DR. LUCIE ZINGER (Orcid ID : 0000-0002-3400-5825)
DR. HOLLY BIK (Orcid ID : 0000-0002-4356-3837)
DR. ANTHONY A CHARITON (Orcid ID : 0000-0002-5809-3372)
DR. BRUCE E DEAGLE (Orcid ID : 0000-0001-7651-3687)
DR. IAN A. DICKIE (Orcid ID : 0000-0002-2740-2128)
DR. ALEX J. DUMBRELL (Orcid ID : 0000-0001-6282-3043)
DR. GENTILE FRANCESCO FRANCESCO FICETOLA (Orcid ID : 0000-0003-3414-5155)
DR. LUCA FUMAGALLI (Orcid ID : 0000-0002-6648-2570)
DR. SIMON N JARMAN (Orcid ID : 0000-0002-0792-9686)
DR. LEHO TEDERSOO (Orcid ID : 0000-0002-1635-1249)

Article type : Editorial

Corresponding author mail id: lucie@zinger.fr

Editorial note for Molecular Ecology

DNA metabarcoding - need for robust experimental designs to draw sound ecological conclusions

Authors

Lucie Zinger[1], Aurélie Bonin[2], Inger G. Alsos[3], Miklós Bálint[4,5], Holly Bik[6], Frédéric Boyer[2], Anthony A. Chariton[7], Simon Creer[8], Eric Coissac[2], Bruce E. Deagle[9], Marta De Barba[2], Ian A. Dickie[10], Alex J. Dumbrell[11], Gentile Francesco Ficetola[2,12], Noah Fierer[13,14], Luca Fumagalli[15], M. Thomas P. Gilbert[16,17], Simon Jarman[18,19], Ari Jumpponen[20], Håvard Kauserud[21], Ludovic Orlando[22,23], Johan Pansu[7,24,25], Jan Pawlowski[26,27], Leho Tedersoo[28], Philip Francis Thomsen[29], Eske Willerslev[23,30,31], Pierre Taberlet[2, 3]

Affiliations

[1] Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France.

[2] Université Grenoble Alpes, CNRS, Laboratoire d'Ecologie Alpine (LECA), F-38000 Grenoble, France.

[3] UiT – The Arctic University of Norway, Tromsø Museum, NO-9037 Tromsø, Norway.

[4] Senckenberg Biodiversity and Climate Research Centre, Senckenberganlage 25, 60325 Frankfurt am Main, Germany.

[5] LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Senckenberganlage 25, 60325 Frankfurt, Germany.

[6] Department of Nematology, University of California, Riverside, Riverside, CA, USA

[7] Department of Biological Sciences, Macquarie University, NSW 2113 Australia

[8] School of Natural Sciences, Bangor University, Gwynedd, LL57 2UW, UK

[9] Australian Antarctic Division, Kingston, Tasmania, Australia.

[10] BioProtection Research Centre, School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.

[11] School of Biological Sciences, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK

[12] Department of Environmental Science and Policy, Università degli Studi di Milano. Via Celoria 26, 20133 Milano, Italy.

[13] Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA.

[14] Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA.

[15] Laboratory for Conservation Biology, Department of Ecology and Evolution, University of Lausanne, Biophore building, CH-1015 Lausanne (Switzerland).

[16] Section for Evolutionary Genomics, Biological Institute, University of Copenhagen, Øster Farimagsgade 5, 1353 Copenhagen, Denmark

[17] Norwegian University of Science and Technology, University Museum, Trondheim, Norway.

[18] Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, WA, Australia.

[19] Environomics Future Science Platform, CSIRO National Collections and Marine Infrastructure, Crawley, WA, 6009, Australia.

[20] Division of Biology, Kansas State University, Manhattan, KS66506, USA.

[21] Section for Genetics and Evolutionary Biology (EVOGENE), University of Oslo, Blindernveien 31, 0316 Oslo, Norway.

[22] Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, CNRS UMR 5288, Université de Toulouse, Université Paul Sabatier, 31000 Toulouse, France.

[23] Lundbeck Foundation GeoGenetics Center, University of Copenhagen, Øster Voldgade 5-7, 1350K Copenhagen, Denmark.

[24] Station Biologique de Roscoff, UMR 7144 CNRS-Sorbonne Université, 29688 Roscoff, France.

[25] CSIRO Ocean & Atmosphere, Lucas Heights, NSW 2234, Australia.

[26] ID-Gene ecodiagnostics, Campus Biotech, Avenue Sécheron 15, 1202 Geneva, Switzerland.

[27] University of Geneva, Department of Genetics and Evolution, 1211 Geneva, Switzerland.

[28] Institute of Ecology and Earth Sciences, University of Tartu, 14a Ravila, 50411 Tartu, Estonia.

[29] Department of Bioscience, University of Aarhus, Ny Munkegade 116, DK-8000 Aarhus C, Denmark

[30] Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK.

[31] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

ORCID

Lucie Zinger : 0000-0002-3400-5825
Aurélie Bonin: 0000-0001-7800-8609
Inger G. Alsos: 0000-0002-8610-1085
Miklós Bálint: 0000-0003-0499-8536
Holly Bik: 0000-0002-4356-3837
Frédéric Boyer: 0000-0003-0021-9590
Anthony A. Chariton: 0000-0002-5809-3372
Simon Creer: 0000-0003-3124-3550
Eric Coissac: 0000-0001-7507-6729
Bruce E. Deagle: 0000-0001-7651-3687
Marta De Barba: 0000-0002-2979-3716
Ian A. Dickie: 0000-0002-2740-2128
Alex J. Dumbrell: 0000-0001-6282-3043
Gentile Francesco Ficetola: 0000-0003-3414-5155

Noah Fierer: 0000-0002-6432-4261

Luca Fumagalli: 0000-0002-6648-2570

M. Thomas P. Gilbert: 0000-0002-5805-7195

Simon Jarman: 0000-0002-0792-9686

Ari Jumpponen: 0000-0002-6770-2563

Håvard Kauserud: 0000-0003-2780-6090

Ludovic Orlando: 0000-0003-3936-1850

Johan Pansu: 0000-0003-0256-0258

Jan Pawlowski: 0000-0003-2421-388X

Leho Tedersoo: 0000-0002-1635-1249

Philip Francis Thomsen : 0000-0002-9867-4366

Eske Willerslev: 0000-0002-7081-6748

Pierre Taberlet: 0000-0002-3554-5954

Keywords: environmental DNA (eDNA), experimental controls, replication, data quality

DNA metabarcoding, especially when coupled with high-throughput DNA sequencing, is currently revolutionizing our capacity to assess biodiversity across a full range of taxa and habitats, from soil microbes (e.g. Thompson et al., 2017) to large marine fish (e.g. Thomsen et al., 2016), and from contemporary to tens of thousands year-old biological communities (e.g. Willerslev et al., 2003). The breadth of potential applications is immense and spans surveys on the diversity or diet of species native to specific ecosystems to bioindication (Pawlowski et al., 2018). The approach is also especially cost-effective and easy to implement, which makes DNA metabarcoding one of the tools of choice of the 21st century for fundamental research and the future of large-scale biodiversity monitoring programs (reviewed in Bohan et al., 2017; Creer et al., 2016; Taberlet, Bonin, Zinger, & Coissac, 2018; Thomsen & Willerslev, 2015). However, as is often the case with any emerging technology, we feel that the rise of DNA metabarcoding is occurring at a pace and in a manner that often loses sight of the challenges in producing high-quality and reproducible data (Baker, 2016). DNA metabarcoding is by essence a multidisciplinary approach building upon many complementary expertises, including field and theoretical knowledge, taxonomic expertise, molecular biology, bioinformatics, and computational statistics. Combining all these within single studies is necessary, not so much for producing and analyzing the data *per se,* but rather for minimizing and controlling the possible biases that can be introduced at any step of the experimental workflow - i.e. from the sampling to data analysis - and that can lead to spurious ecological conclusions (reviewed in Bálint et al., 2016; Nilsson et al., 2019; Dickie et al., 2018; Taberlet et al., 2018).

Whether the starting material consists of DNA from bulk samples (community DNA) and/or from environmental DNA (eDNA), all DNA metabarcoding studies rely on a deceptively simple succession of core experimental steps: (i) sampling and preservation of the starting material, (ii), DNA extraction, (iii) PCR amplification of a taxonomically-informative genomic region, (iv) high-throughput DNA sequencing of the amplicons, and (v) sequence analysis using bioinformatic pipelines. Despite this apparent simplicity, each step can potentially introduce its own sources of artifacts and biases (Figure 1). For example, the sampling design might not be effective for capturing the full taxonomic diversity or the ecological processes under study, an undesired bias for studies based on species detection. The availability of DNA in the samples is governed by its production rate, transport and persistence, processes which are all largely dependent on the targeted organisms, their biomass, and the ecosystem considered. A correct assessment of an ecological phenomenon based on DNA metabarcoding require not only implementation of standardized standardized, randomized and repeatable sampling designs and procedures (Dickie et al., 2018), but also consideration of DNA dynamics in the underlying matrix (i.e. in gut, feces, water or soil matrices from tropical or boreal organisms/ecosystems; Barnes & Turner, 2016). Likewise, the community under study can be enriched - on purpose or not - with specific taxa depending on how the sample is collected (e.g. filter size for water samples, removal of roots or not for soils), how it is transported/preserved, and how DNA is extracted (differential extraction efficiencies). PCR amplification is also well known to be an important source of biases, that are now fully revealed with high-throughput DNA sequencing techniques. The preferential amplification of certain taxa over other ones due to inappropriate primers provides one such example of potential bias (Clarke, Soubrier, Weyrich, & Cooper, 2014; Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014). Primer biases can both skew abundance profiles and lead to false negatives. PCR amplification can produce false negatives too through the presence of e.g. PCR inhibitors, but also many false positives through the introduction of replication errors by the DNA polymerase or the formation of chimeric fragments (reviewed in Taberlet et al., 2018). False positives can also be introduced at any step of the experimental workflow through the presence of reagent contaminants (Salter et al., 2014), or through samples, extractions or PCR cross-contaminations. An even more insidious source of false positives pertains to the occurrence of "tag jumps", sometimes referred to as "mistagging", "tag-switching", or "cross-talks" (Carlsen et al., 2012; Edgar, 2018; Esling, Lejzerowicz, & Pawlowski, 2015; Schnell, Bohmann, & Gilbert, 2015). PCR amplicons are indeed often tagged with unique short nucleotide sequences added on the 5'-end of the primers (i.e. "tags"), which allow pooling all PCRs within a single sequencing run and reducing sequencing costs. Each sequence obtained resulting in apparent cross-contaminations can then be bioinformatically assigned back to its sample of origin on the basis of its tag (Schnell et al., 2015). However, the procedures underlying the preparation of DNA libraries and/or the sequencing can introduce these "tag jumps", when the tag assigned to one particular sample is in fact recombined to the sequences belonging to another sample (Taberlet et al., 2018). This introduces additional, nonnegligible levels of sample cross-contaminations, which primarily involve the most abundant taxa and can have a disproportionate impact on samples with low DNA concentrations (Esling et al., 2015; Murray, Coghlan, & Bunce, 2015; Schnell et al., 2015). Similarly, the Illumina index located on the P5 sequencing adaptor can be subjected to “index jumps”, resulting in apparent cross-contaminations (Taberlet et al., 2018). This bias happens when several individual Illumina sequencing libraries are pooled and loaded on the same sequencing lane (Kircher, Sawyer, & Meyer, 2012) Finally, high-throughput DNA sequencing instruments have their own error rates (Schirmer et al., 2015). The above list of problems is clearly not exhaustive, and the interested reader will find more complete reviews elsewhere (e.g. Bálint et al., 2016; Nilsson et al., 2019; Taberlet et al., 2018). Still, it illustrates that any potential bias must be considered carefully when designing an experimental protocol and when interpreting the results. This is crucial to limit their impact on downstream analyses, and to ensure that the conclusion drawn from such data are authentic.

There is now an increasingly diverse range of field, laboratory (e.g. Caporaso et al., 2011; Taberlet et al., 2018; Valentini et al., 2009) and bioinformatics (e.g. Boyer et al., 2016; Caporaso et al., 2010; Dumbrell, Ferguson, & Clark, 2016) procedures aiming at reducing the amount of both false negatives (i.e. due to partial sampling, extraction, amplification or sequencing bias) and false positives (i.e. due to contaminations, “tag/index jumps”, or PCR and sequencing errors) in DNA metabarcoding experiments. However, using these protocols does not necessarily guarantee that the problem of false positives or negatives is completely under control. These protocols must continuously be reconsidered, especially alongside the emergence of novel DNA sequencing technologies that provide new opportunities, but also new challenges. Additionally, each individual study and each genomic marker comes with its own specificities, and this often requires customization of the above protocols. The sequence clustering threshold to be used to form molecular taxonomic units relevant to the question addressed (e.g. removing intraspecific marker variability when the species level is desired) provides such an example, and will critically depend on both the marker specificities and PCR/sequencing error rates. Bioinformatics tools can further fail to exclude molecular artifacts when the filtering thresholds are relaxed, which inflates sample diversity estimates. Likewise, they can also generate false negatives, for example when a genuine metabarcode is falsely flagged as an error or chimera, or when it is assigned to an incorrect taxon due to incomplete or inappropriate reference databases (Alsos et al., 2018; Coissac, Riaz, & Puillandre, 2012). This can be especially problematic when the question investigated strongly relies on species detection. It is therefore crucial to include several types of experimental controls so as to facilitate the exclusion of spurious signal and support the reliability of the biological conclusions (Figure 1). Amongst these controls, conducting pilot experiments is particularly helpful to assess how appropriate the sampling design is (Dickie et al., 2018). We also recommend that both biological replicates (i.e. multiple independent samples) and technical replicates (i.e. multiple extractions/PCR of the same sample and/or extract) are included in the experimental workflow to disentangle the effect of both the biological and technical variances (Ficetola et al., 2015). These replications are necessary because both sampling and PCR can introduce biases in a stochastic manner, especially when the concentration of the target DNA is low. It is also essential to analyze a sufficient number of negative controls at the field sampling, DNA extraction, PCR, and sequencing steps, as well as positive controls consisting of mock communities, known DNA samples, or even synthetic sequences reflecting the attributes of the targeted products (Figure 1). All these controls must be sequenced along the biological samples, as they facilitate the detection of sporadic contaminations and tag or index jumps while helping adjusting filtering and clustering thresholds. Ultimately, they will be a token of the reliability of the whole data curation process (De Barba et al., 2014). We also encourage careful consideration of the bioinformatics workflow itself, since the filtering steps necessary to curate the data will critically depend on the experimental design and the ecological question under study. Typically, sequences of low abundance in a given sample may be genuine or artifacts deriving from PCR/sequencing errors or tag/index jumps. The retained filtering threshold for taxon presence is thus dependent on the underlying rates of artifacts, as well as on the sequencing depth. As the different experimental controls provide direct measurements of these artifacts, they will therefore allow better tuning of the filtering thresholds. All of these technical considerations should be precisely reported within publications together with relevant illustrations and statistics characterizing the workflow, as they are necessary to assess the relevance and quality of the data underpinning specific conclusions. A last, a most obvious example of control consists in assessing the plausibility of the taxonomic composition based on *a priori* knowledge of the system or taxa studied. Such knowledge can be derived from data obtained with complementary sensing approaches such as visual observations. In this case, building exhaustive local reference databases of the genomic marker used from local specimens will secure the taxonomic assignment step (e.g. Alsos et al., 2018). When local information is unavailable, typically when studying microorganisms, it remains possible to assess whether the community is composed of clades that are expected to occur in the system surveyed or not, as e.g. soils, sediments, and gut environments harbour highly different bacterial phyla (e.g. Thompson et al., 2017).

As users, readers, referees or editors, we realize that the above-mentioned issues remain too often overlooked. This problematic stance can lead to unsubstantiated claims and undermine scientific advances if not resolved. Inappropriate practices such as estimating species richness from fingerprint profiles (Bent, Pierson, & Forney, 2007), the absence of biological replicates (Prosser, 2010), or that of contaminant controls (Perez-Muñoz, Arrieta, Ramer-Tait, & Walter, 2017) have been repeatedly criticized in the field of microbial ecology, and in the latter case, they contribute to the rising debate about the existence or not of a womb microbiota. Ancient DNA research has also developed rigorous standards to tackle issues related to contamination, sequencing errors, and data reproducibility (Poinar and Cooper (2000). We believe that the community of DNA metabarcoding users has now come of age and learnt from its past errors. At a time when more and more exhaustive guides of best practices on the subject are emerging (Knight et al., 2018; Pollock, Glendinning, Wisedchanwet, & Watson, 2018; Taberlet et al., 2018), and where DNA sequencing costs are rapidly decreasing, we should be always mindful of the adage "better safe than sorry". This note does not mean to imply that the systematic use of the highest technical and analytical standards is reasonable nor the universal remedy for all the challenges associated with DNA metabarcoding. Rather, we strongly encourage researchers and end-users to adopt reflective decision-making when designing their experiment and to critically appraise their results, with the ultimate aim to prove the robustness and reproducibility of their conclusions.

References

Alsos, I. G., Lammers, Y., Yoccoz, N. G., Jørgensen, T., Sjögren, P., Gielly, L., & Edwards, M. E. (2018). Plant DNA metabarcoding of lake sediments: How does it represent the contemporary vegetation. *PLoS One*, *13*(4), e0195403.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*(7604), 452.

Bálint, M., Bahram, M., Eren, A. M., Faust, K., Fuhrman, J. A., Lindahl, B., … Tedersoo, L. (2016). Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. *FEMS Microbiology Reviews*, *40*(5), 686–700.

Barnes, M. A., & Turner, C. R. (2016). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, *17*(1), 1–17.

Bent, S. J., Pierson, J. D., & Forney, L. J. (2007). Measuring species richness based on microbial community fingerprints: the emperor has no clothes. *Applied and Environmental Microbiology*, *73*(7), 2399.

Bohan, D. A., Vacher, C., Tamaddoni-Nezhad, A., Raybould, A., Dumbrell, A. J., & Woodward, G. (2017). Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends in Ecology & Evolution*, *32*(7), 477–487.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: a unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., … Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl 1*, 4516–4522.

Carlsen, T., Aas, A. B., Lindner, D., Vrålstad, T., Schumacher, T., & Kauserud, H. (2012). Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecology*, *5*(6), 747–749.

Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. (2014). Environmental metabarcodes for insects: *in silico* PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, *14*(6), 1160–1170.

Coissac, E., Riaz, T., & Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, *21*, 1834–1847.

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., … Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, *7*(9), 1008–1018.

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, *10*(9), 20140562.

De Barba, M., Miquel, C., Boyer, F., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing for omnivorous diet analysis and validation of data accuracy. *Molecular Ecology Resources*, *14*(2), 306–323.

Dickie, I. A., Boyer, S., Buckley, H. L., Duncan, R. P., Gardner, P. P., Hogg, I. D., … Weaver, L. (2018). Towards robust and repeatable sampling methods in eDNA-based studies. *Molecular Ecology Resources*, *18*(5), 940–952.

Dumbrell, A. J., Ferguson, R. M. W., & Clark, D. R. (2016). Microbial community analysis by single-amplicon high-throughput next generation sequencing: data analysis – from raw output to ecology. In T. J. McGenity, K. N. Timmis, & B. Nogales (Eds.), *Hydrocarbon and lipid microbiology protocols*. Humana Press.

Edgar, R. (2018). *UNCROSS2: identification of cross-talk in 16S rRNA OTU tables*. *bioRxiv*. doi:10.1101/400762

Esling, P., Lejzerowicz, F., & Pawlowski, J. (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*, *43*(5), 2513–2524.

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., … Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, *15*(3), 543–556.

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, *40*. doi:10.1093/nar/gkr771

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews. Microbiology*, *16*(7), 410–422.

Murray, D. C., Coghlan, M. L., & Bunce, M. (2015). From benchtop to desktop: important considerations when designing amplicon sequencing workflows. *PLoS One*, *10*(4), e0124671.

Nilsson, R. H., Anslan, S., Bahram, M., Wurzbacher, C., Baldrian, P., & Tedersoo, L. (2019). Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nature Reviews. Microbiology*, *17*(2), 95–109.

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., … Kahlert, M. (2018). The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *The Science of the Total Environment*, *637-638*, 1295–1310.

Perez-Muñoz, M. E., Arrieta, M.-C., Ramer-Tait, A. E., & Walter, J. (2017). A critical assessment of the ``sterile womb'' and ``in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. *Microbiome*, *5*(1), 48.

Poinar, H. N., & Cooper, A. (2000). Ancient DNA: do it right or not at all. *Science*, *5482*, 1139.

Pollock, J., Glendinning, L., Wisedchanwet, T., & Watson, M. (2018). The madness of microbiome: attempting to find consensus “best practice” for 16S microbiome studies. *Applied and Environmental Microbiology*, *84*(7), e02627–17.

Prosser, J. I. (2010). Replicate or lie. *Environmental Microbiology*, *12*(7), 1806–1810.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., … Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*, 87.

Schirmer, M., Ijaz, U. Z., D’Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, *43*(6), e37.

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*(6), 1289–1303.

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA for biodiversity research and monitoring*. Oxford University Press.

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., … Earth Microbiome Project Consortium. (2017). A communal catalogue reveals Earth’s multiscale microbial diversity. *Nature*, *551*(7681), 457–463.

Thomsen, P. F., Møller, P. R., Sigsgaard, E. E., Knudsen, S. W., Jørgensen, O. A., & Willerslev, E. (2016). Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. *PLoS One*, *11*(11), e0165252.

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA as an emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, *183*(C), 4–18.

Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E., Coissac, E., Pompanon, F., … Taberlet, P. (2009). New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trn*L approach. *Molecular Ecology Resources*, *9*, 51–60.

Willerslev, E., Hansen, A. J., Binladen, J., Brand, T. B., Gilbert, M. T. P., Shapiro, B., … Cooper, A. (2003). Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, *300*, 791–795.

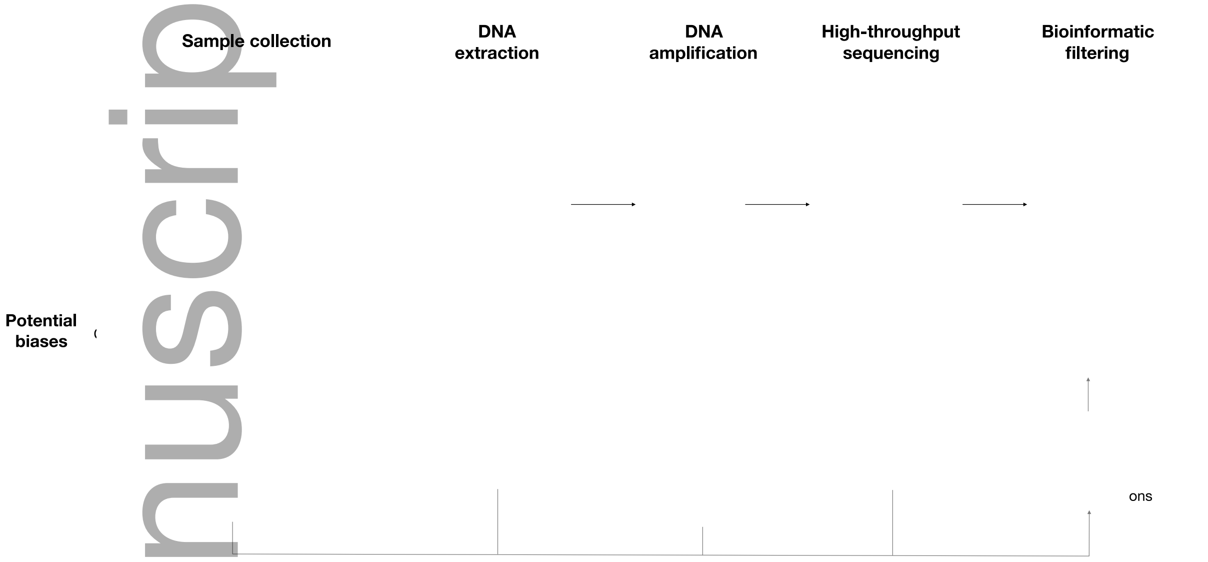

Figure 1 Summarized workflow of DNA metabarcoding and biases in the data production process, with the potential associated controls to assess data quality. Expectations on the local community, either from a priori knowledge on the site or organisms targeted, or obtained through e.g. vizual census, specimen collection, or building of a local reference database, constitute a first assessment of the DNA metabarcoding experiment success. Pilot experiments are essential for optimizing the whole experimental design, from the sampling strategy (e.g. number of biological replicates) to the entire technical approach. Field, extraction, PCR, and tagging-system negative and positive controls should be sequenced along with biological samples. They all aim at identifying (i) potential contaminants that could be introduced at any experimental step, and (ii) potential experimental artifacts due to the DNA extraction, PCR, and sequencing steps. Field negative controls consists of extracting DNA from storage/extraction buffers brought to the field or used to clean sampling instruments. Tagging-system negative controls can only be implemented when amplicons are identified by a unique combination of tags attached to the 5' end of each amplification primer, and where one or several tag combinations remain unused in the experimental design. In such conditions, tagging-system controls can be performed at the bioinformatics analysis step, by monitoring the number of sequences harboring unexpected tag combinations. This number is actually a direct measurement of the tag-jump rate. “Index jumps” are more difficult to evaluate, and can be controlled either by indexing both library adapters (P5 and P7) or when the libraries sequenced together have

identifiable sequences that could indicate their origin. The positive controls (constructed using either synthetic DNA with the primer target sequences on both sides, DNA extracted from a mock community, or known environmental samples), as well as prior expectations on the taxa that should occur in the system can be used to evaluate the effectiveness of the data production process, the impact of contaminants on the retrieved ecological signal and the adequacy of bioinformatics filtering procedures.

DNA dynamics
(~season, system, organism)
reagents/aerosols contaminants
N

Sample collection

DNA extraction

DNA amplification

High-throughput sequencing

Bioinformatic filtering

Potential biases

Undersampling
Contamination from past/neighbouring events
Experimental contamination

Undersampling
Taxon-specific inefficiency
Experimental contamination

+ Polymerase errors/chimeras
+ Inappropriate primers

+ Tag/index jump &
sequencing errors

Inappropriate filtering thresholds
Mis-classifications

Potential controls

- Expected target (or non-target) taxa
- Building of a local reference database
- Pilot experiment
- Biological replicates
- Field negative controls

- Technical replicates
- Extraction negative controls
- Positive controls

- Technical replicates
- PCR negative controls
- Positive controls
- Use of multiple primer set or *in silico* pre-evaluation of primers

- Tagging system negative controls

- Filtering/clustering criteria and threshold adjustments based on all controls and replicates
- Taxonomic congruence with *a priori* expectations