Quantifying and reducing cross-contamination in single- and multiplex hybridization capture of ancient DNA

Elena I. Zavala | Ayinuer Aximu-Petri | Julia Richter | Birgit Nickel | Benjamin Vernot | Matthias Meyer

Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Correspondence
Elena I. Zavala and Matthias Meyer, Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. Emails: elena_zavala@eva.mpg.de and mmeyer@eva.mpg.de

Abstract
The use of hybridization capture has enabled a massive upscaling in sample sizes for ancient DNA studies, allowing the analysis of hundreds of skeletal remains or sediments in single studies. Nevertheless, demands in throughput continue to grow, and hybridization capture has become a limiting step in sample preparation due to the large consumption of reagents, consumables and time. Here, we explored the possibility of improving the economics of sample preparation via multiplex capture, that is, the hybridization capture of pools of double-indexed ancient DNA libraries. We demonstrate that this strategy is feasible, at least for small genomic targets such as mitochondrial DNA, if the annealing temperature is increased and PCR cycles are limited in post-capture amplification to avoid index swapping by jumping PCR, which manifests as cross-contamination in resulting sequence data. We also show that the reamplification of double-indexed libraries to PCR plateau before or after hybridization capture can sporadically lead to small, but detectable cross-contamination even if libraries are amplified in separate reactions. We provide protocols for both manual capture and automated capture in 384-well format that are compatible with single- and multiplex capture and effectively suppress cross-contamination and artefact formation. Last, we provide a simple computational method for quantifying cross-contamination due to index swapping in double-indexed libraries, which we recommend using for routine quality checks in studies that are sensitive to cross-contamination.

Keywords
ancient DNA, automation, environmental DNA, hybridization capture, index hopping, index swapping, sediment DNA

1 | INTRODUCTION

Advancements in DNA sequencing techniques have led to tremendous improvements in the analysis of degraded DNA. This is most apparent in the progress that has been made in ancient DNA studies, where complete or partial genomes have been recovered from thousands of ancient human individuals (Racimo et al., 2020). There are two methods that are used for the generation of these data. The first is shotgun sequencing (or the direct sequencing of DNA libraries) where a random subset of DNA molecules that were retrieved...
from a sample is sequenced. As molecules from the taxa of interest (e.g., humans) often constitute only a small proportion of the total molecules in a sample, this method is not economically feasible for many samples. The second strategy is hybridization-based capture of library molecules using synthetic DNA or RNA probes of the sequences of interest (Burbano et al., 2010; Carpenter et al., 2013; Cruz-Davalos et al., 2017; Fu et al., 2013; Suchan et al., 2021), which enriches libraries for molecules with similarity to predefined genomic targets, thereby greatly lowering sequencing costs.

As hybridization capture has become the dominant strategy for data acquisition in population-scale ancient DNA studies, sample sizes continue to increase (Mathieson et al., 2015; Narasimhan et al., 2019; Vernot et al., 2021; Wang et al., 2021; Zavala et al., 2021). However, the consumption of time, consumables and reagents associated with all current capture protocols are beginning to turn hybridization capture into a serious bottleneck in sample preparation. One possible solution to this problem lies in the development of automated protocols for hybridization capture on liquid handling systems (Slon et al., 2017; Suchan et al., 2021). Automation has been shown to be particularly important for applications that necessitate high-throughput sample screening, such as the recovery of ancient human and faunal DNA from sediments. For example, fully automated sample preparation (DNA extraction, library preparation, and hybridization capture) was recently used to screen more than 700 sediment samples from Denisova Cave for the preservation of hominin and other mammalian mitochondrial (mt) DNA to reconstruct the occupational history of the site (Zavala et al., 2021).

Yet the development of higher throughput protocols, especially for hybridization capture, remains critical for further increasing the resolution of ancient DNA studies using sediments and other material. One possible strategy for achieving this goal is the capture of multiple libraries in a single reaction – that is, multiplex capture. Multiplex capture has been successfully used for exome capture of modern samples in order to reduce costs for consumables, reagents and hands-on time while increasing throughput (Filier et al., 2021; Neiman et al., 2012; Ramos et al., 2012; Rohland & Reich, 2012; Shearer et al., 2012; Wesolowska et al., 2011), but has not been explored in the context of ancient DNA research.

One main concern regarding the use of multiplex capture with ancient DNA is the elevated risk of cross-contamination between samples. Specifically, it has been proposed that the parallel amplification of libraries tagged with sample-specific indices may lead to index swapping via jumping PCR during post-capture amplification (MacConaill et al., 2018; van der Valk et al., 2020) (see Figure 1 for a proposed mechanism). Index swapping is also known to occur on the sequencer itself when libraries from different samples are sequenced together: on Illumina platforms, it has mainly been linked to signal spreading between neighbouring/mixed clusters on the flow cell and/or to the incorporation of residual oligonucleotides remaining from library preparation during cluster generation (especially on Illumina’s newer HiSeqX, HiSeq4000 and NovaSeq platforms) (MacConaill et al., 2018; van der Valk et al., 2020). Double-indexing, that is, using index sequences in both library adapters (or inline barcodes immediately adjacent to the insert sequence (Rohland & Reich, 2012)), has been found to be an efficient means to decrease misassignments of index reads (Kircher et al., 2012; MacConaill et al., 2018) and has become a widely used strategy for sequencing ancient DNA. Yet, it remains unclear whether double-indexing is sufficient in preventing cross-contamination between libraries if ancient DNA libraries are combined prior to hybridization capture and sequencing. This question is critical, as cross-contamination has been shown to be a substantial threat to data integrity in ancient DNA studies (Prüfer & Meyer, 2015; van der Valk et al., 2020). In particular, for samples where the DNA of interest represents a miniscule fraction of the total DNA – for example, human DNA in sediments (Slon et al., 2017) or pathogens from ancient skeletal remains (Bos et al., 2015) – even a small amount of cross-contamination could mimic a true signal.

In this study, we present laboratory protocols and analysis tools that facilitate the hybridization capture of large numbers of ancient DNA libraries. Specifically, we describe a simple and straightforward computational strategy to monitor cross-contamination in ancient DNA data sets and show that multiplex capture can safely be performed with double-indexed ancient DNA libraries if post-capture amplification conditions are optimized to suppress index swapping during PCR. We also show that nonoptimal amplification conditions can lead to the formation of artefacts even if hybridization capture is performed individually for each library. Last, we provide an automated capture protocol in 384-well plate format, suitable for both single- and multiplex capture, that minimizes cross-contamination and artefact formation by adjusting the PCR cycle number individually for each well on the plate.

2 | MATERIALS AND METHODS

2.1 | Sampling, DNA extraction, library preparation

DNA was extracted from various sediment samples from Palaeolithic sites (Supplementary Data File 3). Approximately 50 mg of material from each sample was transferred in 2 ml tubes. DNA extraction was performed following the protocol described in (Rohland et al., 2018) with purification on a Bravo NGS Workstation B (Agilent Technologies) using binding buffer “D”. The resulting DNA extract was then converted into a single-stranded double-indexed library on the Bravo NGS Workstation B following the protocol described in Gansauge et al. (2020). DNA extraction, library preparation and indexing PCR setup were performed in a dedicated clean room at the Max Planck Institute for Evolutionary Anthropology (Leipzig, Germany).

2.2 | Hybridization capture

Two successive rounds of hybridization capture with human mtDNA probes (Slon et al., 2017) were performed on the Bravo
NGS Workstation B, either in-solution, as described in Fu et al. (2013), or on beads, as described in Slon et al. (2017). One microgram library was used as input for the first round of capture in the singleplex experiments. For multiplex capture, library pools were used containing 500 ng of each library. After each round of capture, the yield of library molecules was determined by qPCR using 1 µl of the capture eluate (from a total volume of 25 µl) as template (with a 1:10 dilution for the second capture round) and following the protocol outlined in Gansauge et al. (2020). Ten µl of the capture eluate was then amplified using Herculase II Fusion DNA polymerase (Agilent Technologies) in 20 µl reactions containing 1x Herculase II reaction buffer, 250 nM each dNTP, 1 µM each primer (primer pair IS5/IS6 [Meyer & Kircher, 2010] or primer pair IS105/IS109 [see Figure S9 for primer sequences]), and 0.2 µl enzyme. Thermal cycling included a denaturation step at 95°C for 2 min, followed by various numbers of PCR cycles with denaturation at 95°C for 30 s, annealing at 60°C for 30 s and extension at 72°C for 30 s, or annealing and extension at 68°C for 1 min. In some experiments, cycling was performed to PCR plateau for all samples (30 cycles), in other experiments, the qPCR results were used to determine a cycle number that avoids PCR plateau as described in detail in Box 2 of Gansauge et al. (2020). Cycling was performed in a thermal cycler, or if cycle numbers differed among the reactions on a plate, on the Bravo NGS Workstation after adding 15 µl of mineral oil to avoid evaporation (see Supporting Information S4 for a summary of 384-well capture and library amplification on the Bravo NGS Workstation and the Supporting Information video for a visual summary; the electronic protocol files have been deposited on Zenodo [10.5281/zenodo.5656691]). The amplified libraries were purified using 38% PEG solution SPRI beads (Fu et al., 2013) and the purified DNA eluted in 22 µl EBT buffer (10 mM Tris-HCl, 0.05% Tween-20 pH 8.0) for the first capture round and 15 µl EBT for the second round. The second capture round was performed using 10 µl purified library as input. Post-capture quantification of library yield, library amplification, and purification were performed exactly as after the first capture round. Pools
containing 5 µl of each library (or library pool) were then created for sequencing on the Illumina MiSeq platform in 2 x 76 bp paired-end configuration with two 8-bp index reads. The exact parameters used for capture and library amplification in each experiment are provided in Supplementary Data File 3. One capture experiment was performed manually, following the detailed protocol provided in Supporting Information S3.

2.3 | Data processing

Bustard (Illumina) was used for base calling and the resulting sequencing data was processed using the pipeline described in Slon et al. (2017). In short, leeHom (Renaud et al., 2014) was used to merge overlapping paired-end reads into full-length molecule sequences, which were then mapped to the revised Cambridge reference sequence. Next, bam-rmdup (https://github.com/mpievabiohazard-tools) was used to remove sequences shorter than 35 basepairs, along with PCR duplicates. The remaining sequences were assigned to different mammalian families using BLAST (Altschul et al., 1990) and MEGA’s lowest common ancestor algorithm (Huson et al., 2007) as described in Slon et al. (2017).

2.4 | Identification of index-swapping and cross-contamination

Although index swapping has been extensively studied (Griffiths et al., 2018; MacConaill et al., 2018; Owens et al., 2018; Vodak et al., 2018), no broadly applicable method has been proposed, to our knowledge, that aims to directly quantify the level of cross-contamination it causes on a per-library level. Here we suggest a simple and straightforward computational method for detecting and quantifying events of cross-contamination among libraries carrying sample-specific unique double-indices that is based solely on an analysis of the observed counts of index pairs in a pool of libraries that are sequenced together. The method makes use of the notion that if each index is used only once, two successive index swapping events are required to fully replace one index combination by another. Cross-contamination thus proceeds through a transient state where only one of the indices has been replaced by that of another library (see Figure 1 for an illustration). When the counts for all possible combinations of known index sequences are plotted so that expected index combinations fall on the diagonal, these transient states occupy the corners of a rectangle defined by the two libraries between which index swapping has occurred (see Figure 1 for an illustration). We therefore refer to these transient states as “corner libraries”. It is important to note that corner libraries can also arise through processes that are unrelated to jumping PCR, such as errors in sequencing, misincorporation of nucleotides during PCR or cross-contamination of indexing primers. However, these processes are likely to lead to an exchange of only one of the indices. This is in contrast to index swapping, which due to the stochasticity of jumping PCR is expected to result in the symmetrical formation of two corner combinations for each pair of libraries that are amplified in the same reaction.

Our method recognizes possible cases of cross-contamination by identifying all pairs of corner combinations in the data that are compatible with index swapping between one of the expected index pairs and a contaminating library with a different pair of indices (see Supporting Information S1 for further details on the implementation). We then leverage the expectation of symmetry in any true cross-contamination and assume that the number of molecules occupying the transient state is closest to the smaller of the two corner library counts. The ratio between the number of sequences carrying this corner combination and the number of sequences carrying the expected index combination is then used as a proxy for the frequency at which one of the indices in the sample library is replaced by an index originating from the putative contaminant library. As this process has to occur twice in order for both indices to be exchanged and a misassignment to occur, we estimate the probability of this event by squaring the frequency at which one index is exchanged. The probability of cross-contamination can then be multiplied by the number of sequences carrying the expected index pair to estimate the number of sequences that may have originated from a specific contaminant library, but were assigned to the sample library due to index swapping. In addition to estimating individual cross-contamination events, the sum of all events affecting a specific sample library is computed by adding all events that have been determined to contribute at least 0.5 sequences. It should also be noted that the search space of possible cross-contamination events is not confined to libraries carrying indices that were used in the current library pool. Instead, every observed combination of indices that were not used in the sample library pool but are “known” (i.e., are physically available in the laboratory and/or were used in previous experiments) is evaluated as a possible source of cross-contamination. This strategy in principle enables the detection of cross-contamination across experiments and library pools, although limited to libraries with uniquely sample-specific indices (i.e., not between libraries that share one or both of their indices).

3 | RESULTS

3.1 | Exploring the feasibility of multiplex capture for ancient DNA libraries

To determine the feasibility of multiplex capture for the enrichment of genomic targets in ancient DNA libraries, we created two pools of double-indexed libraries prepared from ancient sedimentary DNA. All libraries had previously been found to contain small traces of human mtDNA, six of them with damage patterns indicative of ancient DNA (Zavala et al., 2021). Each pool contained eight libraries that had been combined in equal amounts, and was subjected to hybridization capture using human mtDNA probes. For comparison, we enriched the same libraries individually for human
molecules. hybridization, as well as in their total contents of human mtDNA thus differed in the number of copies of each molecule available for tent of unique DNA molecules in each library. Libraries in the pool by combining equal volumes of each library, irrespective of the con - not unexpected, because the multiplex library pools were created

capture. However, in contrast to singleplex capture, where a simi -

table 3), showing that two capture rounds were necessary for
efficient enrichment of target DNA in the libraries used here. Similar numbers were observed in the multiplex reactions (between 0.17 and 11.53% mapped sequences after the first capture round, with a median of 0.37%, and between 6.04 and 43.44% after the second round, with a median of 15.27%; Figure 2a), indicating successful capture. However, in contrast to singleplex capture, where a similar number of sequences were recovered from each library, multiplex capture led to a less uniform distribution of sequences across libraries after two rounds of capture (Figure 2b). This observation is not unexpected, because the multiplex library pools were created by combining equal volumes of each library, irrespective of the content of unique DNA molecules in each library. Libraries in the pool thus differed in the number of copies of each molecule available for hybridization, as well as in their total contents of human mtDNA molecules.

A further investigation of the index sequences obtained from the single- and multiplex experiments revealed that 22.8% of the sequence molecules in the multiplex pool showed unexpected combinations of indices after the first capture round, and 35.6% after the second (Figure 3a, Supplementary Data File 3). These numbers are much higher than in the singleplex pool (1.3 and 2.1%, respectively, Figure 3b, Supplementary Data File 3), suggesting that jumping PCR, that is, the mispriming of incomplete extension products during PCR (Figure 1), may have led to index swapping among molecules from different libraries during the amplification step. To test this hypothe -

capture. Sequencing of the amplified pool on a MiSeq lane revealed the presence of 33.6% unexpected index combinations (Supplementary Data File 3), consistent with the observations from multiplex capture.

As sequences with unexpected index combinations can be re -

moved in down-stream analyses, index swapping by PCR jumping leads to cross-contamination between double-indexed sample li -

braries only if it occurs repeatedly, that is, if both indices of a mole -
cule are replaced. To determine the impact of cross-contamination on the data, we developed a computational method that detects and quantifies cross-contamination events using the transient products of index swapping, that is, pairs of unexpected indices that are compatible with index swapping between two sample libraries (or between a sample library and another library carrying an unexpected index combination) (Figure 1; see Materials and Methods 2.4 and Supporting Information S1 for further details and Table S1 for an example). Applying this method to the multiplex experiment, we estimate that between 7.0 and 57.9% of sequences assigned to each expected index pair are a product of cross-contamination (or between 137,589 and 269,805 sequences per library), compared to

![Figure 2](image-url)
values between 0.0008 and 0.021% in the singleplex experiment (between 9 and 246 sequences) (Figure 3c,d). Cross-contamination is thus unacceptably high in the multiplex experiment under the conditions used here.

We next investigated whether jumping PCR may introduce additional artefacts during library amplification, namely molecules with chimeric insert sequences. These molecules may arise from the hybridization and extension of DNA strands that contain only partial insert sequences (see Figure 1e for an illustration). A visual inspection of sequence alignments indeed indicated an accumulation of molecules with identical alignment start or end coordinates in some captured libraries (Figure S4), as would be expected if chimeras were created. To investigate this possibility more formally, we determined the proportion of sequences that are expected to share an alignment start or end coordinate by chance in mtDNA data sets of varying sizes by downsampling mtDNA sequences from previously published Neandertal low-coverage genomes (Supporting Information S2).

When comparing the sharing of unique start and end coordinates in the capture data with these values, we find an excess of sharing in some of the second round singleplex captures (Figure 4), indicating that up to ~15% of the sequences in some libraries may be derived from chimeras, artificially inflating the mtDNA coverage in these libraries. The absence of evidence for chimera formation after the first round of capture suggests that these artefacts are predominantly formed when the sequence complexity is low during amplification, that is, when libraries are already enriched for sequences that are similar to each other. It should be noted that only a single jumping PCR event is needed to create a chimeric sequence when amplifying library molecules carrying the same index sequences after singleplex capture. The absence of evidence for chimeric sequences in the multiplex capture is compatible with that most jumping PCR events occur between molecules with different index sequences, leading...
to the formation of chimeras with unexpected index combinations which are subsequently removed from further analysis.

3.2 Determining the impact of PCR cycle number on artefact formation

To reduce index swapping and the formation of molecules with chimeric inserts in hybridization capture, we explored possible options for decreasing artefacts formation in the library amplification steps. As jumping PCR is known to occur most frequently when amplification reactions reach plateau (Paabo et al., 1990), we performed another capture experiment with eight libraries prepared from ancient sedimentary DNA, again in parallel using single- and multiplex capture, where we used qPCR to determine the maximum number of PCR cycles that could be performed before the richest library reaches PCR plateau (12 cycles after the first and 15 cycles after the second capture round, respectively; see Supplementary Data File 3). Under these experimental conditions, 12%–66% (median 29%) of the sequences obtained after two rounds of singleplex capture mapped to the human mtDNA genome, compared to 3%–63% (median 27%) of the sequences from the multiplex capture pool (Supplementary Data File 3, Figure 5a), again indicating successful enrichment of mtDNA. However, in contrast to the previous experiment, where concentration differences between libraries after singleplex capture were mitigated by amplification to PCR plateau, we observed a highly uneven distribution of sequences across libraries not only for the multiplex, but also the singleplex captures (Figure 5b).

Estimates of the contribution of cross-contamination to sequences with expected index combinations ranged from 0.01 to 2.13% (median 0.05%) in the multiplex capture pool and were thus substantially lower than in the initial multiplex capture experiment, indicating that limiting the number of PCR cycles in amplification successfully reduced index swapping (Figure 5c). Likewise, we found no evidence for the formation of chimeric molecules in either single- or multiplex capture (Figure 5d). Although for most libraries, estimates of cross-contamination were close to zero (median 0.016%), we found one library with evidence of moderate cross-contamination (2.43%, with 113 out of 4,671 sequences estimated to be affected) (Figure 5e). This and the similarity in the patterns of libraries affected by cross-contamination in the single- and multiplex captures, points to the possibility that cross-contamination may have been introduced into the libraries already prior to capture. This hypothesis seems plausible here as all libraries used in this and the previous capture experiment had been reamplified in parallel in order to generate large enough DNA quantities for repeated capture experiments. As these amplifications were performed to PCR plateau to maximize DNA yield, sporadic cross-contamination between indexed libraries, for example during bead-based clean-up, may have caused index swapping analogous to what was observed during plateau amplification of multiplex-captured libraries. It is promising that our method was able to identify this single instance of cross-contamination – a process that would otherwise have been challenging to detect.

3.3 Library amplification by two-step PCR

We next investigated whether PCR conditions can be further optimized to limit index swapping in pre- and post-capture amplification of indexed libraries. As jumping PCR is initiated by cross-hybridization of incomplete extension products, we hypothesized that it may be less likely to occur under more stringent annealing conditions. We therefore extended the length of the primers that were previously used for library amplification (IS5 and IS6: Meyer & Kircher, 2010), which resulted in an increase in the annealing temperature from 60 to 68°C (primers IS105 and IS109; see Figure S9), enabling a two-step thermal profile in PCR. We tested these PCR conditions by amplifying the same pool of 92 libraries that had been previously used to evaluate the possibility of PCR-induced index

FIGURE 4 Proportion of shared alignment start and end coordinates in human mtDNA sequences obtained after the first and second round of single- and multiplex capture. The black dots represent the expected proportion of shared start and end coordinates in the absence of chimeric molecules, determined from downsampled shotgun data (in black, regression line calculated using the loess method). Data from the capture experiments are coloured by library. Values for the sharing of start and end coordinates are plotted independently.
swapping (Section 3.1), again omitting capture and amplifying the libraries to PCR plateau. In contrast to the earlier experiment, where unexpected index pairs contributed to 33.6% of the data and cross-contamination was estimated to 10.4% on average per library, MiSeq data generated from the amplification product of the new primer pair and two-step cycling produced only 8.85% unexpected index pairs and cross-contamination reduced to less than 0.25% per library (Supplementary Data File 3). Increasing the annealing temperature in PCR is thus another potential strategy for reducing index swapping during library amplification pre- and post-capture, even if amplification proceeds to PCR plateau.

### 3.4 An optimized protocol for automated hybridization capture

We have shown that an optimal protocol for single- or multiplex hybridization capture of double-indexed libraries should utilize both two-step PCR cycling and a reduction in the number of PCR cycles in post-capture amplification to avoid cross-contamination and the formation of chimeric molecules. However, the adjustment of PCR cycles poses a challenge to the automation of hybridization capture in plate format on liquid handling systems, which is essential for laboratories aiming at high-throughput sample preparation. One solution to this problem could be to limit the cycle number for all reactions on a plate to the maximum number of cycles that can be performed before the richest library reaches PCR plateau (as performed in section 3.2). However, this strategy is problematic in that it produces unequal concentrations of amplification product (as shown in Figure 5a), possibly lowers the efficiency of hybridization in the second capture round due to insufficient amplification of some of the sample libraries, and provides no straight-forward means of pooling libraries in similar concentrations prior to sequencing. It would thus be preferable to individually adjust the PCR cycle number for each well on the PCR plate.

Based on these considerations we developed an automated hybridization capture protocol in 384-well format using the Bravo NGS Workstation B (Agilent Technologies), in which the optimal PCR cycle number for post-capture amplification is determined by qPCR. The protocol then uses two-step PCR for post-capture amplification on the system deck by moving the reaction plate between two permanently heated deck positions. Using pre-sorted pipette tip boxes, amplification products are removed and transferred to a fresh plate as soon as the respective wells reach the desired number of PCR cycles. All reactions are overlain with mineral oil during cycling to prevent evaporation (see Supporting Information for a detailed protocol).

We tested this protocol using eight libraries prepared from ancient sedimentary DNA, which were not reamplified prior to capture to exclude the possibility of precapture indexing swapping. All libraries were then enriched for human mtDNA in two successive rounds of hybridization capture, individually as well as in a multiplex pool of eight, and sequenced on two flow-cells of a MiSeq. The percentage of sequences that mapped to the human mtDNA

---

**FIGURE 5** Characterization of single- and multiplex capture experiments performed with reduced PCR cycle number in library amplification. (a) Number of sequences assigned to each library after two rounds of single- or multiplex capture. (b) Percentage of full-length molecule sequences of at least 35 bp mapped to the human mtDNA genome with a mapping quality of 25 or greater. (c) Log of the number of sequences (log(seqs)) assigned to each expected index pair. (d) Proportion of shared alignment start and end coordinates of unique mtDNA sequences (coloured by library) compared to the downsampled shotgun data (black). (e) The percentages of sequences assigned to each library that were estimated to result from cross-contamination.
reference genome ranged from 25.0 to 65.5% (median 47.3%) for the singleplex captures and from 19.7 to 69.7% (median 48.2%) for the multiplex captures, indicating successful enrichment (Figure 6a, Supplementary Data File 3). While the number of sequences assigned to each library was more heterogeneous for the multiplex than the singleplex capture (Figure 6b), the number of unique hominin mtDNA molecules retrieved from each library was similar between capture types. We also tested a manual version of the optimized capture protocol (Supporting Information S3), performing multiplex capture on two pools of eight libraries and singleplex captures in comparison, and obtained similarly consistent results (Figure S6, Supplementary Data File 3), demonstrating that both the automated and manual protocols are suitable for determining hominin mtDNA preservation in libraries prepared from ancient sedimentary DNA. Importantly, unexpected index combinations contributed to less than 1% of the data (Figure 6d and Supplementary Data File 3) and estimates of cross-contamination did not exceed 0.0004% of the sequences (or not more than 25 sequences in total and at most 1 per cross-contamination event) from any of the libraries in the manual and automated single- and multiplex capture experiments (Figure 6c and Supplementary Data File 3). We also did not detect evidence of the formation of chimeric molecules (Figure 6e, Supplementary Data File 3).

4 | DISCUSSION

High-throughput sample screening can become cost prohibitive in ancient DNA studies. Here we present a strategy to save considerable costs for reagents, consumables and hands-on time by performing hybridization capture of DNA libraries from multiple samples in parallel in a single reaction. To enable multiplex capture without inducing substantial levels of cross-contamination between the sample libraries we developed a post-capture amplification protocol that minimizes index swapping by jumping PCR. This was achieved by adjusting cycle number individually for each reaction so that the amplification is stopped before reaching PCR plateau and by modifying primer sequences to allow for a higher annealing temperature.

**FIGURE 6** Characterization of single- and multiplex capture performed with an automated protocol allowing for individual PCR cycle number adjustment in each position of the 384-well microtitre plate. (a) Percentage of full-length molecule sequences of at least 35 bp mapped to the human mtDNA genome with a mapping quality of 25 or greater. (b) Number of sequences assigned to each library after two rounds of single- or multiplex capture. (c) The percentages of sequences assigned to each library that were estimated to result from cross-contamination for multiplex (purple) and singleplex (green) capture. (d) Log of the number of sequences (log(seqs)) assigned to each expected index pair. (e) Proportion of shared alignment start and end coordinates of unique mtDNA sequences from libraries containing ancient hominin mtDNA (coloured by library) compared to the downsampled shotgun data (black)
Importantly, we also provide a strategy for adjusting PCR cycle number individually for each reaction even if PCR is carried out in plate format on an automated liquid handling platform. Using a method for estimating cross-contamination from the distribution of observed index pairs, we demonstrate that together, both measures reduce cross-contamination to below detectable levels. Although we provide detailed protocols for manual and automated hybridization capture in 384-well format using the Fu et al. (2013) method for capture with single-stranded DNA baits, the amplification protocols provided here can be combined with all other available capture methods.

Despite the proof-of-principle presented here, important considerations have to be made before integrating multiplex capture into sample preparation workflows. First, the content of DNA often varies in ancient biological material. Libraries from different samples may contain vastly different numbers of target molecules and different numbers of copies may have been created from each molecule during initial library amplification. Thus, even if libraries are pooled in equal mass, hybridization capture recovers different numbers of molecules from each sample library, leading to heterogeneity in sequence representation among samples. It may therefore be advisable to avoid pooling of libraries from different types of material, such as skeletal remains and sediments, or from archaeological sites with vastly different DNA preservation. Second, multiplex capture was only attempted here for a relatively small genomic target: human mtDNA. It is unclear if multiplex capture with larger genomic targets can be efficiently performed with ancient DNA, although success with multiplex capture of modern samples (Rohland & Reich, 2012) maintains this possibility. For these reasons, we currently recommend multiplex capture primarily for the recovery of small genomic targets, where high sequencing depth can easily be reached to compensate for differences in sequence representation among samples. Multiplex capture may also be used as a screening tool when working with large sample sizes to determine whether ancient DNA from the taxa of interest is present, possibly followed by a subsequent individual enrichment of libraries for larger genomic targets.

Contamination is a serious threat when working with small quantities of DNA and measures have been implemented to minimize its impact in ancient DNA research. Common strategies for contamination prevention include the use of clean rooms, the separation of laboratory space for pre- and post-amplification work, and the inclusion of negative controls to detect sporadic contamination with exogenous DNA or between samples (Cooper & Poinar, 2000). However, it has been assumed that contamination is unlikely to occur after library preparation, especially when libraries are tagged with sample-specific indices in both adapter sequences. Contrary to this expectation, we detected small levels of sporadic cross-contamination even in sequence data from libraries that were enriched for mtDNA in singleplex captures, indicating that an exchange of fine droplets or aerosols occurred between libraries from different samples before they were amplified to PCR plateau pre- or post-capture. These results highlight that caution is required when reamplifying DNA libraries, even if they are double-indexed. Primers that allow for the highest possible annealing temperature should be used for this purpose to minimize index swapping, such as the IS105/IS109 pair provided here, and PCR plateau avoided whenever possible. These measures also prevent the formation of chimeric insert sequences, which can occur if sequences from small genomic targets are enriched by successive rounds of hybridization as shown here.

We also provide a simple computational method that can be used to examine sequence data to determine if cross-contamination by index swapping has occurred and to quantify its impact. As the method requires only the number of sequences assigned to different index pairs it can be applied across various platforms and be used to test for cross-contamination in various capture protocols and kits. In order to maintain the quality and transparency of results, we encourage labs to include this test in their routine quality control pipeline and publish the results. In summary, our work shows that it remains critical to monitor and suppress the formation of artefacts in ancient DNA sample preparation and sequencing as technological and methodological advances continue to push the limits for DNA recovery.

ACKNOWLEDGEMENTS
This project was funded by the Max Planck Society. We thank B. Schellbach and A. Weihmann for their help with sequencing, J. Kelso and J. Visagie for helpful discussions and assistance with raw data processing.

CONFLICT OF INTEREST
The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS
Elena I. Zavala and Matthias Meyer designed the study. Elena I. Zavala, Ayinuer Aximu-Petri, Julia Richter, and Birgit Nickel performed laboratory experiments. Elena I. Zavala, Benjamin Vernot and Matthias Meyer performed data analysis. Elena I. Zavala and Matthias Meyer wrote the manuscript with input from all authors.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://doi.org/10.5061/dryad.2280gb5t7.

DATA AVAILABILITY STATEMENT
The files containing sequence counts used for calculating cross-contamination have been made available on Dryad (https://doi.org/10.5061/dryad.2280gb5t7).

Code availability: The automated protocol electronic files can be found on Zenodo (10.5281/zenodo.5656691). The code for calculating cross-contamination can be found on github (https://github.com/mpiev a/index_swapping).

ORCID
Elena I. Zavala https://orcid.org/0000-0001-7687-5370
Vernot, B., Zavala, E. I., Gomez-Olivencia, A., Jacobs, Z., Slon, V., Mafessoni, F., Romagné, F., Pearson, A., Petr, M., Sala, N., Pablos, A., Aranburu, A., de Castro, J. M. B., Carbonell, E., Li, B., Krajcarz, M. T., Krivoshapkin, A. I., Kolobova, K. A., Kozlikin, M. B. ... Meyer, M. (2021). Unearthing Neanderthal population history using nuclear and mitochondrial DNA from cave sediments. *Science*, 372(6542), https://doi.org/10.1126/science.abf1667

Vodak, D., Lorenz, S., Nakken, S., Aasheim, L. B., Holte, H., Bai, B., Myklebost, O., Meza-Zepeda, L. A., & Hovig, E. (2018). Sample-Index misassignment impacts tumour exome sequencing. *Scientific Reports*, 8(1), 5307. https://doi.org/10.1038/s41598-018-23563-4

Wang, Y., Pedersen, M. W., Alsos, I. G., De Sanctis, B., Racimo, F., Prohaska, A., Coissac, E., Owens, H. L., Merkel, M. K. F., Fernandez-Guerra, A., Rouillard, A., Lammers, Y., Alberti, A., Denoeud, F., Money, D., Ruter, A. H., McColl, H., Larsen, N. K., Cherezova, A. A., ... Willerslev, E. (2021). Late Quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature*, https://doi.org/10.1038/s41586-021-04016-x

Wesolowska, A., Dalgaard, M. D., Borst, L., Gautier, L., Bak, M., Weinhold, N., ... Gupta, R. (2011). Cost-effective multiplexing before capture allows screening of 25 000 clinically relevant SNPs in childhood acute lymphoblastic leukemia. *Leukemia*, 25(6), 1001-1006. https://doi.org/10.1038/leu.2011.32

Zavala, E. I., Jacobs, Z., Vernot, B., Shunkov, M. V., Kozlikin, M. B., Derevianko, A. P., Essel, E., de Filippo, C., Nagel, S., Richter, J., Romagné, F., Schmidt, A., Li, B., O’Gorman, K., Slon, V., Kelso, J., Pääbo, S., Roberts, R. G., & Meyer, M. (2021). Pleistocene sediment DNA reveals hominin and faunal turnovers at Denisova Cave. *Nature*, 595(7867), 399–403. https://doi.org/10.1038/s41586-021-03675-0

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Zavala, E. I., Aximu-Petri, A., Richter, J., Nickel, B., Vernot, B., & Meyer, M. (2022). Quantifying and reducing cross-contamination in single- and multiplex hybridization capture of ancient DNA. *Molecular Ecology Resources*, 22, 2196–2207. https://doi.org/10.1111/1755-0998.13607