Inline index helped in cleaning up data contamination generated during library preparation and the subsequent steps

Ying Wang1,2 · Hao Yuan1,2 · Junman Huang1,2 · Chenhong Li1,2

Received: 23 August 2021 / Accepted: 23 September 2021 / Published online: 29 October 2021
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract
Background High-throughput sequencing involves library preparation and amplification steps, which may induce contamination across samples or between samples and the environment.
Methods We tested the effect of applying an inline-index strategy, in which DNA indices of 6 bp were added to both ends of the inserts at the ligation step of library prep for resolving the data contamination problem.
Results Our results showed that the contamination ranged from 0.29 to 1.25% in one experiment and from 0.83 to 27.01% in the other. We also found that contamination could be environmental or from reagents besides cross-contamination between samples.
Conclusions Inline-index method is a useful experimental design to clean up the data and address the contamination problem which has been plaguing high-throughput sequencing data in many applications.

Keywords Contamination · Cross-contamination · Indexing · Next-generation sequencing · Environmental DNA · Phylogenomics · Meta genomics · Multiplexed sequencing

Introduction
High-throughput sequencing has become a dominant and efficient data collection method for researches, such as phylogenomics, environmental DNA, and ancient DNA analyses. However, contamination from the environment or cross-contamination from other samples has become a headache in many studies involving high-throughput sequencing [1–5]. Contaminated data may deeply bias results of phylogenetic analysis [2, 6], over-estimate branch length and affect orthology detection [7]. Contamination also can confound metagenomic studies [1]. For example, distinct microbial community of placenta was proven to be an artifact due to contamination [8]. In ancient DNA analysis, pollution of modern human DNA can bias the estimates [9], whereas unexpected amplification of negative control samples was often reported in eDNA analyses, due to the complex nature of the eDNA samples [10–12].

Contamination can originate from environmental sources, such as extraction kits, plastic consumables and reagents, or accidentally transferred from other samples of the same batch [1, 13]. Synthesizing barcodes or outsourcing sequencing can also result in cross-contamination which is beyond the control of the researcher’s lab [14]. Many steps of washing and amplification are involved when preparing library for high-throughput sequencing which may increase the risk of contamination. If further manipulation of the samples is applied, such as target-gene enrichment, higher chance of contamination often occur.

Strict decontamination protocols and guidelines for preventing contamination in eDNA experiment have been proposed [15]. A standard checklist also was suggested for metagenomic lab for avoiding contamination [1]. Furthermore, various methods were designed for detecting contamination [14, 16] and different pipelines and tools were created to clean up the contaminated data [10, 13, 14, 16–19]. Nevertheless, these methods are either using precautions to reduce the chance of contamination or identifying
contaminator posteriorly though comparing the reads to references or other samples.

Rohland et al. designed incomplete adapters composing P5/P7 barcodes (inline index) and partial Illumina adapters to increase efficacy of target enrichment [20]. They also suggested that the P5/P7 barcodes can be used to trace cross-contamination, however, they did not test this idea. The subsequent research using their method also focused on the uracil-DNA-glycosylase treatment rather than resolving contamination issue [21, 22]. Here, we applied the “inline-index” strategy on two representative datasets, one from a phylogenomic study and the other for environment DNA data. Our objective is to test with empirical data on how the inline-index strategy can be used to solve the problem of cross-contamination.

Materials and methods

Designing and synthesizing the inline indices

Indices of 6 bp were designed using scripts downloaded from the website (https://bioinf.eva.mpg.de/multiplex/, Accessed on May 2018). The following filters were applied: (1) editing distance was three, i.e., at least three substitutions should occur for one index changing to another, so to avoid index swap due to amplification or sequencing errors; (2) no polymer index was used, so to distinguish index signal from dust or chemical particles; (3) exclude the following sequences, ‘AAA’, ‘ACA’, ‘CCC’, ‘CAC’, ‘GAC’, ‘GGG’, ‘GTG’, ‘TTT’, and ‘TGT’ to avoid using the nucleotides that can be illuminated with the same laser consecutively. The inline indices were added to the 3’ end of the IS1 or IS2 sequences of Meyer and Kircher [23]. Reverse and complementary oligos to the inline indices were added to the 5’ of their IS3 sequences (Fig. 1A). Twenty-four pairs of oligos (IS1_IndL and IS3_IndL’) with inline indices were designed and synthesized for making P5 adapter and 24 pairs (IS2_IndR and IS3_IndR’) were done for P7 adapter (Table 1). The inline indices were synthesized at Sangon Biotech (Shanghai, China). Protocol of Meyer and Kircher [23] was followed to mix the oligos to make P5 and P7 adapters. It should be noticed that each IS1_IndL and IS2_IndR should be paired with its corresponding reverse and complementary IS3. Extreme precaution should be taken to avoid contamination during index preparation. The prepared indices can be aliquoted to small portions and stored under −20 °C.

Fig. 1 Scheme of inline index and library preparation. A Scheme of inline index; B Steps of library preparation

© Springer
| TCTGCC IS1_Ind1 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind1 | ggagacAGATCCGAAA*G*A*G*C |
| GTCTCT IS1_Ind2 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind2 | ggaacAGATCCGAAA*G*A*G*C |
| ATATTG IS1_Ind3 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind3 | caatagATCCGAAA*G*A*G*C |
| TGGAGG IS1_Ind4 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind4 | etccagATCCGAAA*G*A*G*C |
| TCTAGT IS1_Ind5 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind5 | acctagATCCGAAA*G*A*G*C |
| AGAGTA IS1_Ind6 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind6 | tctctagATCCGAAA*G*A*G*C |
| GGCACA IS1_Ind7 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind7 | tgggcaAGATCCGAAA*G*A*G*C |
| TATCCT IS1_Ind8 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind8 | gagataAGATCCGAAA*G*A*G*C |
|TTATGC IS1_Ind9 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind9 | geataaAGATCCGAAA*G*A*G*C |
| AGTGGG IS1_Ind10 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind10 | ccacctagATCCGAAA*G*A*G*C |
| GTCAAG IS1_Ind11 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind11 | cttcagATCCGAAA*G*A*G*C |
| CAGCAA IS1_Ind12 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind12 | ttgcctagATCCGAAA*G*A*G*C |
| TCGCCG IS1_Ind13 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind13 | cgcctagATCCGAAA*G*A*G*C |
| CTAAGA IS1_Ind14 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind14 | tctctagATCCGAAA*G*A*G*C |
| CCGCTT IS1_Ind15 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind15 | aagcctagATCCGAAA*G*A*G*C |
| AATGTA IS1_Ind16 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind16 | taacctagATCCGAAA*G*A*G*C |
| GTTACC IS1_Ind17 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind17 | ggtacctagATCCGAAA*G*A*G*C |
| CCAGGT IS1_Ind18 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind18 | acctgctagATCCGAAA*G*A*G*C |
| AATCGA IS1_Ind19 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind19 | tctctagATCCGAAA*G*A*G*C |
| AACCA IS1_Ind20 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind20 | tcgcctagATCCGAAA*G*A*G*C |
| GACGAC IS1_Ind21 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind21 | gtcctagATCCGAAA*G*A*G*C |
| CGCGTT IS1_Ind22 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind22 | acgcctagATCCGAAA*G*A*G*C |
| CCGTAC IS1_Ind23 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind23 | tctcctagATCCGAAA*G*A*G*C |
| GTAACG IS1_Ind24 | A*C*A*C*TCTTTTCTTATACACAGCGCTTCTCCAGACTCtt*g*t*g*c*c IS3_Ind24 | gattacagATCCGAAA*G*A*G*C |
| IS2_IndR IS2_Ind25 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind25 | aagtcctagATCCGAAA*G*A*G*C |
| TCATAA IS2_Ind26 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind26 | ttgatctagATCCGAAA*G*A*G*C |
| CAAGAG IS2_Ind27 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind27 | ctcctgctagATCCGAAA*G*A*G*C |
| CGATCA IS2_Ind28 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind28 | tgatctagATCCGAAA*G*A*G*C |
| TTGATT IS2_Ind29 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind29 | aatcctagATCCGAAA*G*A*G*C |
| TCCGAG IS2_Ind30 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind30 | ctgcgttagATCCGAAA*G*A*G*C |
| CCTGAA IS2_Ind31 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind31 | ttcgtagATCCGAAA*G*A*G*C |
| ATTCCT IS2_Ind32 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind32 | aagactgctagATCCGAAA*G*A*G*C |
| GCGACT IS2_Ind33 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind33 | agtcctagATCCGAAA*G*A*G*C |
| GGCTTC IS2_Ind34 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind34 | gtcctagATCCGAAA*G*A*G*C |
| AATACG IS2_Ind35 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind35 | gatcctagATCCGAAA*G*A*G*C |
| TACGCT IS2_Ind36 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind36 | acctgatctagATCCGAAA*G*A*G*C |
| ACCGTC IS2_Ind37 | G*T*G*A*CTGGAGTTCAAGACTGTTGCTTCTCCAGATCT tc*t*g*t*t IS3_Ind37 | gctccctagATCCGAAA*G*A*G*C |
Library preparation and sequencing

Two batches of sequencing libraries were prepared. One was for a phylogenomic study of eight chondrichthyan species, *Himantura uarnak*, *Chiloscyllium griseum*, *Rhinobatos annandalei*, *Tetronarce nobiliana*, *Rhina ancylostoma*, *Neoharriotta pinnata*, *Pristiophorus cirratus*, and *Heterodontus francisci*. The genomic DNA was extracted from tissue samples using Ezup DNA extraction kits (B518251, Sangon Biotech, Shanghai, China). The other batch of samples was 11 eDNA samples collected from the Red River, Yunnan, China. The water samples were filtered with 3.0 µm PCTE membrane on site and DNA was extracted following Lin et al. [24]. The extracted DNA samples were sheared with a Covaries M220 focused-ultrasonicator to ~250 bp (Covaris Inc., Woburn, Massachusetts). The protocol of Meyer and Kircher [23] was followed to prepare the sequencing libraries. For each sample, a unique pair of P5 and P7 adapters with inline index was added, while the P7 adapters also have their own classic P7 index (Fig. 1B). Two different persons prepared the two batches of samples to represent handling variation in the lab. The libraries were sequenced on an Illumina HiSeqX 10 platform by Genewiz (Genewiz, Inc, Shanghai, China). We chose these two batches of samples to represent the common high-throughput sequencing applications where contamination has been problematic.

Data analyses

The raw reads were trimmed with trim_galore v0.6.4 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to remove the adapter sequences and low quality reads. The duplicated reads due to PCR amplification were excluded using the “-fastx_uniques” command in USEARCH v10.0.240 [25]. A custom Perl script (demultiplex_inline.pl, supplementary files) was used to extract the first 6 bp of the sequences, compared to the sequences of the inline index pairs used for samples. The reads with both end matching the inline index pairs were considered as “true reads” and the rest reads were considered as “others”. The “others” were further counted and separated into four parts: (1) “cross-contamination” (contamination_cnt.pl, supplementary files), DNA fragments had inline index but instead of the correct one, they were used in other samples in the same batch; (2) “inline index mutation” (kmer_summary_1sub.py, supplementary files), the inline index sequences had one mutation to the correct one; (3) “inline index contamination” (kmer_summary.py, supplementary files), the sequences matched inline index used in our lab but not for the same batch of samples; (4) “unknown”, all the rest reads cannot trace back to a source.
Results

Composition of the reads

The “true reads” that have correct inline index at both ends accounted for 38.16 to 80.04% of the total reads of each sample in the eDNA experiment (Fig. 2A). Samples 2 and 3 had substantial amount of contamination, probably due to operational errors dealing with these two samples. True reads consisted of 87.81 to 92.36% of total reads of each sample in the chondrichthyan study (Fig. 3A).

Correction of reads with index mutation

Because reads with inline index mutation accounted for most part of the “other” reads, we corrected the index sequences with one mutation to their closet indices used in our lab. After correction, the true reads increased to 48.67 to 90.51% for the eDNA experiment (Fig. 2B) and 91.27 to 95.63% for the chondrichthyan experiment (Fig. 3B).

Cross-contamination

Contamination from other samples accounted for 0.83–27.01% in the eDNA experiment and 0.29–1.25% in

---

**Fig. 2** Percentage of reads assigned correctly in the eDNA experiment. A before correction on reads with inline index mutation; B after correction on reads with inline index mutation

**Fig. 3** Percentage of reads assigned correctly in the chondrichthyan experiment. A before correction on reads with inline index mutation; B after correction on reads with inline index mutation

**Fig. 4** Composition of the others reads after correction of inline index with one substitution. A data from the eDNA experiment; B data from the chondrichthyan experiment
the phylogenomic study of chondrichthyan (Fig. 4). The results showed that cross-contamination ubiquitously exists no matter who run the experiment and whether the tubes were close to each other or not on the rack, so the contaminator did not necessarily jump from adjacent tubes (Fig. 5).

**Inline index contamination**

Some reads had at least one inline index that was not used in the same batch of samples but can be found in other inline indices used in our lab. Those could be due to cross-contamination between indices when they were synthesized or contamination through common reagents. The eDNA samples had 1.21 to 12.1% contamination due to inline index mixing, whereas the chondrichthyan samples had 1.16 to 2.71% (Fig. 4).

**Unknown contamination**

Besides the cross-contamination among samples and inline index contamination, there were a large portion of reads with unknown sources, which accounted for 6.9 to 20.3% in the eDNA samples and 4.07% to 7.47% in the chondrichthyan samples (Figs. 3 and 4). The unknown contamination was randomly distributed among samples.

**Discussion**

The results of “true reads” and “others” showed that the baseline contamination in our lab is around 10% to 20%, but an accidental handling can increase the number to almost 40%. Allio et al. [7] found very low cross-contamination (0.26%) in general in their genome shotgun sequencing data of swallowtail butterfly, but one sample, *Parnassius imperator* had a significant higher value, 26.71% of the contigs were contaminated, suggesting that human errors can have sporadic but severe impact on high-throughput sequencing data.

We found that the “others” had similar composition between the two experiments. If we disregard samples 2 and 3 of the eDNA experiment, the composition of the “others”, ranked from high to low was “inline index mutation”, the “unknown”, “index contamination”, and “cross-contamination”.

Because we applied a substitution distance of three base pairs, it should be safe to assign the reads with one base-pair mutation in their inline indices back to the “correct” sample. Indeed, we found few inline index had two substitutions. The index mutation could be caused by two reasons. One is sequencing error, which should be minimal because of the low error rate in sequencing the first six base pairs. The other error could happen while synthesizing the inline indices, so we should test the accuracy of synthesized indices from different provider in future.

The results of contaminated reads showed that the source of contamination was from sample 10, 7 and 1 in the eDNA experiment and from sample 8, 7 and 1 in the chondrichthyan study. The degree of cross-contamination also was not uniform. For example, samples 2 and 3 of the eDNA experiment had 24.93% and 27.01% contamination mainly from sample 10, suggesting handling errors may be involved. The route of cross-contamination may be through tube lids, gloves, pipette tips or aerosol.

The eDNA sample 7 had 12.1% inline index contamination, and mostly from one pair of inline indices, suggesting those may be contaminated by other library, prepared in our lab. We suggest that the index should be synthesized once at a time, so inline index contamination could be reduced. Extreme care should be taken when making the adapters in the lab, and the adapters should be divided into small aliquot for storage.

There was at least one unknown “inline index” sequenced for those reads. Our lab also has been using classic protocol of library prep [23], which has no inline index applied. We
suspect those libraries that made with the classic protocol may have polluted common reagents, which in turn contaminated the samples of this project. Therefore, the first 6 bp base pairs of the reads cannot match any inline indices.

In the “others”, reads with one mutation in the index can be corrected into its original sample, but reads with mixed inline index or unknown index sequence cannot be assigned to any sample and should be excluded from further analyses.

Most analytical approaches are based on data filtering. For example, Croco [26] is a database independent method that can be used to trace cross-contamination from divergent species, but difference between closely related organisms cannot be recognized. ConFindr [17] and be used to identifies contaminated samples, if it contains more than one allele of core single-copy ribosomal protein genes. Dickins et al. [14] proposed a two-part pipeline to identify the contaminated samples based on unexpected number of variants and a phylogenetic approach.

Inline-index method, however, is independent on the sequence of samples, so there is no requirement on the similarity between samples to deduce the composition of the reads. The reads can be recognized using inline index even the sample contain DNA from unknown species, such as the eDNA data, but the analytical methods may not work in this situation. Kircher et al. [27] designed a double-index technique to detect jumping PCR by adding indices to both P5 and P7 of Illumina adapters. Peyrégne et al. [28] implied that the double-index method may be used to monitor contamination. However, the double-index are added before sequencing, there are a lot of chances that contamination can happened before that. Rohland et al. [20] invented the inline-index method that adding a pair of unique barcodes to both ends of the DNA insert to trace the contamination. Because they aimed at ancient DNA research and paid most of their attention to the effects of unique barcode on damage rate, but did not test whether the inline-index approach can be used to mitigate problems of cross-contamination. Our research results showed that contamination occurred ubiquitously and the unique inline barcode can be used to trace the source of read contamination.

Furthermore, because samples are labeled with inline-index before sending them to sequencing facility, cross-contamination resulted in sequencing center can be controlled. Cross-contaminated reads from other samples can be assigned back to their origin based on their inline-index pair to rescue the data. Finally, with 24 pairs of inline indices, plus the P7 index, more samples can be multiplexed and sequenced in the same sequencing lane to save the cost.

Conclusion

The inline-index method is experimental, so it could be a good complementary to the analytical approach, such as CroCo and ConFindr. The widely use of the combination of these methods should be a routine in near future. Environment DNA samples have a nature of mixed source of DNA with various concentrations, so it is impossible to know if the reads are genuine or from other samples by using the computational approach. In this scenario, inline index method or other experimental method probably is the only way to address the contamination problem. In future, we can test whether the “blunt end” and “ligation” step of library prep can be combined together without washing steps in between, so to add the inline index at the very first step of library prep to further eliminate the chance of cross-contamination.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11033-021-06884-y.

Acknowledgements We are grateful to Ms. Lifang Peng for helping with collecting water samples.

Author contributions YW and CL conceived the research plan. YW and JH did the experiments and collected data. YW, HY and JH analyzed the data. YW and CL wrote the draft. All authors edited and approved the final version of the manuscript.

Funding This work was supported by “Science and Technology Commission of Shanghai Municipality (19050501900)” and “Shanghai Academy of Environmental Sciences”.

Data availability The raw reads have been uploaded to NCBI under Bioproject Numbers PRJNA749868, PRJNA750567.

Declarations

Conflict of interest CL, YW and HY are authors on a patent applied for by Shanghai Ocean University that covers the inline index technology (201811406204.X), and the remaining author (JH) has no financial conflict of interest.

Ethical approval We followed the guidelines approved by the Ethical Committee of Shanghai Ocean University, China.

Consent to participate All authors consent to participate.

Consent for publication All authors consent for publication.
1. Eisenhofer R et al (2019) Contamination in low microbial biomass microbiome studies: issues and recommendations. Trends Microbiol 27(2):105–117
2. Laurin-Lemay S, Brinkmann H, Philippe H (2012) Origin of land plants revisited in the light of sequence contamination and missing data. Curr Ecol Evol. https://doi.org/10.3389/fevo.2020.609973
3. Sepulveda AJ et al (2020) The elephant in the lab (and Field): contamination in aquatic environmental DNA studies. Front Ecol Evol. https://doi.org/10.3389/fevo.2020.609973
4. Ballenghien M, Faivre N, Galtier N (2017) Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. BMC Biol 15(1):25
5. Koutsovoulos G et al (2016) No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. Proc Natl Acad Sci 113(18):5053–5058
6. Philippe H et al (2011) Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 9(3):e1000602
7. Allio R et al (2020) Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. Syst Biol 69(1):38–60
8. Perez-Muñoz ME et al (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
9. Racimo F, Renaud G, Slatkin M (2016) Joint estimation of contamination, error and demography for nuclear DNA from ancient humans. PLoS Genet 12(4):e1005972
10. Hänfling B et al (2016) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. Mol Ecol 25(13):3101–3119
11. Sepulveda AJ et al (2019) Adding invasive species biosurveillance to the pioneer infant microbiome. Microbiome 5(1):48
12. Serrao NR, Reid SM, Wilson CC (2018) Establishing detection thresholds for environmental DNA using receiver operator characteristic (ROC) curves. Conserv Genet Resour 10(3):555–562
13. Glassing A et al (2016) Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathogens 8(1):24
14. Dickins B et al (2014) Controlling for contamination in resequencing studies with a reproducible web-based phylogenetic approach. Biotechniques 56(3):134–141
15. Goldberg CS et al (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species. Methods Ecol Evol 7(11):1299–1307
16. De Simone G et al (2020) Contaminations in (meta) genome data: an open issue for the scientific community. IUBMB Life 72(4):698–705
17. Low AJ et al (2019) ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. PeerJ 7:e6995
18. Straube N et al (2018) A phylogenomic approach to reconstruct interrelationships of main clupeocephalan lineages with a critical discussion of morphological apomorphies. BMC Evol Biol 18(1):158
19. Flickinger M et al (2015) Correcting for sample contamination in genotype calling of DNA sequence data. Am J Hum Genet 97(2):284–290
20. Rohland N et al (2015) Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos Trans R Soc Lond B Biol Sci 370(1660):20130624
21. Wang CC et al (2019) Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. Nat Commun 10(1):590
22. Lipson M et al (2018) Population turnover in remote Oceania shortly after initial settlement. Curr Biol 28(7):1157-1165 e7
23. Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. https://doi.org/10.1101/pdb.prot5448
24. Lin M, Zhang S, Yao M (2019) Effective detection of environmental DNA from the invasive American bullfrog. Biol Invasions 21(7):2255–2268
25. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19):2460–2461
26. Simion P et al (2018) A software tool ‘CroCo’ detects pervasive cross-species contamination in next generation sequencing data. BMC Biol 16(1):28
27. Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 40(1):e3
28. Peyrégne S, Prüfer K (2020) Present-day DNA contamination in ancient DNA datasets. BioEssays 42(9):e2000081

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.