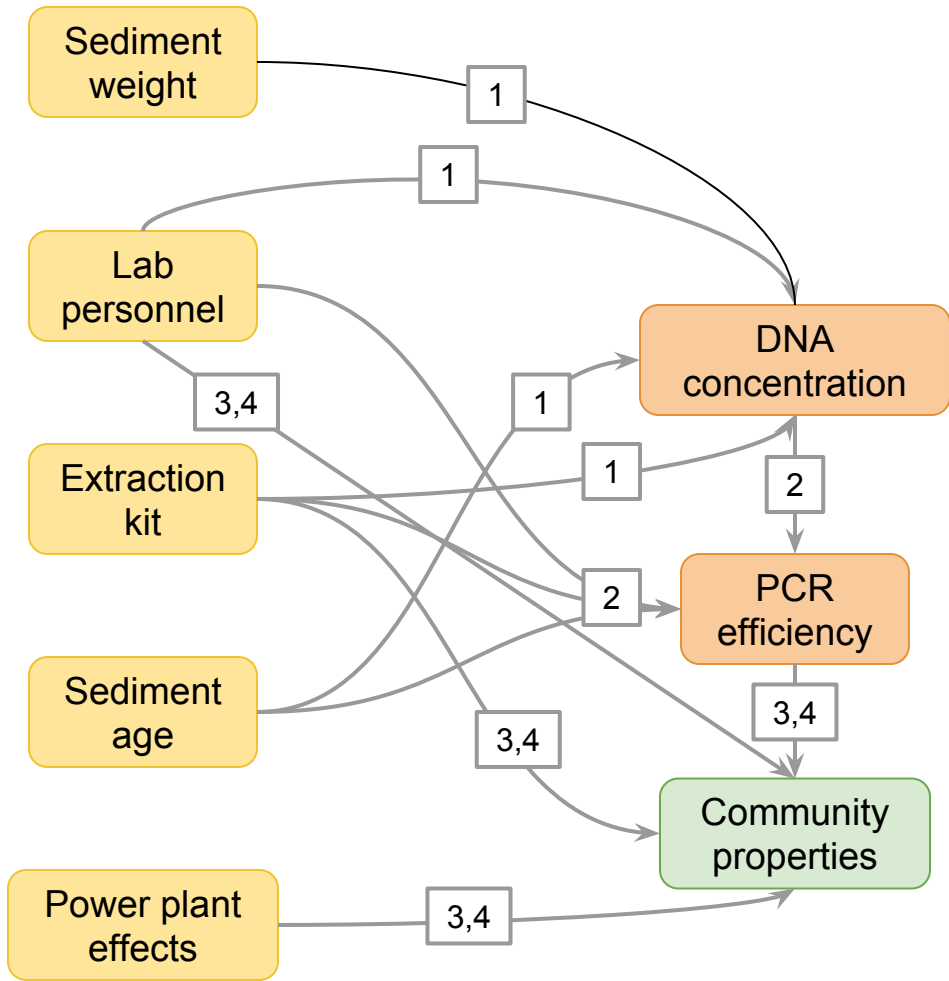
495 Lab\_LV\_model\_40000-iter.RData: ordination results with a latent variable model (40 000

496 iterations)

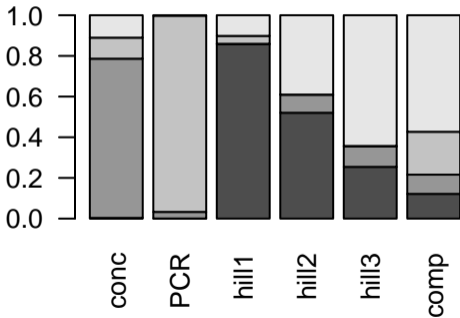


497 **Author contributions statement**

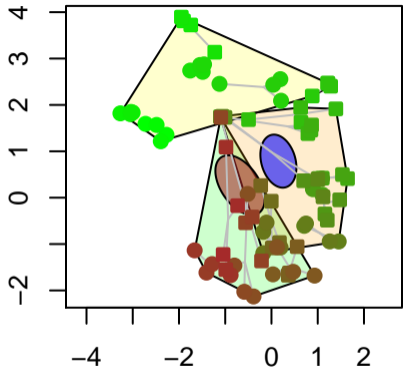
498 MB conceived the ideas. MB and HPG designed the methodology and obtained the  
499 cores. MB and OM performed the molecular laboratory work. MS and RAD performed the  
500 organohalogene measurements. MB processed the sequences, analysed the data and lead  
501 the writing of the manuscript. All authors contributed critically to the drafts and gave final  
502 approval for publication.



Biological signal      Expected bias  
Unexpected bias      Other factors

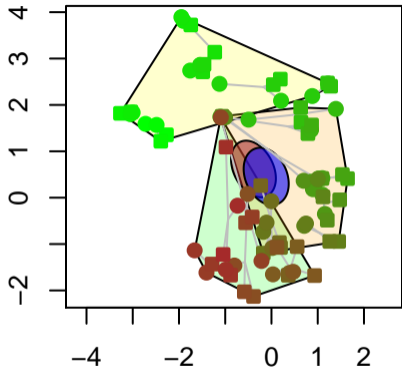


Community axis 2



Community axis 1

Community axis 2



Community axis 1