Raw reads were quality checked with FASTQC v0.11.8 (Andrews, [2010](https://resjournals.onlinelibrary.wiley.com/doi/full/10.1111/syen.12503#syen12503-bib-0002)). Sequencing reads resulting from samples with highly degraded DNA were treated from this point as single-end reads. This approach was chosen, as degraded DNA is likely to randomly ligate together during the adapter ligation stage of library preparation, resulting in chimeras of different genomic regions (Willerslev & Cooper, [2005](https://resjournals.onlinelibrary.wiley.com/doi/full/10.1111/syen.12503#syen12503-bib-0080)). Nevertheless, the sequencing information contained in the reads is still reliable, as chimera formation typically results in DNA inserts larger than read length; therefore, more reliable results are obtained by treating data as single-end (Rowe *et al*., [2011](https://resjournals.onlinelibrary.wiley.com/doi/full/10.1111/syen.12503#syen12503-bib-0062)). For the sample that underwent sonication (*H. puera*), reads were carried forward as paired-end. Reads with ambiguous bases (N's) were removed from the dataset using Prinseq 0.20.4 (Schmieder & Edwards, [2011](https://resjournals.onlinelibrary.wiley.com/doi/full/10.1111/syen.12503#syen12503-bib-0063)). Trimmomatic 0.38 (Bolger *et al*., [2014](https://resjournals.onlinelibrary.wiley.com/doi/full/10.1111/syen.12503#syen12503-bib-0010)) was used to remove low-quality bases from their beginning (LEADING:3) and end (TRAILING:3), by removing reads below 30 bp and by evaluating read quality with a sliding window approach. Quality was measured for sliding windows of four base pairs and had to be greater than PRHED 25 on average. The resulting cleaned reads were used for *de novo*genome assembly with spAdes v3.13.0 (Bankevich *et al*., [2012](https://resjournals.onlinelibrary.wiley.com/doi/full/10.1111/syen.12503#syen12503-bib-0004)) with kmer values of 21, 33 and 55. The completeness of each assembly was assessed using BUSCO 3.0.2b (Simão *et al*., [2015](https://resjournals.onlinelibrary.wiley.com/doi/full/10.1111/syen.12503#syen12503-bib-0065)) using the Insecta lineage set. However, due to the fragmented nature of the genomes, BUSCO has trouble identifying orthologs; therefore, the genomes were searched for the Insecta lineage set using a tblastn approach (e-value threshold 1e-5, minimum identity of 60%) with standalone BLAST 2.9.0 (Camacho *et al*., [2009](https://resjournals.onlinelibrary.wiley.com/doi/full/10.1111/syen.12503#syen12503-bib-0013)).

**Raw sequence processing**

Raw sequence quality was inspected with the program FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). All reads were quality-trimmed and filtered using Trimmomatic-0.36 (Bolger et al. 2014). Bases were trimmed when the average quality of 4-base sliding windows were below 15 and when bases at the start and end of reads had a quality below 10. Subsequently, all trimmed reads shorter than 25 bp were filtered out. Finally, all reads were trimmed by their first 6 bp (index sequence) and cut to a maximum length of 75 bp (fragment size of degraded DNA isolate). Processed reads were loaded into the program Geneious 2019.2.1 (<https://www.geneious.com>) and an additional 4 bases were trimmed from the 3’ end of every sequence to remove the artificial poly-G tail of many sequences. All trimmed sets of reads were again checked for quality with FastQC.

2.2. Phylogenomics analyses

Raw paired-end reads were first filtered and trimmed using *fastp* v0.20.1 ([Chen et al., 2018](https://www.sciencedirect.com/science/article/pii/S1055790323001227#b0060)) with the following parameters: --correction --detect\_adapter\_for\_pe --cut\_front 3 --cut\_tail 3 --qualified\_quality-phred 20 --average\_qual 20 --length\_required 35 --unqualified\_percent\_limit 10 --n\_base\_limit 0, and then processed following the Genome Analysis Toolkit (GATK v3.8) workflow ([Van der Auwera et al., 2013](https://www.sciencedirect.com/science/article/pii/S1055790323001227#b0360)). Reads were merged with Picard toolkit (<https://broadinstitute.github.io/picard/>) and then mapped onto the chromosome-level [genome assembly](https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/genome-assembly) of [*Melopsittacus undulatus*](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/budgerigar) (bMelUnd1.mat.Z; GenBank accession no.: GCA\_012275295.1), which is grouped together with *Agapornis*under Loricoloriinae ([Schweizer et al., 2010](https://www.sciencedirect.com/science/article/pii/S1055790323001227#b0320)). Reads were mapped to this *M. undulatus*reference genome using BWA-MEM v.0.7.13 ([Li and Durbin, 2009](https://www.sciencedirect.com/science/article/pii/S1055790323001227#b0245)) and marked for duplicates using Picard toolkit. Reads were further realigned around [indels](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/indel) using GATK IndelRealigner. The base quality scores of the samples were rescaled using mapDamage2 ([Jónsson et al., 2013](https://www.sciencedirect.com/science/article/pii/S1055790323001227#b0195)), with the *M. undulatus* genome as the reference. SNP calling and genotyping was done on the rescaled BAM files using ANGSD v0.933 ([Korneliussen et al., 2014](https://www.sciencedirect.com/science/article/pii/S1055790323001227#b0215)) with the following parameters: *-minMapQ 30 -minQ 30 -uniqueOnly 1 -remove bads 1 -trim 0 -only\_proper\_pairs 0 -baq 1 -C 50 -doMaf 1 -doMajorMinor 1 -skipTriAllelic -doCounts 1 -minMaf 0.05 -SNP\_pval 1e-6 -minInd 10 -setMinDepth 24 -maxDepth 480 -dumpCounts 2 -doGeno 23 -doPost 1 -postCutoff 0.34 -GL 1 -doGlf 3*. BAQ computation (*-baq 1*) was shown to greatly reduce false SNP calls around misaligned indels ([Li, 2011](https://www.sciencedirect.com/science/article/pii/S1055790323001227#b0240)). To test for the effect of sample type on the number and quality of SNP calls, we compared the number of mapped reads, the average mean per-site depth of coverage, and the proportion of missing genotype data between sample type using two-sided t-tests in R. These metrics were computed using samtools idxstats and coverage (samtools v.1.9; [Danecek et al., 2021](https://www.sciencedirect.com/science/article/pii/S1055790323001227#bib386)). The significance of the tests was assessed with the Bonferroni correction for multiple pairwise comparisons.

### 3.1 Cleaning of raw reads

Nextflow is a scalable and reproducible scientific workflow programming model that can easily be implemented in local environments or expanded to take advantage of high-performance computing cluster environments (Di Tommaso et al., [2017](https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13660#men13660-bib-0015)). We have developed a custom pipeline to process hDNA sequencing libraries and recently converted the entire cleaning workflow to a Nextflow pipeline (nf-polish; <https://github.com/MozesBlom/nf-polish>). nf-polish processes each sequencing library separately, rather than by individual, and we only merge libraries once they have been mapped to a reference. Prior to any library modification, nf-polish quantifies the degree of (adapter) contamination and sequencing quality using fastqc (version 0.11; Andrews, [2010](https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13660#men13660-bib-0001)). Adapter “read-through” is commonplace due to the short size of the DNA fragments and the sequencing set-up frequently used by sequencing facilities (e.g., 150-bp paired-end sequencing). Following an initial round of quality control, nf-polish uses superdeduper (Petersen et al., [2015](https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13660#men13660-bib-0050), Now part of htsstream, version 1.3) to remove PCR duplicates. Deduplication can be done by comparing each read in a library to each other (e.g., superdeduper; computationally highly intensive) or by comparing the start coordinates of read-pairs following mapping (e.g. picard-dedup; computationally less demanding). However, for hDNA, coordinate-based deduplication probably leads to a high proportion of false positives since only the start position of each read(−pair) is used to identify duplicates. Paired-end reads have two starting coordinates (the first position in the 5′ direction of both the forward and reverse read) but cleaned hDNA libraries have a high proportion of merged reads (leading to an effectively single-end data set with a single starting coordinate at the 5′ position). With duplicates largely removed, nf-polish then trims away frequently used Illumina sequencing adapters (trimmomatic version 0.39; [Bolger et al., [2014](https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13660#men13660-bib-0005)]) and merges read-pairs with a substantial overlap between forward and reverse read (pear version 0.9; Zhang et al., [2014](https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13660#men13660-bib-0068)). Merging (and deduplication) is done because each mapped read should correspond to a unique DNA molecule to avoid over-inflation of coverage or a nonbiological skew in coverage for one haplotype. Moreover, merging improves the quality score of that fragment and is therefore done prior to quality trimming (trimmomatic). The final cleaning step includes the removal of low-complexity reads (>50% of one nucleotide type) since these often stem from highly repetitive regions, are difficult to reconstruct and are prone to mapping error. Following each processing step, nf-polish calculates processing statistics with seqkit (version

1. Data processing and mapping

Raw FASTQ files are trimmed with fastp [[24](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR24)], which automatically detects adapter sequences from the read data, under the assumption that only one adapter is present in the reads. Fastp also automatically enables poly-G trimming for Illumina NovaSeq or NextSeq samples by checking the flow cell identifier, so it handles read trimming for multiple Illumina platforms. For historical samples, fastp simultaneously adapter- and quality-trims reads and merges overlapping paired-end reads. By default, merged reads below a threshold of 30 bp are discarded. However, this threshold can be modified using the configuration file. Most ancient and historical DNA studies discard reads below a threshold of length between 30 and 35 bp, depending on the preservation of the samples (e.g. [[25](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR25), [26](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR26)]). The default merging settings are recommended in order to exclude modern-day contaminating sequences for which paired-end reads will not overlap since they are typically longer. Merged reads are then mapped to the reference genome using BWA aln with settings optimized for ancient/historical DNA (-l 16500 -n 0.01 -o 2) [[27](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR27)]. For modern samples, FASTQ files are adapter- and quality-trimmed with fastp and then mapped to the reference genome assembly with BWA mem [[28](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR28)] using default settings. For ancient/historical samples, multiple sequencing libraries are commonly generated per sample to avoid overrepresentation of specific PCR duplicates in the read data, using different index(es) for each sequencing library. The alignments of each index of a sample that were sequenced on different lanes are therefore merged to generate one BAM file per index. Next, PCR duplicates are identified and excluded from ancient/historical data using both read start and end mapping coordinates using a custom Python script. For modern samples, the same merging algorithm is run if applicable and PCR duplicates are identified and marked using Picard MarkDuplicates (). For both historical and mode<https://broadinstitute.github.io/picard/>rn data, the alignments of each sample from different indices are then merged to generate one BAM file per sample, and reads around indels are realigned with GATK IndelRealigner [[29](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR29)] to improve mapping accuracy. Basic mapping statistics are reported for each processing step using samtools and Qualimap [[30](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR30)], and are summarized using MultiQC [[31](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR31)]. Histograms depicting the depth per site as well as the genome-wide average depth and minimum and maximum depth thresholds for downstream analyses are also generated. The user can decide if genome-wide average depth should be calculated including or excluding sites with zero coverage. By default, depth thresholds are set to 1/3 and ten times the average genome-wide depth of each sample, which should be adjusted by the users according to their data characteristics. However, the pipeline uses an absolute minimum depth of three reads per site that cannot be changed by the user. Although the pipeline can be run with samples sequenced to lower depths, an average genome-wide depth of at least 6X per sample is recommended to have enough statistical power for detecting heterozygous sites. MultiQC reports and depth histograms based on BAM files after indel realignment are included into an automatically generated GenErode pipeline report.

There are two main optional steps to further process the BAM files before running any downstream data analyses: (1) base quality rescaling with MapDamage2 [[32](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR32)] for selected ancient/historical samples that have not been treated with the uracil-DNA glycosylase (UDG) enzyme and show post-mortem damage; and (2) subsampling of the selected BAM files from ancient/historical and/or modern samples using samtools to a target genome-wide depth to avoid biases introduced when comparing samples at different coverages. Additionally, a mitochondrial contamination check can be run for selected ancient/historical samples, in which the trimmed and merged reads are aligned to the mitogenomes of a set of species for which contamination may be present in the laboratory (i.e. from the laboratory reagents [[33](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR33)]). The user is asked to include a mitogenome FASTA file of the target species at this step. The GenErode pipeline produces BAM file statistics and a table listing the ratio of mapped reads to the mitochondrial genome of a potentially contaminating species and of the target species to help identify sequencing libraries with more reads mapping to the mitogenome from a different species. The results from this mitochondrial contamination check are not further used in the pipeline but the BAM files containing the sequences mapping to the mitochondrial genome of the target species are kept so that they can be used for downstream analyses outside of the pipeline. After each of these optional steps, basic mapping statistics are reported using samtools and Qualimap, and are summarized using MultiQC.

1. 3.

Genotyping and variant filtering

This part of the pipeline is designed to perform variant calling on a per-sample basis. The rationale being that ancient, historical, and modern samples typically come from disparate time periods and locations, so using the information from some samples to predict variation in others can lead to undesired biases. Therefore, variants are called in each sample using bcftools [[34](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR34)] mpileup and call*.* Before proceeding to downstream analyses, the variant calls are subjected to several filtering steps. Methylated CpG sites are protected from UDG enzyme treatments and so post-mortem damage might remain in such sites [[35](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR35), [36](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR36)]. Ideally, a variant filter should therefore capture sites in ancient/historical samples that have a basepair change from CpG due to post-mortem damage. GenErode offers three optional methods to identify CpG sites: (1) identifying them in the reference genome assembly, (2) identifying them in selected samples once genotyped; and (3) combining both strategies. In all three cases, these CpG sites are automatically excluded from the final VCF files and downstream analyses. For datasets with only one or a few samples, we recommend identifying CpG sites in the reference genome assembly (option 1), which will remove damaged (or mutated) sites in the samples that are CpG in the reference. In larger datasets composed of modern and ancient/historical samples that are mapped to a more distantly related reference genome assembly (e.g. from a different species than the samples), it is recommended to identify CpG sites using all genotyped samples (option 2). Finally, we recommend a combination of both strategies (option 3) when a more stringent CpG filter is desired. After CpG site removal from VCF files, the pipeline moves on to other filtering steps. Only sites with mapping and base qualities of at least 30 are kept. The pipeline will exclude all variants that are located within 5 bp of indels (insertions or deletions), and will subsequently remove all indels. Also, all sites falling outside the depth thresholds specified by the user (as described above) are removed. An allelic imbalance filter removes heterozygous sites in which less than 20% or more than 80% of reads support each allele to avoid erroneous genotypes caused by contamination or misalignments [[37](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04757-0#ref-CR37)]. Variants falling within repetitive regions, earlier identified in the reference genome assembly, are also excluded in this step.

Finally, VCF files from all samples are merged and sites that are not biallelic as well as sites with more than 10% missing genotypes across all samples are removed. This missingness threshold can be adjusted by the user to suit each particular dataset, for example by relaxing it when many samples are to be analyzed at the same time. A BED file is created of all the genomic locations of sites remaining after the filtering and the merged VCF file is split into one VCF file of historical and one VCF file of modern samples.