An open-sourced bioinformatic pipeline for the processing of Next-Generation Sequencing derived nucleotide reads: Identification and authentication of ancient metagenomic DNA.

*THOMAS C. COLLIN[1], KONSTANTINA DROSOU[2,3], JEREMIAH DANIEL O'RIORDAN[4], TENGIZ MESHVELIANI[5], RON PINHASI[6], ROBIN N. M. FEENEY[1]

[1]School of Medicine, University College Dublin, Ireland

[2]Division of Cell Matrix Biology & Regenerative Medicine, University of Manchester, United Kingdom

[3]Manchester Institute of Biotechnology, School of Earth and Environmental Sciences, University of Manchester, United Kingdom

[4]j.daniel.oriordan@icloud.com

[5]Institute of Paleobiology and Paleoanthropology, National Museum of Georgia, Tbilisi, Georgia

[6]Department of Evolutionary Anthropology, University of Vienna, Austria

*Correspondence: thomas.c.collin@icloud.com

Abstract

Bioinformatic pipelines optimised for the processing and assessment of metagenomic ancient DNA (aDNA) are needed for studies that do not make use of high yielding DNA capture techniques**.** These bioinformatic pipelines are traditionally optimised for broad aDNA purposes, are contingent on selection biases and are associated with high costs. Here we present a bioinformatic pipeline optimised for the identification and assessment of ancient metagenomic DNA without the use of expensive DNA capture techniques. Our pipeline actively conserves aDNA reads, allowing the application of a bioinformatic approach by identifying the shortest reads possible for analysis (22-28bp). The time required for processing is drastically reduced through the use of a 10% segmented non-redundant sequence file (229 hours to 53). Processing speed is improved through the optimisation of BLAST parameters (53 hours to 48). Additionally, the use of multi-alignment authentication in the identification of taxa increases overall

confidence of metagenomic results. DNA yields are further increased through the use of an optimal MAPQ setting (MAPQ 25) and the optimisation of the duplicate removal process using multiple sequence identifiers (a 4.35-6.88% better retention). Moreover, characteristic aDNA damage patterns are used to bioinformatically assess ancient vs. modern DNA origin throughout pipeline development. Of additional value, this pipeline uses open-source technologies, which increases its accessibility to the scientific community.

Introduction

Optimised bioinformatic pipelines are of particular importance in the broad study of metagenomics, which consist of large datasets of multi-origins, and the associated complexities of large-scale computational processes such as comparative sequence alignment and multiple taxa identifications. The emerging field of ancient metagenomics adds to these processing complexities with the need for additional steps in the separation and authentication of ancient sequences from modern sequences. Currently, there are few pipelines available for the analysis of ancient metagenomic DNA (aDNA)[1–4] The limited number of bioinformatic pipelines for aDNA metagenomics can be attributed to a low yield of DNA compared to that achieved in modern metagenomic DNA extraction. This results in the use of high-cost DNA capture techniques to improve aDNA yields to levels suitable for bioinformatic assessment[3–7]. In parallel, the lack of bioinformatic pipelines optimised for the processing of lower aDNA yields (i.e. those studies which do not use high-cost DNA capture techniques) deters researchers from exploring and developing alternative methods of aDNA extraction for metagenomic purposes.

Those metagenomic studies performed using DNA capture techniques, necessary for existing bioinformatic pipelines, allow for the implementation of “quick” bioinformatic comparisons using RefSeq or Blastn megablast options due to higher yields of DNA[5–7]. Due to higher yields, however, these pipelines, by their nature, often compromise between read conservation and processing time, thus losing sequences throughout individual computational steps. In doing so these pipelines do not account for the nature of metagenomic aDNA which, having originated from multiple sources

(e.g. bone[8], soil[3,9]), can vary in aDNA yields and in the degree of damage over time leading to a potential loss and underrepresentation of aDNA sequences.

The damage pattern of aDNA serves as a useful tool in the distinction between ancient and modern DNA sequences[10,11]. Characteristically, ancient sequences should consist of heavily fragmented DNA strands[6,12,13], depurination, depyrimidination and deamination events[14,15] and miscoding lesions[16,17]. These characteristics therefore form the central damage pattern for aDNA authentication in the development of this bioinformatic pipeline.

In addition to the high cost associated with DNA capture techniques, a further issue arises from the direct selection of target taxa for probe generation prior to sequencing. This action inevitably introduces a selection bias into a study and could prevent the metagenomic analysis of all aDNA present within a sample with targeted DNA yields overshadowing untargeted yields[7,18,19]. The additional cost associated with computation and long processing times for comparative analysis acts as a barrier to more widespread use of aDNA metagenomics in fields such as archaeology, bioanthropology and paleoenvironmental sciences. Bioinformatic costs associated with computation to achieve faster processing speeds or the use of third-party interface platforms are usually unavoidable due to the large demands on processing time and the steep learning curve needed to gain proficiency in the open-source alternatives, which are often less user-friendly and lack the benefit of customer support and easily accessible manuals.

Using a method developed for the extraction of metagenomic aDNA from anthropogenic sediment (sediment that has come into contact with past human activity) without the use of DNA capture techniques[9,20,21], we present a bioinformatic pipeline optimised for the identification and authentication of metagenomic aDNA that can be applied to studies yielding comparatively lower yields of aDNA. The development of the bioinformatic pipeline involves four fundamental underpinnings, in which it must:

1. Cater to low yielding metagenomic aDNA and conserve reads wherever possible
2. Be able to process data within a reasonable timeframe and allowing for multiple taxa identifications

3. Be developed using open-source technologies and software wherever possible to facilitate universal access and to bypass financial barriers
4. Be accompanied by a step-by-step user-friendly manual, to facilitate its use without coding and programming expertise (supplementary; SI)

Results and Discussion

Identifying the smallest retrievable aDNA fragment and establishing minimum sequence length threshold

The smallest length of retrievable aDNA fragments were identified through the digital visualisation of aligned DNA sequences using UGENE[22] and Geneious R10[23] software. To test for the smallest retrievable aDNA fragment, adapter sequences were removed using Cutadapt[24], with a minimal sequence length threshold (cut) of 15bp. This threshold was chosen after a cut of 0 was initially used and quality analysis of sequences using FastQC software[25] showed the absence of DNA sequences below 15bp (Figure 1). The resultant file sequences were then compared to those of the National Center for Biotechnology Information's (NCBI) genomic nucleotide database using Basic Local Alignment Search Tool (BLAST)[26–28]. BLAST results were imported into MEtaGenome ANalyzer (MEGAN)[1] for visualisation of genomic assignments. *Mammalia* and *Plantae* assignments at the taxonomic genus level were assigned a unique identification number. Using the random number generator specified in the SI (step 9), unique identification numbers were selected and used for further in-depth alignment to its associated reference genome using Burrows-Wheeler Aligner (BWA)[29]. Alignments were converted to SAM format[30] and imported into DNA visualisation software. This visualisation allowed for the manual identification of (C>T, G>A) damage patterns at the terminal ends of aligned sequences against a reference genome. Fragments without these characteristics were deemed modern in origin.

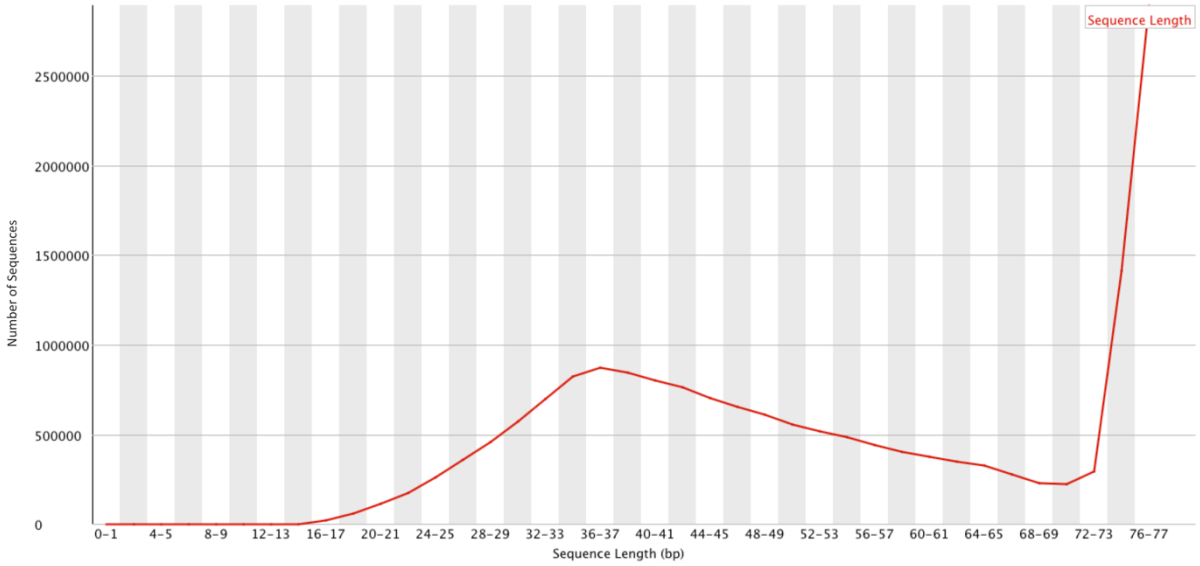

Figure 1. Distribution of Sequence Length After Adapter Removal Using FastQC. The number of sequences are plotted against their respective sequence lengths. The red line represents the peak number of sequences for each length. The plot shows the absence of DNA sequences below 15bp.

This procedure resulted in the identification of ancient sequences ranging from 22-28bp in length using the extraction method outlined by Collin[9,20]. For comparison, using Dabney's method which was developed for aDNA extraction from bone[31], Slon[3] found that the lowest extractable reads were 35bp when using a ladder that mimicked aDNA. This suggests that the method used in this study is capable of extracting shorter aDNA fragments and as such may achieve a wider range of representative sequences. This is particularly important when considering the nature of DNA damage over time which not only results in increased fragmentation through oxidative strand breakages[10,11], but also the potentially disproportionate lesion damage to genomes with high cytosine content[16,17,32] which could otherwise be overlooked. It should be noted however that the shorter the DNA sequence, the more prone it is to misalignment errors[33]. For this reason, a cut of 28bp was selected as a conservative threshold for the purpose of this study.

Creation of a non-redundant, 10% representative sample file for comparative analysis

One of the most common issues with the bioinformatic assessment of DNA sequences is the processing power required (corresponding to speed) and time it takes to process data. This is especially true for metagenomic data, where multiple taxa identifications are sought within a single sample. The size of a file is proportional to the time required for a process to take place. Therefore, reducing the size of a file will reduce the time required for processing.

A non-redundant sample file was created using FastX Toolkit[34]. FastX's collapser program merges all sequence repetitions (duplicates) for a region of coding into a single representative sequence, while maintaining read count data. This is performed for all duplicates until only unique sequences remain (Figure 2A). This reduction in sequences reduces the file size considerably (1.8GB – 956MB) thus reducing associated processing time.

This pipeline further improves the processing time of comparative analysis by splitting the non-redundant sample file into 10 representative files of equal size, each file being representative of 10% of the total sequences within an entire sample. This was achieved using pyfasta[35]. This action resulted in a further reduction of file size (956Mb – 96Mb per 10% file). A representative file is made possible due to two properties of DNA extraction. Firstly, multiple sequences representing the various coding regions for taxa are extracted, increasing the likelihood of representative distribution between the files. Secondly, NGS platforms sequence DNA fragments in a random order as they enter the flowcell. The corresponding sequence data is saved in this same order, meaning DNA sequences should be randomly distributed throughout the representative files (Figure 2B).

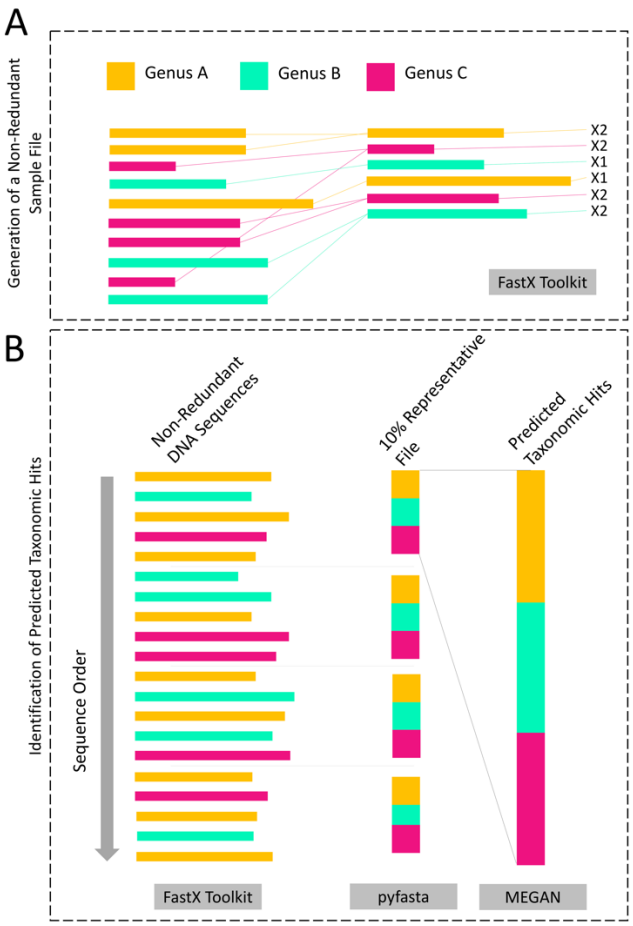

Figure 2. Schematic Overview for the Generation of Non-Redundant Segmented (10%) File for Comparative Analysis and Identification of Predicted Taxonomic Hits. (**A**) Generation of non-redundant file. Identically repeating sequences are merged into a single representative sequence for each occurrence in the generation of a non-redundant sample file. (**B**) Generation of the segmented (10%) file and identification of predicted taxonomic hits. Different taxa are represented by varying colours and the software used for each step is listed in the grey boxes.

To validate that these files are representative of an entire sample's sequences, the full sample identifications were compared to predictions made using the 10% sample files. Samples were processed using a cut of 28bp before being made non-redundant and split into 10 representative files. To eliminate the potential for bias, a randomised number generator (SI, step 9) was used to select one of the representative files for comparative alignment using BLAST. Comparative alignment was performed and resulting BLAST files imported into MEGAN using a bit-score ('min-score') of 40

within the top 10% of best alignments, and the default "naïve" lowest common ancestor (LCA) algorithm as described by Huson[1]. The bit-score measures the similarity of a sample sequence to a comparative sequence through complex computations[36]. Huson[1] used a bit-score of 30 for MEGAN analysis of an ancient mammoth dataset, here we use 40 as a more conservative threshold. The resulting genomic assignments for the 10% representative file were compared to the total genomic assignments achieved from the unsplit originating non-redundant file. The mean percentage difference and standard error between expected hits based on the 10% files and actual hits achieved with the 100% file were calculated. The expected total genomic assignments predicted by the 10% file was accurate to those achieved within the 100% file within −0.007% (±1.101 SEM). The use of a 10% representative non-redundant file drastically reduced the amount of time required for comparative alignment processing: processing time for the representative file was 76.86% less than the time required for the original unaltered file (229 hours to 53).

BLAST parameters in comparative analysis for time conservation

BLAST is a multi-platform algorithm that allows users to query sample sequences against a specified database[26]. In the case of a metagenomics study, we recommend use of NCBI's entire nucleotide database[37] (https://www.ncbi.nlm.nih.gov/). BLAST has a series of options that allow the user to optimise the comparative analysis process for the data being queried[26–28]. The 'task exec' option is the most important of these, allowing users to specify the type of search required. The task blastn and task blastn-short options are best suited for inter-taxon comparisons using short sample sequences, the latter of these being optimised for sequences shorter than 50bp[38]. In the context of aDNA most authentic sequences fall within the range of 30-70bp[6,12,13]. The use of blastn-short is therefore unsuitable for aDNA analyses, and thus blastn is utilised herein.

Specification of the task blastn option automatically sets the amount of base pairs required to confirm a match between a query sequence and a reference sequence as 11bp[38]. This is referred to as word_size. The lower the word_size set, the more homologous sequences will be detected regardless of high fragmentation and DNA damage patterns. While this is beneficial for a study assessing aDNA sequences,

the lower the word_size set the more processing power and time is required. This introduces a point of compromise between highest homology and shortest processing time. We used a word_size setting of 14bp, half the value of the smallest available query sequence after cut threshold (28bp) was accounted for. Using the 10% representative, non-redundant sample file, this improved processing time by 10.42% when compared to use of default settings (53 hours to 48). The resulting BLAST files were compressed to reduce file size and to allow for faster transfer into MEGAN (16 minutes compared to 27).

Confirmation of a taxon's presence using multi-alignment authentication

The use of a multi-alignment authentication approach to metagenomic DNA sequences reduces the likelihood of misalignment errors[3,7,21]. The first authentication is undertaken using BLAST with MEGAN for the comparative analysis and assignment of genomic sequences. This is facilitated through the use of bit-scores, measuring the similarity between a query sequence and a reference[36]. BLAST results were imported into MEGAN and parameters were set using a conservative bit-score of 40 within the top 10% of best alignments, and the default "naïve" LCA algorithm. By using the LCA algorithm, reads are assigned across a taxonomy[1]. Sequences that have a bit-score within the specified percentage of best alignments within a taxonomy are binned into the lowest possible common ancestor position. Those sequences that align to multiple taxa within a grouping are binned into a higher taxonomic level until multiple assignments are no longer occurring. To reduce the chances of false positive identifications of taxa at the family or genus level a minimum of 1% of the total assigned reads was necessary to accept a taxon as present and for use in downstream analyses[3,21]. The resulting taxa identifications are used as the main taxon labels from this point forward, and they inform the user on which reference sequences to download for in-depth sequence alignment using BWA, the second authentication.

BWA is a software package for mapping low-divergent sequences to a large reference genome[29]. Because the software uses low-divergent sequences it is considered more stringent than mass comparative alignment tools such as BLAST. BWA has a variety of modes to select from depending on the data being queried. Typically for ancient DNA either BWA aln or BWA-MEM are utilised[8,39]. While BWA-

MEM is recommended for the use of sequences between 70-100bp[40], we recommend the use of BWA aln for aDNA sequences between 30-70bp for its ability to retain more reads based on the current literature[41] with a disabled seed length (-l 1000). Furthermore, it has been reported that the use of BWA aln over BWA-MEM conserves more single nucleotide polymorphisms (SNPs)[40,41] useful for in-depth analysis of a taxon. Seed length refers to the amount of base pairs within a read required to match sequentially between a query and reference sequence for a match to be made[29]. While the use of seed length can considerably improve the processing time required for alignment to complete[41,42], due to both the damaged nature of aDNA fragments and the multiple-origin sources of DNA, we recommend disabling seed length to allow for the alignment and conservation of damaged reads[43]. Upon completion of alignment a minimum threshold of 250 genomic hits are deemed necessary for a taxon to be processed downstream. This is because alignments with less than 250 reads were often found insufficient for mapDamage to plot damage patterns effectively. In some cases, MGmapper[44] was also utilised for additional authentication of taxa using the default settings and a minimum score of 20. This can, however, add to the overall processing time of genomic fragments and was not deemed a necessary step for the confirmation of a taxon for further analysis. Of additional importance, the use of mult-alignment authentication for the identification of a taxon reduces the potential for human derived selection bias by removing the ambiguity of reference genome selection. The requirement of multiple genomic hits using multiple authentication methods increases confidence in overall metagenomic results.

Identification of optimal mapping quality score for ancient metagenomic authentication

Mapping quality (MAPQ) refers to the degree of confidence that a sequence is correctly mapped to reference genome coordinates. In aDNA research a MAPQ of between 25-30 is typically used to extract aligned reads from poorly and non-aligned reads[21,43,45,46]. A MAPQ of 25-30 corresponds to a map accuracy of 99.68-99.9% (-10 log10 P), while allowing for representation of damaged sequences which typically score lower than their modern counterparts owing to deamination events at the terminal ends of sequences.

To identify the optimum mapping quality, aligned sequences were imported into DNA visualisation software[22,23] after in-depth alignment using BWA. Visualisation of DNA sequences allowed for the manual identification of mapping qualities associated with reads possessing misincorporation events as a consequence of deamination and lesion damage[10,11,47]. The majority of sequences possessing this damage fell into the MAPQ range of 25-30. Of note, however, was the aligner's inability to ascertain a read's single point of origin when it fell within a repeat element of a genome. This typically resulted in a MAPQ between 20-30. The heavily fragmented nature of aDNA and the occurrence of conserved sequence regions across taxa increases the likelihood of repeating sequences along a genome in conserved regions. As such, while the use of a higher MAPQ increases confidence in results, it can also result in the loss of authentic aDNA.

To test the percentage difference in authentic ancient reads passing into subsequent steps using a MAPQ of 25 and 30, sequences were processed and taxonomic assignments were identified using MEGAN. All taxonomic assignments were given a unique identification number and using the random number generator (SI, step 9) were selected for in-depth alignment to its associated reference genome using BWA. Aligned sequences were extracted from the originating file using SAMtool's[30] with a MAPQ specification of either 25 or 30. Sequences were sorted by coordinate and duplicates removed before authentication of DNA damage patterns using mapDamage[48,49]. Both MAPQ scores resulted in the identification of authentic aDNA fragments with a damage pattern >10%. Using a MAPQ of 25 resulted in the greater retention of sequences (23.43%) and an increased damage profile (8.33% C>T, 8.22% G>A) of DNA sequences when compared to a MAPQ of 30. As such, 25 was selected as the optimum MAPQ.

Conservation of mapped sequences during PCR duplicate removal

Duplicate sequences arise during the sequencing process when two or more copies of the same DNA molecule cross over onto different primer lawns within a flowcell during bridge PCR amplification. For this reason, we define a duplicate as the presence of two or more identical DNA sequences assigned to a single sample. PCR duplicates can be problematic in the assessment of authentic DNA sequences, most commonly

arising in the proportional over-representation of specific areas of coding[50]. This is referred to as amplification bias or base-composition bias. To ensure the integrity of authentic DNA data, and to mitigate the potential effects of duplicate sequences, they are bioinformatically removed.

The removal of PCR duplicates from NGS-derived data often involves the use of either SAMtools or Picard Tools[30,51]. Both these methods identify duplicate sequences by the external coordinate location of outer mapped reads. In cases where two or more sequences have the exact same 5' position coordinates, the sequence with the highest map quality score is retained and all other sequences are removed. SAMtools[30] differs in that the same function can also be accomplished on the reverse, 3' end of a sequence if specified. Furthermore, SAMtools is not applicable to unpaired sequences or those mapped to different chromosomes[52]. This means that a sequence with the same 5' start coordinate as another sequence, but mapped to a different chromosome, will be marked as a duplicate and removed. Picard Tools[51] avoids this issue by taking into account the intrachromosomal sequences. Additionally, Picard Tools takes into account soft-clipping of bases at the 5' position of mapped reads, performing calculations to locate where the 5' position would be if the entire sequence were mapped to the reference genome[52]. However, the use of external coordinate location as a method for duplicate removal in both commonly used methods cannot account for internal sequence variations such as SNPs, resulting in a potential loss of authentic DNA sequences. SNPs represent one of the most common types of genetic variation that can be used for detailed interpretation of a taxon[53]. The conservation of sequences increases the likelihood of retaining these SNPs. Additionally, these conserved reads play a crucial role in the statistical measurement of aDNA damage patterns, with a better assessment of deamination frequency over base position of reads[49]. This is especially true in the context of an exploratory metagenomic study without use of DNA capture techniques, where the DNA representative of a taxon is often small (1Kbp – 100Kbp) in number.

aweSAM[54] is a SAM assembly collapser, that uses a sequence's coordinates at the 5' and 3' end along with strand information as the unique insert identifiers, while keeping the sequence with the highest MAPQ. The use of multiple unique insert identifiers to locate a duplicate facilitates the conservation of reads that may be lost

through other duplicate removal tools, such as SAMtools and Picard Tools. We found that use of aweSAM resulted in the conservation of between 4.35 – 6.88% of total mapped sequences when compared to SAMtools and Picard Tools. However, the increased complexity of additional unique insert identifiers used has an unfavourable effect on the processing time and the memory required, which represents a trade-off between read conservation and duration of a process. Individual users may wish to adjust this compromise based on their requirements.

Assessment and authentication of aDNA damage patterns

DNA damage patterns were identified using mapDamage. mapDamage is a computation framework that quantifies aDNA damage patters among NGS-derived sequences[48,49], using a statistical model based on the damage profile of aDNA fragments described by Briggs[10].

Taxa are deemed ancient by assessing the frequency of C>T base substitutions at the 5' terminus along with G>A base substitutions at the 3' terminus of queried sequences. Depending on the source of DNA and the age of a context or specimen, different damage frequencies may be set as a minimum threshold to accept a sample as ancient[47]. Here we use two frequencies: a lower threshold of ≥0.05 at both terminal ends (representing 5% damage), and a higher threshold of ≥0.10 for terminal ends (representing 10% damage). As the extraction method[9,20] used in this study is designed for the exploration of ancient metagenomic DNA within anthropogenic sediments without the use of DNA capture techniques, yields of authentic ancient sequences can be lower. As such, ≥0.05 can be used as the lowest threshold for a taxa to be considered potentially ancient for further study using subsequent DNA capture techniques. The reduced potential for selection biases using this exploratory method makes the lower threshold an acceptable compromise. Ultimately, if a taxa reached the higher threshold of ≥0.10 it can be considered definitively ancient in origin.

Conclusion

The bioinformatic pipeline demonstrated here actively conserves reads by identifying the shortest DNA sequences available. This pipeline displays a substantial decrease in the amount of time required for the processing of metagenomic DNA through the

generation of non-redundant 10% representative files and optimisation of BLAST parameters. Multiple-alignment authentication ensures confidence in the authenticity of taxa identifications. Furthermore, the additional conservation of aDNA sequences is achieved through the use of the optimal MAPQ setting, and the use of multiple sequence identifiers within the duplicate removal process.

This bioinformatic pipeline can be used in the exploratory assessment of metagenomic aDNA, when used in conjunction with an extraction method without the use of DNA capture techniques. The use of two damage thresholds allows for the future selection of DNA probes for subsequent in-depth metagenomic studies. The use of open-sourced bioinformatic software throughout the pipeline reduces the cost burden of many bioinformatic software packages, and thus increases the accessibility of metagenomic analyses.

Materials and Methods

Genomic material tested

The genomic material used for the authentication of this bioinformatic pipeline originate from anthropogenic sediments taken from two archaeological sites: Drumclay Crannóg, Co. Fermanagh, Northern Ireland and Satsurblia Cave, Imereti, Georgia.

Drumclay Crannóg is an Irish Early-Medieval to Pre-Industrial archaeological site located in county Fermanagh, Northern Ireland (54°21′33.18″N 7°37′24.18″W). Anthropogenic sediments were secured in bulk from six locations representing a primary occupation layer dating to the CE 10-11th C; The garden (SN4818), the garden pathway to the house (SN3746), the wall/ wall packing of the house (SN4526), the northern compartment within the house (SN4537), the hearth (SN4551), and the southern ‘bed’ compartment (SN4592).

Satsurblia Cave is an Upper-Palaeolithic (29,000 BCE – 14,000 BCE) archaeological cave site located in the geographic region of the Southern Caucasus (42°22′38.05″N 42°36′3.40″E). Anthropogenic sediments were secured in bulk from two areas of anthropogenic activity dating to approximately the same time-frame; one associated with tool processing (Area A; SAT17 LS30-35) and the other associated with hearth use (Area B; SAT17 LS36-40)[55].

All anthropogenic sediments were excavated, sampled and stored in bulk, following conventional archaeological excavation techniques and standards[56,57].

Extraction, preparation and sequencing of genomic material

DNA extraction, library preparation and indexing steps were undertaken in an EU grade B (ISO 5) clean room under EU grade A (ISO 5) unilateral air-flow hoods at a dedicated aDNA laboratory, University College Dublin (UCD), Ireland. The laboratory surfaces were periodically cleaned with 10% sodium hypochlorite solution, DNA-OFF (MPBIO 11QD0500) and 70% ethanol, and all utensils and equipment were treated likewise after use and sterilised by UV irradiation when not in use. Tyvek suits (DuPont TYV217), hair nets (Superior bouffant01), face masks (Superior SKU83524), and nitrile gloves were used to limit contamination. PCR and subsequent steps were undertaken in an EU grade C laboratory (ISO 7) due to increased sample stability upon amplification. Extraction of DNA was performed as outlined by Dabney[31] with optimisations described by Collin[9,20] and libraries prepared using Meyer and Kircher[58]. PCR Amplification was performed as outlined by Gamba[13] at a rate of 15 cycles and cleaned as specified by Collin[9,20]. Analysis of PCR reaction concentrations were performed on an Agilent 2100 Bioanalyser following the instructions of the manufacturer. Based on these concentrations, samples were pooled into a 20ng working solution. Concentration and molarity (nmol/L) of the working solution were ascertained using the Bioanalyser and a Qubit4 for fluorometric quantification following manufacturer guidelines. Sequencing was undertaken at UCD Conway Institute of Biomolecular and Biomedical Research on an Illumina NextSeq 500/550 using the high output v2 (75 cycle) reagent kit (Illumina TG-160-2005).

Computational hardware specifications

Non-BLAST process analyses were performed in UCD using a Mac mini (late 2014) with 3GHz Intel Core i7 processor, 16GB 1600 MHz DDR3 RAM and a total storage of 1TB. BLAST processing was performed using the University of Manchester CSF3 system: BLAST-process v2.9.0 at “Skylake” facility using parallel job functionality on 16 or 32 core configurations using a Single Note Multi-core (SMP) setup with 6GB per core. Skylake facility is made up of 864 cores: 27 nodes of 2×16-core Intel Xeon Gold

6130 CPU @ 2.10GHz + 192GB RAM + 100Gb/s (4X EDR) mlx5 Mellanox InfiniBand. The total storage available for files in the lustre file system (known as scratch) is 692TB. CSF Nodes are based on CentOS Linux 7.4.1708.

Identification of the smallest retrievable aDNA fragment

Adapter sequences were removed from raw sequencing files using cutadapt v1.9.1[24] with a minimum sequence length or "cut" threshold of 0bp and imported into FastQC[25] for smallest sequence length available (15bp). Samples were re-cut using the smallest identified sequence length above (15bp), with a minimum overlap of 1 to reduce the over-cutting of bases that closely match an adapter sequence. Resulting sequences were comparatively analysed to a localised NCBI genomic database (2019) using BLAST[26–28] task_blastn and a word_size of 14. BLAST results were input into MEGAN v6.2.13[1] using a bit-score of 40 within the top 10% of best alignments, and the default "naïve" LCA algorithm. All *Mammalia* and *Plantae* assignments at the genus taxonomic level were assigned a unique identification number. Using the random number generator specified in SI (step 9), unique numbers were selected and aligned to their associated reference genome using BWA v07.5a.r405[29] aln function with a disabled seed (-l 1000) and converted to SAM format using BWA samse function. Aligned sequences were individually imported into UGENE v1.32[22] and Geneious R10 software for visualisation and manual identification of (C>T, G>A) misincorporation events at the terminal ends of sequences. The smallest identifiable fragment with misincorporation characteristics indicative of deamination DNA damage was deemed the smallest retrievable aDNA fragment, and thus considered the minimum sequence length threshold for subsequent applications of cutadapt.

Validation of the non-redundant, 10% representative sample file for comparative analysis

Adapter sequences were removed from raw sequencing files using cutadapt v1.9.1[24] with a cut of 28bp and a minimum overlap of 1. Non-redundant sample files were created using the fastx_collapser function of the Fastx toolkit v0.0.13[34] and split into 10 (-n 10) new files of similar size using the split function of pyfasta v0.5.2[35]. Comparative alignment to a localised NCBI genomic database (2019) was undertaken

using BLAST[26–28] task_blastn and a word_size of 14. BLAST results were input into MEGAN v6.2.13[1] using a bit-score of 40 within the top 10% of best alignments, and the default "naïve" LCA algorithm. Mean percentage difference between the genomic hits achieved from the 10% representative file and the total genomic assignments achieved from the unsplit non-redundant file generated after fastx_collapser was calculated. Standard deviation was further calculated in order to obtain the standard error of the mean.

Identification of optimal mapping quality score

Adapter sequences were removed using cutadapt v1.9.1[24], a cut of 28bp and minimum overlap of 1. Non-redundant sample files were created using the fastx_collapser function of Fastx toolkit v0.0.13[34] and split into 10 (-n 10) files using the split function of pyfasta v0.5.2[35]. Sequences were comparatively analysed to a localised NCBI genomic database (2019) using BLAST[26–28] task_blastn and a word_size of 14. BLAST results were input into MEGAN v6.2.13[1] using a bit-score of 40 within the top 10% of best alignments, and the default "naïve" LCA algorithm. To reduce the chances of false positive identifications of taxa at the family or genus level a minimum of 1% of the total assigned reads was necessary to accept a taxon as present and use for downstream analyses[3,21]. Sequences passing this threshold were aligned to their corresponding genome using the original cut fasta file and BWA v07.5a.r405[29] aln function with a disabled seed (-l 1000). SAI file alignments were converted to SAM format using BWA samse function. Aligned SAM files were imported into UGENE v1.32[22] and Geneious R10[23] software for visualisation of damaged sequences and their aligner-assigned MAPQ.

To compare MAPQ of 25 and 30, taxonomic assignments derived from MEGAN analysis were assigned a unique identifier number and randomly selected using the random number generator (SI, step 9), for BWA alignment and conversion to SAM as specified above. Mapped sequences were extracted using SAMtools v1.3.1[30] view function and a MAPQ of 25 and 30 to create two separate comparative files. Sequences were sorted by coordinate using SAMtools sort function and duplicate sequences removed using aweSAM_collapser.sh shell script[54]. MapDamage2.0[49] was used to quantify DNA damage through the presence of misincorporation events (C>T,

G>A) at the terminal ends of the sequences. Percentage difference of total sequences identified and ancient damage patterns was calculated using the mean variation between MAPQ 25 and 30.

Comparison of duplicate removal tools

Adapter sequences were removed using cutadapt v1.9.1[24] with a cut of 28bp and minimum overlap of 1. Creation of mon-redundant sample files was facilitated by the fastx_collapser function of Fastx toolkit v0.0.13[34] and subsequently split into 10 (-n 10) files using split function of pyfasta v0.5.2[35]. Comparative alignment to a localised NCBI genomic database (2019) was facilitated by BLAST[26–28] using task_blastn function and a word_size of 14. BLAST results were input into MEGAN v6.2.13[1] using a bit-score of 40 within the top 10% of best alignments, and the default "naïve" LCA algorithm. A minimum of 1% of the total assigned reads was necessary to accept a taxon as present and use for downstream analyses[3,21]. Passing sequences were aligned to their corresponding genome using the original cut fasta file and BWA v07.5a.r405[29] aln function with a disabled seed (-l 1000). SAI file alignments were converted to SAM format using BWA samse. Sequences were filtered using a MAPQ of 25 and sorted using SAMtools v1.3.1[30]. A threshold of 250 total aligned reads were required to proceed with downstream analysis.

Duplicate sequences were removed from resulting files using three methods:

1. SAMtools' "rmdup" function with the option for removal of single-end matches at the 5' location only[30]
2. Picard Tools "MarkDuplicates" function with the option for removing duplicates from the output file[51]
3. aweSAM_collapser as specified by developers[54]

Percentage difference was calculated from the variation between reads passing each duplicate removal process.

The fully developed bioinformatic pipeline

Adapter sequences are removed using cutadapt v1.9.1[24] with a minimum sequence length of 28bp based on smallest ancient fragments retrievable. A minimum overlap of 1 is used to reduce over-cutting of bases that closely match an adapter sequence.

Non-redundant samples are created using the fastx_collapser function of the Fastx toolkit v0.0.13[34] and split into 10 (-n 10) files of similar size using the split function of pyfasta v0.5.2[35]. The resulting files are considered representative of 10% of the initial non-redundant sample sequences.

The most commonly occurring genera are identified by cross-referencing trimmed representative non-redundant sequencing data with a localised genomic database downloaded (2019) from NCBI. Cross-referencing for genera identification is facilitated through BLAST[26–28] using task_blastn and a word_size of 14.

Resulting output files are imported into MEGAN Community Edition v.6.2.13[1] for taxonomic assessment . LCA parameters use a bit-score of 40 within the top 10% of best alignments, and the default “naïve” LCA algorithm. To reduce chances of false positive identifications of taxa at the family or genus level a minimum of 1% total assigned reads is necessary to accept a taxon as present and use for downstream analyses[3,21]. MGmapper[44] can be employed in additional identifications of genera using default settings and a minimum score of 20.

Following acceptance of a taxon, samples are aligned to their corresponding genome using the original trimmed fasta file and BWA v07.5a.r405[29] aln function with a disabled seed (-l 1000) allowing damaged sequences to align. Sequences are mapped and filtered for a minimum MAPQ of 25, then sorted using samtools v1.3.1[30]. At this point a minimum threshold of 250 total aligned reads are required to proceed with downstream processes. Duplicates are removed using aweSAM_collapser.sh shell script[54], allowing users to read from both the 5’ and 3’ coordinates and retain the read with highest MAPQ. MapDamage2.0[49] is used to quantify DNA damage through the presence of C to T substitutions on the 5’ end and G to A substitutions on the 3’ end of the sequences. A minimum value of 5-10% on both ends is used for a taxon to be identified as ancient. Read lengths are calculated through cumulative observation and quartile calculation. Phred quality scores and %GC are assessed using FastQC[25] (Figure 3).

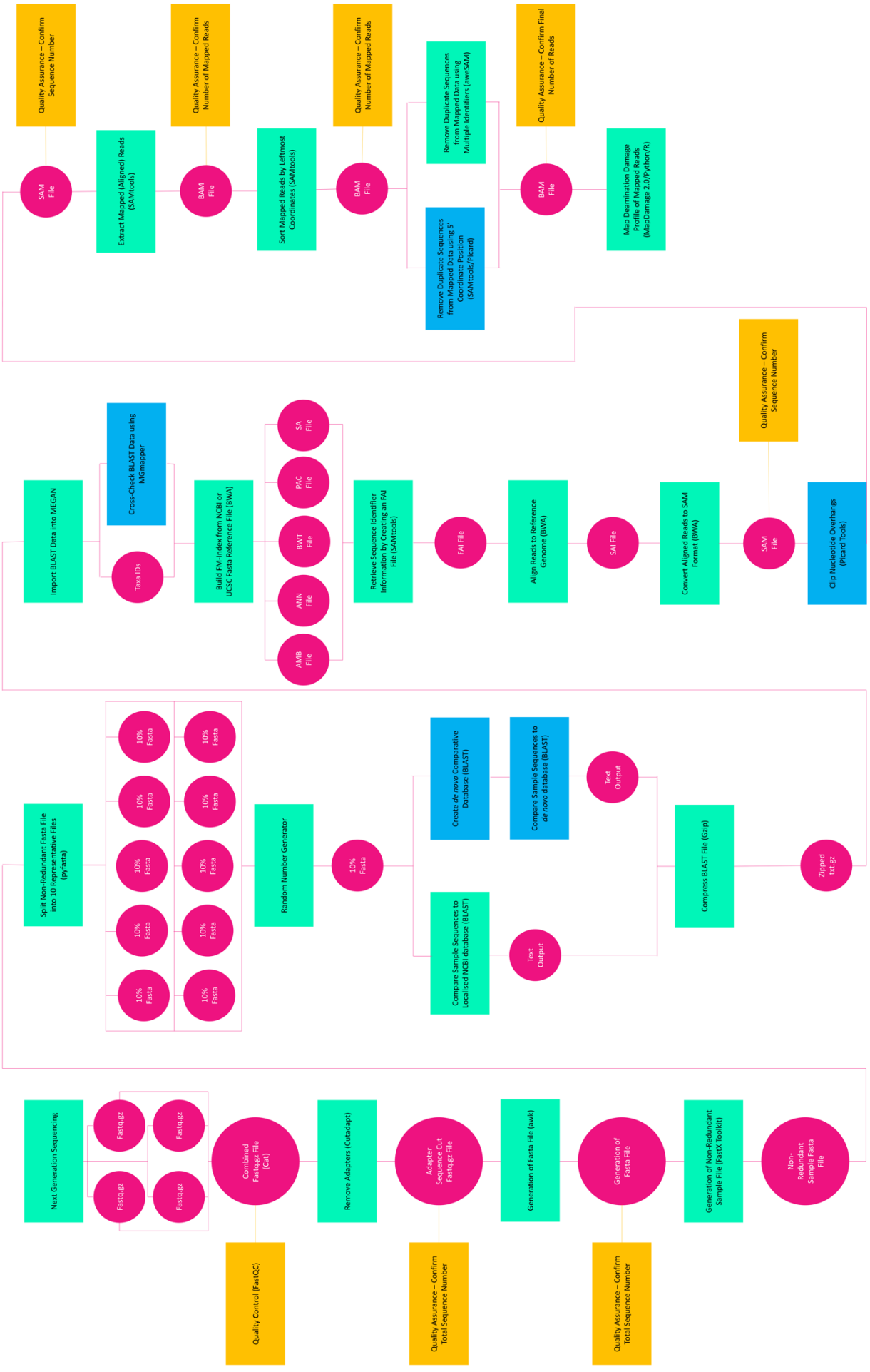

Figure 3. Schematic Overview of the Developed Bioinformatic Pipeline. Processes and their associated software are represented by rectangular boxes. Teal boxes represent processes that form part of the core bioinformatic pipeline. Yellow boxes represent quality control and assurance steps. Blue boxes represent optional processes within the bioinformatic pipeline (see SI). Pink circles represent process output generated (files).

Acknowledgements

We would like to thank Dr Eileen Reilly (Site Entomologist), Dr Nora Bermingham (Site Director) and Jacqueline McDowell (Historic Environment Division, Department for Communities, Northern Ireland) for access to the Drumclay samples used within this study. Likewise, we thank Dr Mareike Stahlschmidt (Site Geoarchaeologist) for providing the Satsurblia samples used within this study. We thank the Computational Shared Facility at the University of Manchester for use of its systems and we would further like to thank George O. T. Merces for his insights into the structure of this paper. This project was funded by the Medical Trainee PhD Scholarship, Anatomy, School of Medicine, UCD awarded to T.C.C.

Author Contributions

T.C.C. conceived the project. R.N.M.F. and R.P. supervised the project. T.C.C. wrote the paper with contributions from K.D., J.D.O.R., R.N.M.F and R.P. Ancient DNA analysis was performed by T.C.C. Bioinformatic pipeline development performed by T.C.C. Bioinformatic training provided by K.D. and J.D.O.R. Samples bioinformatically processed in University College Dublin by T.C.C. and in University of Manchester CSF3 by K.D. and T.C.C. Bioinformatic assistance provided by J.D.O.R. and K.D. Sample access facilitated by T.M.

References

1. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN Analysis Of Metagenomic Data. Genome Res. 2007;17(3):377–86.

2. Pratas D, Hosseini M, Grilo G, Pinho AJ, Silva RM, Caetano T, et al. Metagenomic Composition Analysis of an Ancient Sequenced Polar Bear Jawbone from Svalbard. Genes (Basel). 2018;9(9):445.

3. Slon V, Hopfe C, Weiß C, Mafessoni F, Rasilla M, Lalueza-Fox C, et al. Neandertal and Denisovan DNA from Pleistocene sediments. Science. 2017;356:eaam9695.

4. Parducci L, Bennett K, Ficetola GF, Alsos I, Suyama Y, Wood J, et al. Ancient Plant DNA in Lake Sediments. The New phytologist. 2017;214.

5. Pratas D, Pinho A. Metagenomic Composition Analysis Of Sedimentary Ancient DNA from The Isle of Wight. In 2018. p. 1177–81.

6. Parducci L, Alsos I, Unneberg P, Pedersen M, Han L, Lammers Y, et al. Shotgun Environmental DNA, Pollen, and Macrofossil Analysis of Lateglacial Lake Sediments from Southern Sweden. Frontiers in Ecology and Evolution. 2019;7.

7. Harbert RS. Algorithms And Strategies In Short-Read Shotgun Metagenomic Reconstruction of Plant Communities. Appl Plant Sci. 2018;6(3):e1034–e1034.

8. Hansen HB, Damgaard PB, Margaryan A, Stenderup J, Lynnerup N, Willerslev E, et al. Comparing Ancient DNA Preservation in Petrous Bone and Tooth Cementum. PLoS One. 2017;12(1):e0170940.

9. Collin TC, Stahlschmidt MC, Pinhasi R, Feeney RMN. Metagenomic Study of Anthropogenic Sediments: Insights into Public Health and Lifestyle. In Hinxton, Cambridge, UK: Wellcome Genome Campus; 2017.

10. Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, et al. Patterns Of Damage In Genomic DNA Sequences from a Neandertal. Proc Natl Acad Sci USA. 2007;104(37):14616.

11. Dabney J, Meyer M, Pääbo S. Ancient DNA Damage. Cold Spring Harb Perspect Biol. 2013;5(7):a012567.

12. Shapiro B, Hofreiter M. A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA. Science. 2014;343(6169):1236573.

13. Gamba C, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, et al. Genome Flux and Stasis in a Five Millennium Transect Of European Prehistory. Nature Communications. 2014;5(1):5257.

14. O'Rourke DH, Hayes MG, Carlyle SW. Ancient DNA Studies in Physical Anthropology. Annu Rev Anthropol. 2000;29(1):217–42.

15. Höss M, Jaruga P, Zastawny TH, Dizdaroglu M, Pääbo S. DNA Damage and DNA Sequence Retrieval from Ancient Tissues. Nucleic Acids Res. 1996;24(7):1304–7.

16. Hansen A, Willerslev E, Wiuf C, Mourier T, Arctander P. Statistical Evidence for Miscoding Lesions in Ancient DNA Templates. Mol Biol Evol. 2001;18(2):262–5.

17. Gilbert MTP, Hansen AJ, Willerslev E, Rudbeck L, Barnes I, Lynnerup N, et al. Characterization of Genetic Miscoding Lesions Caused by Postmortem Damage. Am J Hum Genet. 2003;72(1):48–61.

18. Ziesemer KA, Mann AE, Sankaranarayanan K, Schroeder H, Ozga AT, Brandt BW, et al. Intrinsic Challenges in Ancient Microbiome Reconstruction Using 16S rRNA Gene Amplification. Scientific Reports. 2015;5(1):16498.

19. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into Biases and Sequencing Errors for Amplicon Sequencing with the Illumina MiSeq Platform. Nucleic Acids Res. 2015;43(6):e37.

20. Collin TC, Pinhasi R, Feeney RMN. Optimisation of Metagenomic Next Generation Sequencing Shotgun Techniques for the Study of Ancient Anthropogenic Sediments. American Journal of Physical Anthropology Supplement. 2016;S62:119.

21. Stahlschmidt MC, Collin TC, Fernandes DM, Bar-Oz G, Belfer-Cohen A, Gao Z, et al. Ancient Mammalian and Plant DNA from Late Quaternary Stalagmite Layers at Solkota Cave, Georgia. Scientific Reports. 2019;9(1):6628.

22. Okonechnikov K, Golosova O, Fursov M. Unipro UGENE: A Unified Bioinformatics Toolkit. Bioinformatics. 2012;28(8):1166–7.

23. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data. Bioinformatics. 2012;28(12):1647–9.

24. Martin M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. EMBnet.journal. 2011;17(1):3.

25. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Internet]. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol. 1990;215(3):403–10.

27. Altschul SF, Gish W. Local Alignment Statistics. Methods Enzymol. 1996;266:460–80.

28. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Res. 1997;25(17):3389–402.

29. Li H, Durbin R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. Bioinformatics. 2009;25(14):1754–60.

30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

31. Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, et al. Complete Mitochondrial Genome Sequence of a Middle Pleistocene Cave Bear Reconstructed from Ultrashort DNA fragments. Proc Natl Acad Sci USA. 2013;110(39):15758.

32. Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Pääbo S. DNA Sequences from Multiple Amplifications Reveal Artifacts Induced by Cytosine Deamination in Ancient DNA. Nucleic Acids Res. 2001;29(23):4793–9.

33. Lovett S. Encoded errors: Mutations and Rearrangements Mediated by Misalignment at Repetitive DNA Sequences. Molecular microbiology. 2004;52(5):1243–53.

34. Hannon GJ. FASTX-Toolkit [Internet]. 2010. Available from: http://hannonlab.cshl.edu/fastx_toolkit

35. Pedersen B. Pyfasta [Internet]. 2010. Available from: https://pypi.org/project/pyfasta

36. Xiong J. Essential Bioinformatics. Cambridge: Cambridge University Press; 2006.

37. Welcome to NCBI [Internet]. National Center for Biotechnology Information. [cited 2019 Jul 30]. Available from: https://www.ncbi.nlm.nih.gov/

38. BLAST Command Line Applications User Manual. National Center for Biotechnology Information; 2008.

39. Weiß C, Dannemann M, Prüfer K, Burbano H. Contesting the Presence of Wheat in the British Isles 8,000 Years Ago by Assessing Ancient DNA Authenticity from Low-Coverage Data. eLife. 2015;4:e10005.

40. Li H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. arXiv. 2013;1303:e1303.3997.

41. Robinson KM, Hawkins AS, Santana-Cruz I, Adkins RS, Shetty AC, Nagaraj S, et al. Aligner Optimization Increases Accuracy and Decreases Compute Times in Multi-Species Sequence Data. Microb Genom. 2017;3(9):e000122.

42. Luo R, Wong Y-L, Law W-C, Lee L-K, Cheung J, Liu C-M, et al. BALSA: Integrated Secondary Analysis for Whole-Genome and Whole-Exome Sequencing, Accelerated by GPU. PeerJ. 2014;2:e421.

43. Schubert M, Ginolhac A, Lindgreen S, Thompson JF, Al-Rasheid KAS, Willerslev E, et al. Improving Ancient DNA Read Mapping Against Modern Reference Genomes. BMC Genomics. 2012;13:178.

44. Petersen TN, Lukjancenko O, Thomsen MCF, Maddalena Sperotto M, Lund O, Møller Aarestrup F, et al. MGmapper: Reference Based Mapping and Taxonomy Annotation of Metagenomics Sequence Reads. PLOS ONE. 2017;12(5):e0176469.

45. Star B, Boessenkool S, Gondek A, Nikulina E, Hufthammer A, Pampoulie C, et al. Ancient DNA Reveals the Arctic Origin of Viking Age Cod from Haithabu, Germany. Proceedings of the National Academy of Sciences. 2017;114:1–6.

46. Pinhasi R, Fernandes D, Sirak K, Novak M, Connell S, Alpaslan-Roodenberg S, et al. Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone. PLOS ONE. 2015;10(6):e0129102.

47. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. PLOS ONE. 2012;7(3):e34131.

48. Ginolhac A, Rasmussen M, Gilbert M, Willerslev E, Orlando L. mapDamage: Testing for Damage Patterns in Ancient DNA Sequences. Bioinformatics (Oxford, England). 2011;27:2153–5.

49. Jonsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters. Bioinformatics. 2013;29(13):1682–4.

50. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and Minimizing PCR Amplification Bias in Illumina Sequencing Libraries. Genome Biol. 2011;12(2):R18.

51. Picard Tools [Internet]. Broad Institute; [cited 2019 Jul 30]. Available from: http://broadinstitute.github.io/picard/

52. Ebbert MTW, Wadsworth ME, Staley LA, Hoyt KL, Pickett B, Miller J, et al. Evaluating the Necessity of PCR Duplicate Removal From Next-Generation Sequencing Data and a Comparison of Approaches. BMC Bioinformatics. 2016;17(7):239.

53. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, et al. A General Approach to Single-Nucleotide Polymorphism Discovery. Nat Genet. 1999;23(4):452–6.

54. Enk J, Devault A. aweSAM_collapser [Internet]. 2013. Available from: https://gist.github.com/jakeenk/

55. Pinhasi R, Meshveliani T, Matskevich Z, Bar-Oz G, Weissbrod L, Miller CE, et al. Satsurblia: New Insights of Human Response and Survival Across the Last Glacial Maximum in the Southern Caucasus. PLoS One. 2014;9(10):e111271.

56. Standard and Guidance for Archaeological Excavation [Internet]. Chartered Institute for Archaeologists; 2014 [cited 2019 Jul 1]. Available from: https://www.archaeologists.net/sites/default/files/CIfAS&GExcavation_1.pdf

57. Policy and Guidelines on Archaeological Excavation [Internet]. Department of Arts, Heritage, Gaeltacht and the Islands. Ireland; 1999 [cited 2019 Jul 1]. Available from: https://www.archaeology.ie/sites/default/files/media/publications/excavation-policy-and-guidelines.pdf

58. Meyer M, Kircher M. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb Protoc. 2010;2010(6):pdb.prot5448.