Data in Brief 61 (2025) 111770

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

An extensive archaeological dental calculus dataset spanning 5000 years for ancient human oral microbiome research

Francesca J. Standeven[a,1], Gwyn Dahlquist-Axe[a,1], Jessica Hendy[b], Sarah Fiddyment[b], Malin Holst[b,c], Krista McGrath[d], Matthew Collins[e,f], Amy Mundorff[g], Anita Radini[h], Josef Wagner[i,j], Conor J. Meehan[k], Andrew Tedder[a], Camilla F. Speller[l,*]

[a] School of Chemistry and Biosciences, University of Bradford, UK
[b] BioArCh, Department of Archaeology, University of York, York, UK
[c] York Osteoarchaeology Ltd, UK
[d] Department of Prehistory and Institute of Environmental Science and Technology (ICTA-UAB), Universitat Autònomade Barcelona, Bellaterra, Spain
[e] Section for Evolutionary Genomics, the GLOBE Institute, University of Copenhagen, København, Denmark
[f] Department of Archaeology, University of Cambridge, Cambridge, UK
[g] Department of Anthropology, College of Arts and Sciences, University of Tennessee, Knoxville, TN, USA
[h] UCD School of Archaeology, College of Social Sciences and Law, University College Dublin, Dublin, Republic of Ireland
[i] Enteric Diseases, Murdoch Children's Research Institute, Parkville, Australia
[j] Department of Paediatrics, The University of Melbourne, Parkville, Australia
[k] Department of Biosciences, Nottingham Trent University, UK
[l] Department of Anthropology, University of British Columbia, Canada

ARTICLE INFO

Article history:
Received 7 November 2024
Revised 7 March 2025
Accepted 3 June 2025
Available online 12 June 2025

ABSTRACT

Archaeological dental calculus can provide detailed insights into the ancient human oral microbiome. We offer a multi-period, multi-site, ancient shotgun metagenomic dataset consisting of 174 samples obtained primarily from archaeological dental calculus derived from various skeletal collections in the United Kingdom. This article describes all the materials used including the skeletons' historical period and burial

* Corresponding author.
E-mail address: camilla.speller@ubc.ca (C.F. Speller).
Social media: @fstandeven193 (F.J. Standeven)
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.dib.2025.111770

Dataset link: PRJEB1716 - Sequencing_ancient_DNA_calculus_samples (Original data)
Dataset link: PRJEB26093 - A plaque on both your houses: Exploring the history of urbanisation and infectious diseases through the study of archaeological dental tartar (Original data)
Dataset link: Supplementary information for An extensive archaeological dental calculus dataset spanning 5000 years for ancient human oral microbiome research (Original data)

Keywords:
Dental calculus
Metagenomics
Ancient DNA
Bioinformatics
Oral microbiome
Bioarchaeology

location, biological sex, and age determination, data accessibility, and additional details associated with environmental and laboratory controls. In addition, this article describes the laboratory and bioinformatic methods associated with the dataset development and discusses the technical validity of the data following quality assessments, damage evaluations, and decontamination procedures. Our approach to collecting, making accessible, and evaluating bioarchaeological metadata in advance of metagenomic analysis aims to further enable the exploration of archaeological science topics such as diet, disease, and antimicrobial resistance (AMR).

Specifications Table

Subject	Archaeology and computational biology
Specific subject area	Archaeology and computational biology combine archaeological methods with advanced bioinformatics to explore ancient biological data such as ancient DNA (aDNA). This interdisciplinary field involves the study of aDNA from archaeological remains, and in the context of this data, from historical dental calculus to contribute to research that is interested in reconstructing past human populations and the evolutionary processes of microorganisms in the oral cavity. Computational biology involves using bioinformatic methods to efficiently process large genomic datasets to analyse the genetic data extracted from ancient and modern samples. By merging computational biology with archaeological data, we can gain a deeper understanding of the interactions between humans and microorganisms over time.

Type of data	• Raw paired-end fastq files
	• Filtered (trimmed) paired-end fastq files
	• Raw and filtered FastQC reports
	• mapDamage plots
	• Large main metadata table
	• Tables
	  ○ Decontamination information
	  ○ Quality control information
	  ○ Damage authentication information
	• Figures
	  ○ Map of materials
	  ○ Trimming and decontamination pipeline
	  ○ Box and whisker plot for trimming and decontamination statistics
	  ○ Multi-fastQC report plot
	• Example of a damage authentication C-T transition frequency plot

Data collection	Dental calculus was removed from the teeth of individuals excavated from various archaeological sites across the UK. aDNA was extracted and sequenced in clean aDNA laboratories (see **Laboratory method** under **Experimental design and methods** section), and raw fastq files were collected from the European Nucleotide Archive (ENA) linked to several projects (see **Data description** section). Sequences were then checked for quality, decontaminated, and screened for damage (see under **Experimental design and methods** section).

(*continued on next page*)

Data source location	Ancient dental calculus came from individuals located across several archaeological excavations across the UK (see map in **Figure 1**). Modern samples came from the University of Tennessee donated skeletal collection within the Forensic Anthropology Centre Anthropology Research Center. Laboratory procedures were carried out at the University of York, York, UK and the University of British Columbia, Vancouver, Canada. Digital files were downloaded using the high-performance computing service at the University of Bradford, Bradford, UK.
Data accessibility	The multiple analogue and digital repositories, reports and references related to skeletal individuals, and ENA projects associated with digital data in this publication can be found in Table S1. Raw metagenomic data has been archived on the ENA as fastq files in projects PRJEB1716, PRJEB12831, and PRJEB75938. All bioinformatic coding scripts are available on GitHub (https://github.com/DrATedder/dental_calculus_dataset/tree/main) for all pre-processing (quality control, decontamination) and post-processing (mapDamage) protocols. Quality control (FastQC), quality filtering (Fastp) and post-processing (mapDamage) reports can be found on Dryad (DOI: 10.5061/dryad.jdfn2z3mk).
Related research articles	

1. Value of the Data

- This dataset is designed for research utilising microbial aDNA identified in ancient dental calculus, enabling the study of oral microbiomes from both burial populations and individuals, contributing valuable insights to archaeological science.
- This data supports the analysis of microbial interactions in the oral cavity and includes metagenomic data from tooth roots and bone to serve as controls, raising awareness to mitigate contamination risks.
- By making this data accessible, we aim to help advance the field of ancient DNA research and enable the exploration of archaeological topics such as diet, disease, and AMR.

2. Background

This dataset was compiled for research capitalising on the abundant microbial aDNA encapsulated in ancient dental calculus [1,2] — a common archaeological material recovered from skeletons spanning most pre-historic and historical periods [3] — to gain potential insights into millennia of oral microbial evolution. Our sizable dataset allows for the comparison of oral microbiomes across multiple historic periods in England spanning 2500 BCE to 1900 CE (and the 20th – 21st century). This period encompasses significant historical events such as the agricultural revolution, industrialisation, and medical discoveries, which permit the analysis of changes in dietary, demographic, environmental, and socioeconomic aspects over time. Therefore, this dataset focuses on expanding the availability of metagenomes dating to the period of cultural and ecological transition during the Industrial Revolution in England and samples from earlier and later periods that contextualise the Industrial Era. In addition to ancient dental calculus, our dataset also includes metagenomic data retrieved from associated archaeological tooth roots and skeletal elements which act as controls to ensure that reliable data can be obtained, as the inclusion of these materials allows for ancient and modern environmental contaminants to be identified and removed [4].

3. Data Description

All our samples have been archived in the European Nucleotide Archive [5] as fastq files in projects PRJEB1716, PRJEB12831, and PRJEB75938 (see Table S1). They encompass 174 samples (348 fastq files using paired end reads) from modern ($n = 10$) and archaeological dental calculus ($n = 133$) and tooth ($n = 2$) samples, environmental controls (bone; $n = 7$), and laboratory

controls (extraction and library blanks; $n = 22$). The ancient samples were excavated from multiple archaeological sites across the UK spanning 5000 years, including the Bronze Age (ca. 2300 BCE–700 BCE), Iron Age (ca. 700 BCE–43 CE), Roman Period (ca. 43 CE–410 CE), Anglo-Saxon Period (ca. 410 CE–1066 CE), Viking Age (ca. 793 CE–1066 CE), Medieval Period (ca. 1066–1485), Industrial (ca. 1750–1850), Post-industrial (ca. 1837–1901). All archaeological sites and periods are displayed in Fig. 1. Table S1 is a summary table with all the necessary information on the individuals in our dataset including: sample code and type; skeletal ID, age, and biological sex; ENA project and number; type of tooth and weight of calculus sampled; adapter sequence; location and time period of archaeological site; and current repository and skeletal report/reference associated with individual skeletons. Samples from modern human remains are derived from the University of Tennessee donated skeletal collection within the Forensic Anthropology Centre Anthropology Research Center (20th – 21st century). Proteomic analysis of a subset of these samples have also been published in Hendy et al. (2018) [6] (see Table S1).

In addition to the use of raw fastq files and the compilation of Table S1, this work also produced raw and filtered FastQC reports (**Supplementary Material 1_FastQC_reports_raw.zip** and **Supplementary Material 2_FastQC_reports_trimmed.zip**) and fastp reports (**Supplementary Material 3_Fastp_reports.zip**) to quality check the data, mapDamage plots (**Supplementary Material 4_mapdamage.zip**) to authenticate ancient DNA; these reports have been archived in the Dryad data repository (DOI: 10.5061/dryad.jdfn2z3mk). The relevant data from these files were then put into tables: the decontamination information on samples and their matching laboratory/environmental blanks into Table 1; quality control results showing percentage of mismatched sequences and reads past filters, and bp and read length pre and post-trimming and pre and post-decontamination, into Table 2; and cytosine-thymine frequency transition values for damage authentication interpretation into Table 3. Figures include a main map of archaeological sites (Fig. 1), a trimming and decontamination pipeline (Fig. 2), a box and whisker plot for trimming and decontamination statistics (Fig. 3), a multi-FastQC report plot for overall quality scores across the dataset (Fig. 4), and an example of a damage authentication C-T transition frequency plot (Fig. 5).

4. Experimental Design, Materials and Methods

4.1. Laboratory method

4.1.1. Sample preparation and contamination controls

Samples of dental calculus were removed from the teeth using sterilised dental picks and stored in individual 2.0 mL Eppendorf tubes. Sample preparation and DNA extractions were conducted in the BioArCh ancient DNA laboratory at the University of York, and the Ancient DNA and Proteins facility at the University of British Columbia. Both labs are dedicated to the analysis of ancient biomolecules, and the introduction of contamination into the workspace is minimised by the use of protective clothing, including Tyvek suits, gloves, masks, and hairnets. The labs are also equipped with UV filtered ventilation and positive airflow, as well as with dedicated equipment and bench UV lights; countertops and other surfaces in the lab are routinely wiped down with dilute sodium hypochlorite. All reagents and equipment in the ancient DNA laboratory are dedicated solely to the study of degraded DNA. Multiple blank DNA extractions and negative PCR controls are run alongside the ancient samples to identify potential contamination at each stage of the procedure.

4.1.2. aDNA extraction

DNA was extracted in batches, with two extraction blanks prepared alongside each batch. Samples of dental calculus and bone were UV-sterilised for 1 min on each side. After crushing to a powder, samples were pre-digested for 5 min with 1 mL of 0.5 M EDTA to remove possible surface contamination. This pre-digestion supernatant was removed, and a further 1.1 mL of 0.5 M EDTA added and rotated at room temperature for seven days to fully demineralize.

Fig. 1. Map of dated archaeological sites across the UK from which the samples were obtained. More information on the exact number of samples from each site, skeletal ID, ENA accession number, and publication reference can be accessed in the supplementary data. The map was adapted from Hendy *et al.* 2018 [7] and the map outline was provided by Twinkl's free education resources: https://www.twinkl.co.uk/search?q=free.

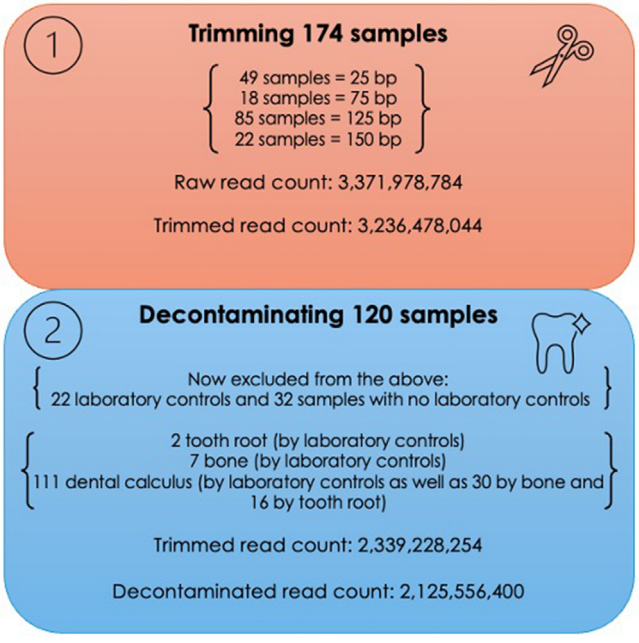

Fig. 2. Flow chart of raw, trimmed, and decontaminated read counts.

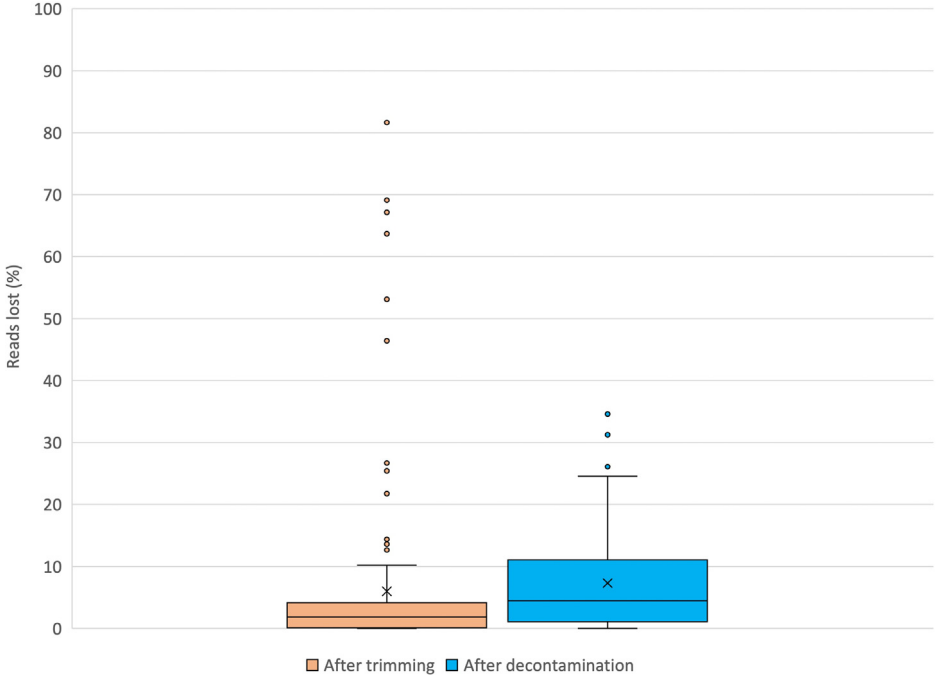

Fig. 3. Box and whisker plots showing proportion of reads lost following trimming and decontamination.

Table 1
Decontamination information on samples and their matching laboratory/environmental blanks.

Sample ID	Run accession number (European Nucleotide Archive)	Sample type	Site	Extraction/library blank decontaminated	Bone blank decontaminated
SK1203, FAO1	ERR1659119	calculus	Lower St Bride's Churchyard, London	no	no
SK1207, FAO2	ERR1659110	calculus	Lower St Bride's Churchyard, London	no	no
SK1215, FAO3	ERR1659111	calculus	Lower St Bride's Churchyard, London	no	no
SK1244.1, FAO4	ERR1659112	calculus	Lower St Bride's Churchyard, London	no	no
SK1526, FAO5	ERR1659113	calculus	Lower St Bride's Churchyard, London	no	no
SK1558, FAO6	ERR1659120	calculus	Lower St Bride's Churchyard, London	no	no
SK1641, FAO7	ERR1659114	calculus	Lower St Bride's Churchyard, London	no	no
SK1653, FAO8	ERR1659117	calculus	Lower St Bride's Churchyard, London	no	no
SK1655, FAO9	ERR1659115	calculus	Lower St Bride's Churchyard, London	no	no
SK1683, FAO10	ERR1659118	calculus	Lower St Bride's Churchyard, London	no	no
SK1785, FAO11	ERR1659116	calculus	Lower St Bride's Churchyard, London	no	no
SK1799, FAO12	ERR1681541	calculus	Lower St Bride's Churchyard, London	no	no
SK1872, FAO13	ERR1681532	calculus	Lower St Bride's Churchyard, London	no	no
SK1932, FAO14	ERR1681533	calculus	Lower St Bride's Churchyard, London	no	no
SK1999, FAO15	ERR1681542	calculus	Lower St Bride's Churchyard, London	no	no
SK2049, FAO16	ERR1681538	calculus	Lower St Bride's Churchyard, London	no	no
SK2134, FAO17	ERR1681534	calculus	Lower St Bride's Churchyard, London	no	no
SK2296, FAO18	ERR1681535	calculus	Lower St Bride's Churchyard, London	no	no
RLH208	ERR1681523	calculus	Royal London Hospital	eBK686	no
RLH349	ERR1681531	calculus	Royal London Hospital	eBK686	no
RLH365	ERR1681525	calculus	Royal London Hospital	eBK686	no
RLH367	ERR1681526	calculus	Royal London Hospital	eBK686	no
RLH386	ERR1681527	calculus	Royal London Hospital	eBK686	no
RLH397	ERR1681530	calculus	Royal London Hospital	eBK686	no
RLH103	ERR9638308	calculus	Royal London Hospital	eBK686	no
RLH131	ERR9638303	calculus	Royal London Hospital	eBK686	no
RLH135	ERR9638304	calculus	Royal London Hospital	eBK686	no
RLH340	ERR9638305	calculus	Royal London Hospital	eBK686	no
RLH421	ERR9638307	calculus	Royal London Hospital	eBK686	no
RLH572	ERR9638300	calculus	Royal London Hospital	eBK686	no
FW68C	ERR9638278	calculus	Fewston, North Yorkshire	no	no
FW88C	ERR9638285	calculus	Fewston, North Yorkshire	no	no
FW98C	ERR9638274	calculus	Fewston, North Yorkshire	no	no
FW192C	ERR9638281	calculus	Fewston, North Yorkshire	no	no
FW217C	ERR9638286	calculus	Fewston, North Yorkshire	no	no
FW268C	ERR9638275	calculus	Fewston, North Yorkshire	no	no
FW303C	ERR9638276	calculus	Fewston, North Yorkshire	no	no
FW331C	ERR9638283	calculus	Fewston, North Yorkshire	no	no
FW450C	ERR9638287	calculus	Fewston, North Yorkshire	no	no
STB12B	ERR9638259	bone	West Kensington	eBK497	n/a

(*continued on next page*)

Table 1 (*continued*)

Sample ID	Run accession number (European Nucleotide Archive)	Sample type	Site	Extraction/library blank decontaminated	Bone blank decontaminated
STB16B	ERR9638262	bone	West Kensington	eBK497	n/a
STB26C	ERR9638263	calculus	West Kensington	eBK497	STB16B, STB12B
STB2C	ERR9638253	calculus	West Kensington	eBK497	STB16B, STB12B
STB8C	ERR9638254	calculus	West Kensington	eBK497	STB16B, STB12B
STB10C	ERR9638255	calculus	West Kensington	eBK497	STB16B, STB12B
STB11C	ERR9638256	calculus	West Kensington	eBK497	STB16B, STB12B
STB43C	ERR9638257	calculus	West Kensington	eBK497	STB16B, STB12B
STB27C	ERR9638258	calculus	West Kensington	eBK497	STB16B, STB12B
STB41C	ERR9638261	calculus	West Kensington	eBK497	STB16B, STB12B
STB9C	ERR9638267	calculus	West Kensington	eBK497	STB16B, STB12B
STB54C	ERR9638270	calculus	West Kensington	eBK497	STB16B, STB12B
STB12C	ERR9638277	calculus	West Kensington	eBK497	STB16B, STB12B
STB16C	ERR9638279	calculus	West Kensington	eBK497	STB16B, STB12B
STB45C	ERR9638282	calculus	West Kensington	eBK497	STB16B, STB12B
VG6C	ERR9638265	calculus	Victoria Gate, Leeds	eBK679	no
VG11C	ERR9638266	calculus	Victoria Gate, Leeds	eBK679	no
VG12C	ERR9638268	calculus	Victoria Gate, Leeds	eBK679	no
VG15C	ERR9638269	calculus	Victoria Gate, Leeds	eBK679	no
SF2300B	ERR9638312	bone	Christ Church, Spitalfields	eBK682	n/a
SF2484B	ERR9638302	bone	Christ Church, Spitalfields	eBK682	n/a
SF01 (SP2182, SK2182, SF1)	ERR9638296	calculus	Christ Church, Spitalfields	eBK682	SF2300B, SF2484B
SF02 (SP2295, SK2295, SF2)	ERR9638297	calculus	Christ Church, Spitalfields	eBK682	SF2300B, SF2484B
SF03 (SP2300, SK2300, SF3)	ERR9638289	calculus	Christ Church, Spitalfields	eBK682	SF2300B, SF2484B
SF04C (SP2301, SK2301, SF4C)	ERR9638280	calculus	Christ Church, Spitalfields	eBK682	SF2300B, SF2484B
SF05 (SP2369, SK2369, SF5)	ERR9638310	calculus	Christ Church, Spitalfields	eBK682	SF2300B, SF2484B
SF06 (SP2468, SK2468, SF6)	ERR9638298	calculus	Christ Church, Spitalfields	eBK682	SF2300B, SF2484B
SF07 (SP2477, SK2477, SF7)	ERR9638290	calculus	Christ Church, Spitalfields	eBK682	SF2300B, SF2484B
SF08 (SP2484, SK2484, SF8)	ERR9638291	calculus	Christ Church, Spitalfields	eBK682	SF2300B, SF2484B
SF09 (SP2647, SK2647, SF9)	ERR9638309	calculus	Christ Church, Spitalfields	eBK682	SF2300B, SF2484B
SF10 (SP2748, SK2748)	ERR9638292	calculus	Christ Church, Spitalfields	eBK682	SF2300B, SF2484B
SF11 (SP2860, SK2860)	ERR9638293	calculus	Christ Church, Spitalfields	eBK682	SF2300B, SF2484B
ML1032C	ERR1329861	calculus	Melton	eBK695	no
ML1823C	ERR1329866	calculus	Melton	eBK695	no
ML3890C	ERR1329871	calculus	Melton	eBK695	no
ML4075C	ERR1329872	calculus	Melton	eBK695	no
ML1483C	ERR1329865	calculus	Melton	eBK695	no
OX12C	ERR1329870	calculus	Oxford St, Leicester	eBK691	OX12B, OX04B
OX09C	ERR1329862	calculus	Oxford St, Leicester	eBK691	OX12B, OX04B
OX04C	ERR1329853	calculus	Oxford St, Leicester	eBK691	OX12B, OX04B
OX05C	ERR1329857	calculus	Oxford St, Leicester	eBK691	OX12B, OX04B
OX12B	ERR1329867	bone	Oxford St, Leicester	eBK691	n/a
OX04B	ERR1407510	bone	Oxford St, Leicester	eBK691	n/a
3DT26C	ERR1329842	calculus	Driffield Terrace, York	eBK691	no
3DT21C	ERR1329834	calculus	Driffield Terrace, York	eBK691	no
6DT7C	ERR1329825	calculus	Driffield Terrace, York	eBK691	no
6DT21C	ERR1329829	calculus	Driffield Terrace, York	eBK691	no
3DT54C	ERR1329846	calculus	Driffield Terrace, York	eBK691	no

(*continued on next page*)

Table 1 (*continued*)

Sample ID	Run accession number (European Nucleotide Archive)	Sample type	Site	Extraction/library blank decontaminated	Bone blank decontaminated
JV15548C	ERR1329827	calculus	Coppergate, York	eBK695	no
JV30944C	ERR1329830	calculus	Coppergate, York	eBK695	no
NEM18C	ERR1329851	calculus	Norton-on-Tees, East Mill, Durham	no	no
NEM093C	ERR1329841	calculus	Norton-on-Tees, East Mill, Durham	no	no
NEM099C	ERR1329847	calculus	Norton-on-Tees, East Mill, Durham	no	no
NBS410C	ERR1329838	calculus	Norton-on-Tees, East Mill, Durham	eBK695	no
NBS262C	ERR1329832	calculus	Norton-on-Tees, East Mill, Durham	eBK695	no
NBS325C	ERR1329836	calculus	Norton-on-Tees, East Mill, Durham	eBK695	no
TKAC	ERR1329826	calculus	St Mary's Church, Tickhill,South Yorkshire	BL711	no
TKDC	ERR1329831	calculus	St Mary's Church, Tickhill,South Yorkshire	BL711	no
TKEC	ERR1329835	calculus	St Mary's Church, Tickhill,South Yorkshire	BL711	no
TKFC	ERR1329839	calculus	St Mary's Church, Tickhill,South Yorkshire	BL711	no
WG1688B	ERR1422702	bone	Wighill, North Yorkshire	n/a	n/a
WG1705C	ERR1329868	calculus	Wighill, North Yorkshire	no	no
WG1252C	ERR1329848	calculus	Wighill, North Yorkshire	BL711	no
WG1566C	ERR1329856	calculus	Wighill, North Yorkshire	BL711	no
WG1082C	ERR1329843	calculus	Wighill, North Yorkshire	BL711	no
WG1585C	ERR1407511	calculus	Wighill, North Yorkshire	BL711	no
WG1688C	ERR1422716	calculus	Wighill, North Yorkshire	BL711	no
FW283T	ERR1329878	tooth	Fewston, North Yorkshire	BL708	n/a
FW435C	ERR1329875	calculus	Fewston, North Yorkshire	BL708	FW283T
FW177C	ERR1329840	calculus	Fewston, North Yorkshire	BL708	FW283T
FW378C	ERR1329874	calculus	Fewston, North Yorkshire	BL708	FW283T
FW130C	ERR1329833	calculus	Fewston, North Yorkshire	BL708	FW283T
FW238C	ERR1329844	calculus	Fewston, North Yorkshire	BL708	FW283T
FW348C	ERR1407505	calculus	Fewston, North Yorkshire	BL708	FW283T
FW366C	ERR1329869	calculus	Fewston, North Yorkshire	BL708	FW283T
FW077C	ERR1329828	calculus	Fewston, North Yorkshire	BL708	FW283T
FW156C	ERR1329837	calculus	Fewston, North Yorkshire	BL708	FW283T
FW319C	ERR1422707	calculus	Fewston, North Yorkshire	BL708	FW283T
FW339C	ERR1329855	calculus	Fewston, North Yorkshire	BL708	FW283T
FW241C	ERR1329845	calculus	Fewston, North Yorkshire	BL708	FW283T
FW283C	ERR1329849	calculus	Fewston, North Yorkshire	BL708	FW283T
FW351C	ERR1329864	calculus	Fewston, North Yorkshire	BL708	FW283T
FW53C (FW53C)	ERR1329824	calculus	Fewston, North Yorkshire	BL708	FW283T
M116	ERR1343019	calculus	University of Tennessee	LBL589	n/a
M20	ERR1343010	calculus	University of Tennessee	LBL589	n/a
M56	ERR1343011	calculus	University of Tennessee	LBL589	n/a
M85	ERR1343012	calculus	University of Tennessee	LBL589	n/a
M103	ERR1343013	calculus	University of Tennessee	LBL589	n/a
M111	ERR1343014	calculus	University of Tennessee	LBL589	n/a
M31	ERR1343015	calculus	University of Tennessee	LBL589	n/a
M73	ERR1343016	calculus	University of Tennessee	LBL589	n/a
M89	ERR1343017	calculus	University of Tennessee	LBL589	n/a
M112	ERR1343018	calculus	University of Tennessee	LBL589	n/a
LC26C	ERR1407497	calculus	Leicester	eBK695	LC26T
LC26T	ERR1422701	tooth	Leicester	eBK695	n/a

(*continued on next page*)

Table 1 (*continued*)

Sample ID	Run accession number (European Nucleotide Archive)	Sample type	Site	Extraction/library blank decontaminated	Bone blank de-contaminated
C028	ERR13173382	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C029	ERR13173383	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C030	ERR13173384	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C031	ERR13173385	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C034	ERR13173386	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C035	ERR13173387	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C036	ERR13173388	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C037	ERR13173389	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C038	ERR13173390	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C039	ERR13173391	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C040	ERR13173392	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C041	ERR13173393	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C042	ERR13173394	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C043	ERR13173395	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C044	ERR13173396	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C045	ERR13173397	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C046	ERR13173398	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no
C047	ERR13173399	calculus	Cross St Cemetery, Manchester	eBK1, eBK2, LBL_439, LBL_695	no

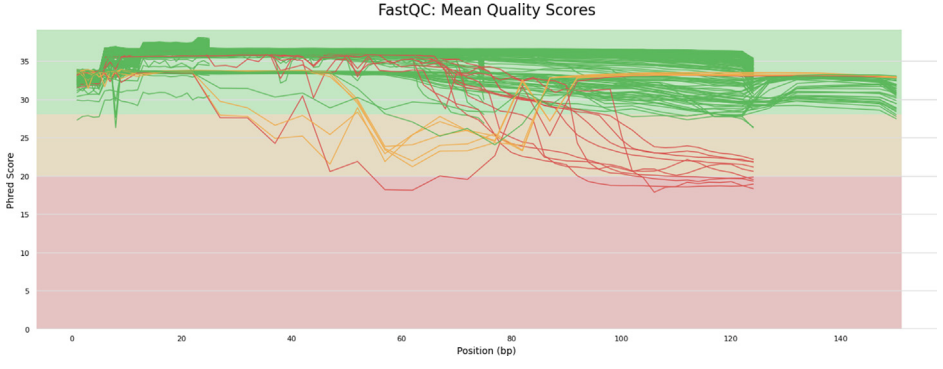

Fig. 4. FastQC sequence quality histograms showing the mean quality value across each base position in the read for all trimmed samples. The green line represents high-quality scores (≥Q25) across all bases, the orange line indicates moderate quality scores (≥Q20-Q25), suggesting some uncertainty in the base calls, and the red line signifies low-quality scores (<Q20).

Table 2
Quality control results showing percentage of mismatched sequences and reads past filters, and bp and read length pre and post-trimming and pre and post-decontamination.

Run accession number (European Nucleotide Archive)	Previously trimmed	Mismatched sequences (%) as calculated by FastP	Reads passed filters (%)	Mean bp before filtering	Mean bp after filtering	Read length before trimming	Read length after trimming	Read length after decon-tamination
ERR1329824	Y	N	99.97	25	25	2,583,158	2,582,416	2,487,082
ERR1329825	Y	N	99.96	25	25	879,712	879,380	876,832
ERR1329826	Y	N	99.97	25	25	409,780	409,684	409,684
ERR1329827	Y	N	99.96	25	25	644,972	644,748	639,120
ERR1329828	Y	N	99.96	25	25	461,822	461,682	447,704
ERR1329829	Y	N	99.93	25	25	605,826	605,460	604,424
ERR1329830	Y	N	99.96	25	25	1,369,538	1,369,022	1,348,176
ERR1329831	Y	N	99.97	25	25	476,316	476,216	476,214
ERR1329832	Y	N	99.97	25	25	892,672	892,432	882,258
ERR1329833	Y	N	99.97	25	25	548,974	548,856	514,522
ERR1329834	Y	N	99.96	25	25	388,638	388,502	387,788
ERR1329835	Y	N	99.98	25	25	239,084	239,058	239,058
ERR1329836	Y	N	99.89	25	25	391,314	390,916	386,706
ERR1329837	Y	N	99.97	25	25	210,862	210,818	201,148
ERR1329838	Y	N	99.94	25	25	198,406	198,304	196,176
ERR1329839	Y	N	99.96	25	25	247,038	246,944	246,944
ERR1329840	Y	N	99.97	25	25	467,224	467,108	451,178
ERR1329841	Y	N	99.93	25	25	74,130	74,084	No match to blank
ERR1329842	Y	N	99.97	25	25	809,948	809,716	806,872
ERR1329843	Y	N	99.97	25	25	897,988	897,792	897,790
ERR1329844	Y	N	99.97	25	25	152,184	152,148	147,296
ERR1329845	Y	N	99.97	25	25	883,494	883,276	843,806
ERR1329846	Y	N	99.97	25	25	721,958	721,780	720,434
ERR1329847	Y	N	99.87	25	25	125,702	125,540	No match to blank
ERR1329848	Y	N	99.96	25	25	402,624	402,492	402,492
ERR1329849	Y	N	99.82	25	25	251,528	251,084	243,366
ERR1329851	Y	N	99.97	25	25	491,376	491,238	No match to blank
ERR1329853	Y	N	99.64	25	25	231,618	230,790	229,678
ERR1329855	Y	N	99.97	25	25	171,824	171,780	162,654
ERR1329856	Y	N	99.95	25	25	78,242	78,210	78,210
ERR1329857	Y	N	99.96	25	25	136,876	136,824	135,554
ERR1329861	Y	N	99.96	25	25	261,800	261,700	258,868
ERR1329862	Y	N	99.96	25	25	181,530	181,462	179,992
ERR1329864	Y	N	99.96	25	25	111,654	111,616	108,182
ERR1329865	Y	N	99.68	25	25	275,252	274,374	271,342
ERR1329866	Y	N	99.85	25	25	321,792	321,326	319,042
ERR1329867	Y	N	99.93	25	25	850,018	849,476	849,112
ERR1329868	Y	N	99.98	25	25	725,652	725,516	No match to blank
ERR1329869	Y	N	99.97	25	25	269,398	269,336	256,532
ERR1329870	Y	N	99.9	25	25	416,958	416,556	413,234
ERR1329871	Y	N	99.96	25	25	450,618	450,452	445,752
ERR1329872	Y	N	99.96	25	25	287,424	287,336	283,892
ERR1329874	Y	N	99.46	25	25	19,568	19,464	19,006
ERR1329875	Y	N	99.97	25	25	194,014	193,966	186,070
ERR1329878	Y	N	99.97	25	25	394,528	394,416	392,586
ERR1329879	Y	N	99.98	25	25	52,032	52,026	Blank sample
ERR1329880	Y	N	99.99	25	25	20,898	20,896	Blank sample
ERR1329881	Y	N	100	25	25	2770	2770	Blank sample

(*continued on next page*)

Table 2 (*continued*)

Run accession number (European Nucleotide Archive)	Previously trimmed	Mismatched sequences (%) as calculated by FastP	Reads passed filters (%)	Mean bp before filtering	Mean bp after filtering	Read length before trimming	Read length after trimming	Read length after decontamination
ERR1329882	Y	N	100	25	25	322	322	Blank sample
ERR1343010	N	19.32	98.2	75	65	42,560,574	41,797,270	40,603,602
ERR1343011	N	16.23	97.08	75	67	22,883,998	22,216,802	21,063,068
ERR1343012	N	17.59	98.48	75	67	34,442,652	33,920,804	32,940,382
ERR1343013	N	20.46	97.87	75	67	30,978,568	30,320,904	29,360,946
ERR1343014	N	13.71	96.69	75	64	34,561,856	33,419,504	32,140,738
ERR1343015	N	18.66	97.23	75	66	33,328,916	32,407,552	31,712,270
ERR1343016	N	23.89	94.21	75	68	31,635,170	29,804,612	28,942,244
ERR1343017	N	26.46	98.03	75	68	27,706,734	27,162,150	25,949,236
ERR1343018	N	15.39	98.29	75	66	30,454,984	29,936,826	28,909,058
ERR1343019	N	12.94	98.3	75	66	30,692,120	30,172,286	29,213,134
ERR1343020	N	32.08	94.99	75	65	13,175,352	12,516,206	Blank sample
ERR1343021	N	96.03	18.35	75	72	9,036,794	1,658,910	Blank sample
ERR1407493	N	83.87	36.28	75	68	359,280	130,368	Blank sample
ERR1407497	N	10.96	97.01	75	68	58,423,618	56,682,086	54,393,516
ERR1407505	N	28.94	94.83	75	70	58,589,354	55,565,860	51,893,282
ERR1407510	N	8.28	97.91	75	53	47,001,744	46,022,816	44,928,798
ERR1407511	N	4.74	97.25	75	58	48,838,650	47,499,126	47,498,426
ERR1407514	N	41.84	74.56	75	55	1,540,026	1,148,350	Blank sample
ERR1422701	N	17.66	96.56	125	109	61,646,806	59,530,712	58,995,200
ERR1422702	N	26.46	93.98	125	104	43,051,330	40,460,558	No match to blank
ERR1422707	N	0.88	97.69	125	70	50,642,508	49,474,228	46,352,006
ERR1422716	N	10.8	95.52	125	102	52,182,752	49,847,030	49,846,178
ERR1659110	N	1.46	98.23	125	83	29,512,290	28,992,366	No match to blank
ERR1659111	N	1.01	98.47	125	79	26,875,932	26,467,410	No match to blank
ERR1659112	N	4.8	98.49	125	91	26,969,988	26,565,434	No match to blank
ERR1659113	N	2.5	97.83	125	85	29,055,384	28,427,660	No match to blank
ERR1659114	N	15.01	85.62	125	86	24,805,462	21,238,676	No match to blank
ERR1659115	N	2.05	98.28	125	85	25,821,896	25,378,038	No match to blank
ERR1659116	N	0.8	98.41	125	76	47,146,778	46,400,438	No match to blank
ERR1659117	N	1.98	98.38	125	76	28,774,712	28,310,064	No match to blank
ERR1659118	N	4.63	98.31	125	86	32,731,048	32,180,872	No match to blank
ERR1659119	N	1.38	98.27	125	78	24,592,154	24,166,868	No match to blank
ERR1659120	N	2.66	98.26	125	82	35,486,302	34,872,316	No match to blank
ERR1681523	N	3.26	96.22	125	72	49,411,398	47,545,688	47,156,130
ERR1681525	N	1.26	97.75	125	70	36,920,280	36,093,176	34,991,188
ERR1681526	N	1.3	97.38	125	76	57,625,826	56,120,936	53,873,302
ERR1681527	N	3.26	96.04	125	84	46,975,236	45,116,958	41,743,542
ERR1681530	N	1.63	97.38	125	78	65,612,894	63,896,898	62,081,334

(*continued on next page*)

Table 2 (*continued*)

Run accession number (European Nucleotide Archive)	Previously trimmed	Mismatched sequences (%) as calculated by FastP	Reads passed filters (%)	Mean bp before filtering	Mean bp after filtering	Read length before trimming	Read length after trimming	Read length after decon-tamination
ERR1681531	N	2.05	96.59	125	65	42,871,320	41,409,526	39,056,832
ERR1681532	N	1.42	98.16	125	63	47,032,120	46,171,356	No match to blank
ERR1681533	N	1.11	98.17	125	71	35,660,320	35,009,952	No match to blank
ERR1681534	N	3.51	98.2	125	82	33,354,436	32,756,626	No match to blank
ERR1681535	N	2.69	97.55	125	81	24,320,390	23,726,436	No match to blank
ERR1681538	N	3.44	97.98	125	82	30,385,320	29,773,404	No match to blank
ERR1681541	N	4.35	97.99	125	85	37,920,560	37,159,360	No match to blank
ERR1681542	N	2.39	98.13	125	83	24,833,088	24,370,008	No match to blank
ERR9638253	N	3.2	99.01	125	78	24,674,054	24,431,924	21,710,018
ERR9638254	N	12.35	98.89	125	102	28,260,842	27,947,766	19,212,448
ERR9638255	N	7.18	98.81	125	90	27,501,276	27,174,018	21,299,116
ERR9638256	N	8.77	98.75	125	95	25,797,650	25,487,318	16,667,600
ERR9638257	N	5.7	98.71	125	82	29,957,450	29,572,208	27,105,134
ERR9638258	N	4.51	98.68	125	84	24,253,730	23,935,210	21,708,146
ERR9638259	N	7.26	98.14	125	82	24,643,382	24,185,674	23,720,532
ERR9638260	N	90.18	53.44	125	117	724,146	387,022	Blank sample
ERR9638261	N	6.51	98.59	125	91	27,624,128	27,235,894	20,130,754
ERR9638262	N	14.97	86.89	125	77	26,588,560	23,103,962	22,367,356
ERR9638263	N	7.45	98.32	125	93	27,257,322	26,799,464	18,403,542
ERR9638264	N	15.13	92.23	125	79	4,642,754	4,282,398	Blank sample
ERR9638265	N	7.77	98.52	125	89	27,644,478	27,236,206	25,596,208
ERR9638266	N	2.11	98.28	125	76	25,996,920	25,550,050	23,611,344
ERR9638267	N	8.94	98.05	125	83	25,031,646	24,543,538	21,289,618
ERR9638268	N	2.09	98.21	125	78	25,575,492	25,117,978	22,180,726
ERR9638269	N	1.88	98.22	125	79	28,529,412	28,024,144	24,710,314
ERR9638270	N	8.27	98.14	125	85	30,716,356	30,146,808	26,628,622
ERR9638271	N	33.67	73.3	125	87	15,952,534	11,694,636	Blank sample
ERR9638272	N	92.6	17.65	125	108	8,723,184	1,539,914	Blank sample
ERR9638274	N	1	98.48	125	71	27,892,400	27,470,640	No match to blank
ERR9638275	N	5.49	98.59	125	87	25,283,884	24,928,344	No match to blank
ERR9638276	N	4.49	98.3	125	84	30,013,726	29,503,986	No match to blank
ERR9638277	N	4.93	97.88	125	81	26,023,962	25,473,438	22,699,452
ERR9638278	N	0.61	98.97	125	70	26,956,626	26,681,156	No match to blank
ERR9638279	N	0.55	99.06	125	57	25,132,870	24,896,936	22,752,928
ERR9638280	N	1.99	98.49	125	77	25,666,712	25,280,914	22,733,888
ERR9638281	N	4.28	98.66	125	84	28,059,552	27,684,056	No match to blank
ERR9638282	N	21.46	85.44	125	90	26,409,456	22,564,398	20,161,446
ERR9638283	N	2.98	97.69	125	77	29,453,596	28,774,998	No match to blank

(*continued on next page*)

Table 2 (*continued*)

Run accession number (European Nucleotide Archive)	Previously trimmed	Mismatched sequences (%) as calculated by FastP	Reads passed filters (%)	Mean bp before filtering	Mean bp after filtering	Read length before trimming	Read length after trimming	Read length after decontamination
ERR9638284	N	9.02	92.18	125	80	18,608,440	17,154,190	Blank sample
ERR9638285	N	3.6	98.92	125	81	27,242,600	26,924,196	No match to blank
ERR9638286	N	10.83	89.97	125	84	22,855,790	20,564,950	No match to blank
ERR9638287	N	6.21	98.09	125	85	26,873,758	26,362,722	No match to blank
ERR9638288	N	77	46.89	125	104	924,370	433,452	Blank sample
ERR9638289	N	2.32	96.19	125	67	44,031,332	42,353,852	35,461,384
ERR9638290	N	2.57	96.72	125	56	32,380,974	31,321,288	24,602,766
ERR9638291	N	3.39	96.66	125	74	33,505,890	32,388,386	28,026,322
ERR9638292	N	3.4	95.72	125	80	39,558,620	37,867,974	29,699,936
ERR9638293	N	1.9	97.03	125	63	35,925,288	34,858,864	22,648,130
ERR9638294	N	73.42	32.86	125	95	2,888,654	949,470	Blank sample
ERR9638295	N	87.54	377.77	125	106	1,222,796	377,772	Blank sample
ERR9638296	N	87.54	30.89	125	106	41,936,516	40,682,502	33,970,538
ERR9638297	N	2.02	97	125	77	34,318,286	33,304,306	28,638,038
ERR9638298	N	3.93	97.04	125	77	34,510,834	33,516,594	25,277,366
ERR9638300	N	3.64	94.59	125	66	33,321,126	31,521,220	30,483,986
ERR9638301	N	90.36	53.6	125	115	1,428,434	765,698	Blank sample
ERR9638302	N	13.32	96.2	125	104	33,809,420	32,526,006	32,438,428
ERR9638303	N	2.15	96.39	125	65	34,867,684	33,609,404	31,310,986
ERR9638304	N	5.01	96.89	125	85	38,289,660	37,099,046	33,449,992
ERR9638305	N	2.18	96.82	125	77	35,541,750	34,414,316	32,141,250
ERR9638306	N	12.14	91.74	125	81	3,605,416	3,307,704	Blank sample
ERR9638307	N	3	97.04	125	77	36,222,640	35,152,254	32,784,300
ERR9638308	N	1.62	97.81	125	76	37,429,770	36,613,212	35,064,998
ERR9638309	N	3.33	95.3	125	69	36,701,602	34,979,784	28,690,092
ERR9638310	N	1.9	96.85	125	65	34,990,762	33,889,492	26,642,234
ERR9638312	N	1.86	97.25	125	52	37,248,422	36,226,502	36,085,818
ERR13173382	N	1.75	13.45	151	78	13,540,920	11,801,754	10,129,902
ERR13173383	N	4.11	9.59	151	113	9,650,924	9,160,164	8,583,678
ERR13173384	N	13.05	7.23	151	92	7,279,134	6,762,296	4,985,224
ERR13173385	N	1.13	10.5	151	72	10,566,408	9,186,790	7,749,908
ERR13173386	N	2.58	14.29	151	96	14,393,380	13,572,660	12,758,236
ERR13173387	N	1.65	12.01	151	76	12,086,216	10,929,198	9,706,028
ERR13173388	N	2.95	11.03	151	89	11,101,120	10,242,942	9,237,776
ERR13173389	N	1.76	11.92	151	86	11,993,372	10,904,066	9,815,586
ERR13173390	N	1.92	13.11	151	82	13,209,018	11,221,758	9,713,430
ERR13173391	N	2.38	11.64	151	88	11,712,516	10,121,162	8,946,904
ERR13173392	N	6.78	11.33	151	103	11,412,902	9,722,646	8,385,754
ERR13173393	N	6.78	9.12	151	104	9,186,792	8,712,604	7,620,040
ERR13173394	N	7.79	17.64	151	102	17,916,774	16,828,530	15,356,984
ERR13173395	N	2.42	12.66	151	83	12,760,300	11,460,324	10,016,358
ERR13173396	N	3.23	8.64	151	72	8,724,436	7,620,118	6,275,740
ERR13173397	N	2.06	12	151	94	12,083,728	11,220,760	10,547,040
ERR13173398	N	3.14	94	151	20.15	20,271,626	18,933,902	17,435,362
ERR13173399	N	7.82	14.55	151	116	14,657,670	14,051,062	13,095,506

(*continued on next page*)

Table 2 (*continued*)

Run accession number (European Nucleotide Archive)	Previously trimmed	Mismatched sequences (%) as calculated by FastP	Reads passed filters (%)	Mean bp before filtering	Mean bp after filtering	Read length before trimming	Read length after trimming	Read length after decontamination
ERR13173400	N	85.45	98.97	151	140	2,072,654	2,013,310	Blank sample
ERR13173401	N	33.43	99.13	151	101	1,259,734	985,856	Blank sample
ERR13173402	N	99.93	98.67	151	150	4,493,018	4,433,302	Blank sample
ERR13173396	N	99.87	98.32	151	150	667,232	655,950	Blank sample

Samples were centrifuged at 13,000 RPM for 2 min and 1 mL of supernatant transferred into fresh tubes. To the supernatant, 2 mg of Proteinase K was added and rotated at 37 °C for 24 h. For all samples except those from Manchester, Cross Street, DNA was extracted from the dental calculus and bone samples using a protocol based on Dabney et al. (2013) [8] with DNA eluted in 60 µL of EB following a five-minute incubation step. For the Manchester, Cross Street samples (C028–C027), DNA extraction followed a modified silica-spin column protocol [9], with DNA concentrated in Amicon® 10 K Ultra-4 Centrifugal Filter Devices (Millipore) and purified with QiaQuick MinElute kits (QIAGEN, Hilden, Germany) before being eluted in 25 µL of Qiagen EB Buffer. All DNA extracts were quantified via Qubit® 2.0 Fluorometer using a High-Sensitivity DNA Assay.

4.1.3. Metagenomic sequencing

For each DNA extract, double-stranded whole genome shotgun Illumina libraries were prepared using a protocol based on Meyer and Kircher (2010) [10]. Each dental calculus library was built using between 200–400 ng of DNA; extraction blanks were prepared with 25 µL of DNA extract; library blanks were built using 25 µL of nuclease free water. The libraries were constructed using a double-barcoding approach as described in Fortes and Paijman (2015) [11] which serves as an additional means to filter chimeric sequences from the dataset, and thus increase the confidence in assigning the sequences to a particular library. Individual P7 indexes were ligated through an indexing PCR step using a proof-reading taq polymerase (AccuPrime Pfx Supermix) with the following cycling conditions were: 95 °C for 5 min, and cycles of (95 °C for 15 s, 60 °C for 30 s, 68 °C for 30 s), and a final extension of 68 °C for 5 min. Optimal cycle numbers for library indexing were determined through the use of quantitative PCR (qPCR) using Fast SYBR [12]. Amplified libraries were subsequently purified using Qiagen MinElute spin columns, the size distribution of the amplified libraries was determined using an Agilent 2100 Bioanalyzer. The dental calculus libraries were pooled in equimolar concentration and subjected to paired-end sequencing on multiple HiSeq2500 lanes at the Wellcome Trust Sanger Institute (WTSI) or on a NextSeq platform, PE 150 + 150 bp Integrated Microbiome Resource (IMR) at Dalhousie (Manchester, Cross Street). In accordance with WTSI protocols, human DNA sequences were removed from dental calculus datasets prior to deposition in the ENA. For samples sequenced at the WTSI (Table S1), the raw metagenomic reads were mapped against the human reference genome (GRCh37) using *bwa aln* with default settings; reads mapping to the human genome reads were extracted from the dataset and only unmapped reads uploaded to the ENA. Samples sequences at the IMR (Manchester, Cross Street) did not have human reads filtered prior to deposition.

4.2. Quality control

FastQC v0.11.9 [13] was used to assess raw digital data quality. FastQC is a quality control tool for raw sequencing data that provides a modular collection of analyses used to gain insight into

Table 3
Cytosine-Thymine frequency transition values for authentication interpretation.

ENA/sample ID	Sample type	Average bp length	Average bp length after trimming	Frequency
ERR1329825	Calculus	25bp	25bp	Empty
ERR1329827	Calculus	25bp	25bp	Empty
ERR1329829	Calculus	25bp	25bp	Empty
ERR1329830	Calculus	25bp	25bp	Empty
ERR1329832	Calculus	25bp	25bp	Empty
ERR1329834	Calculus	25bp	25bp	Empty
ERR1329836	Calculus	25bp	25bp	Empty
ERR1329838	Calculus	25bp	25bp	Empty
ERR1329842	Calculus	25bp	25bp	Empty
ERR1329846	Calculus	25bp	25bp	Empty
ERR1329853	Calculus	25bp	25bp	Empty
ERR1329857	Calculus	25bp	25bp	Empty
ERR1329861	Calculus	25bp	25bp	Empty
ERR1329862	Calculus	25bp	25bp	Empty
ERR1329865	Calculus	25bp	25bp	Empty
ERR1329866	Calculus	25bp	25bp	Empty
ERR1329867	Bone	25bp	25bp	Empty
ERR1329870	Calculus	25bp	25bp	Empty
ERR1329871	Calculus	25bp	25bp	Empty
ERR1329872	Calculus	25bp	25bp	Empty
ERR1407510	Bone	75bp	75bp	$<0.10 > 0.05$
ERR1329826	Calculus	25bp	25bp	Empty
ERR1329831	Calculus	25bp	25bp	Empty
ERR1329835	Calculus	25bp	25bp	Empty
ERR1329839	Calculus	25bp	25bp	Empty
ERR1329843	Calculus	25bp	75bp	Empty
ERR1329848	Calculus	25bp	25bp	Empty
ERR1329856	Calculus	25bp	25bp	Empty
ERR1407497	Calculus	75bp	68bp	$<0.025 > 0$
ERR1407511	Calculus	75bp	58bp	$<0.026 > 0$
ERR1422701	Tooth	125bp	109bp	$<0.05 > 0.025$
ERR1422716	Calculus	125bp	102bp	$<0.026 > 0$
ERR1329833	Calculus	25bp	25bp	Empty
ERR1329844	Calculus	25bp	25bp	Empty
ERR1329849	Calculus	25bp	25bp	Empty
ERR1329864	Calculus	25bp	25bp	Empty
ERR1329874	Calculus	25bp	25bp	Empty
ERR9638253	Calculus	125bp	78bp	$<0.025 > 0$
ERR9638254	Calculus	125bp	102bp	$<0.025 > 0$
ERR9638255	Calculus	125bp	90bp	$<0.025 > 0$
ERR9638256	Calculus	125bp	95bp	$<0.025 > 0$
ERR9638257	Calculus	125bp	82bp	$<0.025 > 0$
ERR9638258	Calculus	125bp	84bp	$<0.05 > 0.025$
ERR9638259	Bone	125bp	82bp	$<0.025 > 0$
ERR9638261	Calculus	125bp	91bp	$<0.05 > 0.025$
ERR9638262	Bone	125bp	77bp	$<0.025 > 0$
ERR9638263	Calculus	125bp	93bp	$<0.025 > 0$
ERR9638265	Calculus	125bp	89bp	$<0.025 > 0$
ERR9638266	Calculus	125bp	76bp	$<0.025 > 0$
ERR9638267	Calculus	125bp	83bp	$<0.025 > 0$
ERR9638268	Calculus	125bp	78bp	$<0.025 > 0$
ERR9638269	Calculus	125bp	79bp	$<0.025 > 0$
ERR9638270	Calculus	125bp	85bp	$<0.025 > 0$
ERR9638277	Calculus	125bp	81bp	$<0.025 > 0$
ERR9638279	Calculus	125bp	57bp	$<0.025 > 0$
ERR9638282	Calculus	125bp	90bp	$<0.025 > 0$
ERR1329828	Calculus	25bp	25bp	Empty
ERR1329840	Calculus	25bp	25bp	Empty
ERR1329845	Calculus	25bp	25bp	Empty

(*continued on next page*)

Table 3 (*continued*)

ENA/sample ID	Sample type	Average bp length	Average bp length after trimming	Frequency
ERR1329875	Calculus	25bp	25bp	Empty
ERR1407505	Calculus	75bp	70bp	<0.025 > 0
ERR1681523	Calculus	125bp	72bp	<0.025 > 0
ERR1681525	Calculus	125bp	70bp	<0.05 > 0.025
ERR1681526	Calculus	125bp	76bp	<0.025 > 0
ERR1681527	Calculus	125bp	84bp	<0.025 > 0
ERR1681530	Calculus	125bp	78bp	<0.025 > 0
ERR1681531	Calculus	125bp	65bp	<0.05 > 0.025
ERR9638280	Calculus	125bp	77bp	<0.025 > 0
ERR9638289	Calculus	125bp	67bp	<0.025 > 0
ERR9638290	Calculus	125bp	56bp	<0.025 > 0
ERR9638291	Calculus	125bp	74bp	<0.025 > 0
ERR9638292	Calculus	125bp	80bp	<0.025 > 0
ERR9638293	Calculus	125bp	63bp	<0.025 > 0
ERR9638296	Calculus	125bp	106bp	<0.025 > 0
ERR9638297	Calculus	125bp	77bp	<0.05 > 0.025
ERR9638298	Calculus	125bp	77bp	<0.025 > 0
ERR9638300	Calculus	125bp	66bp	<0.025 > 0
ERR9638302	Bone	125bp	104bp	<0.025 > 0
ERR9638303	Calculus	125bp	65bp	<0.05 > 0.025
ERR9638304	Calculus	125bp	85bp	<0.025 > 0
ERR9638305	Calculus	125bp	77bp	<0.025 > 0
ERR9638307	Calculus	125bp	77bp	<0.05 > 0.025
ERR9638308	Calculus	125bp	76bp	<0.025 > 0
ERR9638309	Calculus	125bp	69bp	<0.025 > 0
ERR9638310	Calculus	125bp	65bp	<0.025 > 0
ERR9638312	Bone	125bp	52bp	<0.025 > 0
ERR13173382	Calculus	150bp	78bp	<0.025 > 0
ERR13173383	Calculus	150bp	113bp	<0.025 > 0
ERR13173384	Calculus	150bp	92bp	<0.025 > 0
ERR13173385	Calculus	150bp	72bp	<0.025 > 0
ERR13173386	Calculus	150bp	96bp	<0.025 > 0
ERR13173387	Calculus	150bp	76 bp	<0.025 > 0
ERR13173388	Calculus	150bp	89bp	<0.025 > 0
ERR13173389	Calculus	150bp	86bp	<0.025 > 0
ERR13173390	Calculus	150bp	82bp	<0.025 > 0
ERR13173391	Calculus	150bp	88bp	<0.025 > 0
ERR13173392	Calculus	150bp	103bp	<0.025 > 0
ERR13173393	Calculus	150bp	104bp	<0.025 > 0
ERR13173394	Calculus	150bp	102bp	<0.025 > 0
ERR13173395	Calculus	150bp	83bp	<0.025 > 0
ERR13173396	Calculus	150bp	72bp	<0.025 > 0
ERR13173397	Calculus	150bp	94bp	<0.025 > 0
ERR13173398	Calculus	150bp	20.15bp	<0.025 > 0
ERR13173399	Calculus	150bp	116bp	<0.025 > 0

any flaws in the data before performing further analysis [13]. The preprocessing programmer Fastp v0.23.2 [14] was then utilised with default parameters. Fastp is a tool used to filter and trim poor quality reads, cut adapters, repair mismatched base pairs, and produce overall quality. It also provides results that include both pre- and post-filtering data, allowing for a direct comparison on the filtering impact [14].

4.3. Decontamination

To minimise the impact of laboratory and environmental contamination, BBduk [15] was used with default parameters to decontaminate samples using Kmers. This procedure involved

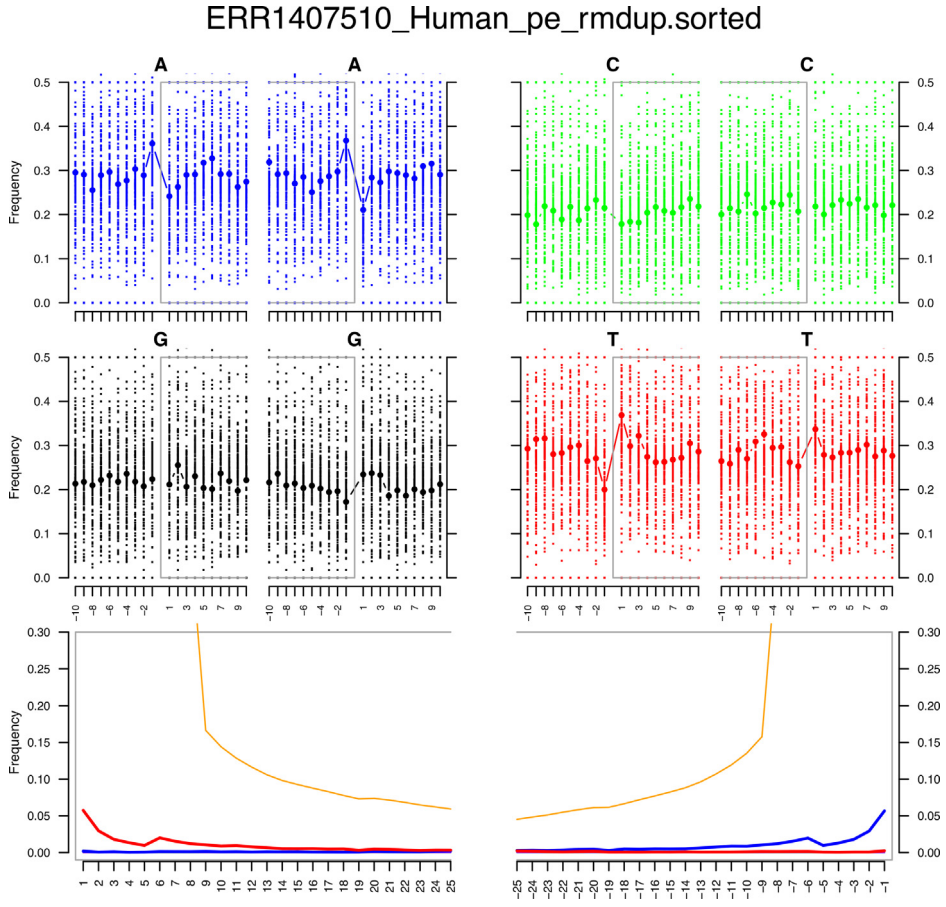

Fig. 5. Example of a mapDamage result of C-T (5′ end) and G-A (3′ end) base substitutions from bone sample OX04B. The orange line shows nucleotides that do not align to the query sequence.

identifying and removing homologous sequences found in extraction and library blanks, and bone blanks when they were available (see Table 1 for which samples were paired with which controls). After decontamination, paired read counts did not match because the tool removed a different number of sequences it considered to be contaminants from forward and reverse reads, resulting in unequal counts. Therefore, the samples were subsequently processed with Trimmomatic v0.39 [16] to re-pair reads (**for trimmomatic commands, refer to GitHub link in specifications table above**).

4.4. aDNA authentication

Only samples (105 samples; refer to Table 3 for this exact list) that were provided with laboratory controls (Table 1) were used to authenticate aDNA. The reasoning for this was to avoid potential confusion arising from modern contamination of otherwise ancient samples. Centrifuge v1.0.3 [7] was used, with default parameters, to assign taxonomic labels by mapping sequences against the human genome, prokaryotic genomes, and viral genomes including 106 SARS-CoV-2 complete genomes. Human reads from Centrifuge outputs were retained and seqtk 'subseq' was used to convert them into fastq files which were then mapped to the human genome (hg38)

[17] using BWA mem v0.7.17 [18]. SAMtools v1.12 [19] (-view -rmdup -flagstat -sort -index) was then utilised for alignment formatting and was sorted into BAM files which were run through mapDamage2 v2.2.2 with default parameters [20].

Limitations

Our aDNA data have some limitations that need to be addressed. These include absent controls, short read length, and damage authentication issues that may cause issues for potential future research projects.

Experimental controls

It was not always possible to obtain environmental controls (such as soil, bone, or tooth root blanks) or laboratory controls (such as a library or extraction blanks) (see Table 1); therefore, depending on the biological question, some researchers may wish to exclude these samples from downstream analysis. It is beneficial to use environmental controls in projects that analyse the microbial composition of human microbiomes through methods such as diversity or AMR analysis. Environmental controls allow researchers to screen for and eliminate environmental contamination from their findings. These environmental operational taxonomic units (OTUs) may include the modern, living microbes at the grave site, or even the DNA from extinct microorganisms, such as ancient soil-dwelling species or those that hail from other animals and plants, that could distort the findings of a human microbiome study.

Quality assessment

Read quality scores (raw and trimmed) as well as read counts after trimming and decontamination can be viewed in Table 2 and Fig. 2. FastQC reports for raw and trimmed datasets are available in **Supplementary Material 1 and 2**, respectively, and Fastp reports are accessible in the **Supplementary Material 3 (zip folders)** in the Dryad repository (DOI: 10.5061/dryad.jdfn2z3mk).

Reads

A subset of 49 samples showed high-quality base calls that had previously been trimmed prior to this research and were 25 base pairs (bp) long. Of the remaining samples, 18 samples were 75 bp, 85 were 125 bp, and 22 were 150 bp (Table 2). Short reads are characteristic of aDNA; however, those samples that are 25 bp are not likely to be practical for some analyses such as damage analysis (see **damage analysis** section below), screening for functional genes, contig assembly and metagenomic assembled genome creation, since they may not provide sufficient quality DNA for these type of analyses. Apart from those 25 bp samples that had been previously trimmed, the remaining reads were trimmed and filtered for post-processing and downstream analysis. These can be viewed in the 'per base sequence' graphs where the X-axis shows the individual bases for reads that have been called, and the Y-axis shows the distribution values (see Fastqc reports in **supporting data**). Samples with good quality scores exhibit a blue line that continually remains high and above a distribution value of 20. Fig. 4 shows overall good mean sequence quality scores (see full multiQC interactive chart in **supporting data - compressed zip folder 3** containing **FastQC reports for trimmed data**).

The dataset initially contained a total of 3371,978,784 raw reads which were reduced to 3236,478,044 reads post-filtering. An average of 6 % reads per sample (median of 1.84 % and interquartile range [IQR] of 4.05 %) were lost in the filtering process (see Table 2 for reads lost per sample). Within the previously trimmed 49 samples, 100 % of reads passed the filtering

process in two library blanks (BL10 and BL11) and the rest passed >99 % of reads. With the exception of laboratory controls, the remaining samples that were not previously trimmed passed >90 % of reads.

Thirty-two samples could not be decontaminated as they were not assigned a laboratory blank. Samples ($n = 89$) that did not have an environmental blank were still subject to decontamination protocols using laboratory controls, but future research should take caution using these samples for further analysis as ancient and modern environmental contaminants may be present. Alternatively, using a tool like Sourcetracker [21] may help understand the environmental composition in these samples. 120 remaining samples (2 tooth root, 7 bone, and 111 dental calculus) comprising 2339,228,254 reads were all decontaminated with laboratory controls and 30 of these calculus samples were decontaminated with bone and 16 with tooth root (Table 1). Following decontamination, 2125,556,400 (an average of 7.3 %, median of 4.47 %, and IQR of 9.99 %) reads were removed (Table 2; Fig. 3).

Damage analysis

Authenticating ancient DNA from complex metagenomic samples can be problematic [22]. Typically, ancient DNA is authenticated through observations of damage-induced deamination profiles within endogenous host DNA using programs such as mapDamage [20] DamageProfiler [23], and PyDamage [24]. However, in the case of dental calculus, low read counts of human host DNA limit the number of alignments which can be evaluated for damage patterns, dramatically reducing the power of the method [22,25,26]. Further, in this study, samples sequenced by the Wellcome Trust Sanger Institute, Cambridge UK, were pre-filtered to remove those reads which map to the Human genome. Despite this limitation, we applied mapDamage [20] to our samples to quantify cytosine deamination in the human DNA component. All 104 mapDamage plots are accessible in the Dryad repository under **Supplementary material 4** (DOI: 10.5061/dryad.jdfn2z3mk) and a summary of C-T frequency transition data is shown in Table 3. Briefly, 36 samples produced empty plots due to short read length (<25 bp); 59 samples show misincorporation curves below 0.025; eight showed above 0.025 but below 0.05; and one showed >0.05 but <0.10, which showed the highest levels of deamination in our sample set (Fig. 5).

One potential solution for low endogenous DNA in the samples presented here, or in other whole-metagenome samples, could be to identify a highly abundant OTU within a given sample, and use this taxon for damage verification [27]. Note, this would potentially require a unique background genome comparison for each sample, based on abundance. Nevertheless, issues with low read abundance of even the 'most abundant' taxa for any given sample may still persist.

Ethics Statement

The authors confirm that they have read and followed the ethical requirements for publication in Data in Brief and that the current work does not involve animal experiments or any data collected from social media platforms. Regarding the involvement of human subjects, as an ectopic growth, dental calculus is not considered a human tissue, and analysis is not subject to the legislation of the Human Tissue Act. All samples have received ethical approval for destructive analysis and data publication from relevant repositories.

Credit Author Statement

Conceptualization: C.S., S.F., M.C. Supervision: C.S., A.T., C.J.M., Funding acquisition: C.S., A.T., C.J.M. Resources: M.H., A.M., C.S., A.R., M.C.. Investigation: F.J.S, G D-A, J.H., A.R.,C.S., S.F., K.M.

Formal Analysis: F.J.S, G. D-A., J.H., C.S., A.T. and J.W. Data Curation, Visualization and Writing - Original Draft: F.J.S, G. D-A. Writing - Review & Editing: all authors.

Acknowledgements

We thank the following individuals, museums and agencies for providing access to skeletal collections: Forensic Anthropology Center at the University of Tennessee, Knoxville (Dawnie Steadman); John Buglass Archaeological Services; Museum of London (Rebecca Redfern, Jelena Bekvalac); Natural History Museum (Ian Barnes, Heather Bonney); Durham University Department of Archaeology (Anwen Caffell, Rebecca Gowland); University of Leicester Archaeological Services; University of York Department of Archaeology (Cath Neal); Washburn Heritage Centre; York Archaeological Trust for Excavation and Research Ltd (Christine McDonnell); York Osteoarchaeology Ltd (Katie Keefe). We are grateful to Dr Lisa MacKenzie-Davey for helping to sample the calculus from Manchester, Cross Street. Thanks to Julian Parkhill, Josef Wagner and David Jackson from the Wellcome Trust Sanger Institute for assistance with project design and metagenomic sequencing. We are grateful for the assistance of Drs Gavin Thomas and Sandy McDonald (University of York) for their advice on research design, and to Drs Mark Jenner and Sarah Goldsmith (University of York) for input on the historical context of the post-medieval skeletons. The authors acknowledge the use of the University of Bradford High Performance Computing Service in the completion of this work.

Funding statement: This research was supported by the Wellcome Trust (108375/Z/15/Z to C.F.S.), the University of York C2D2 Research Priming Fund grant, part-funded by the Wellcome Trust (097829/Z/11/A to C.F.S.), a White Rose University Consortium collaboration grant (to C.F.S., M.J.C. and J.H.), Nottingham Trent internal research funding (to C.J.M.).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2025.111770.

Data Availability

PRJEB1716 - Sequencing_ancient_DNA_calculus_samples (Original data) (European Nucleotide Archive).

PRJEB26093 - A plaque on both your houses: Exploring the history of urbanisation and infectious diseases through the study of archaeological dental tartar (Original data) (European Nucleotide Archive).

Supplementary information for An extensive archaeological dental calculus dataset spanning 5000 years for ancient human oral microbiome research (Original data) (Dryad).

References

[1] J.L. Metcalf, L.K. Ursell, R. Knight, Ancient human oral plaque preserves a wealth of biological data, Nat. Genet. 46 (4) (2014) 321–323.

[2] J.C. Brealey, et al., Dental calculus as a tool to study the evolution of the mammalian oral microbiome, Mol. Biol. Evol. 37 (10) (2020) 3003–3022.
[3] R. Forshaw, Dental calculus - oral health, forensic studies and archaeology: a review, Br. Dent. J. 233 (11) (2022) 961–967.
[4] A. Kazarina, et al., The postmedieval Latvian oral microbiome in the context of modern dental calculus and modern dental plaque microbial profiles, Genes 12 (2) (2021) 309.
[5] ENA, E.N.A. European Nucleotide Archive. 2023 [cited 2023 22 September]; Available from: https://www.ebi.ac.uk/ena/browser/home.
[6] J. Hendy, et al., Proteomic evidence of dietary sources in ancient dental calculus, Proc. R. Soc. B: Biol. Sci. 285 (1883) (2018) 20180977.
[7] D. Kim, et al., Centrifuge: rapid and sensitive classification of metagenomic sequences, Genome Res. 26 (12) (2016) 1721–1729.
[8] J. Dabney, M. Meyer, S. Paabo, Ancient DNA damage, Cold. Spring. Harb. Perspect. Biol. 5 (7) (2013) a012567-a012567.
[9] D.Y. Yang, et al., Technical note: improved DNA extraction from ancient bones using silica-based spin columns, Am. J. Phys. Anthropol. 105 (4) (1998) 539–543.
[10] M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing, Cold Spring Harb. Protoc. (6) (2010).
[11] G.G. Fortes, J.L.A. Paijmans, Analysis of whole mitogenomes from ancient samples, in: T. Kroneis (Ed.), Methods in Molecular Biology (Clifton, N.J.), Springer New York, New York, NY, 2015, pp. 179–195. Editor.
[12] M.-T. Gansauge, M. Meyer, Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA, Nat. Protoc. 8 (4) (2013) 737–748.
[13] Babraham Bioinformatics. *FastQC*. 2019 [cited 2023 4th January]; Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
[14] S. Chen, et al., FASTP: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (17) (2018) i884–i890.
[15] Joint Genome Institute. *BBduk guide*. 2023 [cited 2023 27 February].
[16] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (15) (2014) 2114–2120.
[17] NCBI. *Genome assembly GRCh38*. 2013 [cited 2024 31 July]; Available from: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/.
[18] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760.
[19] P. Danecek, et al., Twelve years of SAMtools and BCFtools, Gigascience 10 (2) (2021).
[20] H. Jónsson, et al., mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters, Bioinformatics 29 (13) (2013) 1682–1684.
[21] D. Knights, et al., Bayesian community-wide culture-independent microbial source tracking, Nat. Methods 8 (9) (2011) 761–763.
[22] R. Everett, B. Cribdon, MetaDamage tool: examining post-mortem damage in sedaDNA on a metagenomic scale, Front. Ecol. Evol. (2023) 10.
[23] J. Neukamm, A. Peltzer, K. Nieselt, DamageProfiler: fast damage pattern calculation for ancient DNA, Bioinformatics 37 (20) (2021) 3652–3653.
[24] M. Borry, et al., PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA de novoassembly, PeerJ. 9 (2021) e11845.
[25] A.S. Gancz, L.S. Weyrich, Studying ancient human oral microbiomes could yield insights into the evolutionary history of noncommunicable diseases, F1000Res. 12 (2023) 109.
[26] K.A. Ziesemer, et al., The efficacy of whole human genome capture on ancient dental calculus and dentin, Am. J. Phys. Anthropol. 168 (3) (2019) 496–509.
[27] K.A. Ziesemer, et al., Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification, Sci. Rep. 5 (2015) 16498.