Discovering the Origin of Mediaeval Parchment Through the Genetic Analysis of Western European Cattle

Stephan Sie

Student №	2751511
Supervisor	Rutger Vos, PhD
Examiner	Sanne Abeln, PhD
Course	XM_0072
EC	30

Abstract

In recent years, the Maelwael Van Lymborch Studies foundation has compiled a collection of several leaves of parchment — all fragments of the same mediaeval manuscript — containing edge decorations possibly created by the same craftsman that created the border decorations in manuscripts with miniatures by the Lymborch brothers. In order to determine whether these pages indeed have a connection to the Lymborch brothers or their French patron, the Duke John of Berry, we performed a protein and DNA analysis on samples of these leaves. The aim of the DNA analysis we report here was to determine the geographical origin of these and other parchment leaves, and therefore the manuscripts to which they once belonged. Through protein analysis, it was confirmed that nine parchment leaves originated from the species *Bos taurus*, i.e. cattle. For these nine leaves, a genetic analysis was performed in order to determine the breed. The analysis is made up of three distinct steps: firstly, the creation of a reference database containing SNPs from various cattle breeds across Europe; secondly, development and execution of a variant calling pipeline to determine SNPs in the leaves of parchment; and lastly, creation of phylogenetic and haplotype networks using the results from the previous two steps. From the resulting networks, the most closely related breed of cattle could be inferred, and in turn, the approximate geographical origin of the leaves of parchment. With the currently available data, the results suggest that of the nine parchment samples, six were from Germany, two from England, and one cannot be determined. The amount of data this is based on, albeit of satisfactory quality, is low enough to question the validity of this conclusion. For future reference a less stringent method could potentially improve the results, by allowing a higher number SNPs.

Introduction

Circa 2018, art historian Rob Dückers acquired a sheet of parchment that contains border decorations made by the same illuminator that made the border decorations in manuscripts with the miniatures of the Lymborch brothers, such as the *Belles Heures.* The sheet of parchment was part of a book that was unfortunately completely disassembled in the nineteenth century, but was thought to originally consist of circa 200 sheets. Now, these sheets are scattered all around the world, but 15 of them are in private collections in Gelderland, the Netherlands. This incentivized the foundation Maelwael Van Lymborch Studies to start research into the origins of this illuminator, their connection to the Lymborch brothers, and whether these pages could be from a book of hours in possession of — and perhaps ordered by — the duke John of Berry,[1] who also employed the Lymborch brother and ordered the *Belles Heures.*

Duke John of Berry was born in 1340 and was the third of four sons of the future king of France, John II the Good. The Duke had left quite an impression on the French political landscape during his years, partly as a result of his outstanding longevity (when he died in 1416, he was 76, and had by then outlived all his brothers). Although his political achievements are impressive, the Duke was also one of the most influential art patrons in the later Middle Ages. In his earlier years, he collected secular and religious buildings and later moved on to exotic and expensive objects. Few people during his time were as prolific as him in collecting manuscripts, though books were but a small part of his collection of valuables. While not the largest, his collection of around 300 manuscripts was of high quality and contained many religious texts such as bibles, psalters and books of hours.[2] As far as we can recollect from inventories, the Duke had a total of fifteen books of hours in his possession, of which eight have survived, including the *Belles Heures* and *Très Riches Heures.*[1] Both books of hours are regarded as masterpieces of the Middle Ages, as the books are some of the most illuminated manuscripts ever made, and both showcase the masterful illuminations created by the Lymborch brothers, although other illuminators were involved as well, such as the anonymous border illuminator mentioned earlier.

Books of hours were one of the most popular manuscripts in the later mediaeval period, especially after the popularisation of the printing press around 1450, which lowered the cost of producing books, enabling the less wealthy to access them. Before 1450, most manuscripts were made with parchment instead of paper and, as the word manuscript suggests, were written by hand. This meant that the production of a book of hours required quite a diverse array of skills and a significant amount of time to create. Parchment had to be sourced from animals, be it sheep, squirrel or cow, often as a by-product from animals used for food. In Western Europe, most parchment was made of sheep or cow skin. Once the skin was obtained, it would take around two weeks to make it ready for use, which involved cleaning, removing hair, scraping, stretching, and cutting.[3]

Afterwards the parchment could finally be passed on to a scribe, who would fill the pages with whatever text the patron wanted.[3] In the case of books of hours, they were certain to contain the Hours of the Virgin, prayers meant to praise the Virgin Mary that were to be said at regular intervals (hours) throughout the day (hence the name book of hours).[4] The scribes would keep spaces open for initials and miniatures, as these would always be added after the scribe was done. The vast majority of manuscripts that have survived the passage of time were completed. In some books that were found with unfinished miniatures, it can be seen that first sketches were made with lead points, which were then firmed up with a brown ink line.

Once the sketches were done, the next step was then gilding, the technique of applying gold sheets over solid surfaces. This had to be done before the sketches were painted, as this step could spoil painted surfaces. Painting was the last step before the book was bound together. Depending on what kind of pigments were used, the longevity, brightness, and, most importantly, the price, could vary a lot. This made illuminated manuscripts a product for the nobles, especially the more lavishly illuminated ones. Three especially skilled illuminators were the Lymborch brothers.[3]

The artists Herman, Paul, and Johan van Lymborch were born to Arnold van Lymborch and Mechteld Maelwael between 1385 and 1388 in Nijmegen, the Netherlands. It was in their family workshop in Nijmegen that the brothers learned their craft of illumination. Later, in 1400, Paul and Johan got the opportunity to intern at a goldsmith in Paris, because of the influence of their uncle, Johan Maelwael, whose name as a painter was already well known. After the internship, when the brothers returned to Nijmegen, they were jailed due to the ongoing war in the region, before being bailed by their uncle, who managed to get the funds from his employer, the Duke of Burgundy. Two years later, in 1402, the brothers themselves entered into the service of Philip the Bold, Duke of Burgundy, just like their uncle. The brothers served the Duke for two years, working on a *Bible moralisée,* until the Duke's death in 1404. The brothers managed to complete 384 miniatures and to start 128 more in the two years they were employed.

After the death of the Duke, the brothers had to look for a new patron, which ended up being their former patron's brother, Duke John of Berry. The brothers entered his service in 1405, where their first assignment was to finish some pages in the *Très Belles Heures de Notre-Dame,* an unfinished book of hours in the possession of the Duke of Berry. In the same year, the brothers started working on another book of hours, the *Belles Heures.* They finished working on the latter between 1408 and 1409. Between the moment that the brothers finished working on the *Belles Heures,* and then starting in 1411–1412 on the *Très Riches Heures du Duc de Berry,* little is known about their activities. The brothers would work on this new manuscript until their deaths in 1416, leaving it unfinished. It is thought that the brothers might have worked on other projects intermittently (although there are no records of this), based on the pace at which the brothers made their miniatures for *Très Riches Heures du Duc de Berry,* which was considerably lower than their pace for the *Belles Heures.*[5,6]

As the brothers used to work for the Duke of Berry, who had a sizable collection of manuscripts, the assumption of the brothers working on projects for which we lack documentation is not far-fetched. Furthermore, the Duke also employed other illuminators. Hence, there is a hypothesis that the fragments from the disassembled manuscript book with painted borders by an illuminator who also worked on the *Belles Heures,* alongside the Lymborch brothers, was once part of the library of the Duke of Berry. To actually answer the question whether the parchment pages are from the Duke of Berry's collection, the foundation Maelwael Van Lymborch Studies set out to research the pages from an (art-) historical and natural science perspective.

To answer the question from a natural science point of view, DNA and protein analyses were performed on samples of parchment, the first time ever in the Netherlands for parchment of this age. Subjected to the analyses were three leaves from (different parts of) the manuscript that may or may not have been part of the Duke's library, five leaves from various mediaeval manuscripts and a few scraps of modern parchment. The protein analysis was used to determine the species of animal from which the different parchment leaves were made, and the DNA was used to try to determine the breed of the species. This was done to determine the geographical origin of the parchment within Europe, and therefore the manuscript, as the parchment presumably originated from a place close to where the manuscript was created.

Various studies have been performed using ancient DNA extracted from historical, and also mediaeval, parchment.[7–11] As mediaeval manuscripts are often of extraordinary value, destructive sampling is often not an option. Instead, less destructive sampling through the means of rubbing an eraser on the parchment and collecting the dry eraser waste, sometimes called ZooMS, is done.[7] This technique was also performed by the lab at Naturalis Biodiversity Center to collect samples for species identification through protein analysis. A common approach within these previous studies is the identification of species from which the parchment is made, be it by either protein, DNA, or the analysis of both. However, few go as far as to perform further population genetics analysis.[7,11] This may be explained by the insufficient amount of data generated in each study: some studies retrieve only a small amount of endogenous DNA,[10] whereas others retrieve enough to cover 7–9% of the target species genome.[12] The biggest loss in raw reads originates from the filtering of mapped reads, which in a study of 1000-year-old parchment led from having 51.4% to merely 5.6% endogenous DNA.[8] The drop in mapped reads, in this case, is suspected to originate from a bias in DNA preservation or retrieval from parchment, where repetitive regions are favoured over gene-rich euchromatic regions. The bias could derive from the alkaline treatment in the parchment production process. The previously mentioned studies indicate that lower amounts of data are a common occurrence with the ancient DNA analysis of parchment. This can be expected as, after the heavy treatment needed to create parchment, cells containing intact nuclei are a rarity. In the case of proteins, however, there is an abundance of them, as the parchment is made of skin which consists of proteins.

Of the studies that were able to gather enough data for further genetic analysis, to the best of our knowledge, as yet none have gone as far as to use these data to determine the breed. They do, however, perform variant calling on the samples, after which clustering methods are used to either confirm the approximate geographical origin of the species based on continent,[7] or to analyse the similarity within the group of samples.[11] The amount of data needed to perform variant calling and its process do differ greatly between studies: some only used positions in which the alignment matches single nucleotide polymorphisms (SNPs) on the HapMap of the species and needed only one read,[12] whereas another opted for needing a minimum of three individual reads, with full consensus on the identity of the base.[11] The differences in the filtering of mapped data and their approaches towards variant calling clearly indicate that there are no standardised protocols for the downstream analysis of ancient DNA from parchment.

The protein analysis concluded that the leaves of parchment used in this research originate from *Bos taurus* (cattle). To then furthermore determine the breed of cattle from which the parchment was made, we made use of a panel generated from a resequencing project on different known breeds of cattle within Europe. Here we report the outcomes of this analysis. Our approach illustrates the extent to which population genetic analysis is feasible with data derived from whole genome high-throughput sequencing applied to ancient DNA extracted from less-destructive sampling of mediaeval parchment.

Methods

To reconstruct the approximate geographical origin of the animals used for the different leaves of parchment, we required three components. First, a reference panel containing SNPs from different *Bos taurus* (cattle) breeds throughout Europe; second, the SNPs for each piece of parchment; and third, a phylogenetic and haplotype network made from the SNPs from the reference panel as well as a piece of parchment.

To create the first two components, a number of existing tools together with custom Python scripts were aggregated into two Snakemake[13] pipelines, with a few additional scripts needed for the reference database and network analysis. See Figure 1 for a workflow of the analysis.

Data

During this research, two datasets were used. The first one originates from the *1000 Bull Genomes* project.[14] The second dataset consists of the raw sequencing data from the parchment generated by the lab at Wageningen University and Research. All data used for research can be found on Zenodo and NCBI (see Data Availability).

1000 Bull Genomes Data

The *1000 Bull Genomes* project started out as an idea at a conference in Melbourne, in 2011, by a group of researchers, one of whom, Ben Hayes, went on to found the project in 2012. The aim of the project was to gain insight into genetic variation in cattle and causative mutations affecting favourable economic traits in beef and dairy cattle.[14] Together with several institutes across the globe, the first run was done in 2012. Ever since, the consortium performs an additional run every six to twelve months. The whole genome sequencing (WGS) data generated during each run is collectively processed by Agriculture Victoria using their variant calling pipeline. During the variant calling step of the pipeline, joint calling is performed, analysing the samples within the run as a population. This provides more power to find breed-specific variation and rare variants, and produces more confidence for the called SNPs.[15] The latest publicly available results from the *1000 Bull Genomes* project are from run 8, consisting of 1832 cattle. The results are stored in VCF files at the European Nucleotide Archive (ENA, BioProject PRJEB42783), from which a selection of European breeds is used in this research, with a few American and Korean outgroups.

Parchment Data

Initially, small pieces of parchment and gum waste, originating from eight different sheets from six different mediaeval manuscripts, and a small vellum piece (P3_S9), with an age of between 500 and 1000 years old, were sampled in the Ancient DNA lab of Naturalis. For five of these sheets the geographical origin was already known. Parchment P195_S1, P700_S2, P786_S5, P78404_S3, and P78406_S4 originate from Belgium, Germany, The Netherlands, France, and Germany respectively. For the small vellum piece and the other three sheets (P1_S7, P2_S8, and P5_S6) the geographical origin is unknown. The three sheets P1_S7, P2_S8, and P5_S6 originate from the same manuscript, opening up the possibility of the sheets originating from the same animal. Ancient DNA was also extracted from the samples at Naturalis, to prevent further contamination of the samples.

Library prep was done at Wageningen University using the XGen ssDNA & Low-Input DNA Library prep kit, for which the manufacturer's protocol was followed with some small adjustments, such as the fragmentation step at the start of the protocol being skipped. DNA was then sequenced using paired-end sequencing on an illumina MiSeq, resulting in two FASTQ files for each piece of parchment, in the form of a forward (R1) and reverse (R2) read file.

Metadata

In order to create the reference panel database, it is necessary to know certain pieces of information for all the cattle that will be stored within it. Aside from the information on each SNP, such as position within the genome, it is necessary to know what breed the cow is, as well as the country and local region it comes from. The `cow_metadata_pipeline` was therefore created using Python and bash scripts to automatically retrieve and format this information in an easily-accessible way. An overview of the process can be seen in Figure 1.

The `cow_metadata_pipeline` can use both SRA run files or ENA file reports containing the cattle of interest as its input, and then proceeds through four steps. Firstly, BioSample, experiment, and run IDs are extracted from the input files. The BioSample IDs are then used to retrieve descriptions of each sample from the NCBI database. The descriptions contain, in most cases, the breed of the cow, which is extracted using a regular expression. Since information on the country of origin of the cow is not available within databases such as NCBI,[16] ENA[17] or any other easily accessible database, a manually-curated list of breeds and their origins is used to connect every breed to a country. The breed, BioSample ID, experiment ID, run ID, country and region within the country are then stored in a tab-separated file, ready for later use.

Relational Database Construction

With the VCF files from the ENA and the metadata collected using the `cow_metadata_pipeline`, the reference panel database can be populated. The data-

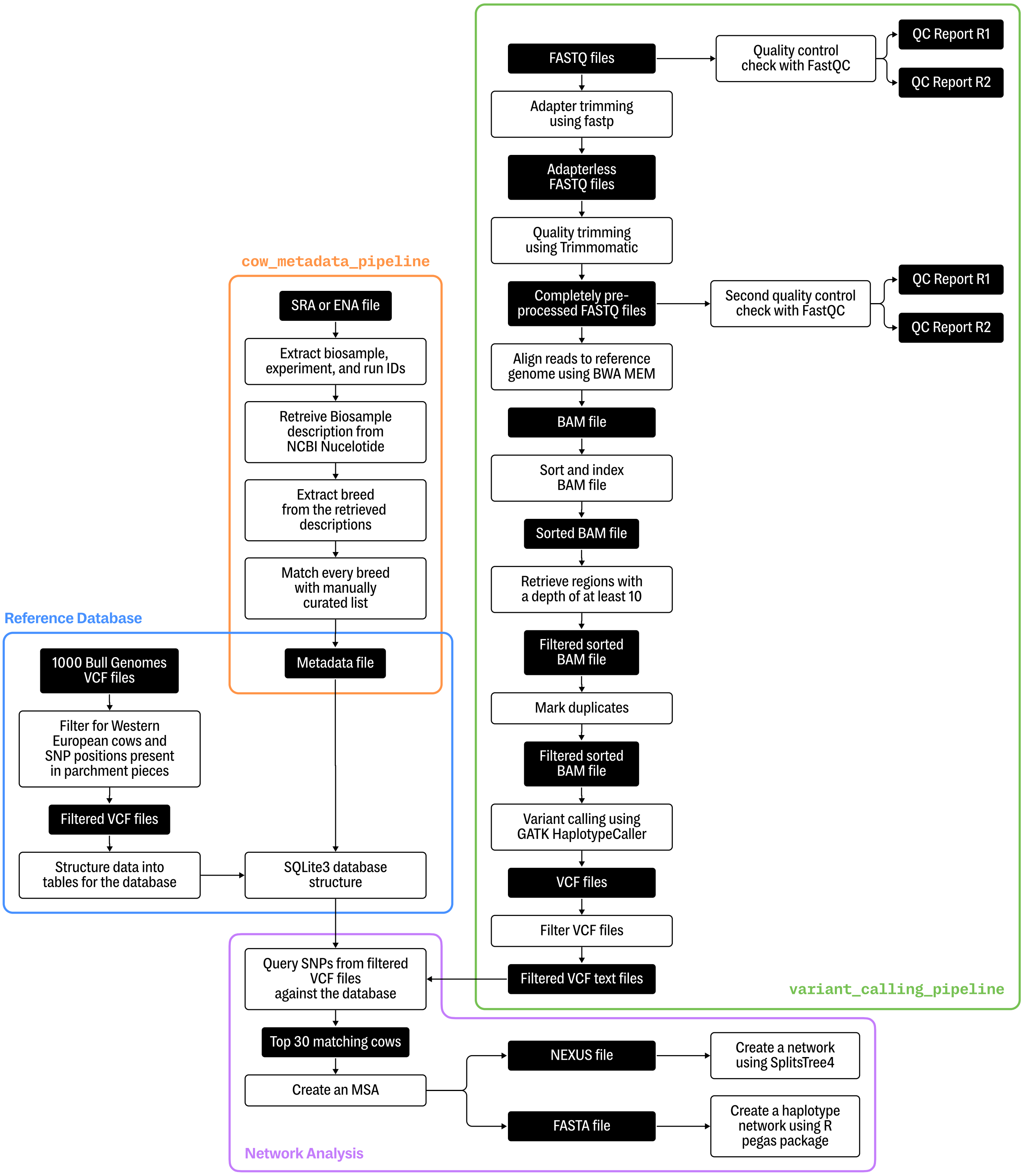

FASTQ files
Quality control check with FastQC
QC Report R1
QC Report R2
Adapter trimming using fastp
Adapterless FASTQ files
Quality trimming using Trimmomatic
Completely pre-processed FASTQ files
Second quality control check with FastQC
QC Report R1
QC Report R2
Align reads to reference genome using BWA MEM
BAM file
Sort and index BAM file
Sorted BAM file
Retrieve regions with a depth of at least 10
Filtered sorted BAM file
Mark duplicates
Filtered sorted BAM file
Variant calling using GATK HaplotypeCaller
VCF files
Filter VCF files
Filtered VCF text files
variant_calling_pipeline
cow_metadata_pipeline
SRA or ENA file
Extract biosample, experiment, and run IDs
Retreive Biosample description from NCBI Nucelotide
Extract breed from the retrieved descriptions
Match every breed with manually curated list
Metadata file
Reference Database
1000 Bull Genomes VCF files
Filter for Western European cows and SNP positions present in parchment pieces
Filtered VCF files
Structure data into tables for the database
SQLite3 database structure
Query SNPs from filtered VCF files against the database
Top 30 matching cows
Create an MSA
NEXUS file
FASTA file
Create a network using SplitsTree4
Create a haplotype network using R pegas package
Network Analysis

Figure 1: Workflow of the analysis. *Figure 1 depicts the workflow used during the research with individual parts such as pipelines grouped and named. With the three main parts being the* `variant_calling_pipeline`, *Reference Database, and Network analysis.*

base is made using SQLite,[18] allowing us to have a serverless and easily-portable database. The database schema consists of three tables: the *accessions* table, containing the metadata on each sample; the *refsites* table, containing genomic coordinates in the reference genome used; and the *snps* table, containing the SNPs from the *1000 Bull Genomes* project (see Figure 2). Each of the unique IDs in the three tables are used for indexing. In total, the database stores the SNPs from 857 cattle, with dozens of individuals per breed. In order to reduce the amount of data that needs to be stored, the database only contains SNPs for the loci for which the leaves of parchment also have a sufficiently covered SNP. The countries from which the breeds originate are spread throughout Western Europe, with the following regions included: Ireland, France, Germany, Belgium, Spain, Switzerland, Austria, Scotland, and England. Aside from the Western European cattle, a set of Korean and American cattle are included as outgroups.

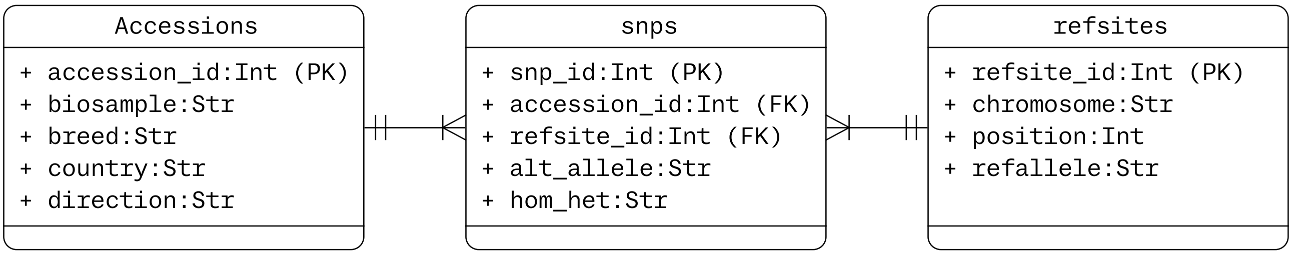
Variant Calling

In order to extract the SNPs from the raw sequencing data, a few steps have to be taken that have been aggregated into a Snakemake pipeline called the `variant_calling_pipeline`. See Figure 1 for an overview of the process.

As the data we received from the lab is raw sequencing data, the first thing that needs to happen is a quality control check, for which we used FastQC.[19] This was then followed by adapter trimming, executed using *fastp.*[20] Alongside the adapters, the first twelve nucleotides of the R2 reads were trimmed, as the kit used for sequencing generates a dinucleotide tail at the start of the R2 reads. With the adapters and dinucleotide tails removed, we then performed quality trimming with the tool Trimmomatic.[21] The tool trimmed the reads based on the following settings: remove bases from the front and back until a base with a quality equal to or higher than 20 is reached; use a sliding window of three, trimming those with an average quality below 15; and drop reads with an average quality below 20 or a length (after trimming) below 35. An additional FastQC check was then performed to see the effects of the preprocessing.

The preprocessed reads from the parchment samples were then mapped against the *Bos taurus* reference genome (ARS-USD1.2), which was done using BWA MEM.[22] As we are dealing with ancient DNA of 500–1100 years old, extracted from chemically and physically heavily treated animal skin, some damage may be expected. This damage can, for example, be in the form of deamination of cytosine to uracil, or DNA fragmentation.[23] Fragmentation of the DNA can be observed in FastQC reports, however, deamination is more complex to identify. To get insight into any damage related to deamination, we used MapDamage2.0[24] to track and quantify DNA damage. MapDamage2.0 takes an alignment (BAM) file as input and visualises the damage with graphs.

Aside from possible damage, another challenge we encountered was that the number of reads, and therefore the coverage in mapped regions, was very low. We do, however, need a certain amount of coverage to be able to call SNPs with a high enough degree of confidence. It was therefore decided to put a threshold coverage of 10 on the mapped regions, as in previous research of ancient DNA at Naturalis Biodiversity Center, this led to satisfactory results. Making use of a custom Python script, regions where the total coverage was equal to or higher than 10 were found and stored in a BED file. The BED file was then used by Samtools[25] to extract reads which map within that region. The extracted reads that passed the coverage threshold were then passed on to Picard to mark duplicates, since duplicates could potentially introduce a bias in the next step,[26] variant calling. The variant calling step was performed using the GATK3.8 HaplotypeCaller[27] tool. The HaplotypeCaller takes as input the reference genome, and the BAM file with duplicates marked. The tool processes the input and outputs a VCF file containing all the found SNPs for the parchment sample. These files, however, contained a lot of SNPs that fall outside the regions with coverage >=10. This was because, when we extracted the reads from the alignment, some reads started within the specified region but ended outside of it. This created the possibility of calling SNPs outside of the regions we want. The results in the VCF file were therefore filtered on these regions using a custom Python script. The script produced a text file containing the data from all the SNPs within the regions, which was the final output of the pipeline.



Accessions
+ accession_id:Int (PK)
+ biosample:Str
+ breed:Str
+ country:Str
+ direction:Str

snps
+ snp_id:Int (PK)
+ accession_id:Int (FK)
+ refsite_id:Int (FK)
+ alt_allele:Str
+ hom_het:Str

refsites
+ refsite_id:Int (PK)
+ chromosome:Str
+ position:Int
+ refallele:Str

Figure 2: Database structure. *For each table, the columns are specified, showing what kind of data is stored within and whether the column functions as a foreign key (FK) or primary key (PK). Crow's foot notation is used to specify relationships between the tables.*

Phylogenetic and Haplotype Networks

We then produced phylogenetic and haplotype networks from the reference panel database and the parchment VCF text files (see Figure 1 for an overview of the process). A custom Python script was used to generate a query that retrieves, from the reference database, all the positions and SNPs that match with a given input text file from the `variant_calling_pipeline`. The script proceeds to: connect to the reference database; use the generated query; rank cattle in the database by number of matches; select the top 30 matches; retrieve the SNPs the top 30 individuals have in the same positions as the parchment sample (and, if non-existent, put the reference allele); filter the SNPs based on the QUAL and DP of the parchment sample; create a multiple sequence alignment (MSA), using IUPAC ambiguity codes; and finally, save the MSA in the NEXUS file format.

The filter applied was either a DP of 5 or 10 and a QUAL of 20 or 30, respectively, depending on the parchment sample. Parchment samples P1_S7, P3_S9, P195_S1, and P78404_S3, were filtered using a DP of 5 and QUAL of 20. Parchment samples P2_S8, P5_S6, P700_S2, P786_S5, and P78406_S4, were filtered using a DP of 10 and QUAL of 30. The filter thresholds were set tailored to the samples because the more stringent thresholds (DP 10 and QUAL 30) reduced the number of SNPs for some samples to a level where further analysis could not be performed. To create the MSA, the IUPAC single nucleotide ambiguity codes were used in order to preserve the heterozygous information of the SNPs for the phylogenetic network.

The NEXUS file was then used as input for SplitsTree4,[28] where we used the Neighbour-Net[29] method, which constructs networks from inferred maximum likelihood distance matrices. The resulting network can then be used to infer the closest related breed of cattle, for which the geographical origin is stored in the database.

In addition to the phylogenetic network, the data was also analysed using a haplotype network. In order to do this an assumption needed to be made, which originates from the fact that we are trying to make a haplotype network using diploid data. The assumption is that all the heterozygous SNPs originate from the same strand. Normally the haplotype can be determined by phasing, the process of determining the haplotype from genotype data, but this requires either closely-related individuals to infer from, or a reference dataset containing common haplotypes,[30] neither of which are, as far as we are aware, available for our data. Therefore, with the low amount of data available to us consisting of mainly heterozygous SNPs, it was decided to assign all the SNPs to one strand.

The haploid sequences were stored in a FASTA file, and then loaded into R, where they were processed using the package *pegas*.[31] The haplotype function was used to extract the different haplotypes from the input sequences, after which the pairwise distances were calculated between the sequences. With the distances calculated, the network was created using the Randomised Minimum Spanning Tree method (RMST). This method was chosen over others, such as Minimum Spanning Network (MSN), as RMST produces fewer links between the haplotypes while taking the ambiguities created by the ordering of the data into account.[32]

Validation

To reduce the chances of drawing false conclusions from the results of the analysis, throughout the analysis of the parchment data, the results from the various processes were validated by using either already-present quality scores, or ones generated by the tools used. With the raw data, FastQC was used to get a first indication of its initial integrity. During preprocessing, the Phred scores within the FASTQ files were used to remove data with low confidence, and artefacts from sequencing were removed. Afterwards, another FastQC run was performed in order to inspect the integrity of the data, ensuring no conclusions were based on faulty data. After the mapping of the raw data to the reference genome, the quality of the mapping was checked. This was done using the Samtools Flagstats tool, which gives a summary of the alignment data.

Once duplicates were marked by Picard, a summary report of the process was automatically generated. In the report, the metric we paid attention to was the *percent_duplication,* where we expected a low percentage that lies at least below 20%. In the case of the variant calling step, two metrics were taken into account (QUAL and DP), which we used to apply a hard filter before generating networks. To validate our method of creating phylogenetic and haplotype networks to determine the closest breed, a test run using a French cow from the reference database was performed. The test run included the same maximum number of the 30 closest hits, except an additional few spots were added for Korean cattle to show some outliers. The aim of this test run was to see whether there is enough breed-specific variation within the Western European cattle population, to find closely-related cattle without specific markers. The five samples for which we know the geographical origin also serve as a benchmark of the aforementioned methods, since if the methods work correctly the right geographical origin should be produced.

Results

Variant Calling

Figure 3 shows some of the plots generated by FastQC for one parchment sample, P786_S5, before and after the preprocessing. Figure 3A shows the quality of the bases in the R2 reads, clearly showing a drop in quality towards the end of the read. This is a common occurrence with Illumina reads.[33] It can also be seen in Figure 3B that the percentage of bases in the first twelve nucleotides is skewed towards two of them, indicating the dinucleotide tail ligated by the library prep kit. The last notable result from the raw data is that there appears to still be a presence of adapters in the data (see the red line in figure 3C). After preprocessing, the quality of the reads is higher on average (Figure 3D); percentages of bases are within reasonable levels and no longer skewed towards two nucleotides, meaning the dinucleotide tail is gone (Figure 3E); and adapters are no longer present within the reads (Figure 3F). Showing that the preprocessing was successful and downstream analysis can be performed. Aside from the aforementioned plots, the sequence length distribution plot in the FastQC report is of great importance, as it will confirm whether the ancient DNA was fragmented or not. For all the parchment samples, the distribution before preprocessing was the same, with all the reads having a length of 201, indicating that fragmentation was not an issue.

The percentage and number of mapped reads for all samples can be seen in Table 1, together with the number of reads after various preprocessing steps in the analysis. On average, 67% of the reads mapped against the reference genome. When looking at the areas of the genome where coverage is at least 10, it becomes apparent that only a small percentage of data is left to work with.

Table 1 also shows that on average merely 0.018% of the *Bos taurus* genome has a coverage that meets our standard of >=10. These regions also still contained possible duplicates, as can be seen in the last column of the table; each piece of parchment has a few hundred to a few thousand duplicates.

Using MapDamage2.0, we show deamination of the bases in plots such as in Figure 4. The two line graphs on the bottom of Figure 4 show in red the C → T substitution based on position from 5′ to 3′, and in blue the G → A substitution. As deamination takes place, at the start and ends of DNA, the C is replaced by U over time, which in turn means that, during polymerisation, an A is built in to match the base. This in turn causes a T to be incorporated to match the A in the next cycle, creating spurious substitutions.[23] In the case of deamination, we would expect the red or blue line to either start high on the left and decrease over the length and/or to start low and increase over the length.

The four plots at the top of Figure 4 show the frequency of each of the four bases in- and outside of the reads, with the read corresponding with the grey box. In the case of deamination it can be seen that the frequencies of C and G are lower within the read than outside it, compared to T and A, which show slight increases. When we look at the results for the sample in Figure 4, it is apparent that none of the telltale signs of deamination are present. The frequencies of all four bases remain mostly constant, and substitutions of C → T and G → A are almost nonexistent. The same results were found for the other parchment samples, for which the MapDamage results can be seen in Supplementary Figure 1–8.

Table 1: Overview of number of reads. *Table 1 contains the amount of reads at different steps throughout the analysis, indicating the limited scope of information available.*

Sample	Total reads before pre-processing	Total reads after pre-processing	Mapped reads (percentage of total reads after pre-processing)	Reads within region of coverage >= 10 (percentage of total reads after pre-processing)	Percentage of reference genome covered with coverage of >= 10	Duplicates within regions
P1_S7	1486160	1454874	970074 (66.68%)	141479 (9.72%)	0.015%	331
2_S8	1993048	1936437	1260250 (65.08%)	112726 (5.82%)	0.008%	201
3_S9	3488412	3386840	2272051 (67.08%)	253253 (7.48%)	0.008%	933
5_S6	1788434	1756179	1411591 (80.38%)	52094 (2.97%)	0.030%	192
195_S1	2941268	2904499	2217571 (76.35%)	698963 (24.06%)	0.021%	3350
700_S2	2599610	2531934	1956739 (77.28%)	743736 (29.37%)	0.038%	2492
786_S5	1567764	1500039	907617 (60.51%)	242688 (16.18%)	0.018%	459
78404_S3	2588818	2475771	1296489 (52.37%)	534740 (21.60%)	0.007%	885
78406_S4	1704308	1696580	1453471 (85.67%)	249130 (14.68%)	0.026%	1739

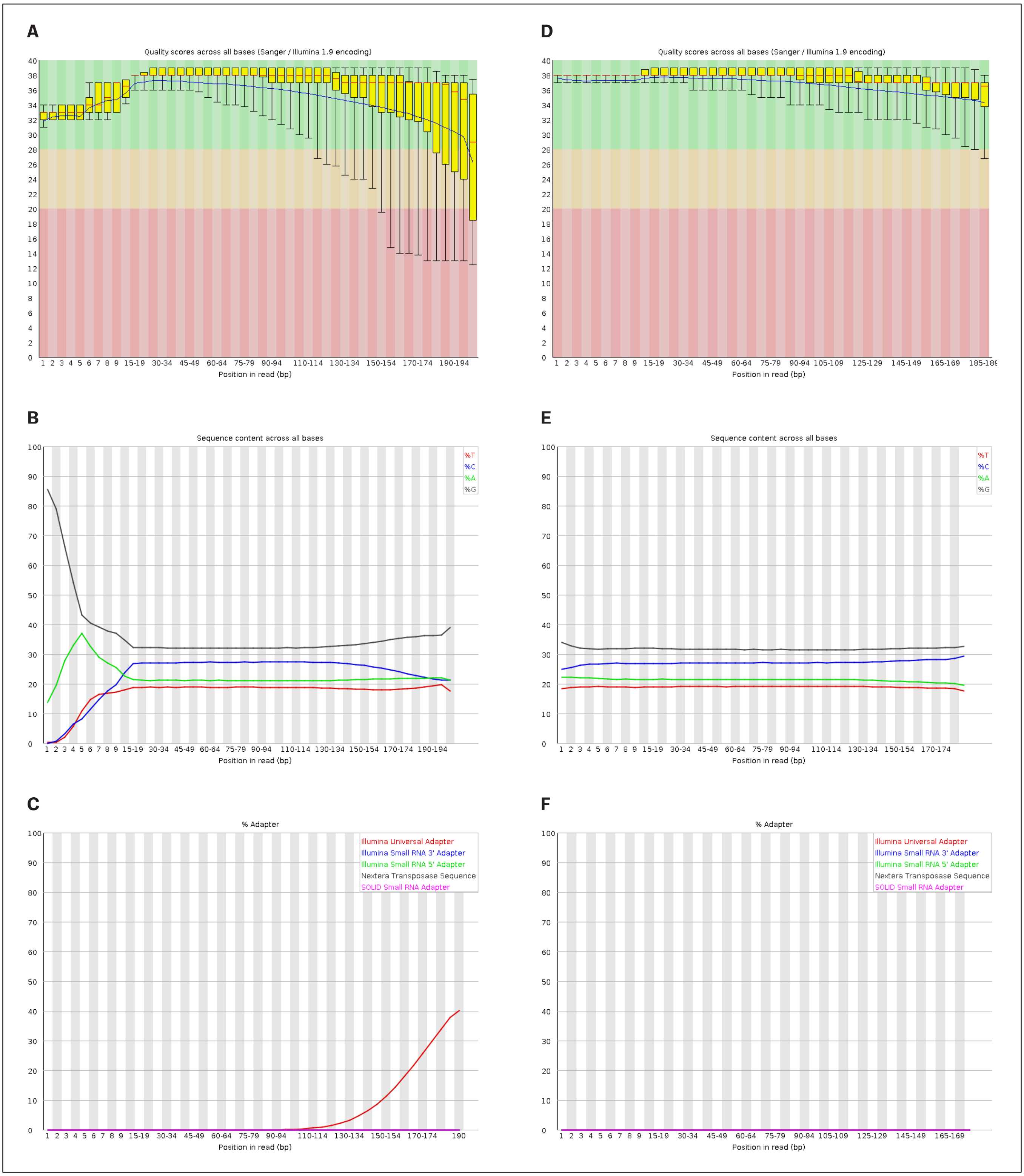


Figure 3: FastQC plots. *The figure shows a small selection of plots created by FastQC, before and after preprocessing, for the R2 reads of parchment P786_S5. Figures 3A and D depict the quality of the nucleotides in each position of the reads; B and E show the percentage of each base in each position of the reads; and C and F depict the presence of adapter sequences in the reads. A clear improvement of the quality can be seen in the differences between figures A, B, and C and figures D, E, and F, indicating the preprocessing was a success.*

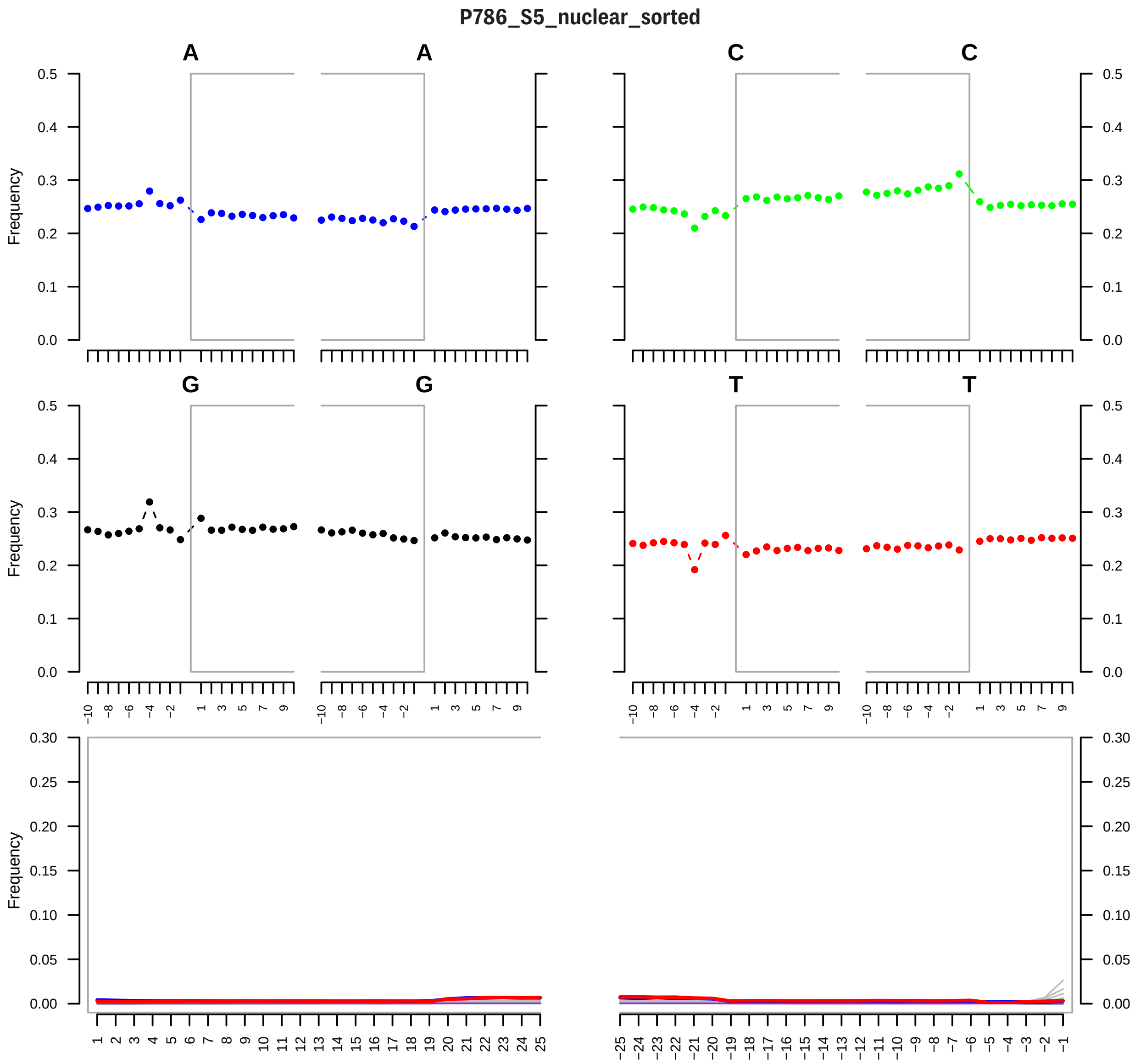


Figure 4: ***MapDamage plots P786_S5.*** *The MapDamage report for parchment piece P786_S5. The plots show no apparent signs of deamination within the sample.*

Table 2: Overview of number of SNPs. *The number of SNPs after different filtering steps. From the raw count until the final number of SNPs used for the network analysis.*

Sample	Total number of raw variants called	Number of variants within alignment region of coverage >= 10	Number of variants matching with the database	Number of SNPs after applying DP and QUAL filters
P1_S7	2500	1597	414	7
P2_S8	1096	126	55	28
P3_S9	3272	1929	479	58
P5_S6	1007	261	92	28
P195_S1	1554	215	85	37
P700_S2	3750	355	116	44
P786_S5	2114	209	68	31
P78404_S3	3091	438	77	20
P78406_S4	3286	2801	721	6

The results from the variant calling step can be seen in Table 2, which shows the number of SNPs after different filtering stages. Starting from the raw number of found SNPs the number of SNPs we consider in further analysis is lowered, besides due to quality filters, by our disregarding of SNPs within the regions with coverage >= 10 lacking a match within the database. The end results of the filtering can be seen in the last column of Table 2; these are the final numbers of SNPs per parchment piece used in the phylogenetic and haplotype networks.

Phylogenetic and Haplotype Networks

Figure 5 depicts two phylogenetic networks created with SplitsTree4,[28] using the database and results from the `variant_calling_pipeline` for parchment piece P786_S5. Figure 5A shows the network created before the QUAL and DP thresholds were applied. The branch for the piece of parchment, numbered 41, is exceptionally long compared to the other cattle. Figure 5B shows the network created when applying the QUAL and DP filter. The network consists of two quite clearly-defined clusters, with the parchment sample, numbered 31, being closest to a cluster containing German, Swiss, and English cattle breeds depicted with the numbers 1–6. The branch of the piece of parchment can be seen to clearly extend further away from the other cattle than any other node in the network. This can be seen for every piece of parchment analysed, in some cases even more apparent than what can be seen in Figure 5B, signifying that the animal used for the leaf of parchment is very different on the genetic level from the cattle in the reference database.

The phylogenetic networks of the other parchment samples can be seen in Supplementary Figures 9–16.

Figure 6 shows the haplotype network generated in R using pegas[31] for parchment sample P786_S5. The network shows a large group of similar samples on the left side, indicated by the size and connections having one difference each. However, P786_S5 is unfortunately not part of it. Our parchment sample can be found in the top right corner of the network, with a large number of differences between itself and the next closest connection, a German cow.

This pattern of having a large number of differences between the parchment and its closest connection repeats itself with the other parchment samples. There is one notable exception, which is the parchment piece P1_S7 (see Supplementary Figure 17).

The haplotype network can also be seen as a translation of what we see in the phylogenetic network in Figure 5B. The branch of the piece of parchment, number 31, can be seen to have a substantial length, indicating a large difference between itself and the other cattle. This large difference is then also notable in the haplotype network, but then depicted with the amount of tangential lines. In Figure 5B, the two clusters are clear to see, which once

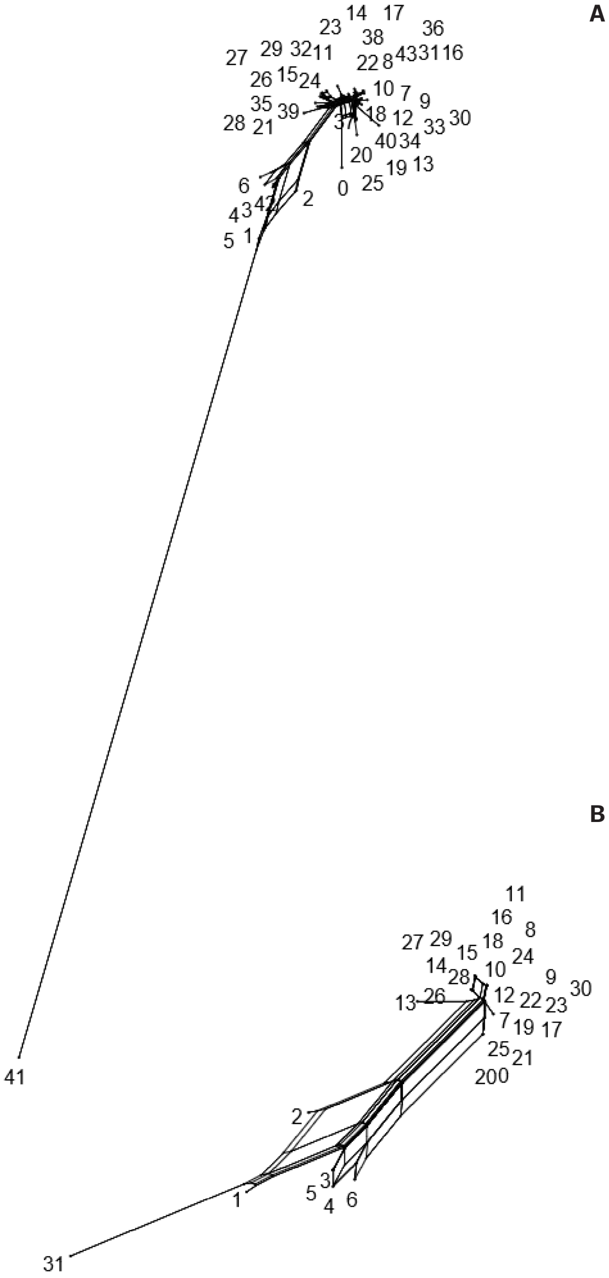


Figure 5: Phylogenetic network of P786_S5. *Figure 5A depicts the phylogenetic network before the QUAL and DP thresholds are applied. The sample (number 41) can be seen to be extremely distant from any other cattle. Figure 5B depicts the phylogenetic network with the QUAL and DP thresholds applied. Within the network, the piece of parchment (31) together with the top 30 hits(1-30) in the reference database and the reference genome (0) are shown. The length of the branch indicates the amount of divergence between samples. The closest tip to the parchment sample in Figure 5B is number 1, which is a German breed of cattle.*

again can then also be seen in the haplotype network of Figure 6. The haplotype networks generated for all the other parchment samples can be found in Supplementary Figures 17–24.

Table 3 shows the geographical origin of the three nearest breeds to the cattle for our parchment samples. The origins were determined by the nearest neighbours of the leaves of parchment in the haplotype and phylogenetic networks. It is apparent that Germany is a common geographical origin, followed by Switzerland and England, although there is never a full consensus of the three closest breeds. For the three sheets originating from the same manuscript (P1_S7, P2_S8, and P5_S6) the table does not show a consensus on the origin. The geographical origins shown for these sheets also do not match our expectation of them originating from France.

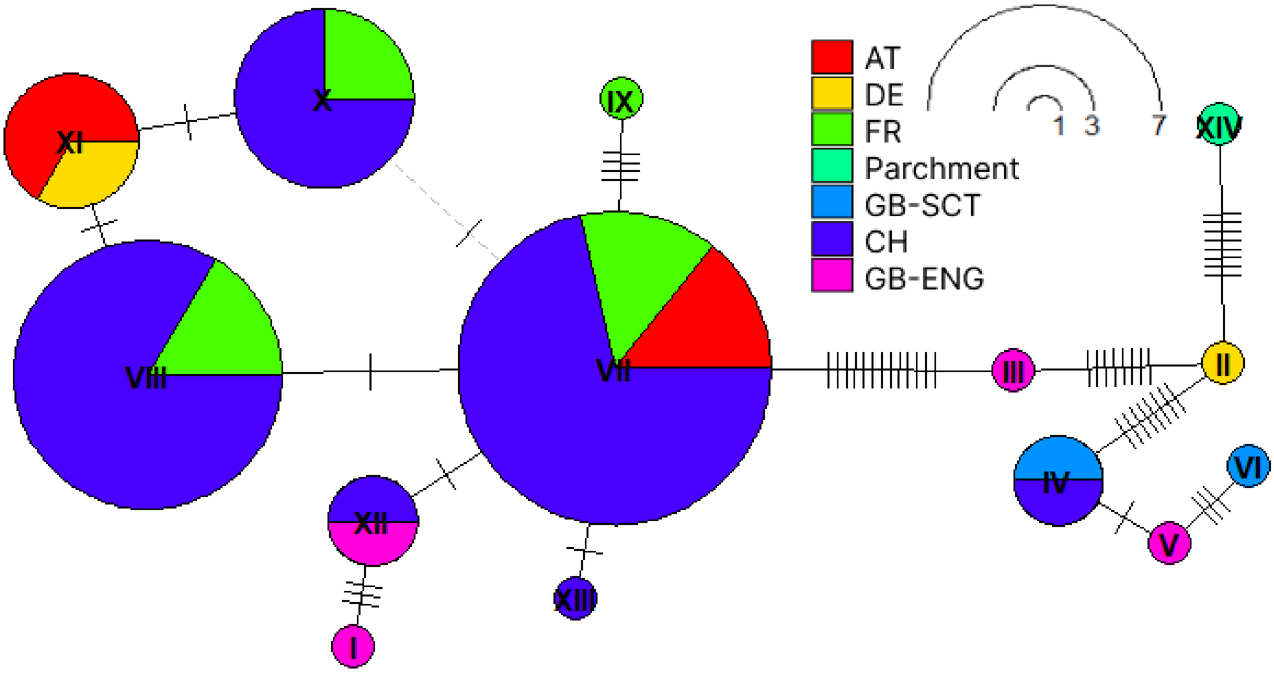


Figure 6: haplotype network P786_S5. *The haplotype network above shows the haplotypes from the top 30 hits of the parchment P785_S5 with the reference database. The size of the circle indicates the number of samples with that specific haplotype. The number of tangential lines on the connecting lines between circles indicate the number of changes between two haplotypes, and the colour indicates the country, as per the legend, of the breed. The network shows that the nearest neighbour to the parchment sample is a German cattle breed.*

Table 3: Geographical origins of nearest neighbours. *Table 3 depicts the results from the haplotype and phylogenetic network analysis. For each of the samples the approximate geographical origin of the three nearest neighbours are shown.*

Sample	Nearest neighbour	Second nearest neighbour	Third nearest neighbour
P1_S7	Non-conclusive	Non-conclusive	Non-conclusive
P2_S8	Germany	Scotland/England/Switzerland	Scotland/England/Switzerland
P3_S9	Germany	England	Scotland/England/France/Spain
P5_S6	England	Switzerland	Scotland
P195_S1	Germany	Scotland	England/Switzerland
P700_S2	Germany	Switzerland	England/Scotland
P786_S5	Germany	England	Scotland/Switzerland
P78404_S3	Germany	Scotland/Switzerland/England	Scotland/Switzerland/England
P78406_S4	England	Switzerland	Switzerland

Validation

The figures from the FastQC reports, in Figure 3, clearly indicate that the preprocessing removed low-confidence reads and sequencing artefacts, like the dinucleotide tails. The Phred scores of the bases are now approximately 37 on average, ensuring a high quality from which no faulty conclusion can be drawn.

The fourth column of Table 1 shows the percentage of mapped reads, with the average being around 67%. The summaries generated from the alignment files also show that, on average, 98% of mapped reads are correctly mapped pairs, with on average only 2% being mapped singletons.

Looking at the sixth column of Table 1 it can be seen that the number of duplicates within each sample is also far below 20%, once again confirming that the sequencing was done successfully and no bias was introduced in the variant calling.

The results from the variant calling seen in Table 2 indicate that the number of low-confidence SNPs in the results is quite high. The filtering of the SNPs results in only a small number, in some cases an extremely small number, of SNPs being used for the network analysis.

The generated phylogenetic network of the test run can be seen in Figure 7. The network shows a clear distinction between two groups, namely the Korean cattle on the left and the European cattle on the right. Circled in red is the French cow used to query against the reference database, seen to match with itself. The query resulted in 423 SNPs being matched. When looking into which countries the 30 closest matches were, only 3 are from France. The three French cattle are all from different breeds, none the same as the queried cow. The closest connection in the network to our queried cow, number 39, is one of these three French cattle. The other French cattle are numbers 52 and 48, with the other cattle in the network being mostly from Switzerland, followed by English, Scottish, Italian, Austrian and Spanish cattle. The haplotype network generated with this data was too densely clustered to read accurately, and is therefore left out.

The results in Table 3 for the five sheets of parchment for which we know the geographical origin show that the closest cattle breed matches that of the actual origin for only two of the five (P700_S2 and P78404_S3).

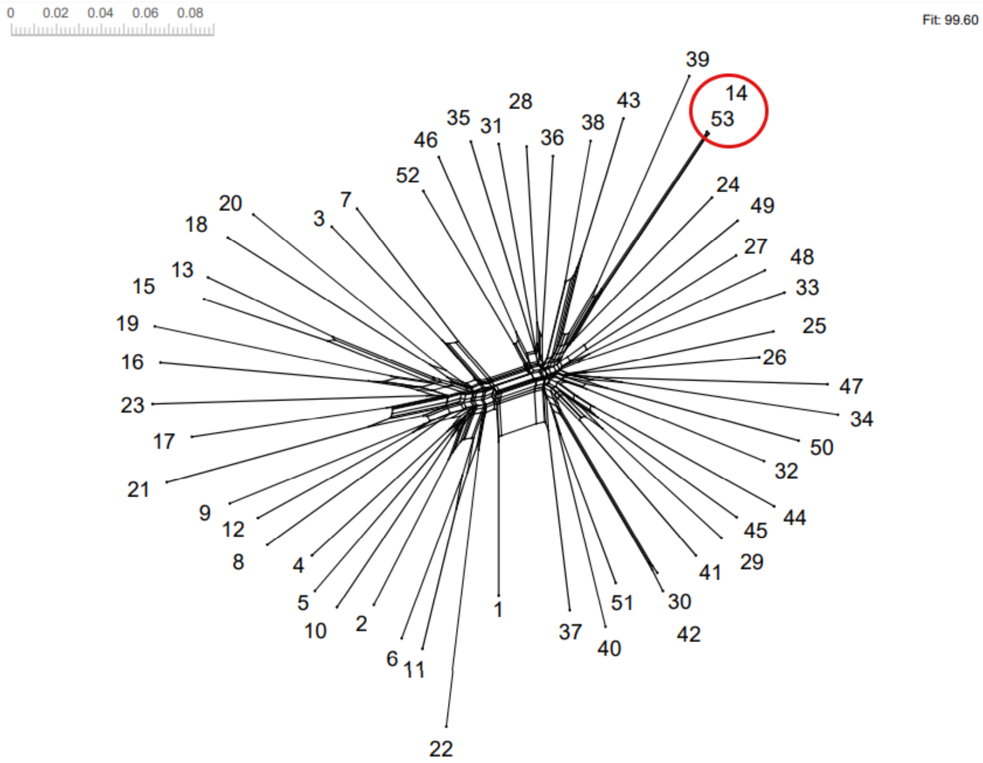


Figure 7: Phylogenetic network of the validation run. *The figure depicts the phylogenetic network generated by the test run. The numbers 1–23 indicate the Korean cows, with 24–53 being European cattle from various countries. Circled in red is the cow used to query against the database, matching with itself. A clear distinction can be seen between the Asian and European cows.*

Discussion

Looking at the figures from the FastQC report in Figure 3, together with the number of reads left after preprocessing, which can be seen in the third column of Table 1, it is apparent that the quality of the data is high. However, the total number of reads is quite low, as we are dealing with the *Bos taurus* genome which has a length of 3 Gb. With about one to three million reads per parchment sample, with a trimmed length of ~185, in the case that all reads were to map to the genome there would still not be enough to cover the genome once. In actuality, not all of the reads map to the reference genome (see the fourth column of Table 1), with on average only 67% of the reads mapping. The average mapping percentage of 67% demonstrates that a sizable portion of the data could still be contaminants, which either originate from the surface of the parchment or were introduced during the handling of the parchment in the lab. Aside from that, there is also the possibility that the parchment was made of bull's hide. With the reference genome not including the Y chromosome, this data would not be mapped, possibly explaining a small part of missing data.

The average mapping percentage of 67% is still a satisfactory rate, particularly considering the age of the DNA. However, since the number of reads is then once again significantly reduced to meet our standard of depth needed for variant calling, the final coverage of the genome is extremely small, averaging around 0.018% (see Column 6 of Table 1). This is not unexpected, as similar studies retrieving ancient DNA from parchment have encountered the same issue.[8]

The alignment was filtered for regions with a depth of at least 10. This threshold could be seen as quite stringent for ancient DNA, as another study has used a lower depth,[11] although it has shown to be fair, as previous studies at Naturalis delivered good results with it. A depth of close to 10 is also a typical depth to see when performing variant calling using WGS data in cattle.[34–36] Applying this threshold did, however, severely limit the scope of regions to be analysed, potentially having led to not identifying breed-specific SNPs in the parchment. For a follow-up study, in order to broaden the scope, it could be beneficial to either reduce the required depth or resequence a piece of parchment aiming for a higher depth.

The results from the variant calling analysis can be seen in Table 2. The raw results seemed promising, however further inspection of the data showed quite a large number of variants called with DP values of far below 10. Many variants also lie in overhanging regions where, as previously explained, reads extend outside of the regions with a depth of 10. This led to filtering of the VCF files and reducing the number of SNPs. Using this filtered data, the database was queried for matches and a phylogenetic network was generated using SplitsTree4. In the resulting network, the length of the branch representing the piece of parchment was tens to hundreds of times longer than branches between cattle in the database. From an evolutionary standpoint, 500 to 1000 years wouldn't be enough time to explain this disparity between the parchment and modern cattle in the network.

Because of the low number of SNPs found per piece of parchment, up until this point we had been lenient by using every SNP within the region of the alignment with a depth of 10. This seemed to introduce an extensive amount of noise in the network analysis (see Figure 5A). It was therefore decided to filter the SNPs that matched with the database, based on their DP and QUAL score. A number of different thresholds were tested, with the combinations of DP 5, QUAL 20 and DP 10, QUAL 30 generating the lowest amount of noise in the networks. Whether these thresholds reduce noise by filtering out likely false positives or only limit the disparity between the parchment and reference cattle to an acceptable level, would have to be researched in more depth.

This point was brought up because we do not know whether the SNPs found in the parchment samples contain known markers for certain breeds, or for that matter any information aside from position and genotype. A good follow up research (which could also validate the research done here) would be to analyse the raw SNPs found for the parchment samples, annotate them, and see whether breed-specific markers were found using the *1000 Bull Genomes* data.

The number of SNPs left over after filtering is extremely small for some parchment samples, as can be seen in Table 2. The resulting networks generated also show this in their relative lack of complexity, such as parchment piece P1_S7; the networks generated using more SNPs are more complex. A common sight in every network is the length of the branch belonging to the parchment sample, indicating the difference between the historic animals and the modern cattle. However, as mentioned before, the amount of time passed doesn't fully explain this genetic difference seen in the network. A possible explanation for this could however be that the DNA in the parchment is more heterozygous compared to modern cattle. This could stem from the way we nowadays carefully monitor and select individuals, that we then selectively breed, to produce the best offspring.

The results from the haplotype networks show a similar story as the phylogenetic networks: instead of long branches, it shows a large number of tangential lines for the parchment sample. The added benefit of these networks is that they give us a clearer insight into the more closely-related breeds and where they are from. The results in Table 3 show us the countries for closely-related breeds for each piece of parchment, with Germany, England and Switzerland appearing frequently. For the five leaves of parchment for which we knew the origin, we only managed to match two of the leaves with their origin

(P700_S2 and 78404_S3). For the other three leaves and the leaves that originate from the same manuscript, the results do not match our expectations. The three leaves from the same manuscript also do not form a consensus on where they do originate from. The most relevant conclusion we can draw from this is that the three leaves do not come from the same animal.

These results do ultimately need to be looked at with a bit of scepticism, as only two of the five benchmark samples were assigned the right origin. Aside from that, the network analysis shows that the differences between the historic animals used for parchment and modern cattle is quite significant. Together with the wide variety of countries, low number of SNPs, and small scope of the genome analysed, it could be possible only SNPs common within the whole Western European population of *Bos taurus* were called. When looking at the phylogenetic network generated from the validation run for the network analysis, it can be seen that, even with 423 high-quality SNPs, getting a close match with cows originating from the same geographical area is hard. This demonstrates once again that breed-specific markers would be beneficial, as simply matching all found SNPs to the database doesn't necessarily allow for a straightforward conclusion. What the network also illustrates, with the clear distinction between the Korean and European cattle and the large variety of geographical origins, is that a potentially large amount of variation within individuals is shared across the whole European population of cows. This could be due to a common ancestor, or because selective breeding of cattle has led to a more homogenous population overall. This would need further research, as there was not enough time to determine the likely cause within the scope of this project.

Another source of this wide variety of breeds appearing in the results could be the fact that we only plot the 30 most similar cattle. Aside from this being a factor limited by hardware, the limit of 30 is only a small number of the reference data we have access to. The way the top 30 were chosen, based on the number of matching SNPs with the parchment, might not be the best way of finding similar cattle as well. Aside from the SNPs that match, the positions within the genome that do not have a SNP are important as well. However, this would again rely on breed-specific markers to effectively investigate further. As we do not have those, trying to match the largest number of SNPs was our best option.

The analysis described in this report was originally optimised towards mitochondrial DNA. However, the number of reads in the raw data which mapped to the mitochondrial genome was extremely low, leading to insufficient data to call SNPs for half of the samples, and too few SNPs to effectively analyse the remaining parchment samples. The mitochondrial genome was first chosen as mitochondrial DNA is commonly used as a marker for population, phylogenetic, and molecular diversity studies in general.[37,38] Aside from that, the rationale was that there would be more mitochondrial DNA present in the raw data than genomic DNA. As this turned out not to be the case, we instead adapted the analysis for nuclear DNA.

Conclusion

The current results of the closest breed, see Table 3, ultimately provide a valid (albeit coarse) indication as to where the animals used for the mediaeval parchment might be from. With the currently available data, we've tried to match a pattern of SNPs found in the leaves of parchment with a reference database of modern day cattle breeds, in order to answer where the parchment geographically originated from. To create the pattern of SNPs, only high-quality reads, with a high enough depth, were used to identify SNPs. Of the identified SNPs only those in which we had a certain amount of confidence were used. This generated the networks found within this research paper and led to Table 3, which shows the geographical origin of the closest related breed to the animals used for our leaves of parchment. From this table we can conclude, based on the current amount of data, that the approximate geographical origin of the parchment sheets are the ones in the first column of Table 3. Of the five leaves of parchment, for which we knew the geographical origins, the analysis managed to match two leaves with a breed of cattle from the same approximate geographical origin (parchment P700_S2 and 78404_S3). However, the other leaves known to originate from the same manuscript (parchment P1_S7, P2_S8, and P5_S6) did not meet our expectations.

Data Availability

The datasets generated and/or analysed during the current study are available in Zenodo, DOI: 10.5281/zenodo.7371345

The scripts, pipelines, and databases created during the current study are available on Github: https://github.com/naturalis/mtdna-reference-database

Acknowledgements

I would like to thank Naturalis and my supervisor Rutger Vos in particular for guiding me through this project. Major thanks go to Rutherford Craze and Hanneke van Asperen for their help with proofreading, spell checking and correcting information about the history of the leaves of parchment.

References

1. Halve ton voor baanbrekend onderzoek naar ontdekt 15e-eeuws getijdenboek uit de kring van de Van Lymborchs. Maelwael van Lymborch Studies. Published August 18, 2022. Accessed January 13, 2023. https://www.maelwaelvanlymborchstudies.com/nieuws/halve-ton-voor-baanbrekend-onderzoek-naar-ontdekt-15e-eeuw-getijdenboek-uit-de-kring-van-de-van-lymborchs/

2. Husband TB. The Art of Illumination: The Limbourg Brothers and the Belles Heures of Jean de France, Duc de Berry.; 2008. Accessed January 13, 2023. http://archive.org/details/TheArtofIlluminationTheLimbourgBrothersandtheBellesHeuresofJeandeFranceDucdeBerr

3. The Making of Medieval Illuminated Manuscripts. Accessed January 13, 2023. https://www.gresham.ac.uk/watch-now/making-medieval-illuminated-manuscripts

4. Stein AWA. The Book of Hours: A Medieval Bestseller | Essay | The Metropolitan Museum of Art | Heilbrunn Timeline of Art History. The Met's Heilbrunn Timeline of Art History. Accessed January 13, 2023. https://www.metmuseum.org/toah/hd/hour/hd_hour.htm

5. De agenda van de Stichting Gebroeders van Limburg. Maelwael van Lymborch Studies. Accessed January 13, 2023. https://www.maelwaelvanlymborchstudies.com/gebroeders-van-lymborch/

6. Bronnen. Maelwael van Lymborch Studies. Accessed January 13, 2023. https://www.maelwaelvanlymborchstudies.com/bronnen/

7. Teasdale MD, Fiddyment S, Vnouček J, et al. The York Gospels: a 1000-year biological palimpsest. R Soc Open Sci. 2017;4(10):170988. doi:10.1098/rsos.170988

8. Piñar G, Cappa F, Vetter W, Schreiner M, Miklas H, Sterflinger K. Complementary Strategies for Deciphering the Information Contained in Ancient Parchment Documentary Materials. Appl Sci. 2022;12(20):10479. doi:10.3390/app122010479

9. Campana MG, Bower MA, Bailey MJ, et al. A flock of sheep, goats and cattle: ancient DNA analysis reveals complexities of historical parchment manufacture. J Archaeol Sci. 2010;37(6):1317-1325. doi:10.1016/j.jas.2009.12.036

10. Piñar G, Tafer H, Schreiner M, Miklas H, Sterflinger K. Decoding the biological information contained in two ancient Slavonic parchment codices: an added historical value. Environ Microbiol. 2020;22(8):3218-3233. doi:10.1111/1462-2920.15064

11. Anava S, Neuhof M, Gingold H, et al. Illuminating Genetic Mysteries of the Dead Sea Scrolls. Cell. 2020;181(6):1218-1231.e27. doi:10.1016/j.cell.2020.04.046

12. Teasdale MD, van Doorn NL, Fiddyment S, et al. Paging through history: parchment as a reservoir of ancient DNA for next generation sequencing. Philos Trans R Soc B Biol Sci. 2015;370(1660):20130379. doi:10.1098/rstb.2013.0379

13. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28(19):2520-2522. doi:10.1093/bioinformatics/bts480

14. Hayes BJ, Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. Annu Rev Anim Biosci. 2019;7(1):89-102. doi:10.1146/annurev-animal-020518-115024

15. Variant detection and genotype calling using Genome Analysis Tool Kit (GATK). figshare. doi:10.26181/5fb75065a6067

16. Sayers EW, Bolton EE, Brister JR, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022;50(D1):D20-D26. doi:10.1093/nar/gkab1112

17. Leinonen R, Akhtar R, Birney E, et al. The European Nucleotide Archive. Nucleic Acids Res. 2011;39(Database issue):D28-D31. doi:10.1093/nar/gkq967

18. SQLite Home Page. Accessed January 13, 2023. https://www.sqlite.org/index.html

19. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Accessed January 13, 2023. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

20. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i890. doi:10.1093/bioinformatics/bty560

21. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170

22. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Published online May 26, 2013. doi:10.48550/arXiv.1303.3997

23. Dabney J, Meyer M, Pääbo S. Ancient DNA Damage. Cold Spring Harb Perspect Biol. 2013;5(7):a012567. doi:10.1101/cshperspect.a012567

24. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013;29(13):1682-1684. doi:10.1093/bioinformatics/btt193

25. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

26. Ebbert MTW, Wadsworth ME, Staley LA, 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. doi:10.1186/s12859-016-1097-3

27. Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Published online July 24, 2018:201178. doi:10.1101/201178

28. Huson DH, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. Mol Biol Evol. 2006;23(2):254-267. doi: 10.1093/molbev/msj030

29. Bryant D, Moulton V. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. Mol Biol Evol. 2004;21(2):255-265. doi:10.1093/molbev/msh018

30. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nat Rev Genet. 2011;12(10): 703-714. doi:10.1038/nrg3054

31. Paradis E. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics. 2010; 26(3):419-420. doi:10.1093/bioinformatics/btp696

32. Paradis E. Analysis of haplotype networks: The randomized minimum spanning tree method. Methods Ecol Evol. 2018; 9(5):1308-1317. doi:10.1111/2041-210X.12969

33. Bioinformatics ecSeq. Why does the per base sequence quality decrease over the read in Illumina? Accessed January 14, 2023. https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina

34. Eck SH, Benet-Pagès A, Flisikowski K, Meitinger T, Fries R, Strom TM. Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. Genome Biol. 2009;10(8):R82. doi:10.1186/gb-2009-10-8-r82

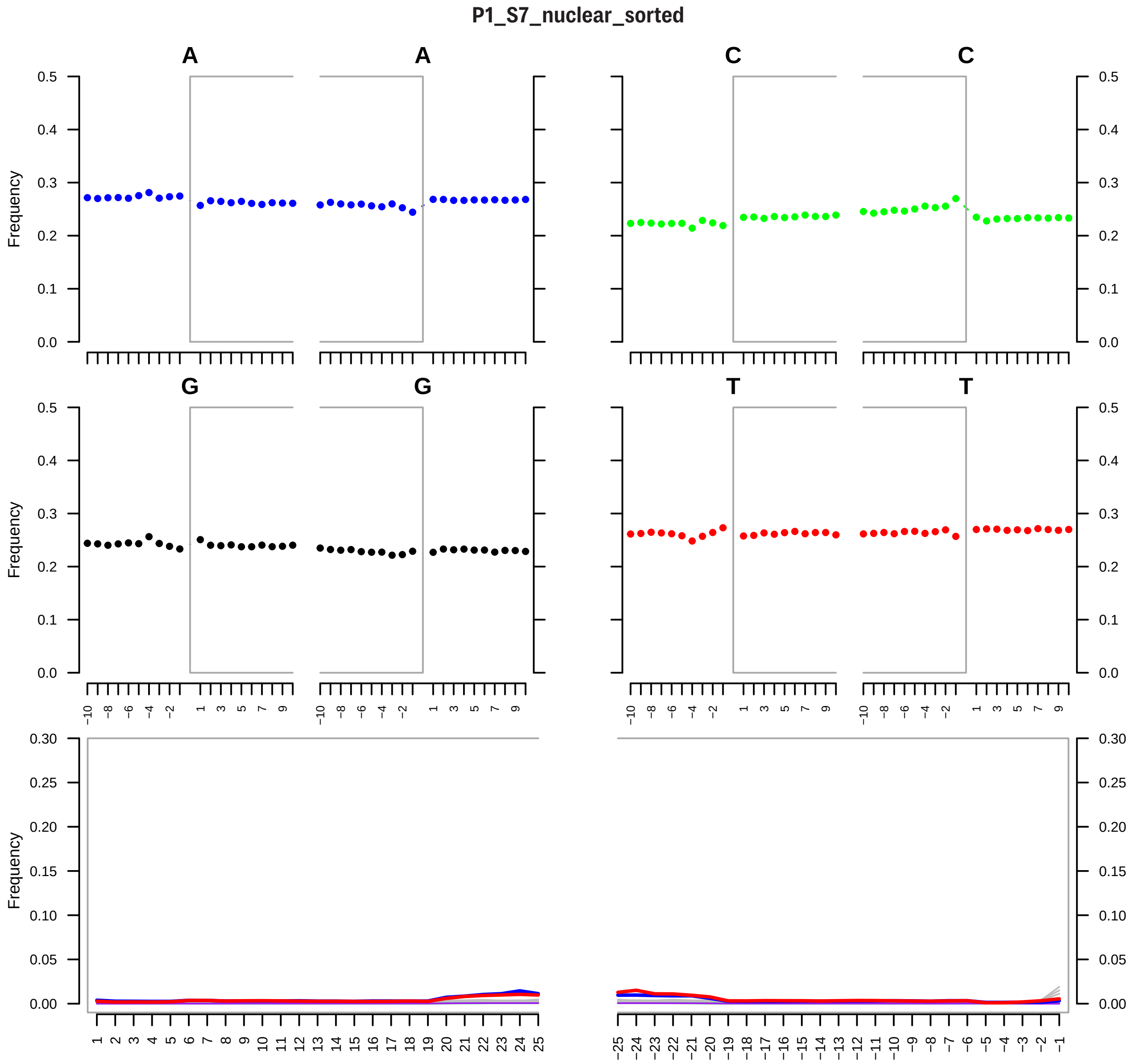
35. Mei C, Wang H, Zhu W, et al. Whole-genome sequencing of the endangered bovine species Gayal (Bos frontalis) provides new insights into its genetic features. Sci Rep. 2016;6(1):19787. doi:10.1038/srep19787

36. Weldenegodguad M, Popov R, Pokharel K, et al. Whole-Genome Sequencing of Three Native Cattle Breeds Originating From the Northernmost Cattle Farming Regions. Front Genet. 2019;9. Accessed January 13, 2023. https://www.frontiersin.org/articles/10.3389/fgene.2018.00728

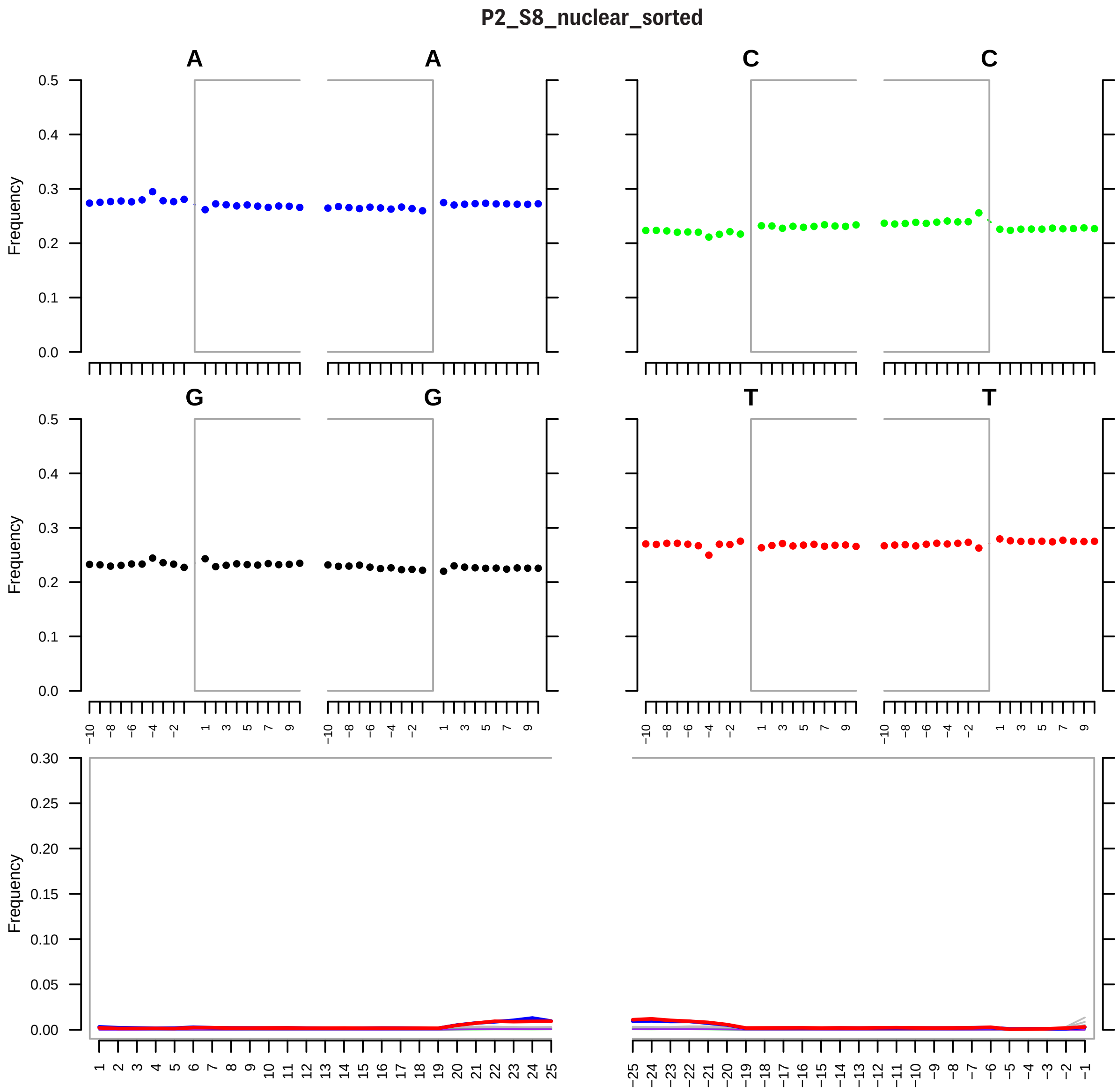
37. Galtier N, Nabholz B, Glémin S, Hurst GDD. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol Ecol. 2009;18(22):4541-4550. doi:10.1111/j.1365-294X.2009.04380.x

38. Merheb M, Matar R, Hodeify R, et al. Mitochondrial DNA, a Powerful Tool to Decipher Ancient Human Civilization from Domestication to Music, and to Uncover Historical Murder Cases. Cells. 2019;8(5):433. doi:10.3390/cells8050433

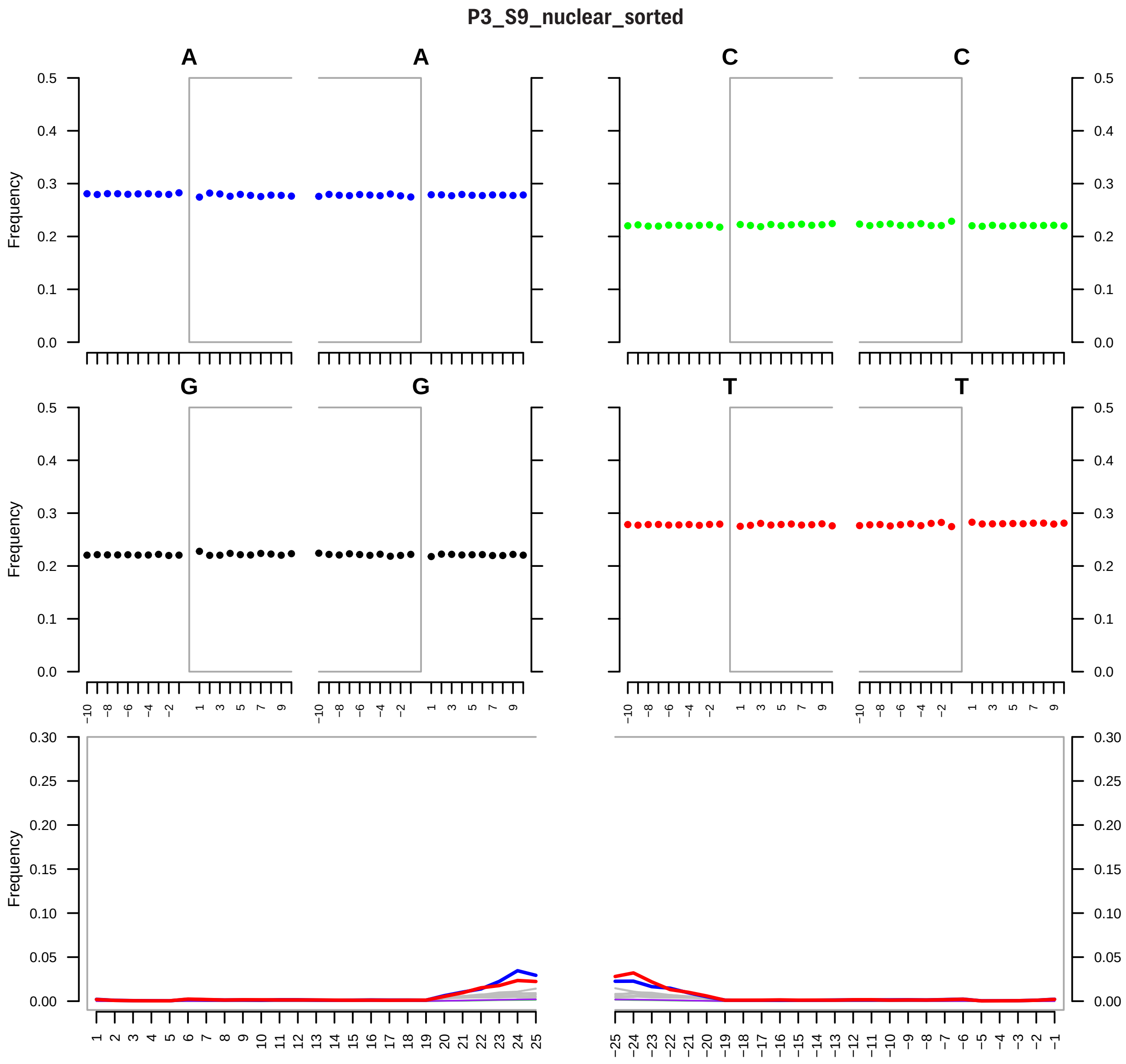
Supplementary Figures



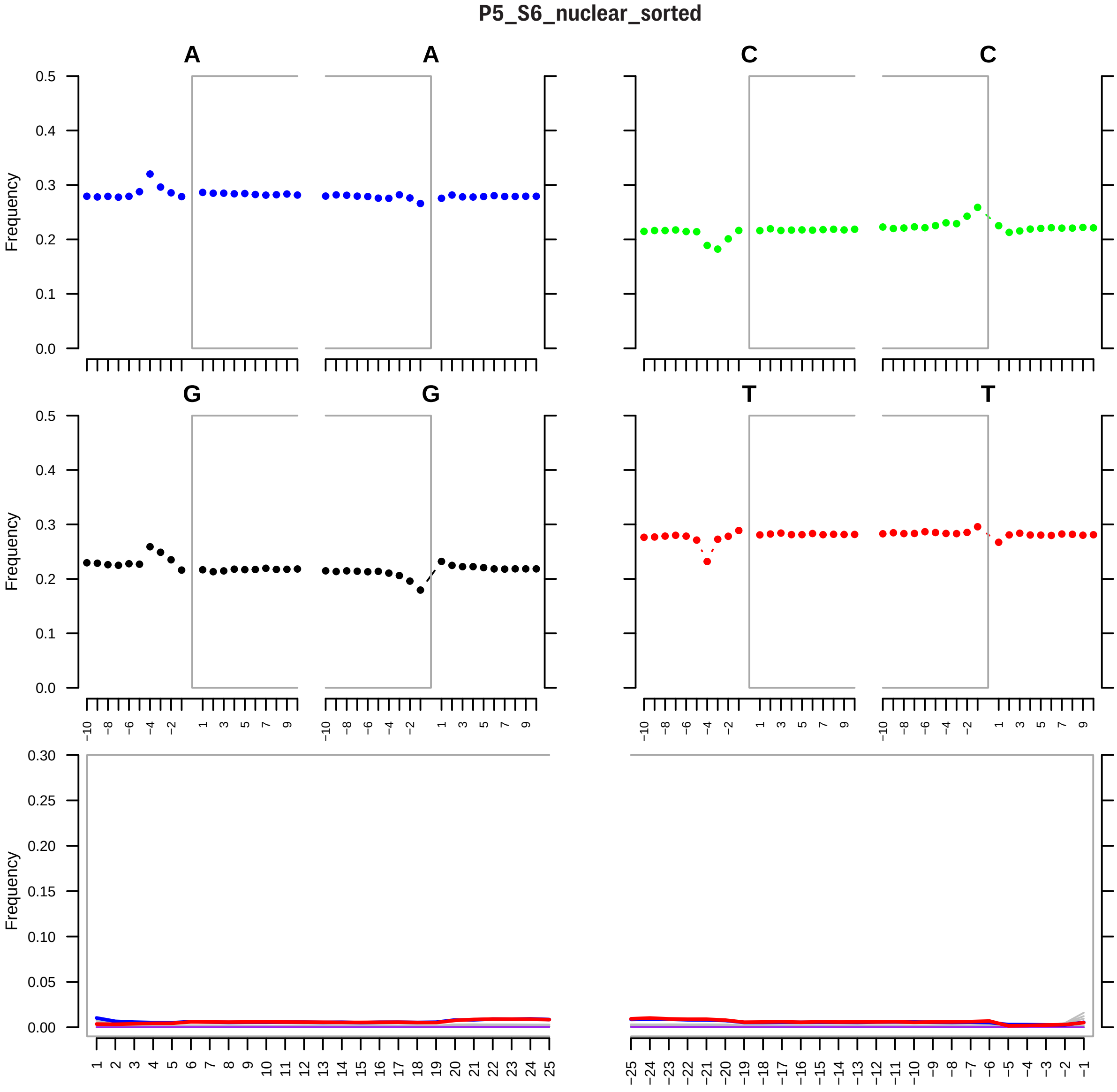
Supplementary Figure 1: MapDamage plots P1_S7. *MapDamage report for parchment sample P1_S7. The plots show no apparent signs of deamination within the sample.*



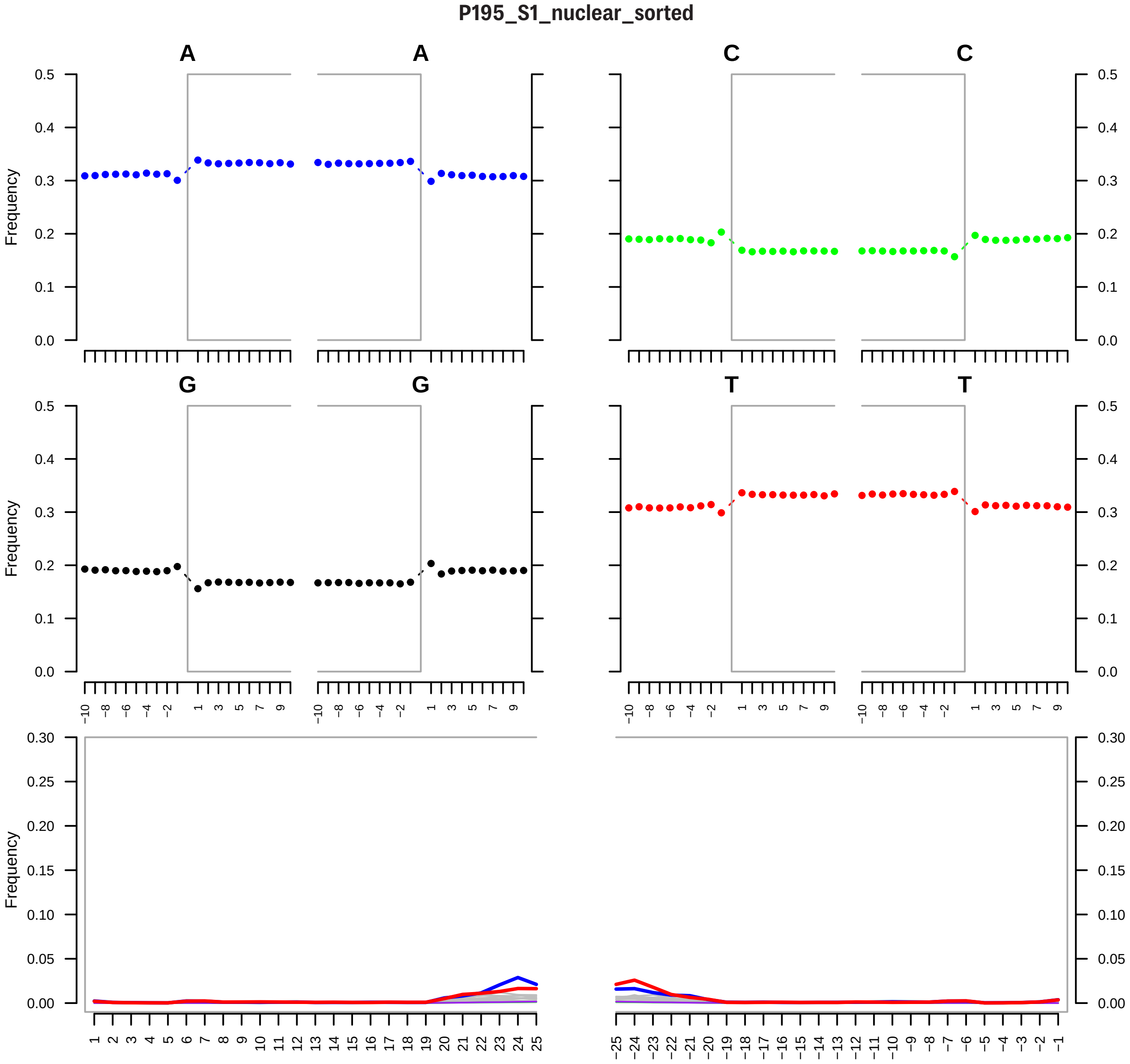
Supplementary Figure 2: MapDamage plots P2_S8. *MapDamage report for parchment sample P2_S8. The plots show no apparent signs of deamination within the sample.*



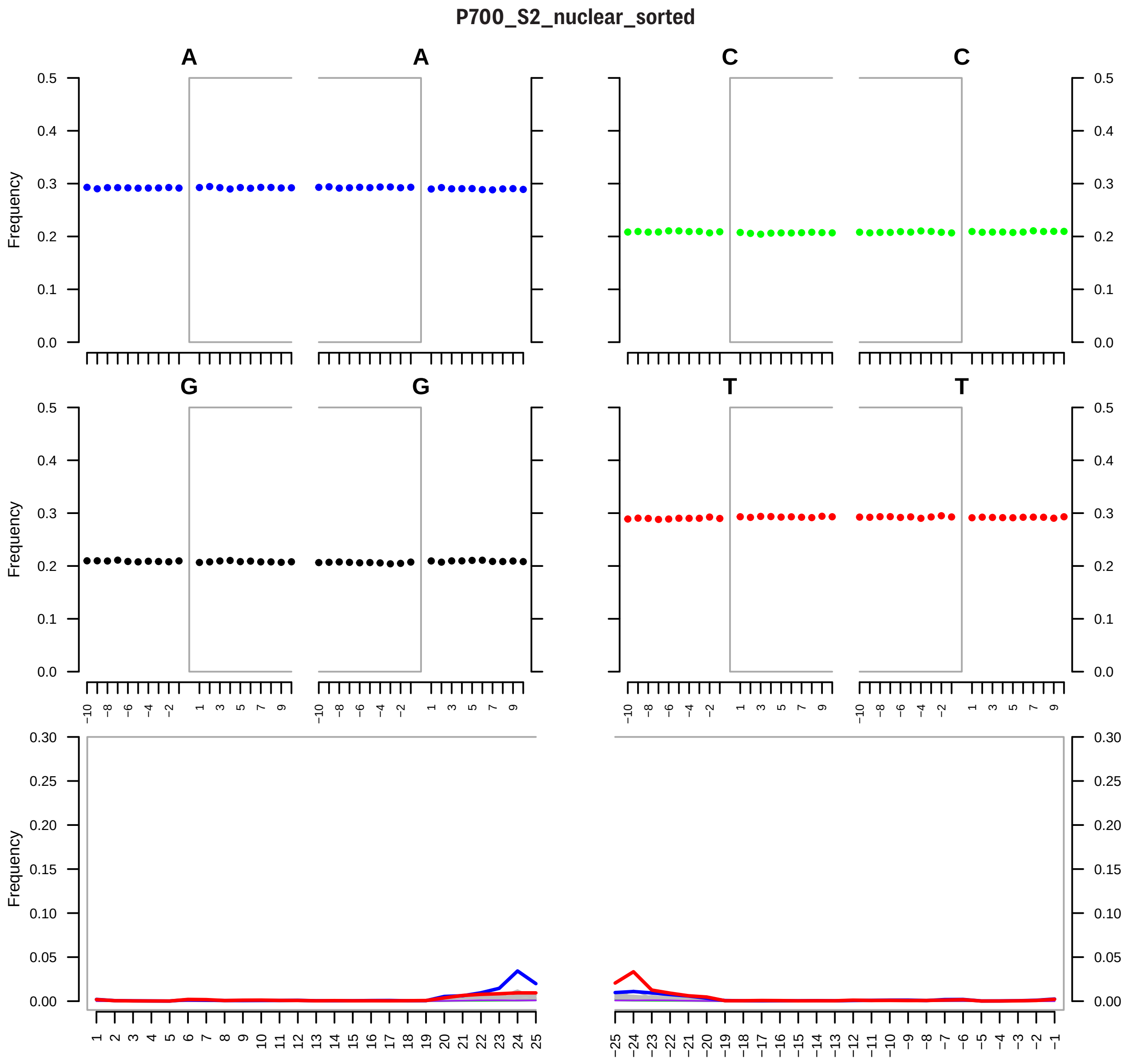
Supplementary Figure 3: MapDamage plots P3_S9. *MapDamage report for parchment sample P3_S9. The plots show no apparent signs of deamination within the sample.*



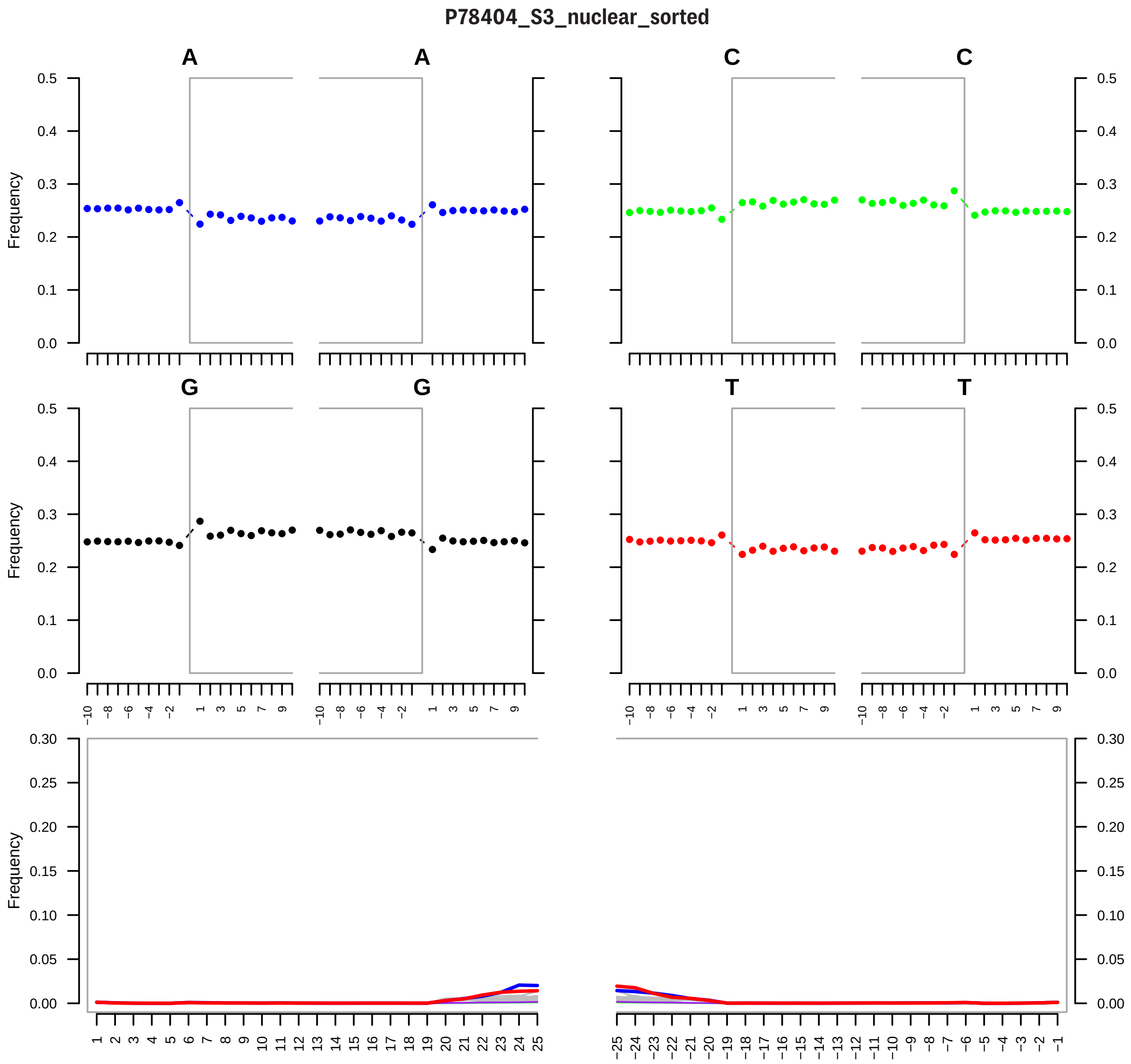
Supplementary Figure 4: MapDamage plots P5_S6. *MapDamage report for parchment sample P5_S6. The plots show no apparent signs of deamination within the sample.*



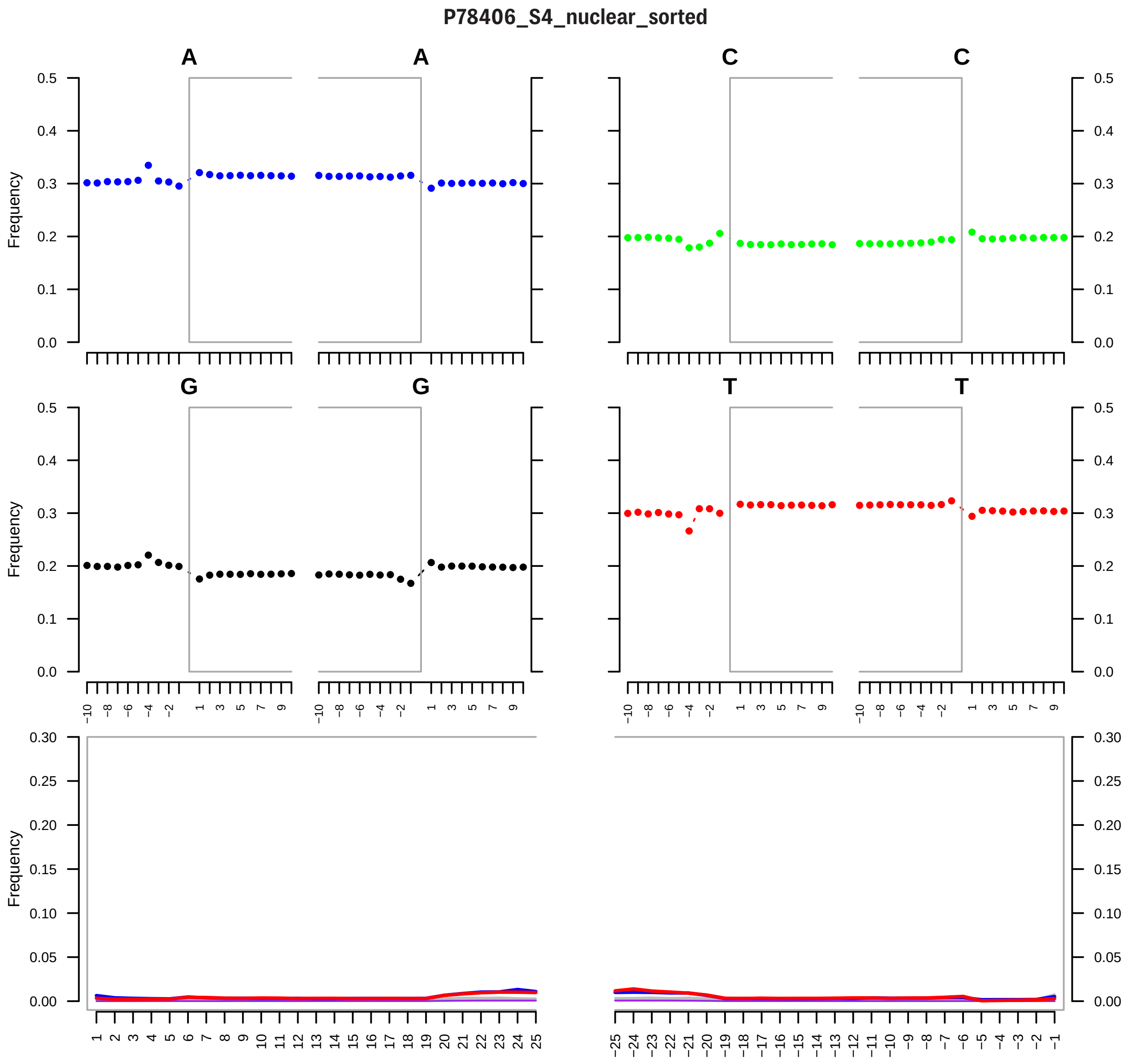
Supplementary Figure 5: MapDamage plots P195_S1. *MapDamage report for parchment sample P195_S1. The plots show no apparent signs of deamination within the sample.*



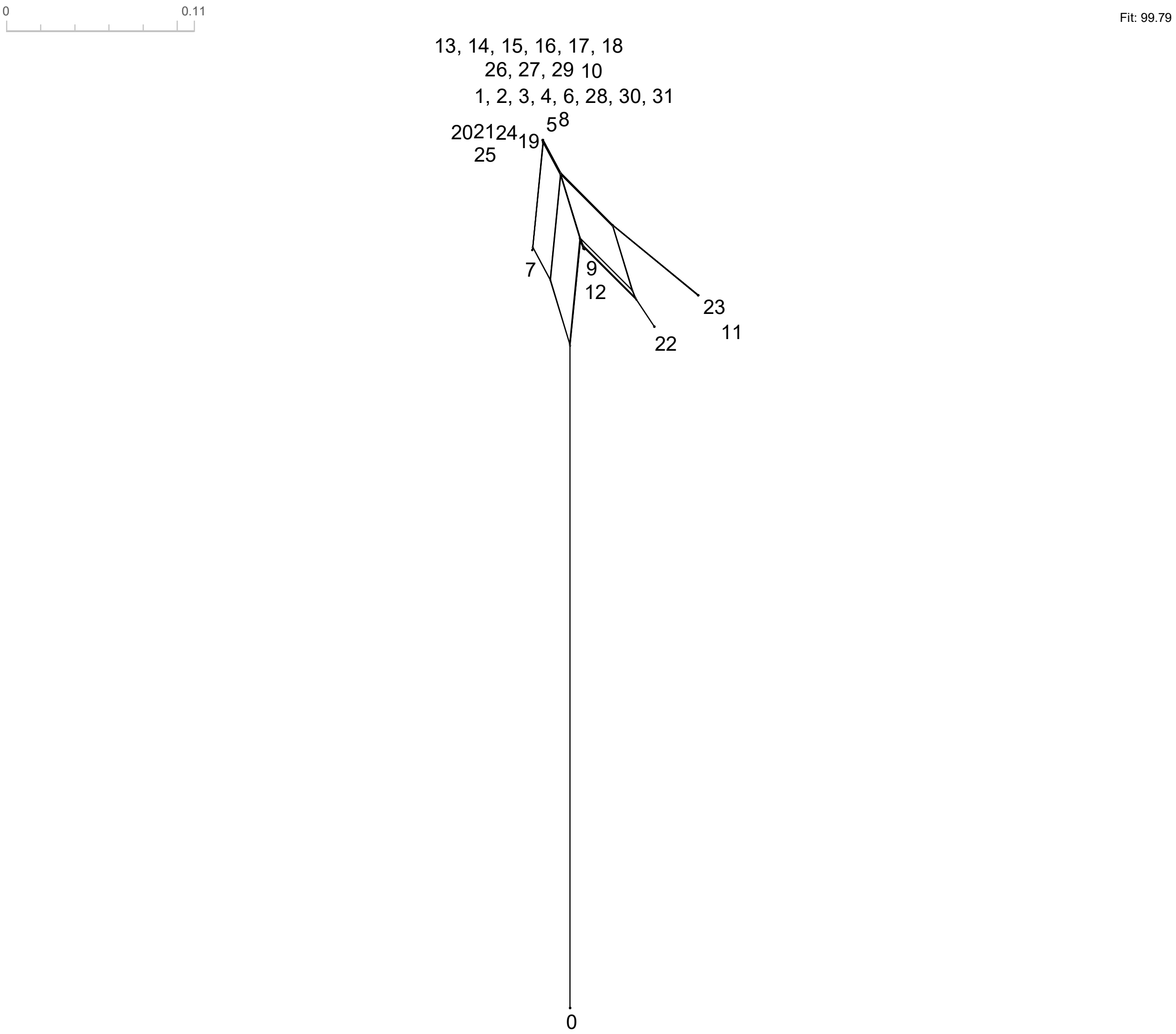
Supplementary Figure 6: MapDamage plots P700_S2. *MapDamage report for parchment sample P700_S2. The plots show no apparent signs of deamination within the sample.*



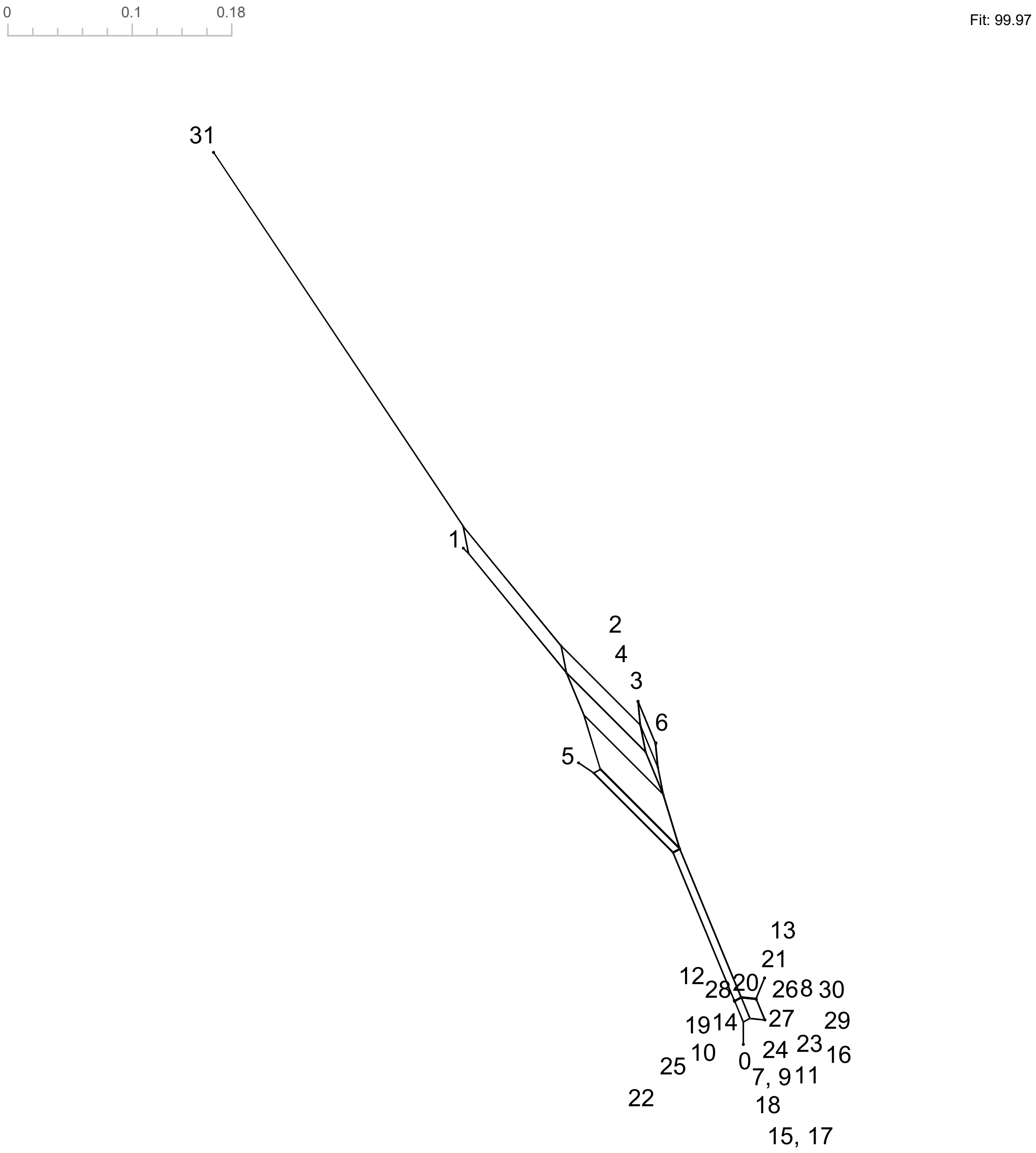
Supplementary Figure 7: MapDamage plots P78404_S3. *MapDamage report for parchment sample P78404_S3. The plots show no apparent signs of deamination within the sample.*



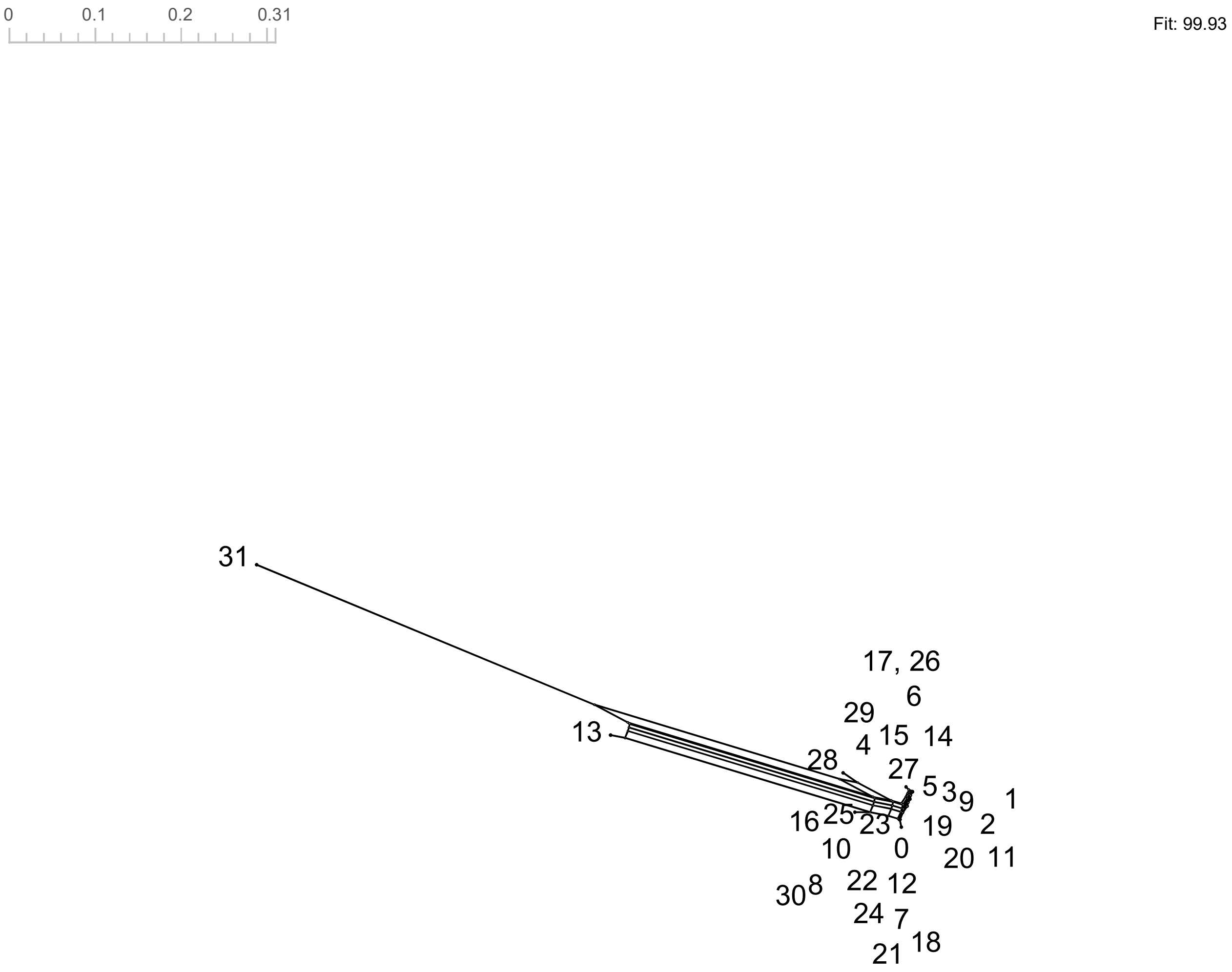
Supplementary Figure 8: MapDamage plots P78406_S4. *MapDamage report for parchment sample P78406_S4. The plots show no apparent signs of deamination within the sample.*



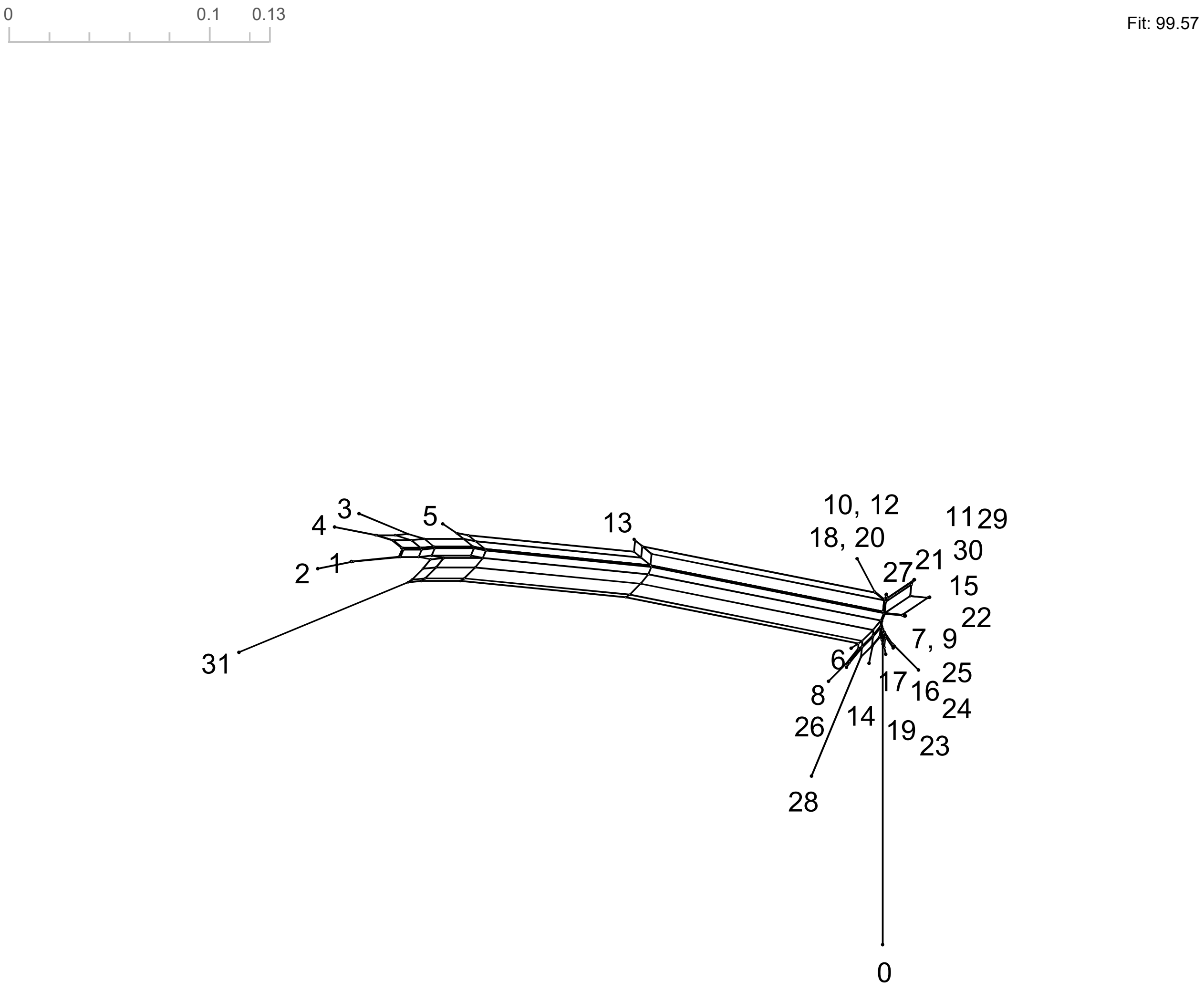
Supplementary Figure 9: Phylogenetic network of P1_S7. *The phylogenetic network with QUAL and DP thresholds of 20 and 5, respectively, applied. Within the network, the piece of parchment (31) together with the top 30 hits (1–30) in the reference database and the reference genome (0) are shown. The length of the branch indicates the amount of divergence between samples. The low number of SNPs for this sample makes it that the majority clusters together, making it impossible to get a conclusive answer on its nearest neighbour.*



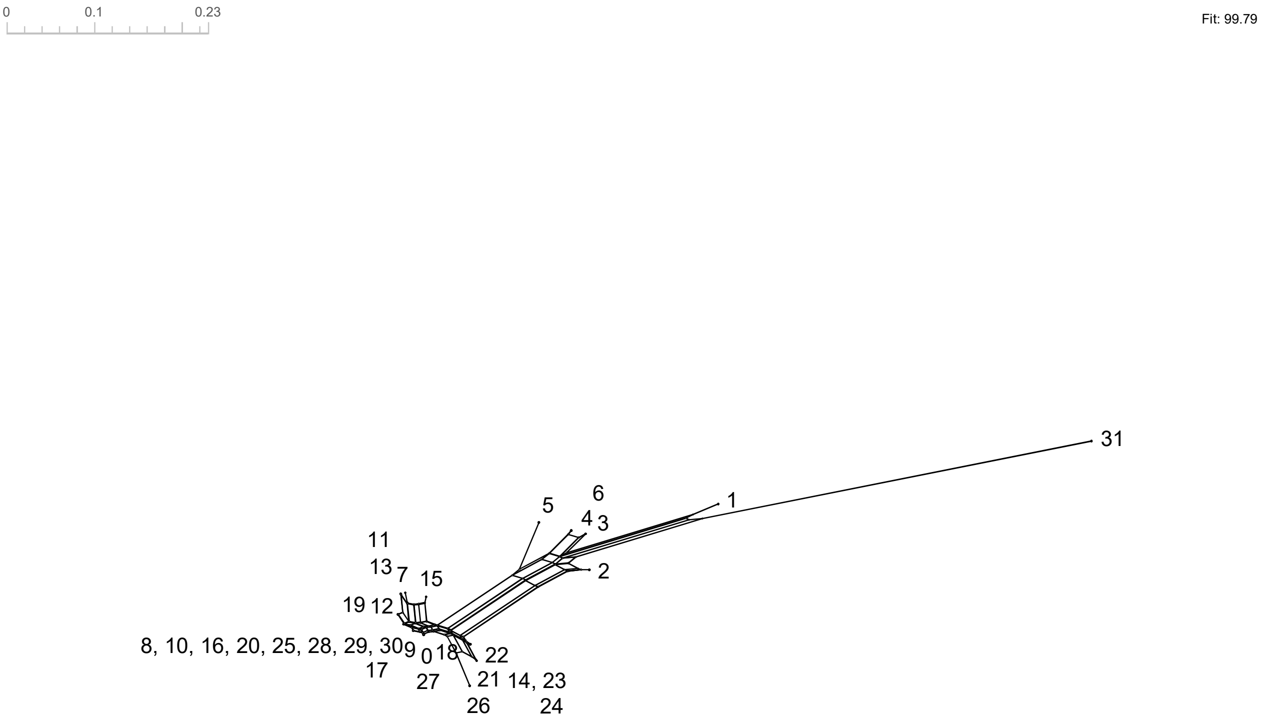
Supplementary Figure 10: Phylogenetic network of P2_S8. *The phylogenetic network with a QUAL and DP thresholds, of 30 and 10 respectively, applied. Within the network the piece of parchment (31) together with the top 30 hits (1–30) in the reference database, and the reference genome (0) are shown. The length of the branch indicates the amount of divergence between samples. The closest tip to the parchment sample is number 1, which is a German cattle breed.*



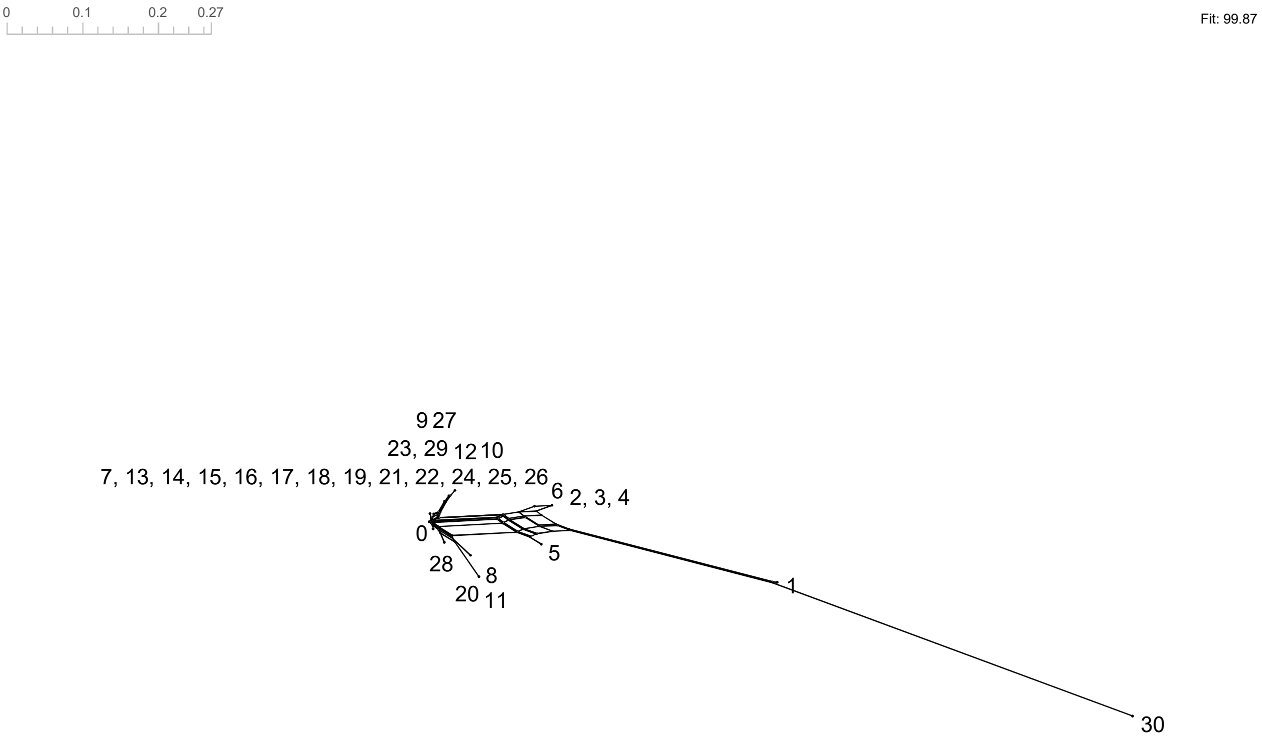
Supplementary Figure 11: Phylogenetic network of P3_S9. *The phylogenetic network with a QUAL and DP thresholds, of 20 and 5 respectively, applied. Within the network the piece of parchment (31) together with the top 30 hits (1–30) in the reference database, and the reference genome (0) are shown. The length of the branch indicates the amount of divergence between samples. The closest tip to the parchment sample is number 13, which is a German cattle breed.*



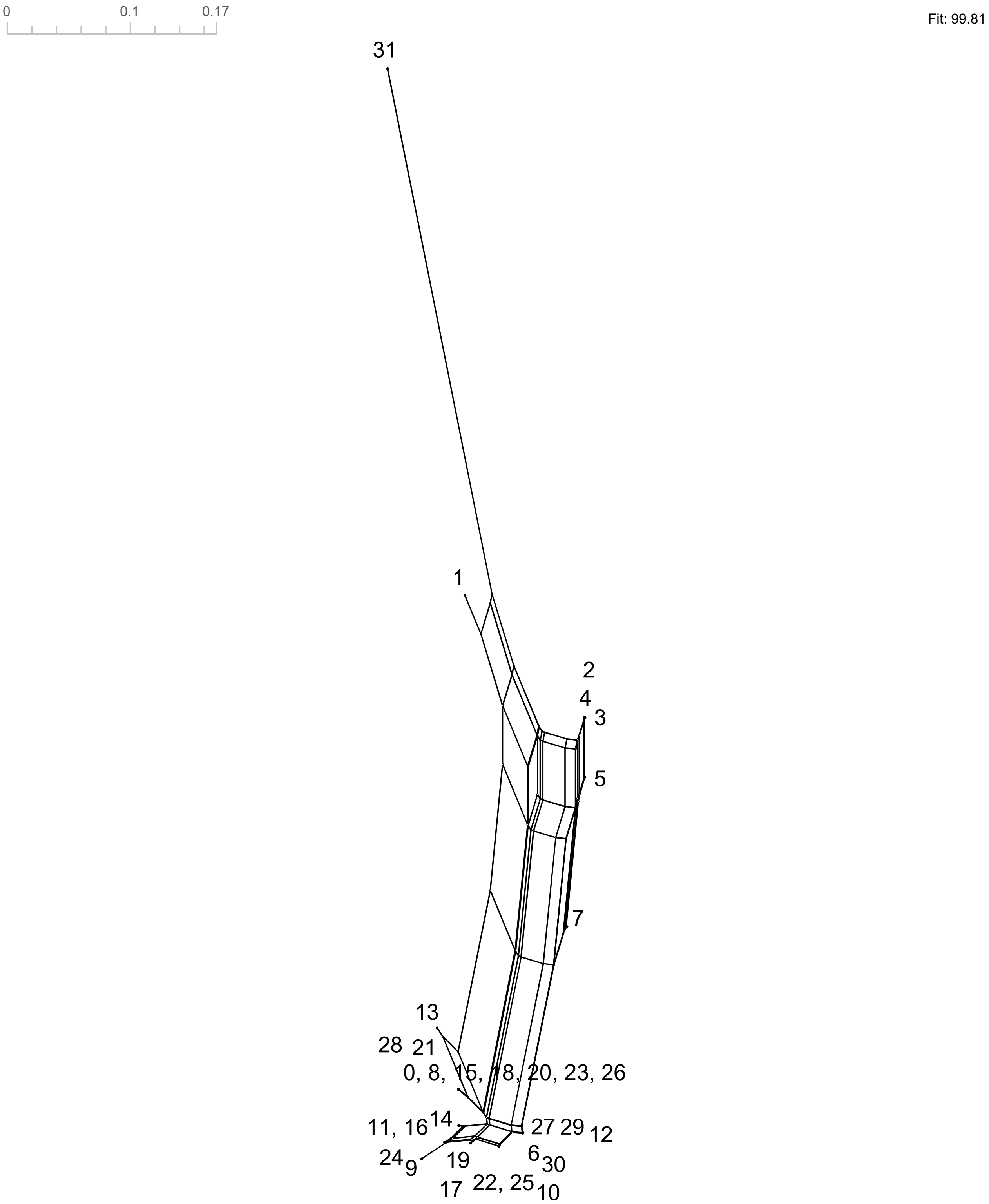
Supplementary Figure 12: Phylogenetic network of P5_S6. *The phylogenetic network with a QUAL and DP thresholds, of 30 and 10, respectively, applied. Within the network the piece of parchment (31) together with the top 30 hits (1–30) in the reference database, and the reference genome (0) are shown. The length of the branch indicates the amount of divergence between samples. The closest tip to the parchment sample is number 1, which is an English cattle breed.*



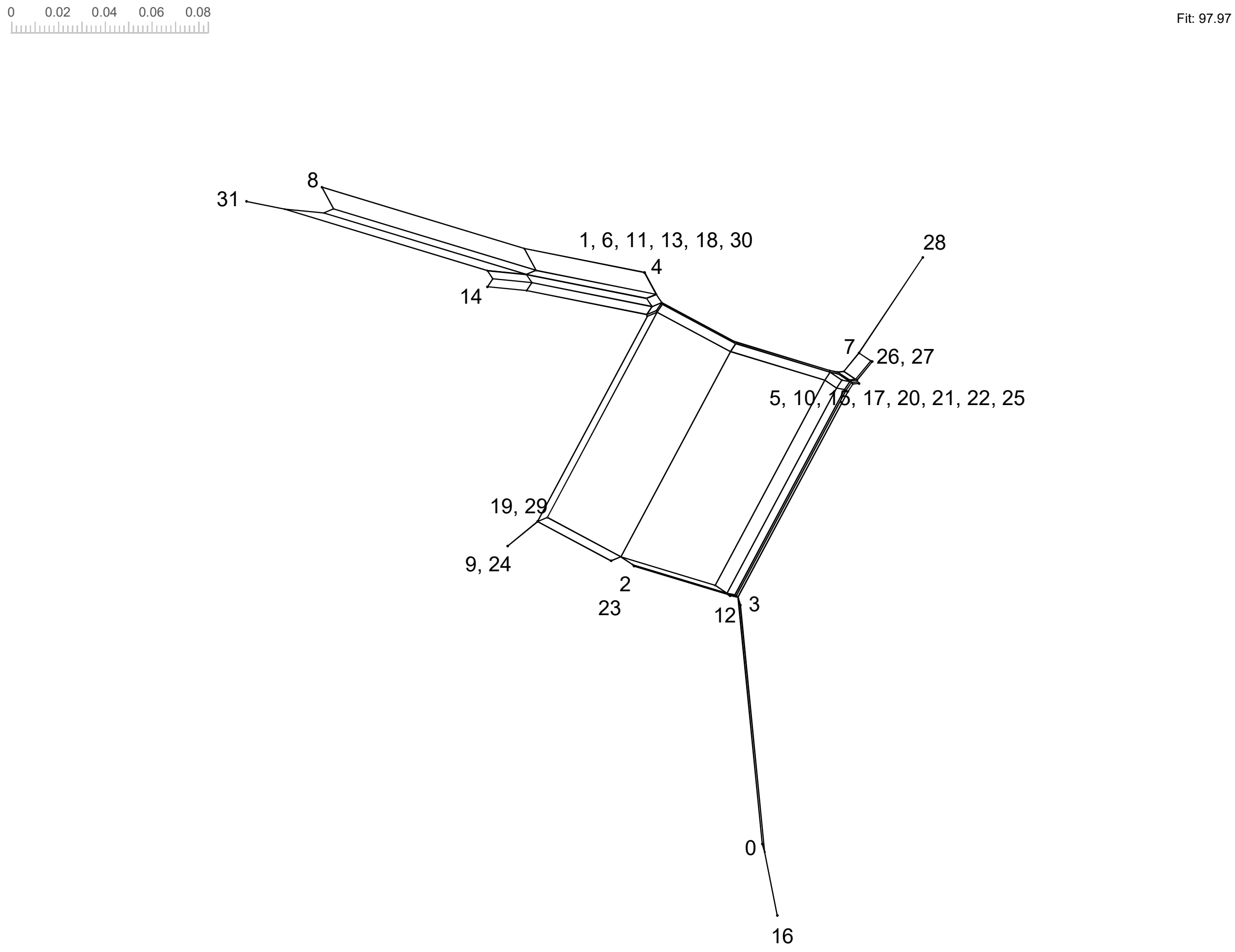
Supplementary Figure 13: Phylogenetic network of P195_S1. *The phylogenetic network with a QUAL and DP thresholds, of 20 and 5, respectively, applied. Within the network the piece of parchment (31) together with the top 30 hits (1–30) in the reference database, and the reference genome (0) are shown. The length of the branch indicates the amount of divergence between samples. The closest tip to the parchment sample is number 1, which is a German cattle breed.*



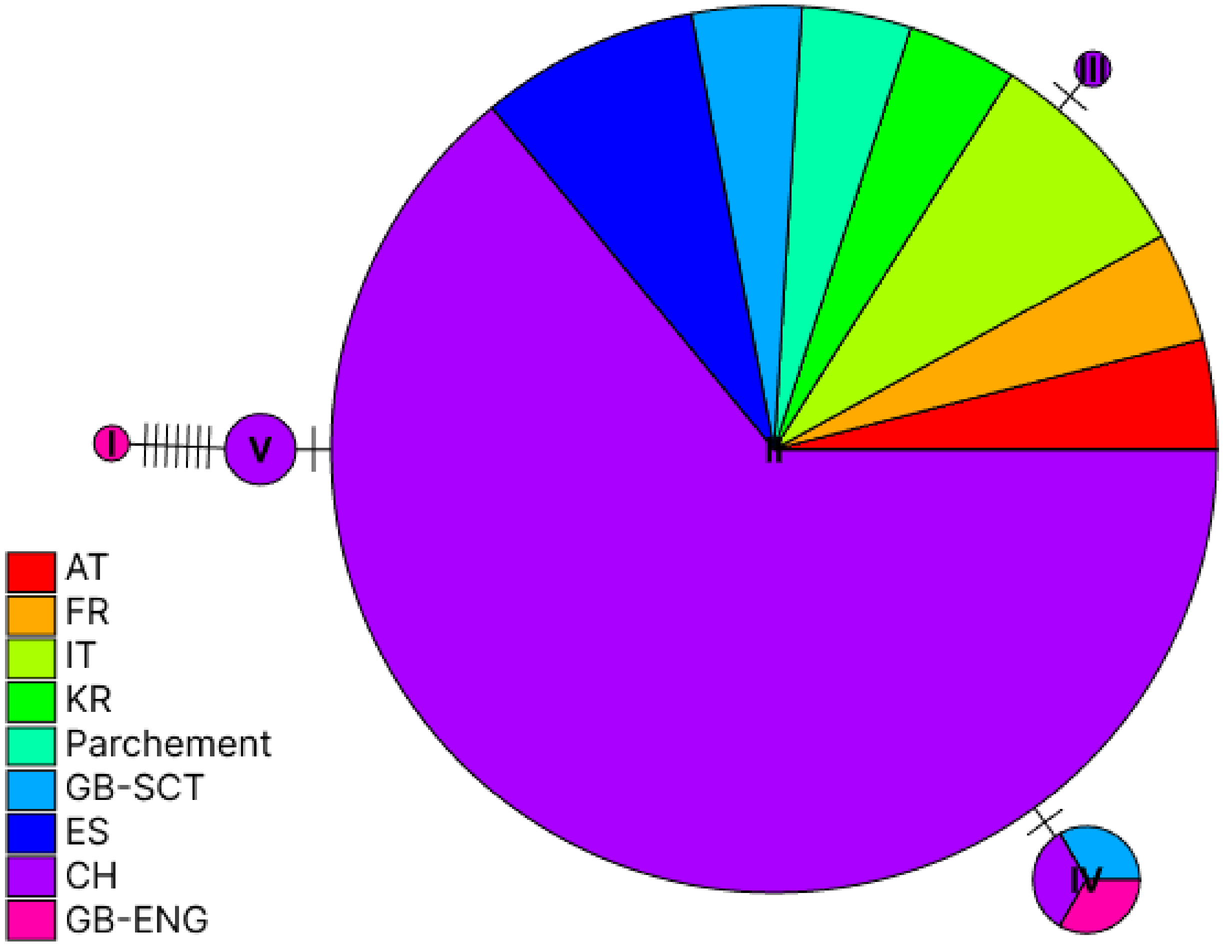
Supplementary Figure 14: Phylogenetic network of P700_S2. *The phylogenetic network with a QUAL and DP thresholds, of 30 and 10, respectively, applied. Within the network the piece of parchment (31) together with the top 30 hits (1–30) in the reference database, and the reference genome (0) are shown. The length of the branch indicates the amount of divergence between samples. The closest tip to the parchment sample is number 1, which is a German cattle breed.*



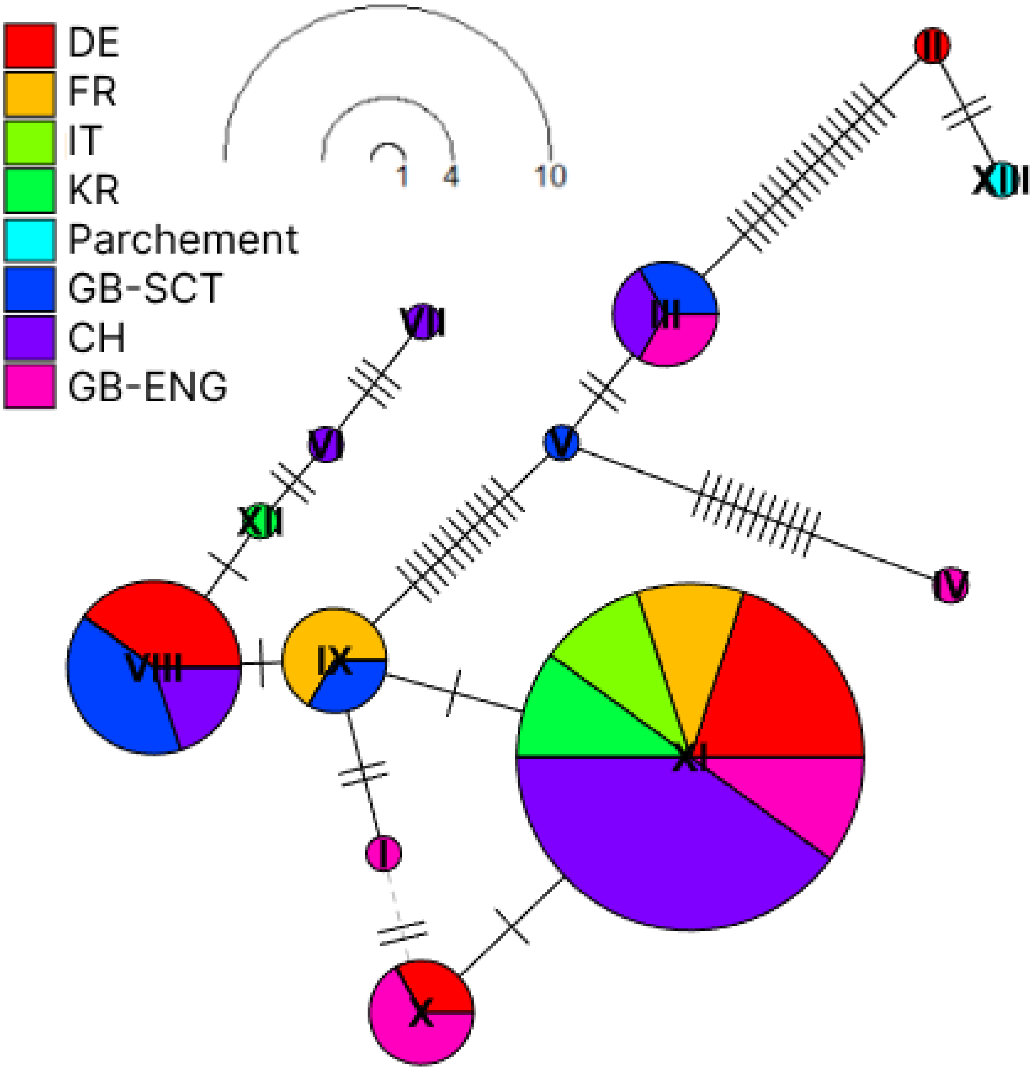
Supplementary Figure 15: Phylogenetic network of P78404_S3. *The phylogenetic network with a QUAL and DP thresholds, of 20 and 5, respectively, applied. Within the network the piece of parchment (31) together with the top 30 hits (1–30) in the reference database, and the reference genome (0) are shown. The length of the branch indicates the amount of divergence between samples. The closest tip to the parchment sample is number 1, which is a German cattle breed.*



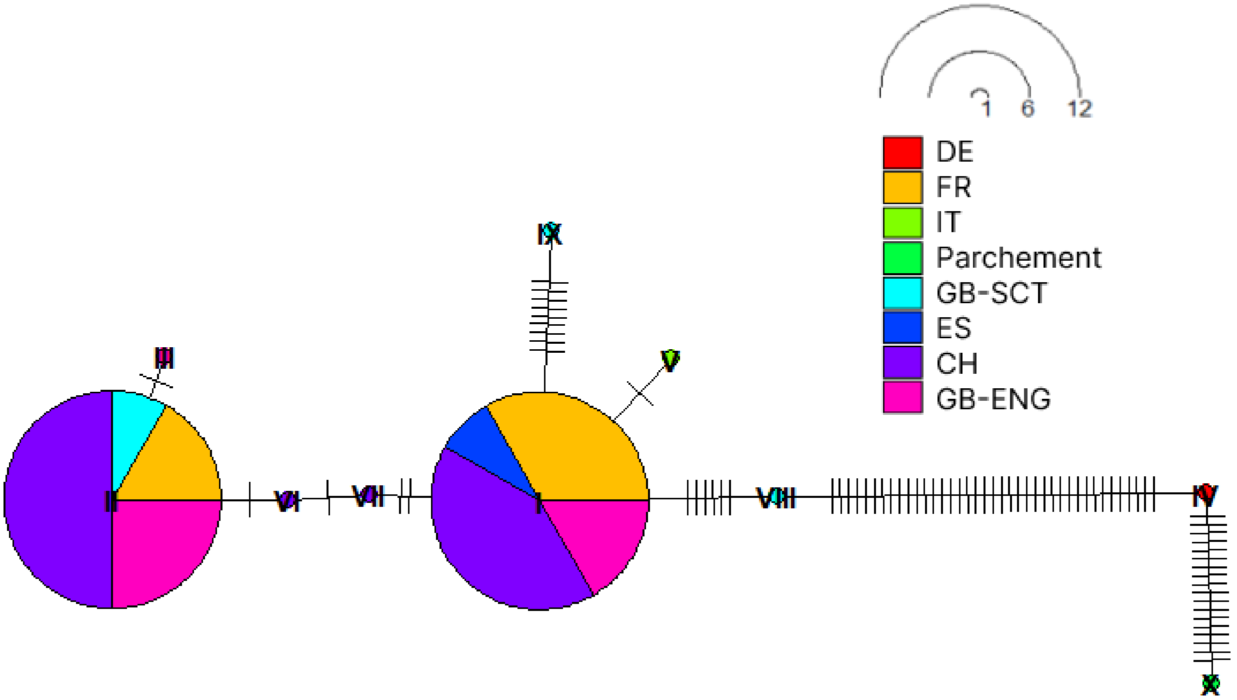
***Supplementary Figure 16: Phylogenetic network of P78406_S4.** The phylogenetic network with a QUAL and DP thresholds, of 30 and 10, respectively, applied. Within the network the piece of parchment (31) together with the top 30 hits (1–30) in the reference database, and the reference genome (0) are shown. The length of the branch indicates the amount of divergence between samples. The closest tip to the parchment sample is number 8, which is an English cattle breed.*



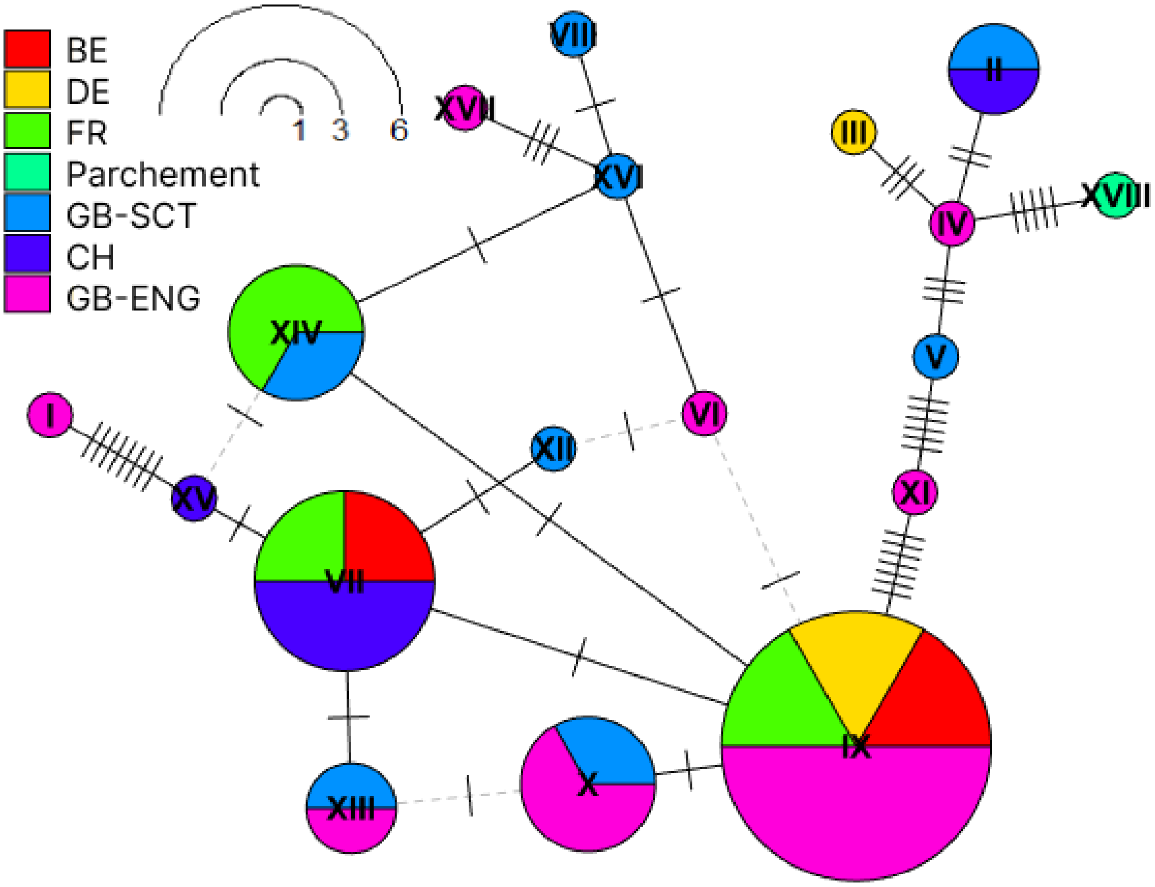
Supplementary Figure 17: Haplotype network P1_S7. *The haplotype network above shows the haplotypes from the top 30 hits of the parchment P1_S7 with the reference database. The size of the circle indicates the number of samples with that specific haplotype. The number of tangential lines on the connecting lines between circles indicate the number of changes between two haplotypes, and the colour indicates the country, as per the legend, of the breed. The network shows that the parchment sample is clustered together with a majority of the samples, making it impossible to find a conclusive nearest neighbour.*



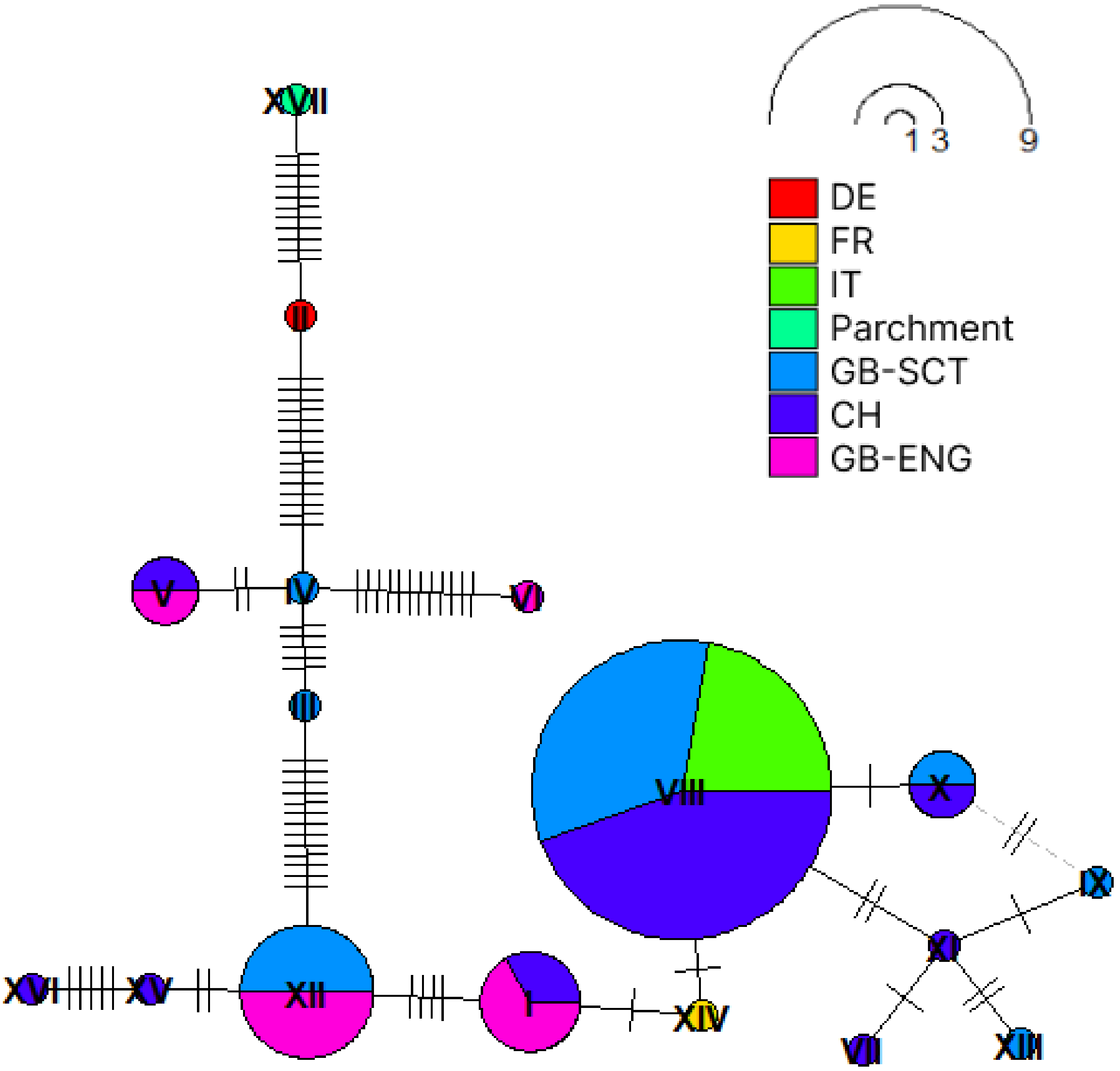
Supplementary Figure 18: Haplotype network P2_S8. *The haplotype network above shows the haplotypes from the top 30 hits of the parchment P2_S8 with the reference database. The size of the circle indicates the number of samples with that specific haplotype. The number of tangential lines on the connecting lines between circles indicate the number of changes between two haplotypes, and the colour indicates the country, as per the legend, of the breed. The network shows that the nearest to the parchment sample is a German breed.*



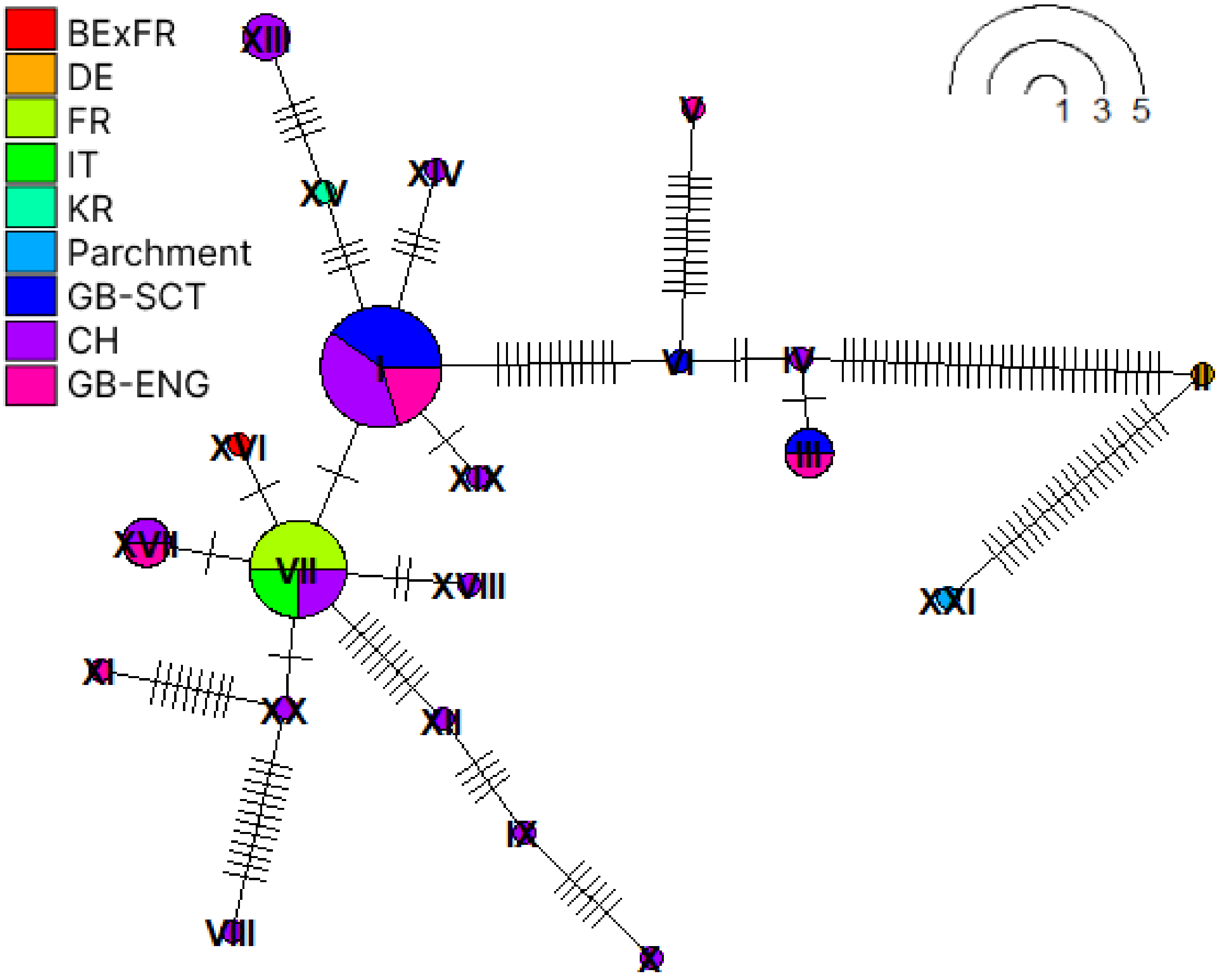
Supplementary Figure 19: Haplotype network P3_S9. *The haplotype network above shows the haplotypes from the top 30 hits of the parchment P3_S9 with the reference database. The size of the circle indicates the number of samples with that specific haplotype. The number of tangential lines on the connecting lines between circles indicate the number of changes between two haplotypes, and the colour indicates the country, as per the legend, of the breed. The network shows that the nearest to the parchment sample is a German breed.*



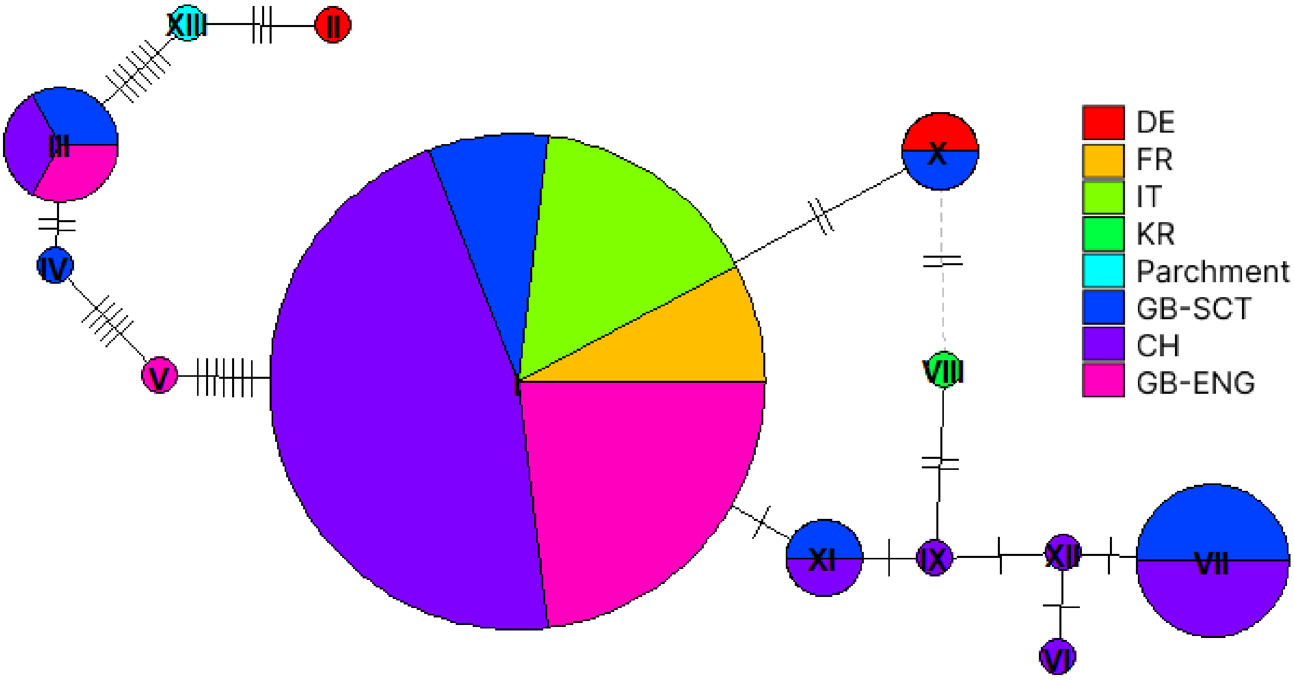
Supplementary Figure 20: Haplotype network P5_S6. *The haplotype network above shows the haplotypes from the top 30 hits of the parchment P5_S6 with the reference database. The size of the circle indicates the number of samples with that specific haplotype. The number of tangential lines on the connecting lines between circles indicate the number of changes between two haplotypes, and the colour indicates the country, as per the legend, of the breed. The network shows that the nearest to the parchment sample is an English breed.*



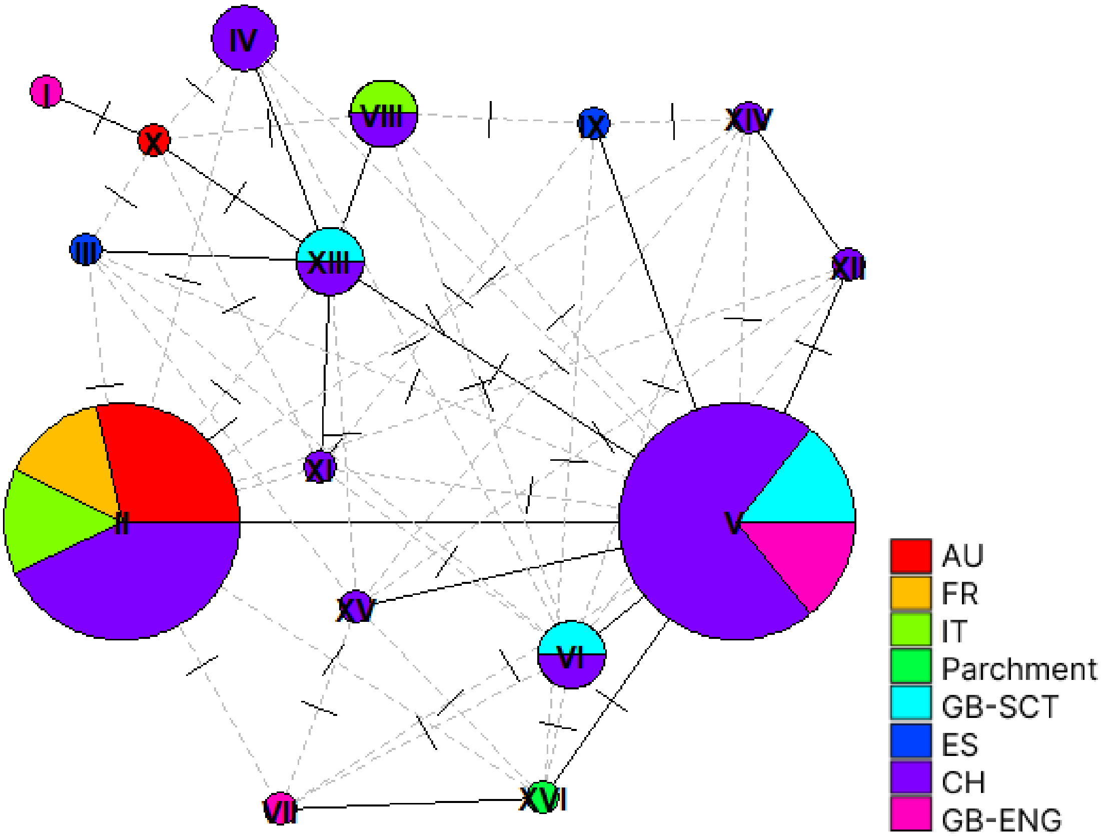
Supplementary Figure 21: Haplotype network P195_S1. *The haplotype network above shows the haplotypes from the top 30 hits of the parchment P195_S1 with the reference database. The size of the circle indicates the number of samples with that specific haplotype. The number of tangential lines on the connecting lines between circles indicate the number of changes between two haplotypes, and the colour indicates the country, as per the legend, of the breed. The network shows that the nearest to the parchment sample is a German breed.*



Supplementary Figure 22: Haplotype network P700_S2. *The haplotype network above shows the haplotypes from the top 30 hits of the parchment P700_S2 with the reference database. The size of the circle indicates the number of samples with that specific haplotype. The number of tangential lines on the connecting lines between circles indicate the number of changes between two haplotypes, and the colour indicates the country, as per the legend, of the breed. The network shows that the nearest to the parchment sample is a German breed.*



Supplementary Figure 23: Haplotype network P78404_S3. *The haplotype network above shows the haplotypes from the top 30 hits of the parchment P78404_S3 with the reference database. The size of the circle indicates the number of samples with that specific haplotype. The number of tangential lines on the connecting lines between circles indicate the number of changes between two haplotypes, and the colour indicates the country, as per the legend, of the breed. The network shows that the nearest to the parchment sample is a German breed.*



Supplementary Figure 24: Haplotype network P78406_S4. *The haplotype network above shows the haplotypes from the top 30 hits of the parchment P78406_S4 with the reference database. The size of the circle indicates the number of samples with that specific haplotype. The number of tangential lines on the connecting lines between circles indicate the number of changes between two haplotypes, and the colour indicates the country, as per the legend, of the breed. The network shows that the nearest to the parchment sample is an English breed.*