



## **Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin**

### **Copyright statement**

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

### **Liability statement**

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

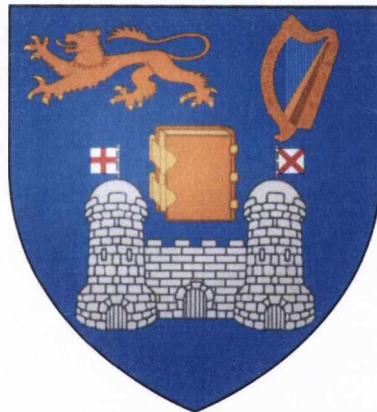
### **Access Agreement**

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

# Next Generation Sequencing of Ancient DNA

Matthew David Teasdale



A thesis submitted to the University of Dublin for the degree of Doctor of  
Philosophy

Department of Genetics

University of Dublin

Trinity College





Thesis 10097

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

ate 13-5-13

## ***Summary***

The application of next generation sequencing (NGS) technologies into the field of ancient DNA has seen a paradigm shift in the level of analysis that can be completed using archaeological samples. However, to fully access the power that these techniques provide, certain methodological hurdles need to be overcome. This thesis presents case studies that describe the implementation of new techniques to the attainment and analysis of aDNA to address these problems, as well as the use of a new substrate (parchment) for the NGS analysis of aDNA.

Chapter 3 presents the results of an mtDNA hybridisation capture experiment in three domestic ungulate species (cow, sheep and goat). This study was a success producing 11 high-coverage genomes (3x coverage greater than 85%), 10 of which were then subsequently used in a Bayesian phylogenetic analyses to try to provide a greater resolution to the analysis of the domestication process in two of the species, sheep and cow.

Chapter 4 presents a population genetic analysis of high coverage SNP genotypes produced from the first draft of a Mesolithic aurochs genome (CPC98). The results from this analysis suggest tentatively that CPC98 has a closer relationship to European taurine animals than African implying that the European aurochs may have introgressed into the *Bos taurus* population post domestication.

Chapter 5 presents the results of an exon capture study of multiple aDNA samples, the analysis highlighting multiple known and novel variants in the aurochs (CPC98). The results from five less well-preserved samples are more modest, suggesting there may be a limit to the preservation state of samples that can be used in these analyses.

Chapter 6 presents the results of a population genetic analysis of two 8,000-year-old cattle samples from western Anatolia using low coverage haploid SNP data, which suggests that they share a greater genetic affinity towards the domestic cattle of Europe than that of present-day admixed Anatolian populations.

In chapter 7 the results of a pilot analysis from a NGS study of parchment are presented. These parchments were found to be amenable to large-scale aDNA analysis with over 8% of the ovine genome recovered in each, an amount of DNA retrieval comparable to that of modern samples. These sequencing results allowed for the placing of the parchment samples within the distribution of modern domestic sheep populations.

In the final chapter an analysis of an Avar skeletal sample with Pott's disease is presented. We found this individual to share a greater affinity to Caucus and central Asian populations than that of central Europe, suggesting that this person may have had an ancestry towards the Avar homeland in Asia. An attempt was also made to retrieve *M. tuberculosis* sequences however; the results of this analysis were less conclusive.



## *Acknowledgments*

Frist I would like to thank my supervisor, Dan Bradley for giving me the opportunity to be part of his group and for his patience and advice.

Secondly I would like to that the members of the bovine lab past and present, Tim, Lilian, Ceiridwen, Emma, Valeria, Yonas, Eppie, Kevin, Ian, Frauke, Olivia, Victoria, Rui, Russell, Mark, Sarah and Abi. For making it such a good place to work. Thanks also to Sue, Dave, Paul, Brenda and Rachel for their help and taking the aDNA labs special requests in their stride.

Thirdly I would to thank all the collaborators who made this thesis possible, everyone at TCD and UCD, especially David, Dave, Paul and Stephen. York especially, Matthew, Michi, Nienke and Chris and everyone who was part of the LeCHE project. Anna and Ron for his seemingly endless supply of enthusiasm and bones and all other archaeological sample providers.

Fourthly I would like to thank Mari Currie actions framework and Trinity College Dublin for funding and a studentship to complete this thesis.

I would like to thank all my friends, especially my flatmates Helen and Eddie and finally my family especially my brother, sister and parents for their continued support.

Table of Contents

Summary ..... i

Acknowledgments ..... iii

List of Tables ..... ix

List of figures ..... x

List of programs used in this thesis ..... xiii

List of samples used in this thesis ..... xiv

**1. General Introduction ..... 1**

1.1 Next Generation Sequencing ..... 1

1.2 Ancient DNA..... 3

1.3 The preparation of next generation sequencing libraries from Ancient DNA... 4

1.3.1 Ancient DNA extraction ..... 4

1.3.2 Quality control of ancient DNA extract..... 6

1.3.3 Next generation sequencing library production..... 7

1.3 Domestication of farm animals in the Near East ..... 11

1.4 Domestic bovine phylogeny..... 14

1.5 Aims of this thesis ..... 16

**2. General methods ..... 19**

2.1 Ancient DNA extraction ..... 19

2.2 Illumina sequencing library preparation ..... 19

2.3 Sequencing library enrichment..... 20

2.4 Initial sequencing quality control ..... 21

2.5 BWA sequence Alignment..... 22

2.6 Haploid SNP calling..... 23

2.7 PLINK analysis..... 24

2.8 Eigensoft Analysis 1x haplotype data..... 24

2.9 Procrustes analysis..... 25

2.10 MAPDAMAGE analysis..... 26

2.11 Depth of Coverage Analysis..... 26

**3. Bioinformatic analysis of domesticate mitochondrial aDNA ..... 27**

3.1 Introduction ..... 27

3.1.1 The Mitochondria as a source of ancient DNA..... 27

3.1.2 Next generation sequencing and whole ancient mitochondrial genomes..... 27

3.1.3 Sequence capture of ancient mitochondrial genomes..... 29

3.1.4 Ancient domesticate mitogenomics..... 31

3.1.5 Objectives for this case study..... 31

3.2 Materials and Methods..... 32

3.2.1 Archaeological Samples..... 32

3.2.2 Ancient DNA libraries preparation and capture..... 32

3.2.3 Initial QC and read trimming ..... 36

3.2.4 Identification of jumping PCR ..... 36

3.2.5 Read filtering and merging ..... 37

3.2.6 SNP calling..... 37

3.2.7 Fasta sequence production..... 37

3.2.8 MEGAN..... 38

3.2.9 Bayesian analysis of the estimated sequence divergence times..... 38

3.3 Results ..... 40

3.3.1 Bovid mtDNA sequencing results..... 40

3.3.2 Estimated contamination rate bovid capture..... 41

3.3.3 Off target sequence as an indicator of shotgun success..... 42

3.3.4 Comparison of the captured and un-captured metagenome ..... 44

3.3.5 Haplotype assignment of cattle samples..... 46



|                                                                                                        |            |
|--------------------------------------------------------------------------------------------------------|------------|
| 3.3.6 Bayesian phylogeny assignment cattle samples .....                                               | 46         |
| 3.3.8 SNP calling and analysis .....                                                                   | 53         |
| 3.3.9 Goat mtDNA phylogenetic Analysis .....                                                           | 54         |
| 3.3.10 Sheep mtDNA haplotype assignment and phylogenetic analysis .....                                | 54         |
| 3.4 Discussion .....                                                                                   | 58         |
| 3.4.1 Assessment of hybridisation capture as a method for ancient domesticate<br>mtDNA retrieval ..... | 58         |
| 3.4.2 mtDNA capture as an indicator for shotgun success .....                                          | 59         |
| 3.4.3 Bayesian phylogenetic analysis of domestic mtDNA genomes .....                                   | 61         |
| 3.5 Conclusion .....                                                                                   | 63         |
| <b>4. Population genetic analysis of a Mesolithic wild aurochs (<i>Bos<br/>primigenius</i>) .....</b>  | <b>64</b>  |
| 4.1 Introduction .....                                                                                 | 64         |
| 4.1.1 The aurochs ( <i>Bos primigenius</i> ) .....                                                     | 64         |
| 4.1.2 Aurochs and domestic cattle - Mitochondria .....                                                 | 65         |
| 4.1.3 Aurochs and domestic cattle – Y Chromosome .....                                                 | 67         |
| 4.1.4 The aurochs genome sequencing project .....                                                      | 68         |
| 4.1.5 PCA analysis of modern and ancient samples .....                                                 | 68         |
| 4.1.6 SNP population genetic analysis of modern cattle populations .....                               | 69         |
| 4.1.7 Objectives for this case study .....                                                             | 70         |
| 4.2 Methods .....                                                                                      | 71         |
| 4.2.1 Bovine SNP50 data set .....                                                                      | 71         |
| 4.2.2 Bovine HD SNP770 data set .....                                                                  | 71         |
| 4.2.3 CPC98 SNP genotyping .....                                                                       | 71         |
| 4.2.4 PLINK analysis .....                                                                             | 75         |
| 4.2.5 Eigensoft Analysis .....                                                                         | 75         |
| 4.2.6 Structure analysis .....                                                                         | 75         |
| 4.2.7 Phylogenetic tree production .....                                                               | 76         |
| 4.2.8 Green et al 2010 D Statistic .....                                                               | 76         |
| 4.3 Results .....                                                                                      | 79         |
| 4.3.1 SNPs genotyped SNP50 .....                                                                       | 79         |
| 4.3.2 50K Eigensoft analysis worlds populations and CPC98 .....                                        | 79         |
| 4.3.3 50K STRUCTURE Analysis world populations .....                                                   | 80         |
| 4.3.4 50K Phylogenetic analysis world populations .....                                                | 81         |
| 4.3.5 50K Eigensoft analysis of taurine Populations and European aurochs .....                         | 86         |
| 4.3.6 50K European taurine and European aurochs .....                                                  | 86         |
| 4.3.7 SNP 770K analysis .....                                                                          | 89         |
| 4.3.8 Green et al 2010, D-Statistic .....                                                              | 90         |
| 4.3.9 Structure Analysis 770K .....                                                                    | 91         |
| 4.4 Discussion .....                                                                                   | 95         |
| 4.4.1 PCA Analyses .....                                                                               | 95         |
| 4.4.2 Structure analysis .....                                                                         | 96         |
| 4.4.3 Utility of current SNP data in the analysis of ancient bovids .....                              | 97         |
| 4.4.4 D-statistic .....                                                                                | 98         |
| 4.5 Conclusion .....                                                                                   | 99         |
| <b>5. Evaluating exon capture for use with ancient genomes .....</b>                                   | <b>102</b> |
| 5.1 Introduction .....                                                                                 | 102        |
| 5.1.1 Selection and the bovine genome .....                                                            | 102        |
| 5.1.2 Nuclear sequence capture and ancient DNA .....                                                   | 103        |
| 5.1.3 Objectives for this case study .....                                                             | 104        |
| 5.2 Methods .....                                                                                      | 105        |
| 5.2.1 Samples chosen for analysis .....                                                                | 105        |
| 5.2.2 Genes targeted for sequencing .....                                                              | 106        |
| 5.2.3 Bait design .....                                                                                | 106        |

|                                                                                                   |            |
|---------------------------------------------------------------------------------------------------|------------|
| 5.2.4 Sequencing library preparation.....                                                         | 107        |
| 5.2.5 RNA capture assay .....                                                                     | 108        |
| 5.2.6 PCR enrichment of RNA captured library.....                                                 | 108        |
| 5.2.7 Sequencing and assembly and SNP calling.....                                                | 109        |
| 5.2.8 Variant Filtering .....                                                                     | 109        |
| 5.2.9 SMART analysis .....                                                                        | 112        |
| 5.2.10 RNA function analysis .....                                                                | 112        |
| 5.3 Results .....                                                                                 | 114        |
| 5.3.1 CPC Results .....                                                                           | 115        |
| 5.3.2 Ardeche2 and Pooled samples.....                                                            | 115        |
| 5.4 Discussion .....                                                                              | 121        |
| 5.4.1 Aurochs capture analysis .....                                                              | 121        |
| 5.4.2 CPC98 TOLL like receptors .....                                                             | 121        |
| 5.4.2 CPC98 other genes and pathways highlighted .....                                            | 123        |
| 5.4.3 CPC98 UTR analysis.....                                                                     | 124        |
| 5.4.4 Off target sequence .....                                                                   | 126        |
| 5.4.5 Contamination .....                                                                         | 127        |
| 5.4.6 Ardeche2.....                                                                               | 128        |
| 5.4.7 Pooled samples.....                                                                         | 129        |
| 5.5 Conclusions .....                                                                             | 130        |
| <b>6. Genomic affinities of two 8,000-year-old <i>Bos taurus</i> animals from Anatolia .....</b>  | <b>133</b> |
| 6.1 Introduction.....                                                                             | 133        |
| 6.1.1 The domestic cattle of Anatolia.....                                                        | 133        |
| 6.1.2 Shotgun sequencing of ancient DNA .....                                                     | 133        |
| 6.1.3 Objectives for this case study.....                                                         | 134        |
| 6.2 Methods .....                                                                                 | 135        |
| 6.2.1 Samples analysed in this study. ....                                                        | 135        |
| 6.2.2 Ancient DNA extraction and library preparation.....                                         | 135        |
| 6.2.5 BWA sequence Alignment.....                                                                 | 135        |
| 6.2.7 Alignment to human genome .....                                                             | 136        |
| 6.2.8 SNP calling, Eigensoft and Procrustes analysis Analysis.....                                | 136        |
| 6.2.7 770K bovine SNP dataset .....                                                               | 136        |
| 6.3 Results.....                                                                                  | 139        |
| 6.3.1 Sequencing Results.....                                                                     | 139        |
| 6.3.2 Mitochondrial genomes .....                                                                 | 139        |
| 6.3.3 MAPDAMAGE analysis.....                                                                     | 140        |
| 6.3.4 PCA analysis .....                                                                          | 143        |
| 6.4 Discussion.....                                                                               | 149        |
| 6.4.1 Test of authenticity .....                                                                  | 149        |
| 6.4.2 PCA Analysis.....                                                                           | 150        |
| 6.5 Conclusion .....                                                                              | 153        |
| <b>7. Evaluating parchment as a reservoir of ancient DNA for next generation sequencing .....</b> | <b>155</b> |
| 7.1 Introduction .....                                                                            | 155        |
| 7.1.1 Parchment and the historical record .....                                                   | 155        |
| 7.1.2 Manufacturing parchment.....                                                                | 155        |
| 7.1.3 Parchment and aDNA .....                                                                    | 156        |
| 7.1.4 Sheep hapmap and breed formation.....                                                       | 157        |
| 7.1.4 Objectives of this case study.....                                                          | 158        |
| 7.2 Methods .....                                                                                 | 158        |
| 7.2.1 Parchment samples chosen for aDNA analysis.....                                             | 158        |
| 7.2.2 Ancient DNA extraction .....                                                                | 159        |
| 7.2.3 Illumina sequencing library preparation.....                                                | 160        |



|                                                                                           |            |
|-------------------------------------------------------------------------------------------|------------|
| 7.2.4 Initial parchment read QC.....                                                      | 160        |
| 7.2.5 FastQ Screen .....                                                                  | 160        |
| 7.2.6 BWA Sequence Alignment .....                                                        | 161        |
| 7.2.7 Alignment to human genome.....                                                      | 161        |
| 7.2.8 SNP calling .....                                                                   | 161        |
| 7.2.9 Whole Mitochondrial genomes.....                                                    | 162        |
| 7.2.10 Sequence alignment and tree drawing .....                                          | 162        |
| 7.2.11 Principal Component Analysis .....                                                 | 162        |
| 7.3 Results .....                                                                         | 165        |
| 7.3.1 Species identification, ZooMS.....                                                  | 165        |
| 7.3.2 Species identification, FASTQ_SCREEN.....                                           | 165        |
| 7.3.3 Sequencing Results .....                                                            | 168        |
| 7.3.4 Mitochondrial Genomes.....                                                          | 168        |
| 7.3.5 SNP analysis.....                                                                   | 172        |
| 7.4 Discussion.....                                                                       | 177        |
| 7.4.1 Suitability of parchment for aDNA analysis .....                                    | 177        |
| 7.4.2 Single species identification.....                                                  | 178        |
| 7.4.3 Sheep aDNA genomics.....                                                            | 179        |
| 7.4.3 Mitochondrial genomes.....                                                          | 180        |
| 7.4.4 SNP analysis.....                                                                   | 180        |
| 7.6 Conclusions .....                                                                     | 181        |
| <b>8. Molecular archaeological analysis of a Croatian Avar with Pott's disease.</b>       | <b>183</b> |
| 8.1 Introduction.....                                                                     | 183        |
| 8.1.1 The Avars .....                                                                     | 183        |
| 8.1.2 CTB6; an Avar individual with Pott's disease.....                                   | 183        |
| 8.1.3 Ancient DNA and pathogens.....                                                      | 184        |
| 8.1.4 Objectives of this chapter.....                                                     | 185        |
| 8.2 Materials and methods.....                                                            | 185        |
| 8.2.1 Archaeological samples.....                                                         | 185        |
| 8.2.2 Ancient DNA extraction.....                                                         | 185        |
| 8.2.3 Illumina sequencing library preparation .....                                       | 186        |
| 8.2.4 Initial sequencing QC.....                                                          | 186        |
| 8.2.5 BWA sequence Alignment.....                                                         | 186        |
| 8.2.6 Bowtie sequence Alignment.....                                                      | 186        |
| 8.2.7 SNP calling and analysis .....                                                      | 187        |
| 8.3 Results .....                                                                         | 187        |
| 8.3.1 CTB6 sequencing results.....                                                        | 187        |
| 8.3.2 PCA analysis.....                                                                   | 187        |
| 8.3.3 Pathogen aDNA analysis.....                                                         | 192        |
| 8.4 Discussion.....                                                                       | 194        |
| 8.4.1. PCA analyses.....                                                                  | 194        |
| 8.4.2. Molecular investigation for <i>M tuberculosis</i> .....                            | 195        |
| 8.5 Conclusion .....                                                                      | 195        |
| <b>9. General conclusions.....</b>                                                        | <b>197</b> |
| 9.1 Summary of results and their significance .....                                       | 197        |
| 9.1.1 Hybridisation capture, mtDNA and exon .....                                         | 197        |
| 9.1.2 Population genetic analysis of a Mesolithic aurochs CPC98 .....                     | 199        |
| 9.1.3 Haploid ancient DNA data.....                                                       | 200        |
| 9.1.4 Parchment as a source of aDNA.....                                                  | 201        |
| 9.2 Future prospects.....                                                                 | 201        |
| 9.2.1 New NGS library making technologies and 3 <sup>rd</sup> generation sequencing ..... | 201        |
| 9.2.1 Further studies.....                                                                | 202        |
| 9.2.2 Population genetics of bovine domestication .....                                   | 203        |



9.2.2 York parchments..... 204

10. References..... 207

11. APPENDIX:1 – Extra tables and figures ..... 229

12. APPENDIX:2 - Code ..... 241

## List of Tables

|                                                                                                                                                                                                                                                                                                                                                                       |     |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3:1 BOVID SAMPLES USED IN THIS ANALYSIS .....                                                                                                                                                                                                                                                                                                                         | 33  |
| TABLE 3:2 OVINE AND CAPRINE SAMPLES USED IN THIS ANALYSIS.....                                                                                                                                                                                                                                                                                                        | 35  |
| TABLE 3:3 RESULTS OF BOVID MTDNA ALIGNMENTS, ALIGNED READS = PERCENTAGE OF READS<br>ALIGNED TO THE REFERENCE GENOME. MEAN COVERAGE = MEAN 'X' COVERAGE OF THE MTDNA<br>GENOME AS CALCULATED BY GATK DEPTHOfCOVERAGE COMMAND, COVERAGE(>=1X,%) = THE<br>PERCENTAGE OF BASES CALLED IN THE MTDNA REFERENCE GENOME FROM EACH SAMPLE<br>GREATER THAN OR EQUAL TO 1X. .... | 39  |
| TABLE 3:4 ESTIMATED CONTAMINATION RATE FOR BOVID SAMPLES >2X GENOTYPES. NOTE:<br>HAPLOGROUP T = PREDOMINATE <i>BOS TAURUS</i> HAPLOGROUP, P = PREDOMINANT <i>BOS</i><br><i>PRIMIGENIOUS</i> HAPLOGROUP, HAPLOGROUP Q = MINOR <i>BOS TAURUS</i> HAPLOGROUP PREVIOUSLY<br>ONLY REPORTED IN ITALY, HAPLOGROUP E LOW FREQUENCY <i>BOS PRIMIGENIOUS</i> HAPLOGROUP. ..     | 42  |
| TABLE 3:5 RESULTS OF SHEEP AND GOAT ALIGNMENTS. ....                                                                                                                                                                                                                                                                                                                  | 51  |
| TABLE 3:6 ESTIMATED CONTAMINATION RATES OF SHEEP AND GOAT SAMPLES SUBJECTED TO<br>PHYLOGENETIC ANALYSIS.....                                                                                                                                                                                                                                                          | 53  |
| TABLE 4:1 GEOGRAPHIC LOCATION AND NUMBERS OF SAMPLES STUDIED 50K. EAT = EAST AFRICAN<br>TAURUS, EUT = EUROPEAN TAURINE, NAT = NORTH AFRICAN TAURINE, SYN = SYNTHETICS,<br>WAH = WEST AFRICAN HYBRIDS, WAT = WEST AFRICAN TAURINE, WAZ = WEST AFRICAN<br>ZEBU AND ZEB = ZEBU/INDICINE.....                                                                             | 73  |
| TABLE 4:2: GEOGRAPHIC LOCATION AND NUMBERS OF SAMPLES STUDIED 770K. EUT = EUROPEAN<br>TAURINE, NEH = NEAR EASTERN HYBRID, WAT = WEST AFRICAN TAURINE,<br>ZEB=ZEBU/INDICINE.....                                                                                                                                                                                       | 74  |
| TABLE 5:1: SAMPLES CHOSEN FOR EXON CAPTURE ANALYSIS .....                                                                                                                                                                                                                                                                                                             | 104 |
| TABLE 5:2 : RAW SEQUENCING ALIGNMENT RESULTS ALL SAMPLES .....                                                                                                                                                                                                                                                                                                        | 113 |
| TABLE 5:3: NON-SYNONYMOUS VARIANTS DISCOVERED DURING THE CPC98 EXON CAPTURE ANALYSIS.<br>WG = MUTATION CHECKED IN CPC98 WHOLE GENOME SEQUENCE WITH CAPTURED READS<br>REMOVED. HET/HOM= SNP IS HETEROZYGOUS OR HOMOZYGOUS AS ASSESSED BY ANNOVAR.<br>MQ = MAPPING QUALITY OF VARIANT ENCODED AS A PHRED QUALITY SCORE AS ASSESSED BY<br>ANNOVAR.....                   | 118 |
| TABLE 5:4: RESULTS OF THE ARDECHE2 EXON CAPTURE ANALYSIS, EXONIC VARIANTS RETRIEVED ≥ 1X.<br>HET/HOM= SNP IS HETEROZYGOUS OR HOMOZYGOUS AS ASSESSED BY ANNOVAR. MQ =<br>MAPPING QUALITY OF VARIANT ENCODED AS A PHRED QUALITY SCORE AS ASSESSED BY<br>ANNOVAR.....                                                                                                    | 119 |
| TABLE 5:5: RESULTS OF THE POOLED EXON CAPTURE ANALYSIS, EXONIC VARIANTS RETRIEVED ≥ 1X.<br>BERCY13 PROVIDED NO EXONIC VARIANT. HET/HOM= SNP IS HETEROZYGOUS OR HOMOZYGOUS<br>AS ASSESSED BY ANNOVAR. MQ = MAPPING QUALITY OF VARIANT ENCODED AS A PHRED QUALITY<br>SCORE AS ASSESSED BY ANNOVAR.....                                                                  | 120 |
| TABLE 6:1: TURKISH SAMPLES CHOSEN FOR SHOTGUN ANALYSIS .....                                                                                                                                                                                                                                                                                                          | 135 |
| TABLE 6:2: SEQUENCING RESULTS FOR THE 2 BOVINE SAMPLES STUDIED. ....                                                                                                                                                                                                                                                                                                  | 138 |
| TABLE 6:3: TURK11 SNPs CAPTURE AND WHOLE SAMPLE SHOTGUN, HAPLOTYPE DISTINCTIONS AS<br>ACHILLI <i>ET AL</i> (2009), LOCAL MUTATIONS MARKED WITH (?). FOR POSITIONS WITH<br>INSUFFICIENT COVERAGE TO CALL SNPs 1X GENOTYPE DATA IS SHOWN IN BRACKETS.....                                                                                                               | 141 |
| TABLE 6:4: TURK13 SNPs CAPTURED AND WHOLE SAMPLE SHOTGUN, HAPLOTYPE ASSIGNMENTS AS<br>ACHILLI <i>ET AL</i> (2009), LOCAL MUTATION MARKED WITH (?). DASHED LINE (-) MARKS SNP THAT<br>AREN'T COVERED AT ALL IN THIS DATASET. FOR POSITIONS WITH INSUFFICIENT COVERAGE TO<br>CONFIDENTLY CALL SNPs GENOTYPES ARE SHOWN IN BRACKETS. ....                                | 142 |
| TABLE 7:1: SEQUENCING RESULTS PA1 AND PA2.....                                                                                                                                                                                                                                                                                                                        | 164 |
| TABLE 8:1: ARCHAEOLOGICAL SAMPLE CTB6 USED IN THIS STUDY. ....                                                                                                                                                                                                                                                                                                        | 185 |
| TABLE 8:2 CTB6 SHOTGUN SEQUENCING RESULTS.....                                                                                                                                                                                                                                                                                                                        | 189 |

List of figures

FIGURE 1:1 AGILENT BIOANALYSER TRACE SHOWING AN ADNA NGS LIBRARY WITH ARTEFACTS INDICATIVE OF A LOW STARTING DNA CONCENTRATION. PEAKS AT 15 AND 1,500 = DNA MARKERS, 100BP – 125BP = ADAPTER DIMER ARTEFACTS CAUSED BY LOW STARTING SAMPLE CONCENTRATION WHICH CAN BE SEQUENCED REDUCING OVERALL SEQUENCING EFFICIENCY, 150BP-250BP = ADNA SEQUENCING LIBRARY, 50BP = PCR PRIMER DIMERS. FU = INTENSITY OF SIGNAL AT BP RANGE. BP = BASE PAIRS. .... 5

FIGURE 1:2 AGILENT BIOANALYSER TRACE SHOWING A GOOD ADNA EXTRACT. PEAKS AT 15 AND 1,500 = DNA MARKERS, ~50-200BP = ADNA, 700BP – OFF SCALE = PROBABLE EXOGENOUS CONTAMINATION..... 7

FIGURE 1:3 MAP SHOWING THE PROPOSED LOCATION OF THE INITIAL DOMESTICATION CENTRE OF SHEEP, GOAT CATTLE AND PIGS (RED), ADAPTED FROM ZEDER (2008). NOTE THE TRANSPORT OF DOMESTICATE ANIMALS TO CYPRUS (GREEN) BY BOAT AT ROUGHLY 10,000YBP PROVIDES CONVINCING EVIDENCE OF HUMAN MANAGEMENT..... 13

FIGURE 1:4 NEIGHBOUR JOINING TREE SHOWING THE UNROOTED BOVINE MITOCHONDRIAL PHYLOGENY, THE MAJOR TAURINE (Q, T, P AND R) AND INDICINE (I1 AND I2) HAPLOROUPS ARE HIGHLIGHTED. THE PREDOMINANT HAPLOGROUPS IN EACH POPULATION ARE ALSO MARKED WITH OVALS GREEN (T AND Q) FOR TAURINE ANIMALS AND RED FOR INDICUS (I). HAPLOGROUP P PREDOMINATES IN THE EXTINCT EUROPEAN ARUOCHS POPULATION AND HAS ONLY BEEN DISCOVERED IN A SINGLE EXTANT INDIVIDUAL. WHEREAS HAPLOGROUP R HAS BEEN DISCOVERED IN A SMALL NUMBER OF MODERN ITALIAN ANIMALS. NOTE THE BROOMSTICK PHYLOGENY WITH DEEP BRANCH LENGTHS PROPOSING GEOLOGICAL TIMESCALES DIVERGENCES BETWEEN POPULATIONS..... 15

FIGURE 2:1 CARTOON OF CRAIG *ET AL* (2008) SAMPLE BARCODING SCHEME. EACH ADNA INSERT IS LIGATED TO SEQUENCING ADAPTERS (DARK GREEN) CONTAINING 6BP BARCODES (BLUE, ORANGE AND LIGHT GREEN) [A], SAMPLES CAN THEN BE POOLED AND SEQUENCED TOGETHER [B] AND THE RESULTANT SEQUENCING READS FROM EACH SAMPLE [C] ARE IDENTIFIED BY THE 6BP BARCODE AT THE START OF THE READ. .... 20

FIGURE 3:1:BAR GRAPH SHOWING THE RAPID INCREASE IN THE NUMBER OF ANCIENT WHOLE MTDNA GENOMES PRODUCED USING BOTH PCR (RED) AND NGS (BLUE) METHODS. THE SIGNIFICANT CONTRIBUTION OF NEXT GENERATION SEQUENCING TO THE TOTAL IS NOTABLE. DATA TAKEN FROM (PAIJMANS *ET AL*. 2012). .... 28

FIGURE 3:2 CARTOON OF TWO METHODS OF HYBRIDISATION CAPTURE ARE SHOWN, ONE IN SOLUTION USING STREPTAVIDIN BEADS [B], THE OTHER USING BOUND PROBES ON A CHIP (THICK BLACK LINE) [A]. GREEN LINES = RNA PROBES BOTH IMMOBILISED TO A CHIP [A] AND BIOTINYLATED (BROWN CIRCLES) [B]. SOLID BLACK LINES = TARGETED DNA WITH LIGATED SEQUENCING ADAPTERS (RED) AND ADAPTER SEQUENCE BLOCKS (BLUE). RED OVAL STREPTAVIDIN COATED BEAD. [C] COMPLIMENTARY BASE PARING OF SEQUENCE BLOCKS (BLUE) AND ADAPTER SEQUENCES (RED)..... 30

FIGURE 3:3 NUCLEAR GENOME ALIGNMENT PERCENTAGE RNA mtDNA CAPTURE (CAP) VS. WHOLE GENOME SHOTGUN (SHOT), REGRESSION LINE FOR ILLUSTRATIVE PURPOSES ONLY. .... 43

FIGURE 3:4 MEGAN ANALYSIS (BLASTN) 100,000 READS TURK13 RNA CAPTURED AND UN-CAPTURED (SHOTGUN). NOTE THE GREATER AMOUNT OF READS ASSIGNED TO THE BACTERIAL CLADE AND OTHER KINGDOMS IN THE RNA CAPTURE ANALYSIS AND A CORRESPONDING INCREASE IN NO HITS IN THE UN-CAPTURE ANALYSIS ..... 45

FIGURE 3:5 BAYESIAN ESTIMATE OF MAXIMUM CLADE CREDIBILITY TREE USING MODERN AND ANCIENT BOVID MITOCHONDRIAL GENOMES OF TAURINE, INDICINE AND AUROCHS ORIGIN. A HKY NUCLEOTIDE SUBSTITUTION MODEL WAS USED. DATES ARE PRESENTED FOR BRANCHES WITH A POSTERIOR PROBABILITY OF  $\geq 0.9$ . MODERN SAMPLES ARE REPRESENTED BY THEIR HAPLOGROUP T ETC. ANCIENT SAMPLES ARE NAMED. MODIFIED WITH PERMISSION FROM FIGURE 5.6 STOCK F, PHD THESIS, TRINITY COLLEGE DUBLIN LIBRARY..... 48

FIGURE 3:6 NUMBER OF READS FROM EACH SAMPLE THAT ALIGN TO THE GOAT (RED)/ SHEEP (BLUE) GENOME. NOTE: BECAUSE OF THE DEGRADED NATURE OF ADNA AND THE SHORT READ LENGTHS USED IN THIS ANALYSIS READS CAN FIND SUFFICIENTLY HIGH ALIGNMENTS TO BE REPORTED AS HITS TO BOTH THE CLOSELY RELATED SHEEP AND GOAT MTDNA GENOME..... 52

FIGURE 3:7 BAYESIAN ESTIMATE OF MAXIMUM CLADE CREDIBILITY TREE USING MODERN AND ANCIENT OVINE SAMPLES. MODERN SAMPLE GENOMES ARE LABELLED BY HAPLOTYPE ANCIENT SAMPLES BY SAMPLE CODE. THE POSTERIOR PROBABILITY OF ALL BRANCHES WITH DATES IS  $\geq 0.9$ . ALL DATES ARE IN YBP, HPD = 95% HIGHEST PROBABILITY DENSITY..... 56



|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |     |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| FIGURE 3:8: BAYESIAN ESTIMATE OF MAXIMUM CLADE CREDIBILITY TREE USING MODERN AND ANCIENT OVINE SAMPLES WITHOUT GUB13. MODERN SAMPLE GENOMES ARE LABELLED BY HAPLOTYPE ANCIENT SAMPLES BY SAMPLE CODE THE POSTERIOR PROBABILITY OF ALL BRANCHES WITH DATES IS $\geq 0.9$ . ALL DATES ARE IN YBP, 95 %HPD = 95% HIGHEST PROBABILITY DENSITY. ....                                                                                                                                                                                                                                                                                                      | 57  |
| FIGURE 4:1: DIAGRAM ADAPTED FROM GREEN <i>ET AL</i> (2010) (SOM 15), <i>D</i> -STATISTIC TO TEST THE NULL HYPOTHESIS THAT H1 AND H2 FORM A CLADE, WITH CPC98 MORE DISTANTLY RELATED. [A] PHYLOGENETIC TREE FOR CATTLE H1 = TEST POPULATION 1, H2 = TEST POPULATION 2, CPC98. [B-D] GENEALOGICAL TREES "A" = ANCESTRAL ALLELE, "B" = ALTERNATIVE ALLELE, BBAA SITES [B] ARE EXPECTED TO OCCUR MORE FREQUENTLY, AS IN THIS CONFIGURATION THE GENEALOGY MIRRORS THE PHYLOGENY. ABBA [C] AND BABA [D] SITES ARE THEN EXPECTED TO OCCUR AT EQUAL FREQUENCIES, IF LINEAGES H1 AND H2 GO BACK TO THE TIME OF DIVERGENCE FROM CPC98 WITHOUT COALESCING. .... | 78  |
| FIGURE 4:2: SMARTPCA ANALYSIS OF CATTLE AND CPC98, BOVINE 50K DATA SET (12,418 SNPs) ROMAGNOLA DISTRIBUTION SHOWN BY "R" ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 82  |
| FIGURE 4:3 STRUCTURE ANALYSIS OF CPC98 (AUR) AND MODERN CATTLE POPULATIONS (17,443 SNPs, $K=3-7$ ). NOTE EACH LINE IN THE BAR GRAPH DENOTES 1 INDIVIDUAL APART FROM ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 83  |
| FIGURE 4:4: NEIGHBOUR JOINING TREE PRODUCED FROM PLINK IBD ANALYSIS OF 1,159 ANIMALS AND 17,443 SNPs. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | 84  |
| FIGURE 4:5: SMARTPCA ANALYSIS OF TAURINE BREEDS FROM THE 50K DATASET. ENCOMPASSING 262 ANIMALS FROM 25 BREEDS AND CPC98 (12,418 SNPs). ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | 85  |
| FIGURE 4:6 EUROPEAN TAURINE ONLY ANALYSIS. [A] PCA SHOWING CPC98 (RED SQUARE), HEREFORD (BLUE CROSSES) AND THE REST OF EUROPEAN TAURINE ANIMALS (GREY TRIANGLES). THIS ANALYSIS HIGHLIGHTS PROBABLE SNP ASCERTAINMENT EFFECTS, WHICH INTERN LIKELY CAUSE AN ARTEFACTUAL POSITIONING OF CPC98. [B] ANALYSIS REPEATED WITH THE HEREFORD ANIMALS REMOVED (12,418 SNPs). ....                                                                                                                                                                                                                                                                            | 88  |
| FIGURE 4:7 770K SMARTPCA ANALYSES OF CATTLE AND CPC98 (161,898 SNPs). ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | 92  |
| FIGURE 4:8: PAIRWISE <i>D</i> -STATISTIC ANALYSIS OF 770K GENOTYPE DATA. NEGATIVE VALUES INDICATE THAT CPC98 IS CLOSER TO POPULATION1 (H1), POSITIVE SCORES INDICATE THAT THE CONVERSE IS TRUE (186,198 SNPs). ....                                                                                                                                                                                                                                                                                                                                                                                                                                  | 93  |
| FIGURE 4:9 770K STRUCTURE ANALYSIS. NOTE ALL BARS INDICATE 1 ANIMAL APART FROM CPC98 WHICH IS PLACED 10 TIMES POST ANALYSIS FOR EASE OF VIEWING (186,198 SNPs). ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | 94  |
| FIGURE 5:1 RNA CAPTURE DESIGN OVERVIEW. A: BAITS PRODUCED FOR TLR4 (EXONS, UTRs, 2KB UPSTREAM) SHOWN BY BLACK BARS ABOVE THE EXON DIAGRAM; REPEAT REGIONS MARKED BY RED BARS WEREN'T COVERED BY BAITS TO TRY TO INCREASE UNIQUE ALIGNMENTS. B: ZOOMED VIEW OF TLR4 EXON 2 A JUSTIFIED BAIT TILING DESIGN WAS SELECTED TO TRY TO IMPROVE THE COVERAGE OF EXONS AND FLANKING SEQUENCE. FIGURE REPRESENTS A MODIFIED UCSC GENOME BROWSER OUTPUT ( <a href="http://genome.ucsc.edu/">HTTP://GENOME.UCSC.EDU/</a> ). RED HIGHLIGHT = 2KB UPSTREAM SEQUENCE, YELLOW = UTRs AND GREEN = EXONS. ....                                                         | 111 |
| FIGURE 5:2: DEPTH OF COVERAGE HISTOGRAM AUROCHS (CPC98) CAPTURE (NON-ZERO POSITIONS). ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | 114 |
| FIGURE 5:3: PERCENTAGE 1X COVERAGE OF TARGETED BASES IN ALL SAMPLES TESTED. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | 117 |
| FIGURE 5:4: CARTOON OF RNA CAPTURE WITH DAISY CHAINED BAITS. GREEN LINE = RNA SEQUENCE, RED LINE = NGS ADAPTER SEQUENCE, BROWN CIRCLE = BIOTIN, BLUE LINE = DNA ADAPTER BLOCKS, RED OVAL STREPTAVIDIN COATED BEAD. [A] RNA PROBE, [B] ADAPTER SEQUENCES (RED) CORRECTLY BLOCKED (BLUE) NO DAISY CHAINING. [C] BAITS INCORRECTLY BLOCKED WITH DAISY CHAINING. [D] BEAD CAPTURE OF BOTH THE CORRECTLY AND INCORRECTLY BLOCKED DNA/RNA MOLECULES. ....                                                                                                                                                                                                  | 127 |
| FIGURE 6:1: 770K SMARTPCA ANALYSIS TURK11, LEFT (TK11, 1,776 SNPs) AND TURK13, RIGHT (TK13, 4,633 SNPs). FULL 770K DATASET. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | 145 |
| FIGURE 6:2: MERGE OF TURK11 AND TURK13 770K SMARTPCA ANALYSES, FULL 770K DATASET. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | 146 |
| FIGURE 6:3: 770K SMARTPCA ANALYSIS TURK11 (TK11, 1,776 SNPs) AND TURK13 (TK13, 4,633 SNPs) EUROPEAN AND NEAR EASTERN SAMPLES ONLY. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | 147 |
| FIGURE 6:4: MERGE OF TURK11 AND TURK13 770K SMARTPCA ANALYSES, EUROPEAN AND NEAR EASTERN DATASET. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | 148 |
| FIGURE 7:1: PARCHMENT SAMPLES CHOSEN FOR ADNA ANALYSIS, LEFT PA1 18 <sup>TH</sup> CENTURY, RIGHT PA2 17 <sup>TH</sup> CENTURY. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 159 |
| FIGURE 7:2: <i>FASTQ SCREEN</i> ANALYSIS OF THE TWO PARCHMENTS (PA1 AND PA2), SHOWING PERCENTAGE OF READS THAT HAD A UNIQUE ALIGNMENT TO A SINGLE GENOME (RED BAR) NON-UNIQUE ALIGNMENT TO A SINGLE GENOME (ORANGE BAR), A UNIQUE ALIGNMENT IN MULTIPLE GENOMES (DARK BLUE BAR), NON-UNIQUE ALIGNMENT TO MULTIPLE GENOMES (LIGHT BLUE BAR).                                                                                                                                                                                                                                                                                                          |     |

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |     |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| N= NUMBER OF MISMATCHES ALLOWED. NOTE THAT AT ALL LEVELS OF STRINGENCY MOST SINGLE SPECIES ALIGNMENTS ARE FOUND TO THE SHEEP (OAR2) GENOME. ....                                                                                                                                                                                                                                                                                                                                                                                                                                               | 167 |
| FIGURE 7:3 NEIGHBOUR JOINING TREE OF PARCHMENT SAMPLES PA1 AND PA2 PLACED WITHIN MODERN SHEEP REFERENCE POPULATION (BOOTSTRAP 1,000 REPLICATES). PARCHMENT, OUTGROUPS AND MOUFLON LABELLED BY NAME ALL OTHER MODERN REFERENCE POPULATIONS BY HAPLOTYPE FULL DATASET LISTED IN APPENDIX:1A. ....                                                                                                                                                                                                                                                                                                | 170 |
| FIGURE 7:4: COVERAGE GRAPHS OF PA1 AND PA2 MITOGENOMES. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 171 |
| FIGURE 7:5: SMARTPCA ANALYSIS OF PA1 (1,834 SNPS) [A] AND PA2 (2,515 SNPS) [B]. NOTE: THE PCA AXES DESCRIBE A SMALL AMOUNT OF VARIATION. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 174 |
| FIGURE 7:6: SMARTPCA ANALYSIS OF PA1 [A] (1,834 SNPS) AND PA2 [B] (2,515 SNPS), EUROPEAN BREEDS ONLY. APD=AUSTRALIAN POLL DORSET, ASU = AUSTRALIAN SUFFOLK, BHM = BLACK HEADED MOUNTAIN, BRL = BORDER LEICESTER, CPW = AUSTRALIAN COOPWORTH, DSH = DORSET HORN, EFB = EAST FRIESIAN BROWN, EFW = EAST FRIESIAN WHITE, FIN = FINN SHEEP, GAL = GALWAY, GTX = GERMAN TEXEL, ICE = ICELANDIC, ISF = IRISH SUFFOLK, NSP = OLD NORWEGIAN SPAELSAU, NTX = NEW ZEALAND TEXEL, PA1 = PARCHMENT 1, ROM = NEW ZEALAND ROMNEY, SBF = SCOTTISH BLACK FACE = STX = SCOTTISH TEXEL AND WIL = WILTSHIRE. .... | 175 |
| FIGURE 7:7 PROCRUSTIS SMARTPCA ANALYSIS OF PA1 (GREEN SQUARE) AND PA2 (RED SQUARE). APD=AUSTRALIAN POLL DORSET, ASU = AUSTRALIAN SUFFOLK, BHM = BLACK HEADED MOUNTAIN, BRL = BORDER LEICESTER, CPW = AUSTRALIAN COOPWORTH, DSH = DORSET HORN, EFB = EAST FRIESIAN BROWN, EFW = EAST FRIESIAN WHITE, FIN = FINN SHEEP, GAL = GALWAY, GTX = GERMAN TEXEL, ICE = ICELANDIC, ISF = IRISH SUFFOLK, NSP = OLD NORWEGIAN SPAELSAU, NTX = NEW ZEALAND TEXEL, PA1 = PARCHMENT 1, ROM = NEW ZEALAND ROMNEY, SBF = SCOTTISH BLACK FACE = STX = SCOTTISH TEXEL AND WIL = WILTSHIRE. ....                   | 176 |
| FIGURE 8:1 CTB6 METATARSAL SAMPLE SUBMITTED FOR ADNA RIGHT AND VERTEBRA SHOWING SIGNS OF POTT'S DISEASE LEFT. ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 184 |
| FIGURE 8:2: SMARTPCA ANALYSIS, CTB6 AND WORLDWIDE POPULATIONS (2,244 SNPs) ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | 190 |
| FIGURE 8:3 PCA ANALYSIS CTB6 AND EUROPEAN CENTRAL ASIAN AND MIDDLE EASTERN POPULATIONS (2,244 SNPs) ....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 191 |
| FIGURE 8:4 SCREENSHOT FROM IGV VIEWER SHOWING STACKING OF CTB6 READS AROUND NCRNA GENES IN THE <i>M. TUBERCULOSIS</i> GENOME. THE TOP LINE OF THE PLOT DEPICTS THE DEPTH OF COVERAGE OVER COMPLETE <i>M. TUBERCULOSIS</i> GENOME THE NEXT TWO LINES POSITIONS OF tRNA AND ALL NCRNA GENES. ....                                                                                                                                                                                                                                                                                                | 193 |



## *List of programs used in this thesis*

| Program                             | Description                                                                                         | Reference                                                                                                                                     |
|-------------------------------------|-----------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| Bowtie                              | Program for mapping NGS reads against a reference genome                                            | (LANGMEAD <i>et al.</i> 2009)                                                                                                                 |
| BWA<br>(Burrows-Wheeler<br>Aligner) | Program for mapping NGS reads against a reference genome                                            | (LI and DURBIN 2009)                                                                                                                          |
| EIGENSOFT                           | A collection of tools for SNP principal component analysis                                          | (PATTERSON <i>et al.</i> 2006)                                                                                                                |
| Fastq Screen                        | A program for aligning raw NGS data to a collection of genomes, for the purposes of quality control | <a href="http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/">http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/</a> |
| FastQC                              | A program for initial quality control of NGS sequencing data                                        | <a href="http://www.bioinformatics.babraham.ac.uk/projects/fastqc/">http://www.bioinformatics.babraham.ac.uk/projects/fastqc/</a>             |
| GATK (Genome<br>analysis tool kit)  | Programs for the manipulation of NGS alignment files                                                | (MCKENNA <i>et al.</i> 2010)                                                                                                                  |
| Igv-viwer                           | Program for visualising NGS sequencing alignments                                                   | (THORVALDSDOTTIR <i>et al.</i> 2012)                                                                                                          |
| MEGAN                               | A program for the metagenomic analysis of next generation sequencing data                           | (HUSON <i>et al.</i> 2011)                                                                                                                    |
| PICARD                              | A further set of utilities for manipulating NGS alignment (BAM/SAM) files                           | <a href="http://picard.sourceforge.net/">http://picard.sourceforge.net/</a>                                                                   |
| PLINK                               | Program for the manipulation of SNP genotype files                                                  | <a href="http://pngu.mgh.harvard.edu/purcell/plink/">http://pngu.mgh.harvard.edu/purcell/plink/</a><br>(PURCELL <i>et al.</i> 2007)           |
| R                                   | Statistical analysis package                                                                        | <a href="http://www.r-project.org/">http://www.r-project.org/</a>                                                                             |
| SAMTOOLS                            | A set of utilities for manipulating NGS alignment (BAM/SAM) files                                   | (LI <i>et al.</i> 2009)                                                                                                                       |
| SeaView                             | Program for the multiple alignment and visualisation of sequence data                               | (GOUY <i>et al.</i> 2010)                                                                                                                     |
| STRUCTURE                           | Software for using multi-locus genotype data to investigate population structure.                   | (PRITCHARD <i>et al.</i> 2000)                                                                                                                |



# *List of samples used in this thesis*

| Sample         | Site                                                                 | Dating                                 | Species                | References for dating and site information                                                                                           |
|----------------|----------------------------------------------------------------------|----------------------------------------|------------------------|--------------------------------------------------------------------------------------------------------------------------------------|
| <b>CPC98</b>   | Carsington Pasture Cave,<br>Brassington, Derbyshire.<br>(Mesolithic) | 6,738 YBP<br>(Direct carbon date bone) | <i>Bos primigenius</i> | <a href="http://capra.group.shef.ac.uk/1/carsing.html">http://capra.group.shef.ac.uk/1/carsing.html</a><br>(TROY <i>et al.</i> 2001) |
| <b>Ard1</b>    | Ardeche, France<br>(Early Neolithic)                                 | 8,630 YBP<br>(Direct carbon date bone) | <i>Bos taurus</i>      | (EDWARDS <i>et al.</i> 2004)                                                                                                         |
| <b>Ard2</b>    | Ardeche, France<br>(Early Neolithic)                                 | 8,640 YBP<br>(Direct carbon date bone) | <i>Bos taurus</i>      | (EDWARDS <i>et al.</i> 2004)                                                                                                         |
| <b>Turk1</b>   | Aktopraklik, Turkey<br>(Neolithic)                                   | 9,000 YBP<br>(Dated by context)        | <i>Bos taurus</i>      | (STOCK <i>et al.</i> 2009)                                                                                                           |
| <b>Turk11</b>  | Mentese, Turkey<br>(Neolithic)                                       | 8,000 YBP<br>(Dated by context)        | <i>Bos taurus</i>      | (GOURICHON and HELMER 2008)                                                                                                          |
| <b>Turk13</b>  | Mentese, Turkey<br>(Neolithic)                                       | 8,000 YBP<br>(Dated by context)        | <i>Bos taurus</i>      | (GOURICHON and HELMER 2008)                                                                                                          |
| <b>UKR07</b>   | Starobels, eastern Ukraine<br>(Mesolithic)                           | 7,000 YBP<br>(Dated by context)        | <i>Bos primigenius</i> | (STOCK <i>et al.</i> 2009)                                                                                                           |
| <b>UKR08</b>   | Starobels, eastern Ukraine<br>(Mesolithic)                           | 7,000 YBP<br>(Dated by context)        | <i>Bos primigenius</i> | (STOCK <i>et al.</i> 2009)                                                                                                           |
| <b>Bercy10</b> | Bercy, France                                                        | 6,000 YBP                              | <i>Bos taurus</i>      | (EDWARDS <i>et al.</i> 2004)                                                                                                         |



|                | (middle Neolithic)                      | (Dated by context)                    |                                            |                                   |
|----------------|-----------------------------------------|---------------------------------------|--------------------------------------------|-----------------------------------|
| <b>Bercy13</b> | Bercy, France<br>(middle Neolithic)     | 6,000 YBP<br>(Dated by context)       | <i>Bos primigenius</i>                     | (EDWARDS <i>et al.</i> 2004)      |
| <b>Ron1</b>    | Bouqras, Syria<br>(Neolithic)           | 9,000 YBP<br>(Dated by context)       | <i>Bos taurus</i>                          | Pinhasi R, Personal communication |
| <b>Ron5</b>    | Bouqras, Syria<br>(Neolithic)           | 9,000 YBP<br>(Dated by context)       | <i>Bos taurus</i>                          | Pinhasi R, Personal communication |
| <b>Afr5d</b>   | West Africa, Mali<br>(Iron Age)         | 4090 YBP (Direct carbon<br>date bone) | <i>Bos taurus</i>                          | (EDWARDS <i>et al.</i> 2004)      |
| <b>Kud3_8</b>  | South Caucasus, Russia<br>(Paleolithic) | 43,900 YBP<br>(Dated by context)      | <i>Capra hircus</i>                        | Pinhasi R, Personal communication |
| <b>Kud1_22</b> | South Caucasus, Russia<br>(Paleolithic) | No date information                   | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication |
| <b>AK2</b>     | Crimea, Ukraine<br>(Bronze Age)         | No date information                   | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication |
| <b>AKT28</b>   | Turkmenistan<br>(Byzantine)             | 330 – 1204 YBP<br>(Dated by context)  | <i>Ovis aries</i>                          | Pinhasi R, Personal communication |
| <b>Kud3_7</b>  | South Caucasus, Russia<br>(Paleolithic) | 43,900 YBP<br>(Dated by context)      | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication |
| <b>Kud1_21</b> | South Caucasus, Russia<br>(Paleolithic) | 43,900 YBP<br>(Dated by context)      | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication |
| <b>Kud3_16</b> | South Caucasus, Russia<br>(Paleolithic) | 43,900 BP<br>(Dated by context)       | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication |
| <b>Kud3_9</b>  | South Caucasus, Russia<br>(Paleolithic) | 43,900 YBP<br>(Dated by context)      | <i>Capra hircus</i>                        | Pinhasi R, Personal communication |



|                   |                                                  |                                          |                                            |                                   |
|-------------------|--------------------------------------------------|------------------------------------------|--------------------------------------------|-----------------------------------|
| <b>Kud1_20</b>    | South Caucasus, Russia<br>(Paleolithic)          | 80,000-360,000 YBP<br>(Dated by context) | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication |
| <b>Crim</b>       | Crimea, Ukraine<br>(Bronze Age)                  | 2,300-2,200 YBP<br>(Dated by context)    | <i>Capra hircus</i>                        | Pinhasi R, Personal communication |
| <b>AKT 29</b>     | Turkmenistan<br>(Byzantine)                      | 330 – 1204 YBP<br>(Dated by context)     | <i>Ovis aries</i>                          | Pinhasi R, Personal communication |
| <b>Lad 24</b>     | St Petersburg<br>(Bronze Age)                    | 2,500 YBP<br>(Dated by context)          | <i>Ovis aries</i>                          | Pinhasi R, Personal communication |
| <b>KK 27</b>      | Rostov, Russia<br>(Bronze Age)                   | 2,500 YBP<br>(Dated by context)          | <i>Ovis aries</i>                          | Pinhasi R, Personal communication |
| <b>Gubs 13</b>    | North Caucasus, Russia<br>(Paleolithic)          | 33,500 YBP<br>(Dated by context)         | <i>Ovis aries</i>                          | Pinhasi R, Personal communication |
| <b>Gubs 25</b>    | North Caucasus, Russia<br>(Paleolithic)          | 33,500 YBP<br>(Dated by context)         | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication |
| <b>Akt'07 89E</b> | Turkey<br>(Neolithic)                            | 8,400 YBP<br>(Dated by context)          | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication |
| <b>CTB6</b>       | Stari Jankovci, Croatia<br>(Middle ages)         | 7th/8th century AD<br>(Dated by context) | <i>Homo sapiens</i>                        | Pinhasi R, Personal communication |
| <b>PA1</b>        | Parchment, York University<br>library collection | 18C<br>(Dated by handwriting)            | <i>Ovis aries</i>                          | Webb C, Personal communication    |
| <b>PA2</b>        | Parchment, York University<br>library collection | 17C<br>(Dated by handwriting)            | <i>Ovis aries</i>                          | Webb C, Personal communication    |



# 1. General Introduction

The application of next generation sequencing (NGS) technologies into the field of ancient DNA has seen a paradigm shift in the level of analysis that can be completed using archaeological samples (MEYER *et al.* 2012). However, to fully access the power that these techniques provide, certain mythological hurdles need to be overcome. This thesis describes the implementation of new technological improvements to the attainment and analysis of aDNA to address these problems, as well as the use of a new substrate (parchment) for the NGS analysis of aDNA.

The introduction to this thesis firstly describes next generation sequencing and its application to aDNA. Then the methodological improvements implemented to the standard Illumina next generation sequencing library preparation protocol used in this thesis to allow for its application to aDNA are outlined. Thirdly, descriptions of the beginnings of agriculture and the bovine phylogeny, which are of common interest to numerous chapters in this thesis, are presented. The introduction then ends with the general aims of this thesis and its relevance to the field of aDNA.

## 1.1 Next Generation Sequencing

Sanger sequencing has been the preeminent DNA sequencing technology for the last 30 years (SANGER *et al.* 1977) with its crowning achievement the sequencing of the first two human genomes (LANDER *et al.* 2001; VENTER

2001). However, at the genomic scale (although improved by major technological advancements in the last 20 years) it has become both cost inhibitive and time consuming (METZKER 2010); providing the demand for more efficient techniques. This niche has been filled by the so-called next or second-generation sequencing technologies (NGS), which following rapid improvements over the last 5 years now have the capacity to sequence up to 600Gb of DNA in a single 11 day run ([http://www.illumina.com/systems/hiseq\\_systems.ilmn](http://www.illumina.com/systems/hiseq_systems.ilmn)). This increase in efficiency has been brought about by the ability of these methods to sequence billions of DNA molecules in parallel on a single machine (METZKER 2010).

The wealth of sequencing data that these new techniques provide has had a profound effect on the way that DNA sequencing studies are conducted (KAHVEJIAN *et al.* 2008), with comparisons being made to introduction of the polymerase chain reaction (PCR) into molecular biology in the 1980s (METZKER 2010). For the first time it is now possible to sequence multiple mammalian genomes/exomes for a relatively low cost, in the time frame of weeks. This increase in data has provided both successes (METZKER 2010; 1000 GENOMES PROJECT CONSORTIUM 2010) and new challenges in the scale of analyses which are now required (KAHVEJIAN *et al.* 2008; FLICEK and BIRNEY 2009; METZKER 2010). The most widely used of the current next generation sequencing technologies is that of the Illumina (previously Solexa) pipeline (BENTLEY *et al.* 2008), which now dominates the market due to its increase capacity compared to rival methods .



## ***1.2 Ancient DNA***

One of the major goals of population genetics is to provide a description of a population or species through time, however this analysis can only be completed indirectly using contemporary genetic diversity (HOFREITER *et al.* 2001). The analysis of ancient DNA (aDNA) although technically challenging, allows for the direct sampling of individuals from the studied population at a range of time depths (HOFREITER *et al.* 2001; HO and GILBERT 2010; STONEKING and KRAUSE 2011; LALUEZA-FOX and GILBERT 2011), greatly improving the resolution of analysis.

Ancient DNA can broadly be defined as the study of degraded DNA from specimens not specifically preserved for genetic analysis (PÄÄBO *et al.* 2004; HO and GILBERT 2010). The analysis of aDNA has achieved a high public profile in the last 25 years, due in part to the technical challenges that the study of archaeological and historical samples presents (HOFREITER *et al.* 2001; PÄÄBO *et al.* 2004; HO and GILBERT 2010; KIRSANOW and BURGER 2012), but it also benefits from the compelling nature of the species under study for example the mammoth (MILLER *et al.* 2008; HO and GILBERT 2010). However, since the field's inception with the independent publishing of DNA sequences from the quagga and an Egyptian mummy (HIGUCHI *et al.* 1984; PÄÄBO 1985) it has not been without controversies, with recurrent debates regarding the contamination of aDNA samples and subsequently the authenticity of results (PÄÄBO *et al.* 2004; WILLERSLEV and COOPER 2005). However, the concerns raised in such articles have led to subsequent technical and theoretical

improvements within the field (HOFREITER *et al.* 2001; GILBERT *et al.* 2003; MALMSTRÖM 2005).

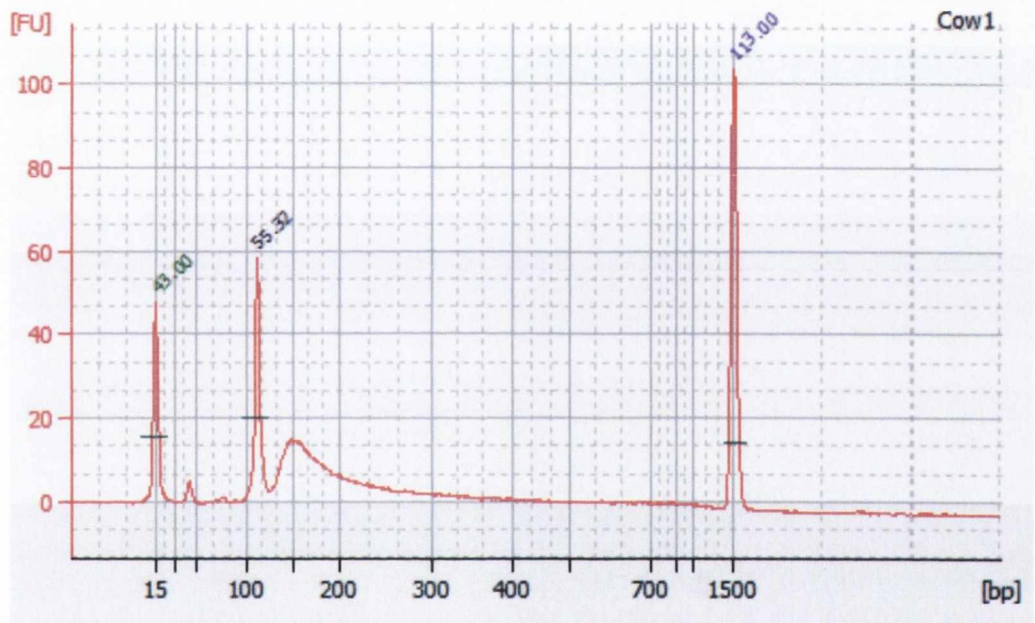
The introduction of next-generation sequencing methods into aDNA research has also lead to profound changes, with the utility of some of these techniques to rapidly sequence billions of small molecules quickly implemented. Early experiments exposed bone samples as having a vast metagenome with large quantities of exogenous contaminants present (POINAR *et al.* 2006). A typical endogenous DNA concentration of a non-permafrost bone sample is between 0.1 and 5% (EDWARDS *et al.* 2010; GREEN *et al.* 2010; ORLANDO *et al.* 2011; SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012), with some notable high profile exceptions (REICH *et al.* 2010). However, in spite of these technical challenges NGS methods have still allowed for the publication of 4 ancient genomes to date (MILLER *et al.* 2008; GREEN *et al.* 2010; RASMUSSEN *et al.* 2010; REICH *et al.* 2010; MEYER *et al.* 2012; MILLER *et al.* 2012).

### ***1.3 The preparation of next generation sequencing libraries from Ancient DNA***

#### **1.3.1 Ancient DNA extraction**

All ancient DNA samples analysed in this thesis from both parchment and bone were subjected to a standard aDNA extraction protocol as outlined in general methods (chapter 2) (YANG *et al.* 1998; EDWARDS *et al.* 2010). This protocol has been used successfully for almost two decades for the extraction

of DNA from ancient specimens for PCR analysis. However, it was found to have one major deficiency when applied to NGS that of a poor DNA sample concentrations. A small starting concentration of DNA in an extract though acceptable for PCR can lead to the production of artefacts in the library making process as shown in Figure 1:1, which have been shown in previous NGS data generated in house to reduce sequencing efficiency (PhD Thesis Frauke Stock, Trinity College Library).



**Figure 1:1 Agilent Bioanalyser trace showing an aDNA NGS library with artefacts indicative of a low starting DNA concentration. Peaks at 15 and 1,500 = DNA markers, 100bp – 125bp = adapter dimer artefacts caused by low starting sample concentration which can be sequenced reducing overall sequencing efficiency, 150bp-250bp = aDNA sequencing library, 50bp = PCR primer dimers. FU = intensity of signal at bp range. bp = base pairs.**

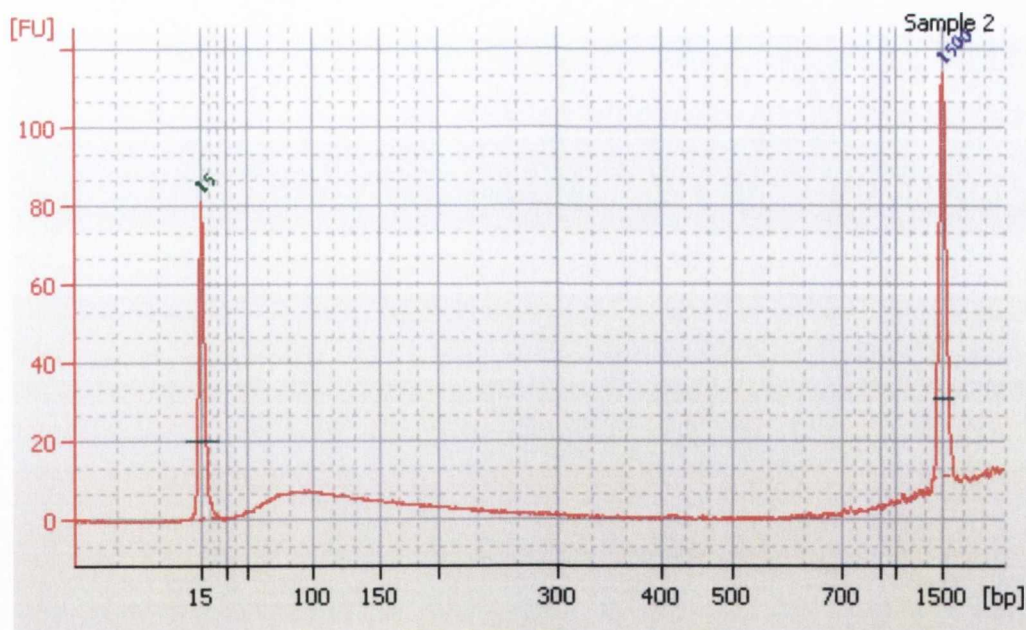
This extraction protocol was therefore modified to increase the final concentration of newly extracted DNA, this was completed via a two-stage process described below. Firstly, the silica spin columns used in the final DNA “clean-up” step were replaced with the Quigen min-elute column to allow for a decrease in the amount of elution buffer required to liberate the DNA from the



column. This elution volume was then reduced from the standard 100µl used in previous extractions to 35 µl. These modifications have the benefit of concentrating the DNA extract, which allows a greater total amount of DNA to be passed to the following library production steps in a smaller volume reducing the artefacts shown in figure 1:1.

### **1.3.2 Quality control of ancient DNA extract**

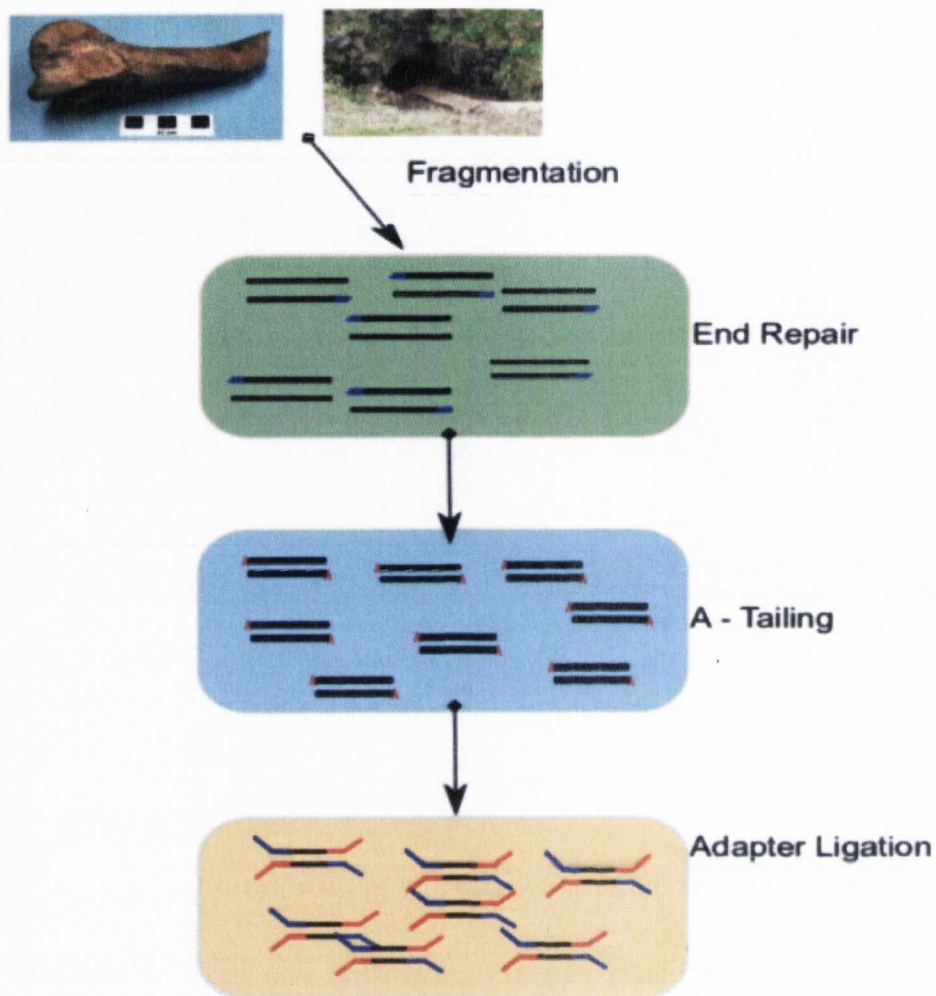
Two further quality control steps were also put into place to assess DNA quality and concentration before library production. The first was the use of the Qubit Fluorometer (Invitrogen), this product allowed for an accurate assessment of the concentration of the DNA sample prior to library preparation. Samples that fell below a total DNA concentration of 10ng/µl were deemed unusable and were not included in the library making process. Secondly an Agilent Bioanalyser (Agilent Technologies) was also used to assess the quality of the DNA extract, an example of a high quality extract is shown in figure 1:2. Both these steps allowed the DNA sample retrieved from each specimen to be assessed and quantified before the expensive library making procedures were completed.



**Figure 1:2 Agilent Bioanalyser trace showing a good aDNA extract. Peaks at 15 and 1,500 = DNA markers, ~50-200bp = aDNA, 700bp – off scale = probable exogenous contamination.**

### 1.3.3 Next generation sequencing library production

Figure 1:3 outlines the library construction used in this thesis. Each stage of this process provides challenges to the construction of aDNA sequencing libraries that can reduce the overall efficiency of analysis. During the completion of this work modifications were introduced to key steps, which addressed these issues and are described below.



**Figure 1:3 Basic outline of the aDNA Illumina library making procedure.** Fragmented ancient DNA is extracted from the specimen under study using standard aDNA techniques (YANG *et al.* 1998; EDWARDS *et al.* 2010). The DNA is cleaned and end repaired (green), an A base overhang is then applied to allow for a more efficient ligation of the sequencing adapter required for cluster formation and sequencing (yellow). The resultant library molecule is then PCR amplified prior to sequencing.



### ***Fragmentation***

The DNA fragmentation and size selection step that is common to all library preparations using modern DNA isn't required for aDNA. This step is routinely bypassed for aDNA library preparation protocols due to the highly degraded nature of the DNA molecules in the sample. However, this stage may be thought of in other terms as the sample selected for analysis. Two main protocols were implement in the selection of samples to try to improve sequencing efficiency. Firstly, only samples that had shown to previously give reliable PCR results were selected for further analysis. Secondly, mtDNA capture as described in chapter 3 was used to highlight bovid samples with preferable endogenous DNA concentrations for further analysis in chapters 5 and 6. Given that some of the research questions asked by this thesis required the analysis of specimens from locations not usually suitable for aDNA analysis, these quality control procedures (though not perfect) taken before the expensive undertaking of library construction and sequencing likely reduced the total cost of their analysis.

### ***End Repair and A-tailing***

Both the end repair and A-tailing steps of the library production protocol were likely improved by the addition of a more concentrated DNA sample provided by the modification of the extraction protocol described above. However, no specific modifications to the library preparation protocol were undertaken at these stages.

### ***Adapter ligation***

The final stage of library production adapter ligation requires the addition of universal sequencing adapters to the extracted DNA molecules. The concentrations of these adapters used in the standard library production protocols were found to be too high for the low amounts of DNA used in the production of aDNA NGS libraries and produced similar artefacts to those seen in figure 1:1. These adapters were therefore diluted 1 in 100 before use to alleviate this problem.

### ***PCR amplification***

The final step of the library making process requires the PCR amplification of the library to sequenceable concentrations. However over amplification of libraries can lead to PCR stacking; the production of clonal amplification products of the same initial library molecules, which then have to be collapsed before analysis and count as wasted sequencing. Previously to this work gel electrophoresis was used to quantify the concentration of DNA within the sequencing library. However, for the library to be visualised using this technique too many rounds of PCR were required leading to over amplified libraries. This issue was addressed by quantifying the PCR products on an Agilent Bionalyzer, which requires far fewer rounds of PCR and allows for these PCR induced artefacts to be reduced.

### ***1.3 Domestication of farm animals in the Near East***

The domestication of plants and animals is one of the major turning points in human evolution as human populations could for the first time control their access to and supply of food (ZEDER *et al.* 2006; ZEDER 2008; VIGNE 2011).

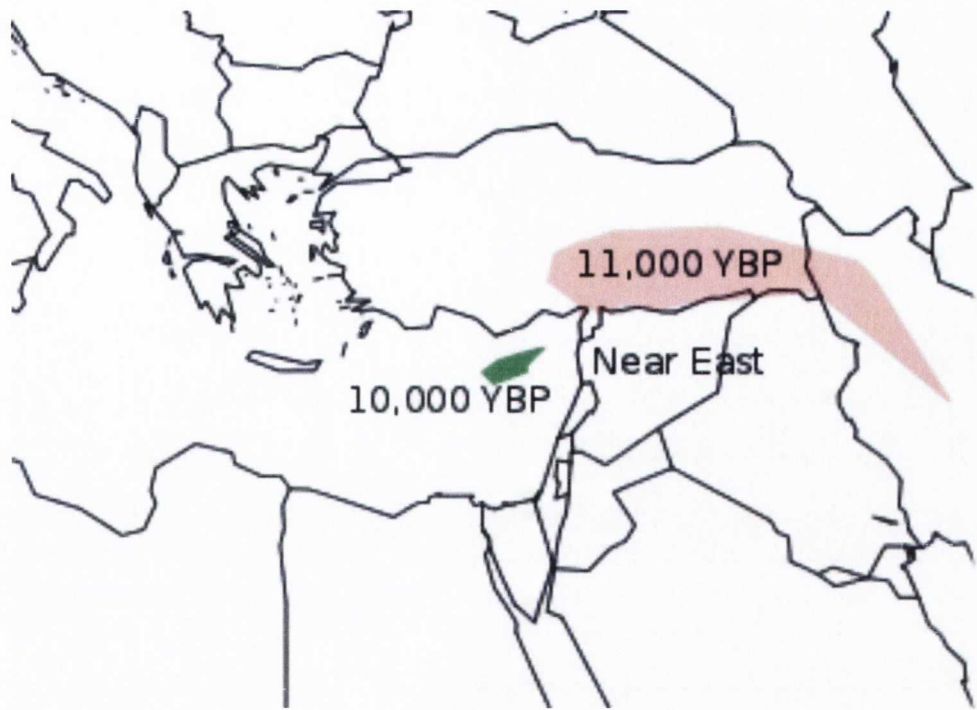
The beginnings of agriculture are thought to have arisen out of the remnants of the Natufian culture which inhabited the Near East from around 12,500 to 11,500 years BP (BAR YOSEF 1998). This culture is thought to be the first sedentary or at least semisendatary population in the Near East and possibly the world (VIGNE 2011). This sedimatism is likely helped by the proposed high carrying capacity of the region at the time, which was has been likened to the garden of Eden (BAR YOSEF 1998).

The two cultures that emerged after the Natufian's the Pre-Pottery Neolithic A and B (PPNA and PPNB) are credited with the domestication of first plants and then animals (ZEDER *et al.* 2006; ZEDER 2008; TRESSET and VIGNE 2011; ZEDER 2011; VIGNE *et al.* 2012). Two major archaeological methods have been applied to the study of these populations. Firstly the study of settlement patterns which show the production of more robust housing (BAR YOSEF 1998), with the novel subsistence strategy of farming proposed to have allowed for a societal diversification of skills and the production of these buildings (CAUVIN and WATKINS 2000).



Secondly the study of plant and faunal remains, this analysis of the specimens usually takes the form of looking at kill off patterns (age at which an animal is killed) (VIGNE and HELMER 2007). Hunter gathers tend to target adult males to increase the returns of their hunting investments, however farmers kill males young apart from a few animals maintained for herd propagation (ZEDER 2008). These differing strategies lead to a dichotomy in the faunal assemblages of these two groups, which can then be used to deduce how ancient societies are using animals (ZEDER 2008; TRESSET and VIGNE 2011). During the Neolithic the faunal assemblages at key sites are seen to change away from ones indicative of a hunting strategy towards those suggesting farming practices are being introduced (VIGNE and HELMER 2007; VIGNE 2011; ARBUCKLE 2013). Evidence of morphological changes indicative of domestication, for example animal size reduction are also seen in the Near East between 10,500 and 10,700 year ago (VIGNE 2011).

Even at this very early stage these animals were well-managed with goat, sheep, cattle, pig and deer all being transported 80 kilometres by boat to start the new Neolithic economies on Cyprus around 10,000 years ago (ZEDER 2008; VIGNE 2011; ZEDER 2011; VIGNE *et al.* 2012). These domestic animals then moved with migrating farmers out of the Neolithic core zone into Europe and Africa (TRESSET *et al.* 2009; TRESSET and VIGNE 2011).



**Figure 1:3 Map showing the proposed location of the initial domestication centre of sheep, goat cattle and pigs (red), adapted from Zeder (2008). Note the transport of domesticate animals to Cyprus (green) by boat at roughly 10,000YBP provides convincing evidence of human management.**

A second major domestication centre in Asia has also been proposed in the Indus valley. Archaeological evidence suggests that this region experienced both the arrival of domesticated plants from the Near East and independent domestications maybe through knowledge transfer between these two groups (FULLER 2006). One of the key centres of domestication in this region is proposed to be the Baluchistan region, with the cultures inhabiting this area first seeing the arrival of domestic crops from the Near East around 9000 BP. This influx is thought to have prompted the domestication of animal species (CHEN *et al.* 2009), notably *Bos primigenius namadicus* the proposed ancestral population of *Bos indicus* (MEADOW 1989; 1991; 1996). These animals are thought to have ranged throughout the Indian subcontinent during the

Pleistocene and Holocene periods and these surviving populations gave rise to the domestic *Bos indicus* populations (VAN VUURE 2005).

## ***1.4 Domestic bovine phylogeny***

Global cattle populations can generally be thought to split into two morphologically distinguishable groups, firstly, the hump less *Bos taurus* breeds and secondly, the humped *Bos indicus* animals of Asia; this distinction did not go un-noticed by Charles Darwin who wrote.

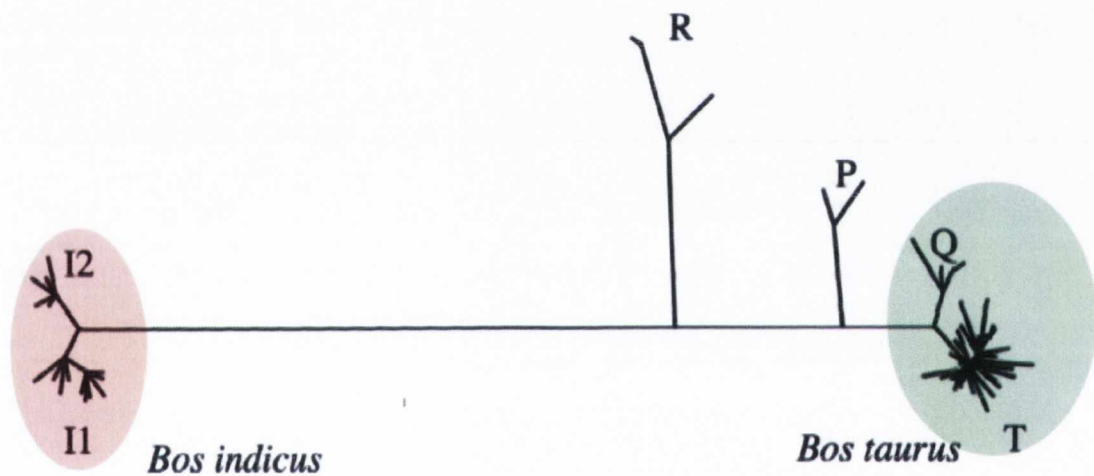
*"I should think, from facts communicated to me by Mr Blyth, on the habits, voice and constitution etc. of the humped Indian cattle, that these had descended from a different aboriginal stock from our European cattle; and several competent judges believe these latter may have had more than one wild parent."*

(Charles Darwin. 'On the Origin of Species', p.28)

The initial population genetic analysis of global cattle populations confirmed this distinction with mitochondrial D-loop sequences showing a deep bifurcation between the two groups (Figure 1:4), suggestive of an estimated divergence time of roughly 200,000 years (LOFTUS *et al.* 1994). This genetic data was used along with archaeological evidence to propose two independent domestications of cattle from two morphometrically and genetically distinct ancestral populations (LOFTUS *et al.* 1994). Subsequent studies using Y-chromosome, microsatellite and whole genome SNP data have further confirmed this suggestion (BRADLEY *et al.* 1994; 1996; MACHUGH *et al.* 1997; TROY *et al.* 2001; FREEMAN *et al.* 2006a; BOVINE HAPMAP CONSORTIUM *et al.*



2009). Recent studies of whole mitochondrial genomes from both modern and ancient cattle suggest that the present-day bovid mtDNA phylogeny represents a restricted sampling by domestication of a more diverse wild ancestral population (TROY *et al.* 2001; BOLLONGINO *et al.* 2006; EDWARDS *et al.* 2007b; ACHILLI *et al.* 2008; 2009; BONFIGLIO *et al.* 2010).



**Figure 1:4 Neighbour joining tree showing the unrooted bovine mitochondrial phylogeny, the major taurine (Q, T, P and R) and indicine (I1 and I2) haploroups are highlighted. The predominant haplogroups in each population are also marked with ovals green (T and Q) for taurine animals and red for indicus (I). Haplogroup P predominates in the extinct European aruochs population and has only been discovered in a single extant individual. Whereas haplogroup R has been discovered in a small number of modern Italian animals. Note the broomstick phylogeny with deep branch lengths proposing geological timescales divergences between populations.**

## ***1.5 Aims of this thesis***

The general aim of this thesis was one of the technological development to allow the efficient next generation sequencing of ancient samples. As described above these improvements are greatly needed as aDNA enters the genomics age, to save both time and money and to increase the quantity of data available for analysis. Some specific aims of this thesis are described below.

Firstly improvements to a standard aDNA extraction protocol (YANG *et al.* 1998; EDWARDS *et al.* 2010) and the Illumina library preparation protocol were implement as described in the introduction above to improve their use for next generation sequencing.

Secondly, the application of RNA-capture technologies to multiple samples targeting multiple genomic regions. These techniques have the potential to greatly improve aDNA analysis by allowing the targeting of specific DNA molecules out of the samples metagenome. This technique was firstly applied to mtDNA data as described in the bioinformatics analysis in chapter 3. With the aim to show that using these methods the production of whole mtDNA genomes from ancient samples is possible. This would allow for the increased resolution of analysis of aDNA mtDNA genomes away from the 160-400bp of mitochondrial control region previously targeted.

This method is also applied to targeted regions of the nuclear genome in an exon capture analysis of multiple bovid samples in chapter 5. This chapter

aims to demonstrate the utility nuclear RNA capture for aDNA in general and then for samples of varying preservation states.

Chapter 4 aims to show the kinds of population genetic analyses that can be applied to samples with excellent DNA preservation. This chapter applies methods that have only been previously used on ancient hominids to elucidate the genetic influence of the wild European aurochs on the current domestic European cattle populations.

Chapter 7 provides an in depth analysis of a new substrate for NGS analysis, historical parchments. This chapter tries to answer some of the objections raised to this substrate in previous PCR work and assess parchment as a source of aDNA in relation to bone. It also assess whether 1x genotypes produced for this analysis can be used to locate the parchment samples within modern day sheep populations.

Chapters 5 and 8 provide further implementations of the analysis of 1x genotypes from ancient samples. These newly published low coverage analysis techniques have the opportunity to gather important information from samples from which whole genomes could not be produced cost effectively. This data is then used to answer important questions about the affinities of these samples to modern day populations.



In summary all the case studies presented in this thesis are linked by their desire to improve aDNA methods and to allow for the maximum amount of data to be retrieved from most of the samples submitted for aDNA analysis. These samples are finite sources of information, due to the destructive sampling techniques inherent in aDNA analysis. Advances in the study of these remains will have huge benefits for sample curators and their collections as well as geneticists.

## 2. General methods

This section of the thesis describes general methods, which are used in the following research chapters (chapters 3-8), modifications where necessary from these protocols are described in each of the following chapters.

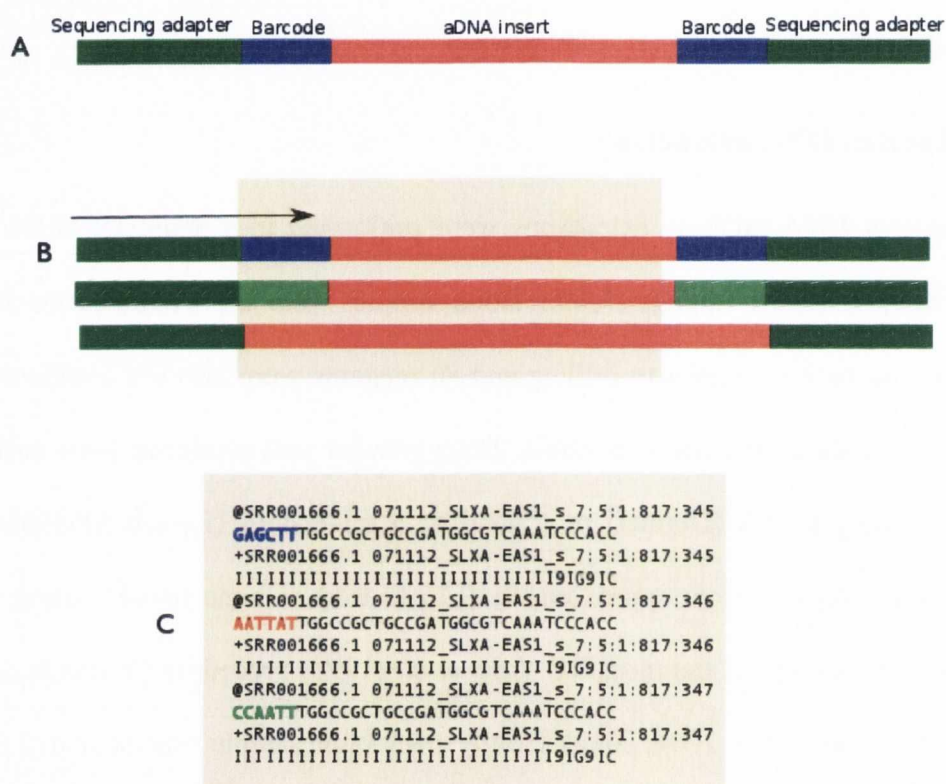
### 2.1 Ancient DNA extraction

All ancient DNA (aDNA) extractions were performed in a dedicated aDNA laboratory at Trinity College Dublin. Bone samples were UV irradiated for 30 minutes on each side prior to drilling and all reagents were also UV irradiated for 30 minutes before use if possible. Bone powder was produced from each sample using a drill (Dermel) or ground in a mixer-mill (Retsch MM301). DNA was prepared from approximately 0.3g to 0.5g of bone powder using a small-scale version of the modified Yang *et al.* (1998) procedure (YANG *et al.* 1998; EDWARDS *et al.* 2004; 2010). Briefly, each bone sample was incubated in extraction buffer (200 $\mu$ M proteinase K, 0.5M EDTA) for 48 hours (55°C for 24 hours and then 37°C for 24 hours). Concentrated to 100 $\mu$ l on a 30kD Amicon filter with two 3ml washes of 10mM Tris EDTA and cleaned of any salt contaminants using a Quiagen min-elute column, following manufacturers instructions apart from the fact that the final elution volume was 35 $\mu$ l.

### 2.2 Illumina sequencing library preparation

Illumina single-end sequencing libraries were produced from samples via the PCR enrichment of end repaired, adaptor-ligated DNA templates, following the

library construction methods of Edwards *et al* (2010). All samples except CPC98 were indexed using the methods of Craig *et al* (2008). This approach to sample barcoding requires the addition of specially designed sequencing adapters that place a 6bp barcode at the beginning of the sequencing reads from each sample (Figure 2:1).



**Figure 2:1 Cartoon of Craig *et al* (2008) sample barcoding scheme. Each aDNA insert is ligated to sequencing adapters (dark green) containing 6bp barcodes (blue, orange and light green) [A], samples can then be pooled and sequenced together [B] and the resultant sequencing reads from each sample [C] are identified by the 6bp barcode at the start of the read.**

### 2.3 Sequencing library enrichment

NGS Library enrichment prior to sequencing was completed using three 20µl PCR amplifications comprising 3µl of end-repaired-adaptor-ligated DNA, 10µl Phusion® high-fidelity PCR master mix with HF buffer (NEB, catalogue no. M0530S) 6.2µl ddH<sub>2</sub>O and 0.4µl each of both the forward and reverse primers (Illumina, catalogue no. FC-102-1003). Amplification reactions per library



consisted of an initial denaturation step of 98°C for 30s, then 12 cycles of 98°C for 10s, 65°C for 30s and 72°C for 30s, followed by a final extension step of 72°C for 5 min.

The products from each PCR reaction (post min-elute clean-up) were pooled (3 per library) and visualized/quantified on an Agilent 2100 Bioanalyzer using a DNA 1000 chip.

## **2.4 Initial sequencing quality control**

Initial quality control of sequencing reads from all case studies were preformed using fastQC at default settings, with the program in GUI mode, to aid in the choice of options for downstream alignment programs.

Reads were then separated from the sequencing data pool by their barcodes using a custom Python script (Appendix 2). This script was designed to search the first 6bp of each read (the positions in the read which encode the Craig *et al* (2008) barcode) for an exact match to a known barcode from one of the samples sequenced. Reads that failed to exactly match any of the barcodes used were then discarded.

Due to the fragmented nature of aDNA (PRUVOST *et al.* 2007), aDNA library inserts can be shorter than the total sequencing read length, for an Illumina Genome Analyser II typically 40-100bp, subsequently a large number of reads will therefore contain adapter sequences at the 3' end of the read. Prior to

downstream analysis these sequences were identified and trimmed using CUTADAPT version 1.0 (MARTIN 2011). CUTADAPT's default settings were modified to require only a 1bp overlap of an adapter sequence into the 3' end of a sequencing read for it to be identified and trimmed. This highly conservative approach will lead to a high proportion of reads being trimmed due to spurious matches. However, this is balanced against subsequent adapter sequence poisoning of downstream analysis, which can lead to poor alignment of reads and the miss-identification of sequence polymorphisms. The reads post trimming were then filtered to a minimum length of 24bp (6bp barcode plus 18bp sequence).

## **2.5 BWA sequence Alignment**

All chapters in this thesis required the alignment of next generation sequencing reads to a reference genome using BWA and these alignments were completed using the following pipeline outline below.

Firstly, a BWA index of the appropriate reference genome to which the reads are then aligned was created using the BWA index command. The `aln` command of BWA was then used to find the suffix array coordinates for the sequencing reads within the indexed genome, with the `-q` flag set to a value of 20 to soft clip the 3' end of reads to an average minimum base quality of 20. This alignment produced SAI (Suffix Array Index) files for all the samples, which were then converted to the human readable Sequence Alignment/Map SAM format using the `samse` command of BWA with default settings (Li *et al.* 2009).

Aligned reads were then further filtered using SAMTOOLS (version 0.1.18) for a mapping quality of 30 and redundant clonal PCR amplified products removed using the SAMTOOLS rmdup command (Li *et al.* 2009; Li 2011). Uniquely aligned reads were then identified using the XT and X1 tags, produced by the BWA alignment. The XT tag refers to the type of match against the reference genome found for each read “Unique/Repeat/N/Mate-sw” (<http://bio-bwa.sourceforge.net/bwa.shtml>). However, these assignments refer to optimal hits only, therefore the X1 tag “Number of suboptimal hits found by BWA” (<http://bio-bwa.sourceforge.net/bwa.shtml>) also needs to be considered to identify and remove all the reads marked by XT as unique but which are non-unique when these suboptimal hits are considered.

## 2.6 Haploid SNP calling

For chapters in which nuclear haploid genotype data was analysed (chapters 6,7 and 8) these SNPs were called following established protocols for 1x aDNA NGS data (SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012). Briefly, SNPs were called for all positions in which the shotgun sequencing of the sample overlapped with the positions of SNPs in the reference SNP datasets, requiring a minimum base quality of 20 and mapping quality of 30. As in Sanchez-Quinto *et al* (2012) the first and last 5bp of reads were classed as being “at risk” for DNA damage and so were masked to the lowest base quality, which removed these regions from the SNP ascertainment pipeline. If multiple reads overlapped a SNP position, one read was then taken at random and used for base calling. SNPs for which there was a G in the reference sequence but an A in the sequencing reads were removed along with reads that



had a C in the reference genome but a T in the sequencing reads. Multi-allelic SNPs were also discarded.

## **2.7 PLINK analysis**

All SNP data management, removal of loci and animals was completed using PLINK v1.07 (<http://pngu.mgh.harvard.edu/~purcell/plink>) with flags applied to select the correct species for example ‘-cow’ or ‘-sheep’. This option allows PLINK to recognise the extra chromosomes present in the bovine and ovine genome as compared to humans

## **2.8 Eigensoft Analysis 1x haplotype data**

To correct for the limitation that only haploid SNP calls for all ancient samples in chapters reporting 1x-sequencing data (chapters 6,7,8), one allele from each of the reference genotypes was then selected at random from each modern individual. All selected bases were then doubled (both modern and ancient) to produce homozygous genotype data (SÁNCHEZ-QUINTO *et al.* 2012) and principal component analysis completed using the SMARTPCA package of EIGENSOFT 4.2 (PATTERSON *et al.* 2006). The parameter file of SMARTPCA was also modified to recognise the differing numbers of autosomes found in the different domestic species. All other parameters were left at their defaults. All pairs of SNPs with an  $r^2 > 0.2$  had one of the pair removed.

## 2.9 Procrustes analysis

One of the central problems when studying ancient populations is that for most individuals only a limited percentage of the genome will be retrieved (typically less than 5% in good samples) (GREEN *et al.* 2010; SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012). This means that for any given dataset only a small subset of the genetic markers typed in the ancient samples will overlap between all individuals, the Procrustes method used by Skoglund *et al* (2012) is designed to counteract this inherent problem and relies on the following assumption that,

*“the genealogical history of each genetic locus—although non-overlapping—represents a random outcome of the same ancestral process, the relationship between non-overlapping data from different individuals from the same population (i.e. individuals with similar ancestry) and individuals from other populations is largely independent of which loci are used.”* (SKOGLUND *et al.* 2012)

This assumption allows for the direct comparison of ancient samples by combining the result of their independent analyses to densely SNP typed modern populations. In this thesis I followed the methods described by Skoglund *et al* (2012) by transforming the results of the PCA analysis (PC1 and PC2) for each ancient sample to match the configuration found when only the modern reference populations are considered (ancient individuals were given the coordinates (0,0)). Taking the mean of the PC1 and PC2 values obtained in each separate analysis for the reference individuals, but maintained

the PC1 and PC2 configuration for the ancient samples (SKOGLUND *et al.* 2012). This method also requires that the ancient individual under study should not too heavily influence the loading of a particular PCA, so that the reference individuals PC1 and PC2 loadings between analyses can be considered as the sampling variance of the number of SNPs used (SKOGLUND *et al.* 2012). This analysis was completed as described in chapters 6,7 and 8

## **2.10 MAPDAMGE analysis**

MAPDAMGE is a PERL script that allows the damage profile of aligned aDNA reads to be assessed, all analysis were completed using standard settings (GINOLHAC *et al.* 2011).

## **2.11 Depth of Coverage Analysis**

Depth of coverage analyses for all alignments were produced using the DepthOfCoverage option from the Genome Analysis Toolkit (GATK) (MCKENNA *et al.* 2010) version 2.0, with default settings.



## **3. Bioinformatic analysis of domesticate mitochondrial aDNA**

### ***3.1 Introduction***

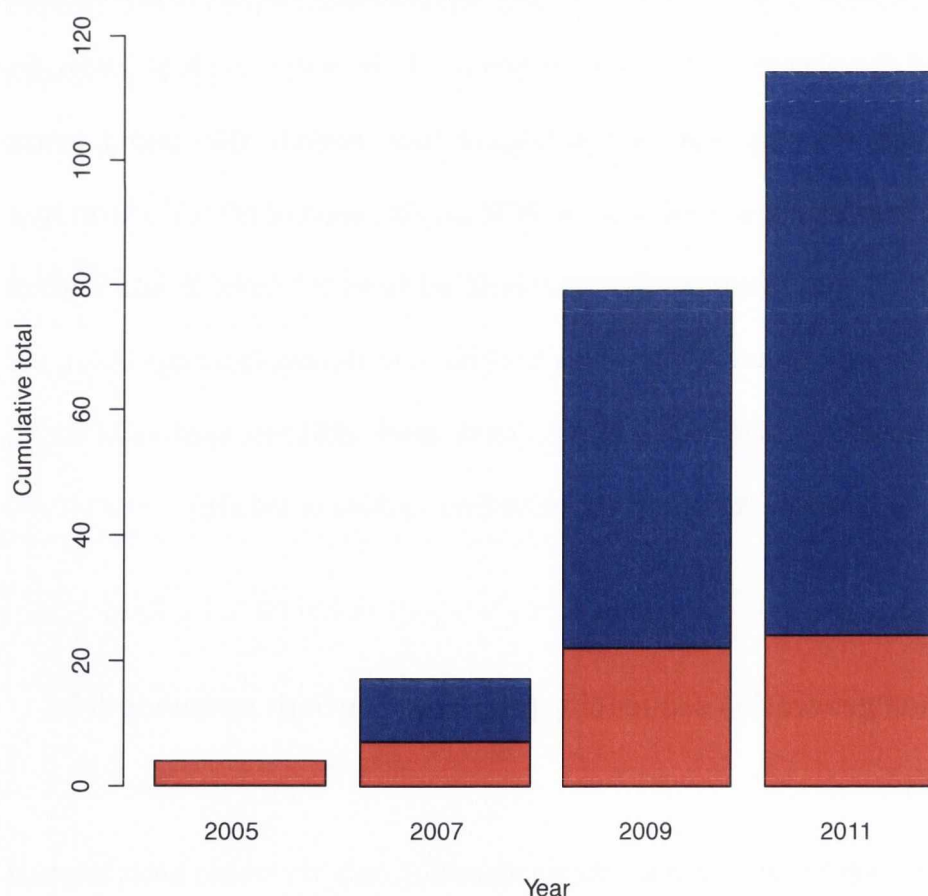
#### **3.1.1 The Mitochondria as a source of ancient DNA**

The mitochondrial genome has formed the basis for the vast majority of ancient DNA studies to date (HOFREITER *et al.* 2001; PAIJMANS *et al.* 2012), with the mitochondrion's high copy number in cells, 3,000 to 6,000 per diploid nuclear genome (MILLER *et al.* 2003), making it an attractive source for aDNA. Early experiments with ancient mtDNA concentrated on the hyper-variable region of the genome, where the mutation rate is fast enough to allow for genetic differentiation over archaeological time periods (HO and LARSON 2006), within a locus amenable to the PCR amplification of aDNA (HOFREITER *et al.* 2001). Successes in the combined analysis of modern and ancient mtDNA sequences have allowed for insights into the evolutionary history of multiple species (COOPER *et al.* 2001; TROY *et al.* 2001; BARNES *et al.* 2002; SHAPIRO *et al.* 2004; BARNES *et al.* 2007; EDWARDS *et al.* 2007b).

#### **3.1.2 Next generation sequencing and whole ancient mitochondrial genomes**

Prior to the advent of next generation sequencing only six whole mitochondrial genomes from ancient specimens had been published (GILBERT *et al.* 2007), four from Moa species (COOPER *et al.* 2001; HADDRATH and BAKER 2001) and two mammoth genomes (KRAUSE *et al.* 2005; ROGAEV *et al.* 2006). The

introduction of NGS methods into the field has seen a sharp increase in the number of mtDNA genomes generated (GILBERT *et al.* 2007; 2008; BRIGGS *et al.* 2009; HO and GILBERT 2010; EDWARDS *et al.* 2010; SCHUENEMANN *et al.* 2011; LALUEZA-FOX and GILBERT 2011; PAIJMANS *et al.* 2012), these studies have provided a wealth of new information and allowed for a greater depth of analyses to be undertaken (LALUEZA-FOX and GILBERT 2011; PAIJMANS *et al.* 2012) (Figure 3:1).



**Figure 3:1:** Bar graph showing the rapid increase in the number of ancient whole mtDNA genomes produced using both PCR (red) and NGS (blue) methods. The significant contribution of next generation sequencing to the total is notable. Data taken from (PAIJMANS *et al.* 2012).

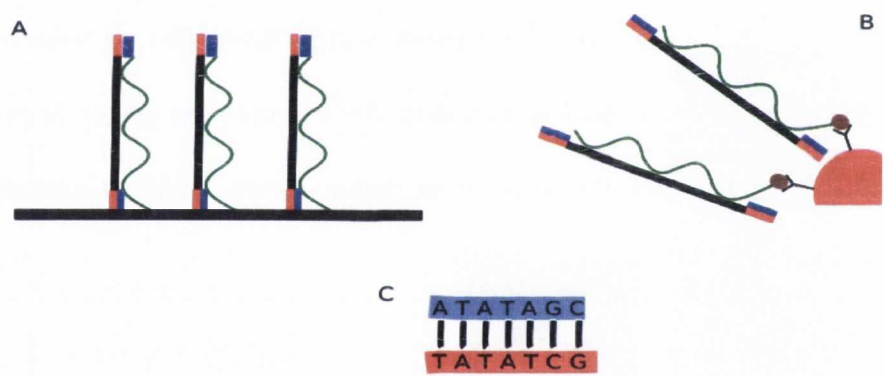
### 3.1.3 Sequence capture of ancient mitochondrial genomes

Although current NGS methods have the ability to sequence up to 600 Gb of DNA in a single 11 day run ([http://www.illumina.com/systems/hiseq\\_systems.ilmn](http://www.illumina.com/systems/hiseq_systems.ilmn)), the power of these techniques for the analysis of aDNA were shown by early studies to be severely hampered by the contamination of samples with exogenous DNA (mainly bacterial), after the animals' death (POINAR *et al.* 2006). This exogenous contamination poses a considerable challenge to the economic reality of sequencing ancient genomes, given that a typical aDNA sample contains on average only 0.1 to 5% endogenous DNA (KNAPP and HOFREITER 2010; ORLANDO *et al.* 2011; KIRSANOW and BURGER 2012). Some notable high-profile exceptions exist to this rule, for example the finger bone which provided the DNA for the Denisovan genome was found to contain 70% endogenous DNA (REICH *et al.* 2010).

This predominance of low endogenous DNA concentrations in archaeological samples has caused the field to look to novel methods, which may increase the amount of useable data in sequencing libraries. The most promising emerging method for aDNA targeting has been that of hybridisation capture and these solutions have taken two main forms. The first, Primer Extension Capture (PEC) relies on the hybridisation of biotinylated primers to regions of interest, these molecules are then made double stranded by a single round of PCR amplification, which allows for the capture of the final molecule with streptavidin coated magnetic beads (BRIGGS *et al.* 2009). The second, in



solution or on chip target enrichment requires the production of DNA or RNA probes usually of 100-120bp in length. These probes are then exposed to aDNA sequencing libraries and the resultant hybrids are purified either with streptavidin beads in the case of in-solution capture (Figure 3:2, panel B) or by the library molecules binding to the immobilised probes in chip based methods (Figure 3:2, panel A). Results from these experiments have lead to the capture of nuclear SNPs from the Neanderthal genome (BURBANO *et al.* 2010) as well as multiple mitochondrial genomes (BRIGGS *et al.* 2009).



**Figure 3:2** Cartoon of two methods of hybridisation capture are shown, one in solution using streptavidin beads [B], the other using bound probes on a chip (Thick black line) [A]. Green lines = RNA probes both immobilised to a chip [A] and biotinylated (brown circles) [B]. Solid black lines = targeted DNA with ligated sequencing adapters (red) and adapter sequence blocks (blue). Red oval streptavidin coated bead. [C] Complimentary base paring of sequence blocks (blue) and adapter sequences (red).

### **3.1.4 Ancient domesticate mitogenomics**

Ancient DNA has provided valuable insights into the domestication and evolution of farm animals (FERNÁNDEZ *et al.* 2006; EDWARDS *et al.* 2007b; LARSON *et al.* 2007), helping to infer centres of domestication (TROY *et al.* 2001; LARSON *et al.* 2007) and predicting the likely number of animals originally domesticated (BOLLONGINO *et al.* 2012). However, studies have centred on the relatively small D-loop region of animal mitochondrial genomes and their resolution could be increased if expanded to the whole chromosome.

### **3.1.5 Objectives for this case study**

Hybridisation capture has been shown to provide an exciting opportunity for the study of ancient genomes (BRIGGS *et al.* 2009; SCHUENEMANN *et al.* 2011). This chapter presents the results of the bioinformatic analysis from two-independent aDNA mitochondrial genome capture experiments using aDNA samples from three domesticate species and the RNA capture technology provided by Agilent Technologies (NG *et al.* 2009). The efficiency and utility of this capture method for use in the targeting of domesticate mitochondrial genomes from archaeological samples of varying time depths is assessed, along with the utility of mtDNA capture as screening tool for further nuclear analysis. The completed mtDNA genomes produced from these studies are then used within a Bayesian phylogenetic framework to try to provide a molecular dating for the divergences of the major haplogroups of sheep and cow and to relate these divergences to the archaeological dates of domestication in these species. This comparison is then related to the probable domestication processes in these animals.

## **3.2 Materials and Methods**

### **3.2.1 Archaeological Samples**

This chapter presents the results of a targeted mtDNA capture analysis of multiple mitochondrial genomes from 3 domestic species shown in table 3:1 and table 3:2. All bovid samples have previously been studied using PCR based methods ovine/caprine had not.

### **3.2.2 Ancient DNA libraries preparation and capture**

All bone preparations, DNA extractions and pre-PCR NGS library making procedures were completed in a dedicated ancient DNA laboratory by Frauke Stock and Anna Linderholm (Stock F, PhD Thesis, Trinity College Dublin Library). Libraries from both sets of samples were made following the standard Illumina protocol using the NEB Next<sup>TM</sup> kit with modifications for ancient DNA (Stock F, PhD Thesis, Trinity College Dublin Library), with Craig *et al* (2008) barcodes (Chapter 2). Mitochondrial DNA was then captured in solution using the Agilent Technologies SureSelect protocol, also by Frauke Stock and Anna Linderholm (Stock F, PhD Thesis, Trinity College Dublin Library), following manufacturers-instructions, except that half reaction volumes were used.



| Sample         | Site                       | Dating                                 | Species                | References for dating and site information |
|----------------|----------------------------|----------------------------------------|------------------------|--------------------------------------------|
| <b>Ard1</b>    | Ardeche, France            | 8,630 YBP<br>(Direct carbon date bone) | <i>Bos taurus</i>      | (EDWARDS <i>et al.</i> 2004)               |
| <b>Ard2</b>    | Ardeche, France            | 8,640 YBP<br>(Direct carbon date bone) | <i>Bos taurus</i>      | (EDWARDS <i>et al.</i> 2004)               |
| <b>Turk1</b>   | Aktopraklik, Turkey        | 9,000 YBP<br>(Dated by context)        | <i>Bos taurus</i>      | (STOCK <i>et al.</i> 2009)                 |
| <b>Turk11</b>  | Mentese, Turkey            | 8,000 YBP<br>(Dated by context)        | <i>Bos taurus</i>      | (GOURICHON and HELMER 2008)                |
| <b>Turk13</b>  | Mentese, Turkey            | 8,000 YBP<br>(Dated by context)        | <i>Bos taurus</i>      | (GOURICHON and HELMER 2008)                |
| <b>UKR07</b>   | Starobels, eastern Ukraine | 7,000 YBP<br>(Dated by context)        | <i>Bos primigenius</i> | (STOCK <i>et al.</i> 2009)                 |
| <b>UKR08</b>   | Starobels, eastern Ukraine | 7,000 YBP<br>(Dated by context)        | <i>Bos primigenius</i> | (STOCK <i>et al.</i> 2009)                 |
| <b>Bercy10</b> | Bercy, France              | 6,000 YBP<br>(Dated by context)        | <i>Bos taurus</i>      | (EDWARDS <i>et al.</i> 2004)               |
| <b>Bercy13</b> | Bercy, France              | 6,000 YBP<br>(Dated by context)        | <i>Bos primigenius</i> | (EDWARDS <i>et al.</i> 2004)               |
| <b>Ron1</b>    | Bouqras, Syria             | 9,000 YBP<br>(Dated by context)        | <i>Bos taurus</i>      | Pinhasi R, Personal communication          |
| <b>Ron5</b>    | Bouqras, Syria             | 9,000 YBP<br>(Dated by context)        | <i>Bos taurus</i>      | Pinhasi R, Personal communication          |
| <b>Afr5d</b>   | West Africa                | 4090 YBP (Direct carbon date bone)     | <i>Bos taurus</i>      | (EDWARDS <i>et al.</i> 2004)               |

3:1 Bovid samples used in this analysis

| Sample         | Site                   | Dating                                             | Species                                    | References for dating and site information |
|----------------|------------------------|----------------------------------------------------|--------------------------------------------|--------------------------------------------|
| <b>Kud3_8</b>  | South Caucasus, Russia | 43,900 YBP<br>(Dated by context)                   | <i>Capra hircus</i>                        | Pinhasi R, Personal communication          |
| <b>Kud1_22</b> | South Caucasus, Russia | No date information                                | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication          |
| <b>AK2</b>     | Crimea, Ukraine        | No date information                                | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication          |
| <b>AKT28</b>   | Turkmenistan           | Byzantine (c 330 – 1204 YBP)<br>(Dated by context) | <i>Ovis aries</i>                          | Pinhasi R, Personal communication          |
| <b>Kud3_7</b>  | South Caucasus, Russia | 43,900 YBP<br>(Dated by context)                   | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication          |
| <b>Kud1_21</b> | South Caucasus, Russia | 43,900 YBP<br>(Dated by context)                   | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication          |
| <b>Kud3_16</b> | South Caucasus, Russia | 43,900 BP<br>(Dated by context)                    | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication          |
| <b>Kud3_9</b>  | South Caucasus, Russia | 43,900 YBP<br>(Dated by context)                   | <i>Capra hircus</i>                        | Pinhasi R, Personal communication          |
| <b>Kud1_20</b> | South Caucasus, Russia | 80,000-360,000 YBP<br>(Dated by context)           | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication          |



| <b>Crim</b>       | <b>Crimea, Ukraine</b> | <b>2,300-2,200 YBP<br/>(Dated by context)</b>      | <b><i>Capra hircus</i></b>                 | <b>Pinhasi R, Personal communication</b> |
|-------------------|------------------------|----------------------------------------------------|--------------------------------------------|------------------------------------------|
| <b>AKT 29</b>     | Turkmenistan           | Byzantine (c 330 – 1204 YBP)<br>(Dated by context) | <i>Ovis aries</i>                          | Pinhasi R, Personal communication        |
| <b>Lad 24</b>     | St Petersburg          | 2,500 YBP<br>(Dated by context)                    | <i>Ovis aries</i>                          | Pinhasi R, Personal communication        |
| <b>KK 27</b>      | Rostov, Russia         | 2,500 YBP<br>(Dated by context)                    | <i>Ovis aries</i>                          | Pinhasi R, Personal communication        |
| <b>Gubs 13</b>    | North Caucasus, Russia | 33,500 YBP<br>(Dated by context)                   | <i>Ovis aries</i>                          | Pinhasi R, Personal communication        |
| <b>Gubs 25</b>    | North Caucasus, Russia | 33,500 YBP<br>(Dated by context)                   | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication        |
| <b>Akt'07 89E</b> | Turkey                 | 8,400 YBP<br>(Dated by context)                    | <i>Ovis aries</i> /<br><i>Capra hircus</i> | Pinhasi R, Personal communication        |

Table 3:2 Ovine and caprine samples used in this analysis



### **3.2.3 Initial QC and read trimming**

The bovine mtDNA sequencing data consisted of two sequencing runs. The first, performed at the Conway Institute, University College Dublin, used 42 cycles and the second completed at TrinSeq, Institute of Molecular Medicine, Trinity College Dublin used 80 cycles. All trimming and alignment steps were completed as described in the general methods section of this thesis. Using the following reference genomes, cow (V00654, downloaded 16-5-12), sheep (N\_001941, downloaded 16-5-12) or goat (NC\_005044, downloaded 16-5-12). One modification was made to the pipeline, the indexed reference genome was produced via the BWA index command with the -a flag set to 'is' as the mtDNA reference genomes are short.

### **3.2.4 Identification of jumping PCR**

As all sequencing libraries (run1 and run2 for cattle, goat and sheep) were subject to multiple rounds of PCR (30 Cycles) from a low starting concentration after being pooled and captured, the resultant libraries are candidates for jumping PCR. This phenomenon results in chimeric PCR products produced by the recombination of amplified molecules in the PCR reactions, CUTADAPT's setting were modified to search for reads that may be the product of jumping PCR. In this case to look for an exact match of 8bp or greater of adapter sequence at the 3' ends of reads with a barcode that didn't match that used to make the library. Reads identified in this way were then filtered from subsequent analysis.

### **3.2.5 Read filtering and merging**

Once PCR duplicated reads were removed from each alignment following the standard methods (Chapter 2) the two BAM files for each sample (1 from each sample per sequencing run) were then merged using PICARD (version 1.73) (<http://picard.sourceforge.net/>).

### **3.2.6 SNP calling**

Three independent SNP calling algorithms were used in this analysis. The first was the SNP calling algorithm of SAMTOOLS mpileup, applied with default settings. The second was the Unified Genotyper from the Genome Analysis Tool Kit (GATK) (McKENNA *et al.* 2010) (version 2.0), using default options and assuming a haploid model. The final SNP calling algorithm was an in-house Python script that called SNPs if the read depth was greater than or equal to 3x and the number of non-reference alleles was greater than or equal to the reference. The results of the combined analysis of all three algorithms were then manually inspected using the genome viewer IGV-VIEWER version 2.1 (THORVALDSDOTTIR *et al.* 2012), before the final consensus sequence for each sample was produced.

### **3.2.7 Fasta sequence production**

FASTA files for each of the samples analysed were then produced using the vcfutils.pl Perl script from SAMTOOLS version 0.1.18.

### 3.2.8 MEGAN

MEGAN is a program which allows for the visualisation and the taxonomic assignment of BLAST results from NGS reads (HUSON *et al.* 2011). Both this program and the blastn analysis of sequencing reads against the NCBI non-redundant nucleotide database used to provide the input BLAST xml files for MEGAN (version 4) were completed using default settings.

### 3.2.9 Bayesian analysis of the estimated sequence divergence times

Representative modern sheep mitochondrial reference genomes were downloaded from GenBank to allow the placement of the ovine aDNA samples within this dataset (Appendix:A). Multiple sequence alignments of both the modern and ancient mtDNA genomes were then completed using version 3.8.31 of the Muscle alignment algorithm (EDGAR 2004), implemented in Seaview version 4.3.4 (GOUY *et al.* 2010). BEAST version 1.7.4 (DRUMMOND *et al.* 2012) was used to provide a phylogenetic assessment of the alignments under the following conditions: HKY+G+I substitution model, a demographic model of exponential growth as one of the coalescence priors and relaxed molecular clock applied to allow rate variation between lineages. The Markov chain Monte Carlo analysis included 80,000,000 steps, samples were drawn every 1,000, generating 80,000 trees from which a maximum clade credibility tree was generated with a burn-in of 10,000 steps (Ho *et al.* 2008).



| <b>Sample</b> | <b>Total Reads</b> | <b>Aligned Reads (%)</b> | <b>Mean Coverage (x)</b> | <b>Coverage(&gt;=1x,%)</b> |
|---------------|--------------------|--------------------------|--------------------------|----------------------------|
| Ard1          | 1,468,505          | 2,445 (0.17)             | 6                        | 99                         |
| Ard2          | 3,433,309          | 35,978 (1.05)            | 118                      | 100                        |
| Turk1         | 2,391,580          | 86 (0.003)               | 0.22                     | 18                         |
| Turk11        | 3,226,760          | 17,513 (0.54)            | 44                       | 100                        |
| Turk13        | 2,259,127          | 7,844 (0.35)             | 22                       | 100                        |
| UKR07         | 3,483,983          | 510 (0.01)               | 1                        | 75                         |
| UKR08         | 1,379,055          | 1,876 (0.14)             | 6                        | 98                         |
| Bercy10       | 3,634,574          | 2,228 (0.06)             | 5                        | 98                         |
| Bercy13       | 2,914,424          | 2,937 (0.10)             | 9                        | 99                         |
| Ron1          | 2,609,469          | 8 (0.0003)               | 0.02                     | 2                          |
| Ron5          | 3,403,546          | 54 (0.001)               | 0.17                     | 11                         |
| Afr5d         | 368,834            | 20 (0.005)               | 0.07                     | 5                          |

**Table 3:3 Results of bovid mtDNA alignments, Aligned reads = percentage of reads aligned to the reference genome. Mean coverage = mean ‘x’ coverage of the mtDNA genome as calculated by GATK DepthOfCoverage command, Coverage(>=1x,%) = the percentage of bases called in the mtDNA reference genome from each sample greater than or equal to 1x.**

### **3.3 Results**

The results of the mtDNA capture experiment will be presented in two parts firstly the result of the cattle analysis, then the sheep/goat.

#### **3.3.1 Bovid mtDNA sequencing results**

The bovid capture analysis provided sufficient coverage to call three complete 1x mitochondrial genomes from the archaeological samples studied (Ardeche2, Turk11 and Turk13). A further four almost complete mtDNA genomes were also produced, two at 99% (Bercy13 and Ardeche1) and two at 98% (UKR08, Bercy10), the remaining 5 samples varied in sequence coverage from 2-75%. These result highlight the power of this RNA capture technique to effectively retrieve mtDNA genomes with over half of the captured samples producing complete or near complete 1x genomes, using only two lanes of an Illumina GA II sequencer.

The haplotypes assigned to the samples falls within the major bovine haplogroups four in the major European bovine haplogroup T (Ardeche1, Ardeche2, Turk11 and Bercy10) the other three samples Turk13, Bercy13 and UKR08 have belong to more divergent haplogroups with only one Turk13 (Q) found in present day cattle populations suggesting that the other two samples (Bercy13 and UKR08) belong to none domestic aurochs populations. This is of particular interest for Bercy13 as it was found in a Neolithic site, which suggests it was hunted from the wild.

### 3.3.2 Estimated contamination rate bovid capture

The stringency of the alignment was then raised to a 3x coverage over 85% of the genome for SNP calling. The SNPs produced from this analysis were used to provide a rough estimation of the contamination rate in these samples; to do this heterogeneities at SNP locations that could be contamination, sequencing error, hetroplasmmy or damage were analyzed. The most probable cause of hetrozygosities is likely to be sequencing error and not contamination from exogenous sequences and therefore these results may well represent an overestimate of the contamination rates. However, this method has been used by other groups to give a rough estimate of contamination (SÁNCHEZ-QUINTO *et al.* 2012) and so was used also in this analysis. The results of these contamination estimates are generally encouraging (Table 3:4) and are comparable to similar study in humans which estimated a rate of 8% (SÁNCHEZ-QUINTO *et al.* 2012). Interestingly, the two samples with the highest contamination rates were from a single archaeological site in France (8.3% Bercy10 and 9% Bercy13).



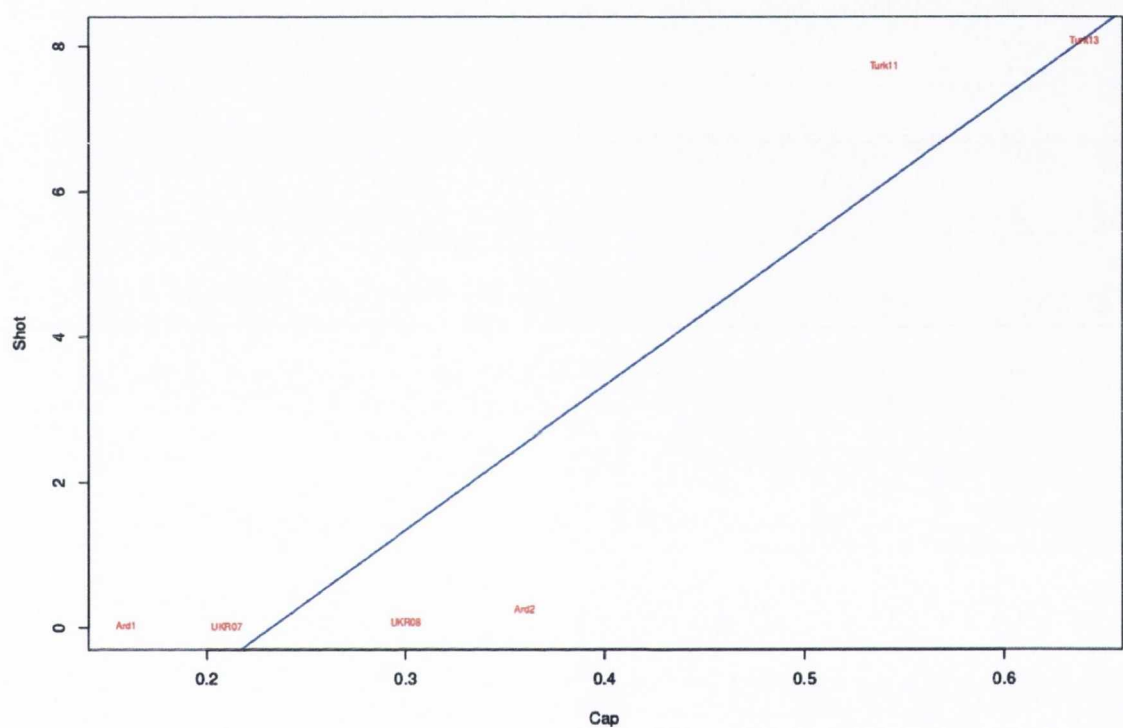
| Sample   | Total consensus base count | Total bases called | Percentage contamination | Percentage coverage 3x | mtDNA haplogroup assigned |
|----------|----------------------------|--------------------|--------------------------|------------------------|---------------------------|
| Ardeche1 | 32                         | 32                 | 0                        | 90                     | T                         |
| Ardeche2 | 642                        | 646                | 0.6                      | 100                    | T                         |
| Bercy10  | 33                         | 36                 | 8.3                      | 85                     | T                         |
| Bercy13  | 426                        | 468                | 9.0                      | 95                     | P                         |
| Turk11   | 254                        | 257                | 1.2                      | 100                    | T                         |
| Turk13   | 582                        | 590                | 1.4                      | 99                     | Q                         |
| UKR08    | 461                        | 474                | 2.7                      | 90                     | E                         |

**Table 3:4** Estimated contamination rate for bovid samples >2x genotypes. Note: Haplogroup T = predominate *Bos taurus* haplogroup, P = predominant *Bos primigenious* haplogroup, haplogroup Q = minor *Bos taurus* haplogroup previously only reported in Italy, haplogroup E low frequency *Bos primigenious* haplogroup.

### 3.3.3 Off target sequence as an indicator of shotgun success

In DNA/RNA capture experiments there is usually a considerable amount of sequence data retrieved in that is not included in the regions under study. The possibility that this off target sequence data could be used to give insights into the potential success of a whole genome shotgun sequencing analysis from mtDNA captured samples was investigated. To do this, the percentage of reads which aligned off their mtDNA target but to the bovine nuclear genome was calculated for seven samples (Figure 3:3) and compared to their results from a shotgun sequencing analysis. This analyses (Figure 3:3) shows that there is a trend, which may be informative for shotgun success. The bone samples that showed relatively high levels of off target nuclear DNA (Turk11 and Turk13) in the mtDNA capture analysis (Cap) also showed high levels of percentage nuclear genome alignment in the shotgun analysis (Shot) (Figure 3:3). An interesting aside to this analysis is that of Ardeche2, though it was the best

performing sample for mtDNA capture, producing a complete ancient mitochondrial genome at a mean coverage of 118x; it only showed a moderate amount of off target nuclear DNA in the mtDNA capture and very little in the shotgun analysis. This disparity between nuclear and mtDNA retrieval rates has been reported previously Allentoft *et al* (2012).



**Figure 3:3 Nuclear genome alignment percentage RNA mtDNA Capture (Cap) vs. whole genome shotgun (Shot), regression line for illustrative purposes only.**

### **3.3.4 Comparison of the captured and un-captured metagenome**

It could be hypothesised that the yield of targeted capture techniques for ancient mtDNA genomes may be affected by the close homology of rRNA and tRNA genes between species. To investigate this effect we compared the results of a metagenomic analyses preformed on reads generated by both RNA capture and whole genome shotgun sequencing from the best sample of those tested above, Turk13 (Figure 3:4). The results of this analysis highlight two interesting effects, firstly the number of reads that can be assigned to the correct kingdom is much greater in the capture analysis (blue) compared to the shotgun sequencing (red), which has a greater number of reads with no hits. Secondly the diversity of sequences retrieved in the capture analysis is also greater, with sequences from six kingdoms identified in the capture reads compared to the three within those from the shotgun approach. The results of this analysis suggest that the mtDNA capture may be causing a concentration of identifiable reads, possible due to the conserved nature of many of mtDNA sequence elements (rRNA, tRNA), which will therefore tend to find more hits to the database.



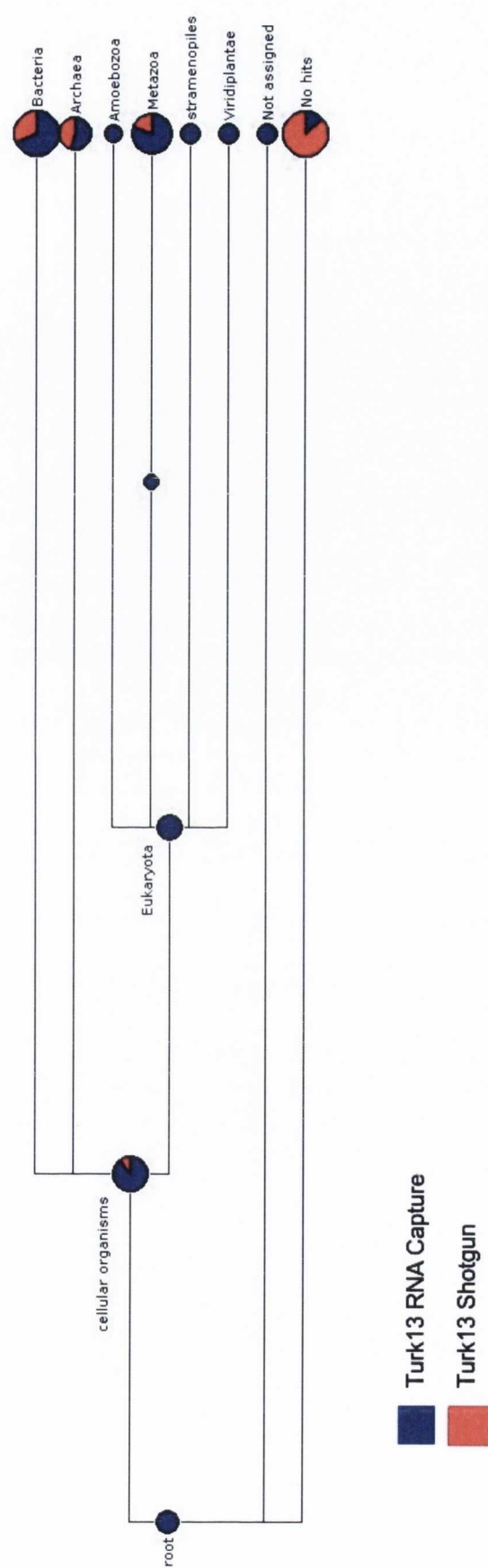


Figure 3:4 MEGAN analysis (blastn) 100,000 reads Turk13 RNA captured and un-captured (shotgun). Note the greater amount of reads assigned to the bacterial clade and other kingdoms in the RNA capture analysis and a corresponding increase in no hits in the un-capture analysis

### 3.3.5 Haplotype assignment of cattle samples

Of the 7 high coverage bovid samples studied in this analysis 4 were found to belong to the predominant domestic taurine cattle mitochondrial haplogroup T (Ardeche1, Ardeche2, Turk11 and Bercy10), reflecting their archaeological assignment as domestic *Bos taurus* cattle (Table 3:1). One sample (Turk13) was found to have a more divergent haplotype Q, which has thus far only been discovered in low frequency in domestic Italian *Bos taurus* populations (ACHILLI *et al.* 2008). The final two samples were found to have haplotypes representative of Eurasian wild ox (*Bos primigenius*): one the rare E haplotype (UKR08) and one with the P haplotype common in western European wild oxen (Bercy13).

### 3.3.6 Bayesian phylogeny assignment cattle samples

Time stamped aDNA samples when included into phylogenetic analyses can be used to calibrate mutation rates and give an estimate of branch time depth. Five of the most complete bovine mtDNA genomes produced above were used in such an analysis implemented by the molecular phylogenetic program BEAST (Frauke Stock, PhD Thesis, Trinity College Dublin), a summary of this work is presented below and in figure 3:5.

The bovine BEAST analysis produced a timing for the divergence of the mtDNA genomes of the two major global cattle populations (taurine/indicine) of 217,487 YBP (107,056 – 383,342 YBP) which is comparable to previously published estimates (335,000 YBP) (ACHILLI *et al.* 2009), however this

analysis is the first to use full ancient mtDNA genomes. The second oldest divergence is of a rare Italian R haplogroup and the rest of the taurine grouping, which was dated to 103,973 YBP (55,787 – 179,793 YBP). The third oldest branch separates the haplogroup P representative of the European wild ox (Chapter 1 and Chapter 3) from the rest of the domestic taurine cattle populations, this separation is dated to 53,461 YBP (29,930 - 91,352). The final separation is that of haplotype T the prominent domestic taurine haplogroup and Q a haplogroup found predominantly in domestic Italian cattle dated to 32,539 YBP (18,368 – 55,191). The major domestic taurine haplogroup T was also found to have an estimated most recent common ancestor at 17,872 YBP.

These datings as mentioned above were produced using time stamped samples, however only two of the samples were directly dated Ardeche 2 and Ardeche1. Bercy13, Turk11 and Turk13 are only dated by context and therefore we have no direct date of the sample of bone that was used to provide the sequencing data. This would cause problems for the analysis if the sample was an older or younger introgressing into that context. This issue is of specific interest to the divergence between the Q and T clade, as Turk13 is the only time stamped sample used to date this divergence. However, both Turkish samples come from an extensively studied and reliably dated site (GOURICHON and HELMER 2008) and are acceptable for use in the analysis. The single P haplotype sample Bercy13 was also not directly dated and therefore the timings of this clade could also be questioned if this sample was incorrectly dated. This sample however also comes from a well-studied and dated site in France



(EDWARDS *et al.* 2004). No further time stamped samples are available for the R or I divergences so these dates are based on the molecular evidence alone.

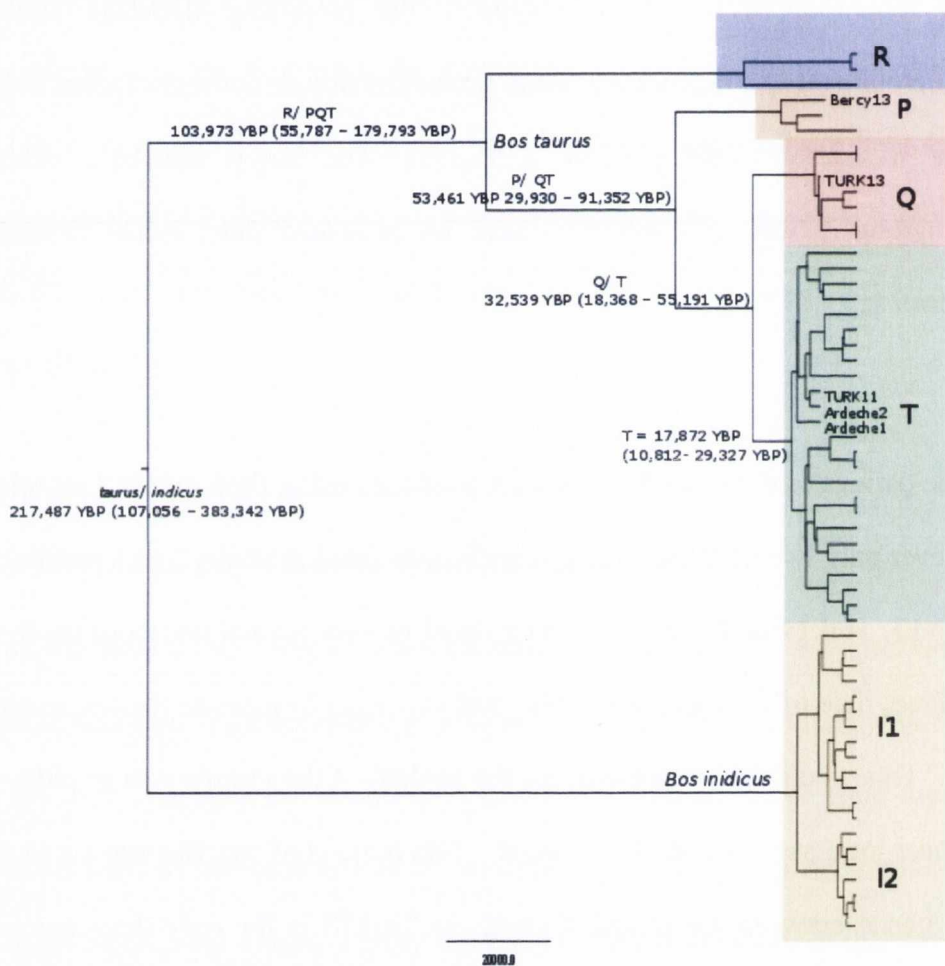


Figure 3:5 Bayesian estimate of maximum clade credibility tree using modern and ancient bovid mitochondrial genomes of taurine, indicine and aurochs origin. A HKY nucleotide substitution model was used. Dates are presented for branches with a posterior probability of  $\geq 0.9$ . Modern samples are represented by their haplogroup T etc. ancient samples are named. Modified with permission from Figure 5.6 Stock F, PhD Thesis, Trinity College Dublin Library.

### 3.3.7 Ovine/Caprine mtDNA sequencing results

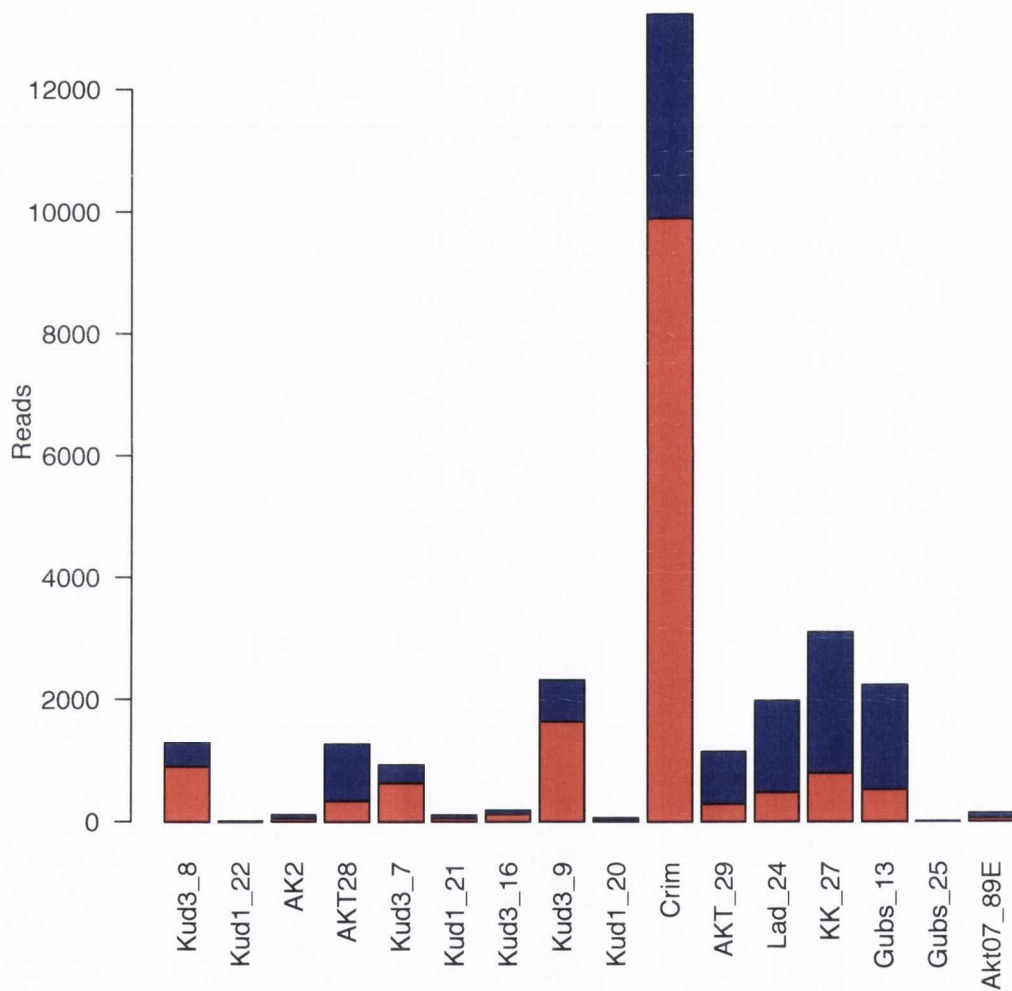
Given the well described difficulties with the taxonomic assignment of sheep and goat assemblages (GILLIS *et al.* 2011) and the lack of any prior PCR analysis of these samples, the first part of this goat/sheep study was therefore centred on trying to provide a species assignment for each sample (Figure 3:6, Table 3:5) To do this all reads from each animal were aligned to both the sheep and goat mtDNA reference genomes and samples assigned to a species using the following conditions. Firstly, for an animal to be called either a sheep or goat the sequencing reads have to cover greater than 80% of the ovine or caprine mtDNA genome. Secondly, for a sample to be assigned as sheep it had to cover 20% more of the sheep mtDNA genome than goat and *vice versa*. The results of this analysis allowed the assignment of 5 samples as sheep (AKT29, Lad24, KK27, Gubs13 and AKT28) and 3 as goat (Kud3\_8, Kud3\_9 and Crim) the other 7 samples had insufficient sequence coverage to allow for a specific species assignment.

| Sample  | Total reads | Aligned<br>reads<br>goat<br>(%) | Mean<br>coverage<br>goat | Coverage<br>(≥1x,%) goat | Aligned<br>reads<br>sheep (%) | Mean<br>coverage<br>sheep x | Coverage (≥<br>1x,%)<br>sheep | Species<br>Assignment |
|---------|-------------|---------------------------------|--------------------------|--------------------------|-------------------------------|-----------------------------|-------------------------------|-----------------------|
| Kud3_8  | 5,167,871   | 892 (0.01)                      | 3                        | 83                       | 394 (0.007)                   | 1                           | 43                            | Goat                  |
| Kud1_22 | 863,065     | 12 (0.001)                      | 0.03                     | 2                        | 9 (0.001)                     | 0.03                        | 2                             | ?                     |
| AK2     | 3,364,356   | 53 (0.001)                      | 0.18                     | 5                        | 67 (0.001)                    | 0.23                        | 10                            | ?                     |
| AKT28   | 1,138,814   | 343 (0.03)                      | 1                        | 40                       | 931 (0.08)                    | 3                           | 91                            | Sheep                 |
| Kud3_7  | 2,637,517   | 632 (0.02)                      | 2                        | 71                       | 303 (0.01)                    | 0.87                        | 33                            | ?                     |
| Kud1_21 | 2,716,735   | 59 (0.002)                      | 0.18                     | 7                        | 57 (0.002)                    | 0.18                        | 7                             | ?                     |
| Kud3_16 | 2,825,000   | 128 (0.004)                     | 0.41                     | 28                       | 67 (0.002)                    | 0.21                        | 12                            | ?                     |
| Kud3_9  | 3,792,383   | 1642 (0.04)                     | 5                        | 91                       | 681 (0.01)                    | 2                           | 48                            | Goat                  |



|           |           |              |       |     |              |      |     |       |
|-----------|-----------|--------------|-------|-----|--------------|------|-----|-------|
| Kud1_20   | 1,449,091 | 26 (0.001)   | 0.08  | 5   | 42 (0.002)   | 0.15 | 11  | ?     |
| Crim      | 2,788,574 | 9,888 (0.35) | 37    | 99  | 3,334 (0.11) | 11   | 72  | Goat  |
| AKT29     | 2,014,377 | 294 (0.01)   | 0.91  | 48  | 858 (0.04)   | 2    | 88  | Sheep |
| Lad24     | 4,604,225 | 483 (0.01)   | 1     | 56  | 1,504 (0.03) | 6    | 98  | Sheep |
| KK27      | 2,935,193 | 797 (0.02)   | 2     | 46  | 2,306 (0.07) | 8    | 100 | Sheep |
| Gubs13    | 1,413,557 | 530 (0.03)   | 1     | 46  | 1,711 (0.12) | 5    | 97  | Sheep |
| Gubs25    | 2,237,902 | 3 (0.0001)   | 0.008 | 0.8 | 9 (0.0004)   | 0.03 | 2   | ?     |
| Akt07_89E | 2,869,325 | 58 (0.002)   | 0.21  | 6   | 92 (0.003)   | 0.33 | 14  | ?     |

Table 3:5 Results of sheep and goat alignments.



**Figure 3:6** Number of reads from each sample that align to the goat (red)/ sheep (blue) genome. Note: because of the degraded nature of aDNA and the short read lengths used in this analysis reads can find sufficiently high alignments to be reported as hits to both the closely related sheep and goat mtDNA genome.

**3.3.8 SNP calling and analysis**

When the stringency of the alignment was then increased (3x  $\geq 85\%$  of the mtDNA genome) to allow for informative SNP typing, 4 samples were retained for further analysis one goat (Crim) and 3 sheep (Lad24 KK27 Gubs13). These samples were also subjected to an estimation of contamination as in the bovid samples above, the results of this analysis produced contaminant estimates for all samples tested under 2.5% (Table 3:6).

| Species | Sample | Genome coverage<br>3x % | Date<br>(YBP)   | Total consensus<br>base<br>count | Total<br>bases<br>called | Percentage<br>contamination |
|---------|--------|-------------------------|-----------------|----------------------------------|--------------------------|-----------------------------|
| Sheep   | Lad24  | 88                      | 2,500           | 542                              | 549                      | 1.3                         |
| Sheep   | KK27   | 97                      | 2,500           | 140                              | 140                      | 0                           |
| Sheep   | Gubs13 | 87                      | 33,500          | 409                              | 411                      | 0.48                        |
| Goat    | Crim   | 98                      | 2,300-<br>2,200 | 2,324                            | 2376                     | 2.2                         |

**Table 3:6 Estimated contamination rates of sheep and goat samples subjected to phylogenetic analysis**



### **3.3.9 Goat mtDNA phylogenetic Analysis**

Given that only one goat sample could be called at sufficient depth and genome coverage for a reliable phylogenetic analysis a BEAST analysis was not undertaken with this sample.

### **3.3.10 Sheep mtDNA haplotype assignment and phylogenetic analysis**

In a phylogenetic analysis of the sheep mtDNA genomes two samples were found to fall within the most populous European haplogroup B (Gubs13 and KK27) and one into the predominant Asian haplogroup A (Lad24) (Figure 3:7). These 3 sheep samples were then incorporated into two BEAST analyses to try to provide information on the timings of divergences of the major sheep haplogroups. The first analysis was completed using all the sheep samples and is shown in figure 3:7. This analysis estimates a coalescence time for the predominant European sheep haplogroup B of 81,811 YBP (46,086-132,560 YBP), and 15,713 YBP (6,178-27,277 YBP) for the major Asian haplogroup A. These two most widely represented sheep haplogroups are estimated to have diverged 95,758 YBP (51,609 – 159,228 YBP). With the most recent common ancestor (MRCA) of all sheep haplotypes estimated at 150,954 YBP (74,505 – 257,176 YBP).

When reviewing the timings presented above it was felt that Gubs13 might be biasing the results of this analysis, as only 87% of the mtDNA genome is covered in this sample and it is dated by context to 33,500 YBP. These two factors may mean that the calibrations produced by this sample are misleading

given its basal position in the B haplogroup. When Gub13 is removed from the analysis (Figure 3:8) the MRCA of haplogroup B is found to date to 5,219 YBP (4,039 -6,551 YBP), the MRCA of A is 3,680 YBP (2,955 – 4,527 YBP), A/B is 13,528 YBP (9,264-18,598 YBP); a date close to the proposed timing of sheep domestication (ZEDER 2008). All the sheep haplotypes converge at 19,638 YBP (13,052 – 27,329). This effect is particularly noticeable when the relationship between the European Mouflon and the domestic sheep haplogroup B is considered. The European Mouflon has been proposed by previous studies to be an early domestic sheep that has become feral (CHESSA *et al.* 2009; MEADOWS *et al.* 2010). With Gub13 included the two clades are predicted to coalesce at 20,612 YBP (9,241 – 36,037 YBP), with only the lowest range of the 95% highest posterior density (HPD) predicting a divergence time consistent with sheep domestication. However with Gub13 removed the time reduces to 4,812YBP (4,039 – 6,511 YBP), comfortably within the time frame of sheep domestication. This result makes more sense if we assume that the mouflon were a domestic sheep that re-adapted to wild life as this timing could indicate the date when the animals went feral.

However these results also suffer from the same inaccuracies as mentioned in the bovine analysis due to the samples only being dated by context and not directly radiocarbon dated bones. Though with Grub13 also being poorly covered inaccuracies in dating may be amplified in this sample.

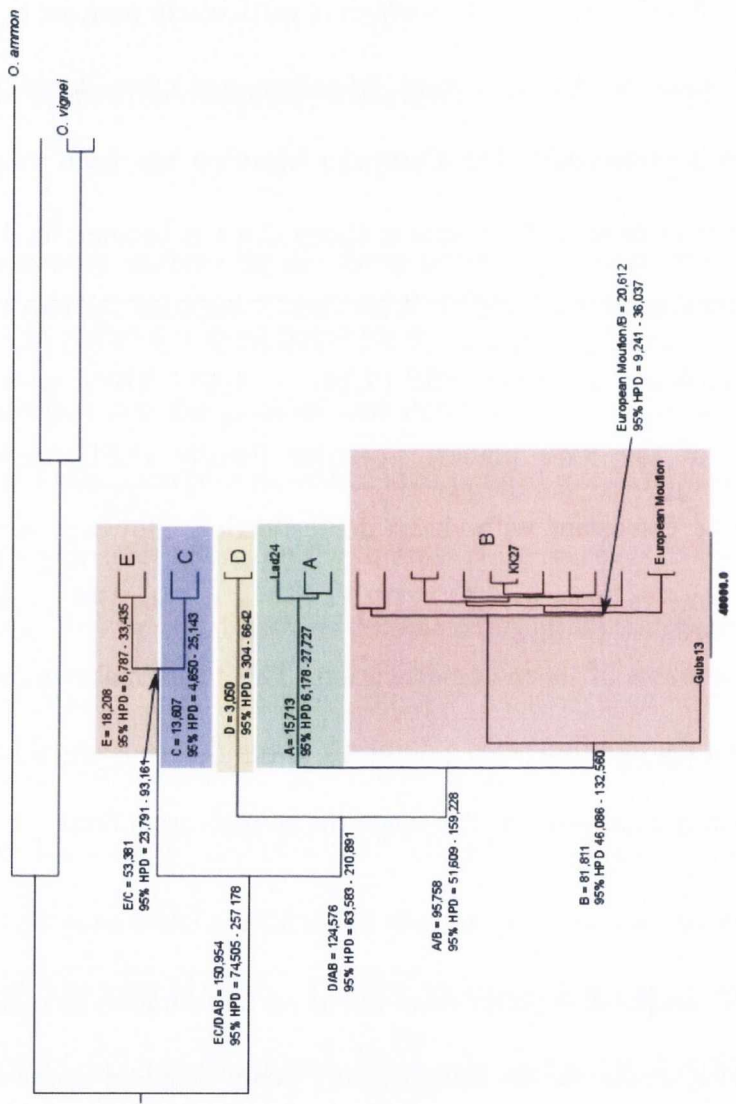


Figure 3:7 Bayesian estimate of maximum clade credibility tree using modern and ancient ovine samples. Modern sample genomes are labelled by haplotype ancient samples by sample code. The posterior probability of all branches with dates is  $\geq 0.9$ . All dates are in YBP, HPD = 95% highest probability density.



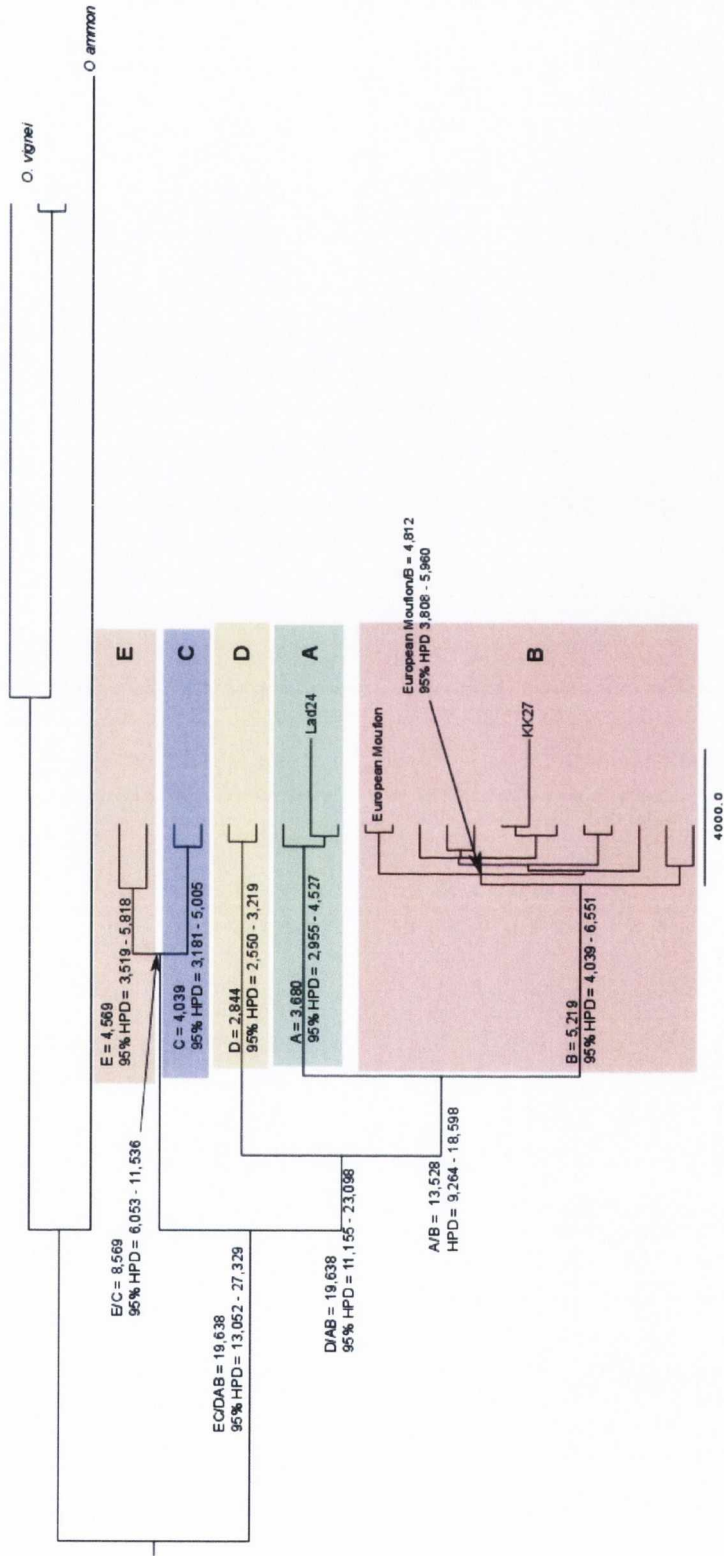


Figure 3-8: Bayesian estimate of maximum clade credibility tree using modern and ancient ovine samples without Gub13. Modern sample genomes are labelled by haplotype ancient samples by sample code the posterior probability of all branches with dates is  $\geq 0.9$ . All dates are in YBP, 95 %HPD = 95% highest probability density.

### **3.4 Discussion**

#### **3.4.1 Assessment of hybridisation capture as a method for ancient domesticate mtDNA retrieval**

Hybridisation capture has been used successfully for the retrieval of both human and Neanderthal ancient mtDNA genomes (BRIGGS *et al.* 2009; SCHUENEMANN *et al.* 2011; SÁNCHEZ-QUINTO *et al.* 2012). This chapter presents the results of these methods applied to ancient domesticate samples from a range of time depths.

Firstly, 12 bovid samples dating from 9,000 to 2,090 YBP were analysed, from which we were able to produce six 1x ancient mitochondrial genomes at greater than or equal to 98% completeness. Secondly, a panel of sheep/goat samples dated from the 3<sup>rd</sup> century BC to roughly 33,500 YBP were examined. The retrieval rate for these samples was less impressive 3/16 (98% complete, 1x genomes) unlike the bovids these samples hadn't previously been subjected to PCR analyses which informed sample choice and were therefore studied blind, which likely skewed the success rate of aDNA retrieval. Taken together however, this analysis shows the power of hybridisation capture and sample pooling for the study of ancient mtDNA genomes, as we were able to sequence 9 almost complete 1x ancient domesticate mtDNA genomes using relatively modest amounts of sequencing from 4 Illumina GA II lanes.

The 11 high-coverage domesticate mtDNA genomes (3x coverage greater than 85%) produced allowed for an estimation of the within species contamination

rate for these samples to be generated. Most of these analyses produced estimates lower than one described for a captured human mitochondrial genome at 8% (average contamination rate for the bovid samples was 3.37% and 1.00% for the ovicaprine samples) (SÁNCHEZ-QUINTO *et al.* 2012). Two sets of samples stand out from this analysis. First, Ardeche1 and KK27, two bones that were shown to have an extremely low contamination rates (estimates of 0%). However, the SNPs called in these analyses are the least well covered of all the samples tested and this low coverage is likely an influence in these reduced estimates, as low frequency contaminants may have not been sequenced. However, a sample of similar coverage to Ardeche1, Bercy10, was found to have a high contamination rate of 8.3%. Bercy10 along with Bercy13 form the second group of outlying samples highlighted in this analysis this time with relatively high contamination rates. As these two bones samples come from a well-described archaeological site in France, the high level of study (and hence handling) of these assemblages may be one reason for this elevated contamination rate (Bercy13, 9.0% Bercy10 8.3%), however, these figures are still comparable to the human sample analysed by Sanchez-quinto *et al* (2012), which was freshly excavated.

### **3.4.2 mtDNA capture as an indicator for shotgun success**

7 bovid samples included in this chapter also formed part of a whole sample shotgun sequencing exercise (the results from two of these samples are presented in more detail in 6), which allowed for the relationship between shotgun sequencing success and off target nuclear alignment in mtDNA capture analyses to be assessed.



The best performing mtDNA capture sample, Ardeche2, which produced a complete bovid mtDNA genome with an average coverage of 118x was found to have only moderate levels of off target bovine nuclear DNA in the mtDNA capture and only 0.26% of reads aligned to the *Bos taurus* nuclear genome when this sample was shotgun sequenced. This results could be attributed firstly to differences in the preservation states of nuclear and mitochondrial DNA in the sample (ALLENTOFT *et al.* 2012) and secondly perhaps to differences in the DNA concentration in the part of the sample analysed, as the results come from two independent extractions and sequencing libraries. However, the same extract that was used for the second of the two mtDNA captures experiments was also found to produced only moderate results in a nuclear DNA capture analysis (Chapter 5) suggesting that this result may indeed be and an inherent feature of the sample.

Two Turkish samples Turk11 and Turk13 were found to contain both high levels of off target nuclear DNA in the mtDNA capture and high raw alignment percentages to the bovid genome shotgun analyses of 7% and 8% respectively. All the other samples tested producing low levels of bovine nuclear DNA in both the capture and shotgun analysis. These results suggest that except in the case of outliers like Ardeche2, off target nuclear DNA retrieved in mtDNA hybridisation capture experiments could be used to assess the chances of success of an aDNA sample in shotgun sequencing. This raises the possibility of using a pre-screen of aDNA samples using mtDNA capture for laboratory pipelines aimed at broader genetic analysis. This could be completed

economically using multiple barcoded samples (~30-50 using the current Illumina HiSeq technology).

The shotgun sequencing described above was also used to provide a comparison of the metagenomes between the captured and uncaptured samples, this analysis provided evidence to suggest that the close homology of the exogenous bacterial and the mitochondrial rRNA and tRNA genes may be influencing the species discovered, with the captured metagenome providing a greater variety and amount of more taxonomic assignments.

### **3.4.3 Bayesian phylogenetic analysis of domestic mtDNA genomes**

The production of ancient mtDNA genomes from these hybridisation capture studies provided the opportunity to calibrate the molecular phylogenetic analyses of two domestic species sheep and cow

The first analysis of the cattle mtDNA genomes summarised in this chapter, produced estimates of divergence times between all cattle clades that fall in a time periods that significantly predate that of the domestication process. Moreover it suggests that the genetic variation within the major domestic taurine cattle T clade cannot all be explained by post domestication mutation, suggesting at least some genetic diversity incorporation from an earlier population expansion is likely.

The sheep BEAST analysis presents a more ambiguous interpretation which is likely due to the limited number of samples used in this analysis and their less

secure dating. When all ancient samples are considered the timings of all the divergences of the major haplogroup are predicted to be ancient, suggesting that the diversity seen in sheep mitochondrial haplogroups is likely explained by population expansions after glaciations. However, when the one outlying ancient sample dated to 33,500 YBP is removed (Gubs13) the analysis presents a different scenario. In this case the coalescence of all the major haplogroups is reduced to 19,638YBP, moreover, the most recent common ancestor of the two most populous haplogroups, A and B, now has a 95% highest probability density that aligns within the time horizon of sheep domestication (ZEDER 2008). This result suggests population expansions following migrations into Asia (haplogroup A) and Europe (haplogroup B) might explain the diversity found between these haplotypes today and necessitate only a single domestication. The divergences between the other less frequent haplotypes (MEADOWS *et al.* 2010) may be a remnant of the last glaciation. This analysis also presents a divergence time between the European Mouflon and the major sheep haplotype B of 4,812YBP (compared to 20,812YBP when Gubs13 is used) that more closely reflects their proposed status as a feral version of an early domestic European sheep. However, the study of further ancient sheep mitochondrial genomes is necessary to confirm these calibrations and elucidate the dynamics of the ovine domestication process.

Finally that one sample (Gubs13) can affect the analysis so much, leads extra caution to the interpretation of BEAST type analyses, which rely on accurate molecular clocks. As this result highlights the interpretation can be greatly influence by single samples.



### ***3.5 Conclusion***

In summary this chapter answers the following question that RNA hybridisation capture can be used to retrieve complete mitochondrial genomes from aDNA samples. This result has important implications for the analysis of aDNA samples by corroborating a new method for the study of ancient mtDNA genomes. This chapter also shows that these genomes can be used within a Bayesian phylogenetic framework (BEAST) to date the divergences between major mtDNA clades, in this case in sheep and cattle. The analysis suggests that two domestications are required to explain bovine mtDNA diversity where as a single domestication is possible in sheep. This chapter also highlights the power of mtDNA to maximise the amount of genetic information retrieved from aDNA samples, as specimens that are unlikely to be used for whole genome sequencing may provide as in this case study useful or novel mtDNA data. Furthermore the off target sequence sequenced during these investigations can also predict the utility of samples in whole genome sequencing experiments. This final point could have major implications for researchers with limited budgets looking for way to screen samples before whole genome NGS analysis.

## 4. Population genetic analysis of a Mesolithic wild aurochs (*Bos primigenius*)

### 4.1 Introduction

#### 4.1.1 The aurochs (*Bos primigenius*)

The wild ox or aurochs (*Bos primigenius*) ranged extensively after the last glacial maximum, covering large parts of Eurasia and North Africa. Both archaeological and genetic studies support the domestication of two geographically and genetically distinct sub-species from this ancestral population (LOFTUS *et al.* 1994; BRADLEY *et al.* 1996; HELMER *et al.* 2005). The first in the Near East, which produced the humpless taurine (*Bos taurus*) cattle (TROY *et al.* 2001; HELMER *et al.* 2005; VIGNE 2011; ZEDER 2011), the second in the Indus Valley leading to the humped indicine/zebu (*Bos indicus*) breeds (MEADOW 1996).

Aurochs are thought to have co-existed with their domestic *Bos taurus* cousins in Europe for roughly 7,000 years (TRESSET *et al.* 2009), before being driven to extinction in the 1600s (VAN VUURE 2005). Contemporary accounts allude to formidable creatures, with Caesar in his record of the Gallic wars stating them to be “below the elephant in size, and of the appearance, colour, and shape of a bull” (CAESAR *et al.* 2005).

#### 4.1.2 Aurochs and domestic cattle - Mitochondria

The past decade has seen a continuing debate, as to the genetic contribution of the aurochs to the *Bos taurus* gene pool, post domestication (TROY *et al.* 2001; GÖTHERSTRÖM *et al.* 2005; BOLLONGINO *et al.* 2006; EDWARDS *et al.* 2007b; ACHILLI *et al.* 2008; BOLLONGINO *et al.* 2008; ACHILLI *et al.* 2009; EDWARDS *et al.* 2010; BONFIGLIO *et al.* 2010). This debate was sparked by the analysis of partial mtDNA D-loop sequences from four British Mesolithic aurochs, which found them to form a haplogroup 'P' that is distinct to that of the monophyletic 'T' haplogroup found predominantly in the extant *Bos taurus* populations (BAILEY *et al.* 1996; TROY *et al.* 2001). This ancient DNA data was used to further support the hypothesis of a Near Eastern domestication centre of taurine cattle from a local aurochs population (TROY *et al.* 2001). Decedents of these animals carrying predominantly T haplotypes (a further closely related haplotype 'Q' has also been found in limited numbers in both ancient (Chapter 3) and modern cattle (ACHILLI *et al.* 2009) of Near Eastern origin) are hypothesised to have spread into Europe with migrating farmers as part of the Neolithic package and displaced the local P carrying aurochs populations (TROY *et al.* 2001).

Further analysis of European Mesolithic specimens (BOLLONGINO *et al.* 2006; EDWARDS *et al.* 2007b; BOLLONGINO *et al.* 2008; STOCK *et al.* 2009; EDWARDS *et al.* 2010), have shown the northern and western European aurochs to be almost exclusively of haplogroup P. A further diverse haplogroup 'E' has also been discovered in limited numbers (EDWARDS *et al.* 2007b)(Chapter 3).



To date only a handful of Mesolithic specimens, discovered in Italy, have been found to have a T haplotype (BEJA-PEREIRA *et al.* 2006; LARI *et al.* 2011). These results raise the possibility that T haplotype carrying wild ox might have had a greater range than previously thought, though this Italian population was likely constrained by the Alps post glaciation (EDWARDS *et al.* 2007b), limiting their impact on the larger European taurine population. Some exceptions to this rule have also been found in the extant *Bos taurus* populations of Italy, where a markedly diverse R haplogroup was discovered (ACHILLI *et al.* 2009; BONFIGLIO *et al.* 2010) (Figure 1:4). These animals however, represent an extremely small subset of domestic cattle in Europe (0.6% of samples studied by Bonfiglio *et al.* (2010) n= 1,747) and it is unclear whether they are indicative of a local domestication, or if they represent introgression of Asian or African aurochs into the Italian taurine populations.

Tantalising evidence for the European aurochs matrilineal survival to the present day is supplied by the discovery of a “living aurochs” a single cow with a P haplotype described as “beef cattle” and sequenced in Korea (ACHILLI *et al.* 2008). However, when all these lines of evidence are taken together they emphasise the major accomplishment that the domestication of cattle in the Near East was (VIGNE 2011) as subsequent independent cattle domestications if they accrued at all were extremely rare. With further evidence by Bollongino *et al.* (2012) suggesting that this initial capture was also of a very small initial population. Moreover, they imply that subsequent local capture of female aurochs in Europe, were on such a small scale that their effect is seen in only vanishingly small numbers in modern day populations (STOCK *et al.* 2009). However, it is tempting to speculate that interactions between the

nascent *Bos taurus* populations and aurochs were likely as these animals spread into Europe.

#### **4.1.3 Aurochs and domestic cattle – Y Chromosome**

The analysis of Götherström *et al* (2005) provided the first piece of genetic evidence for the admixture of European aurochs and domestic cattle by highlighting geographical structure within two taurine Y chromosome haplogroups (distinguished by an A/C SNP in UTY intron 19 and a 2 bp INDEL polymorphism in ZFY intron 5). The first, Y1, was shared with the majority of northern European *Bos taurus* and *Bos primigenius* specimens analysed by Götherström *et al*. The second, Y2, was found in the majority of southern European animals (GÖTHERSTRÖM *et al*. 2005). This cline suggested a hybridisation of northern European domestic cattle and aurochs, as the Anatolian populations sampled by Götherström *et al* also had a Y2 haplotype. This result challenged the conventional view expressed above that domestic cattle have shared little or no ancestry with the European aurochs for the last 70,000 years or so (ACHILLI *et al*. 2008; 2009). However, subsequent sampling of more ancient specimens has called into question the usefulness of these Y chromosome markers for the study of introgression, with the discovery that the Y2 haplotype has a wider distribution in northern European aurochs than previously thought (BOLLONGINO *et al*. 2008).

#### 4.1.4 The aurochs genome sequencing project

In 2010 Edwards *et al* published the complete mitochondrial genome of a British Mesolithic aurochs sequenced using both tradition Sanger and NGS methods. The sample analysed, CPC98 (proximal half of a *humerus*) forms part of a collection of archaeological remains found in Carsington Pasture Cave, Darbyshire (<http://capra.group.shef.ac.uk/1/carsing.html>). The bone has been directly dated to the British Mesolithic with a radiocarbon age of 6,738 $\pm$ 68 calYBP; securely before the arrival of the Neolithic in Britain around 6,000 YBP. CPC98 is remarkably well preserved, carries a *Bos primigenius* (P) mtDNA haplotype and has provided consistent PCR results for the past decade (TROY *et al.* 2001; EDWARDS *et al.* 2007b; STOCK *et al.* 2009; EDWARDS *et al.* 2010). Meta-genomic analysis also indicate a relatively high percentage of endogenous DNA in the sample at roughly 5% (EDWARDS *et al.* 2010). Following the initial success of the mtDNA analysis a complete nuclear genome sequencing project was undertaken, thus far this study has produced 3.4 billion reads and a 6x mean coverage of CPC98s nuclear genome (Stephen Park, personal communication).

#### 4.1.5 PCA analysis of modern and ancient samples

The study of ancient DNA has allowed the genetic relationship between extinct populations and their modern day counterparts to be hypothesised (GREEN *et al.* 2010; RASMUSSEN *et al.* 2010; REICH *et al.* 2010; KELLER *et al.* 2012; SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012; MEYER *et al.* 2012). Studies by Reich *et al* (2010) and Green *et al* (2010) suggest that the genomes



of human populations have been shaped by admixture with two extinct hominids. Green *et al* (2010) found that approximately 1-4% of Eurasian genomes are derived from Neanderthals. With Reich *et al* (2010) providing evidence that the “Denisovan” contributed significantly (4–6%) to the ancestors of present-day Melanesians, a finding that was corroborated by the recent 30x genome sequence of the same sample (MEYER *et al.* 2012). A 2011 study by Skoglund and Jakobsson however, cautions that ascertainment bias for common alleles in modern data sets, can affect the results of these analyses. Both the Neanderthal and Denisovan studies also placed the ancient NGS genotyped sample, within the modern genetic landscape of densely SNP typed populations. Further studies combining modern SNP panels and NGS aDNA data, have allowed the current European populations who most closely resemble Neolithic and Mesolithic Europeans to be elucidated (KELLER *et al.* 2012; SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012; PINHASI *et al.* 2012).

#### **4.1.6 SNP population genetic analysis of modern cattle populations**

The bovine HapMap project typed 37,470 SNPs in 497 animals selected from 19 diverse breeds, to provide a genetic description of global cattle populations (BOVINE HAPMAP CONSORTIUM *et al.* 2009). This study supported the well described separation of the three main cattle populations, European taurine, indicine and African taurine breeds (LOFTUS *et al.* 1994; BRADLEY *et al.* 1996) in both STRUCTURE and PCA analyses (BOVINE HAPMAP CONSORTIUM *et al.* 2009). It also placed two hybrid groups within this global genetic map. The first was the relatively young “synthetic” hybrids, these breeds (represented by Beef Master and Santa Gertrudis in the HapMap analysis) stem mainly from

North America and are the product of selective hybridisation between a small number of *Bos indicus* and *Bos taurus* animals. The second was the older hybrid populations of Africa (represented by Sheko in the HapMap analysis), which were formed out of centuries of trade between East African and Asia. That the two major cattle sub-species can interbreed so successfully after separating approximately 200,000 years ago makes the survival of aurochs/taurine hybrids extremely likely.

#### **4.1.7 Objectives for this case study**

The production of a high quality 6x genome from a British Mesolithic aurochs provides a timely genotype dataset for the analysis of the nuclear genetic affinities of aurochs (*Bos primigenius*) to domestic cattle. This chapter aims to deduce what population genetic analyses are possible with this dataset when it is applied to two modern SNP reference panels with the aim to try to deduce the genetic contribution that these animals have made on the modern domestic *Bos taurus* population. Moreover, it describes the utility of these datasets for the analysis of other ancient samples using this high quality sample as a useful test case. It is also the first such analysis to apply these methods to an ancient domesticate genome.

## **4.2 Methods**

### **4.2.1 Bovine SNP50 data set**

The BovineSNP50 ([http://www.illumina.com/products/bovine\\_snp50\\_whole-genome\\_genotyping\\_kits.ilmn](http://www.illumina.com/products/bovine_snp50_whole-genome_genotyping_kits.ilmn)) data set of Gautier *et al* (2010) (referred to subsequently as 50K) was used for the comparison of CPC98 with a diverse range of modern day cattle populations (Table 4:1, n=1,158). This data set is a merge of sources including the bovine HapMap (BOVINE HAPMAP CONSORTIUM *et al.* 2009), Flori *et al* (2009) and two analyses by Gautier *et al* (2009,2010).

### **4.2.2 Bovine HD SNP770 data set**

SNP Data was also collected from 226 locally held samples, which were genotyped using the new Illumina BovineHD BeadChip (referred to subsequently as 770K). This data set contains relatively few animals as compared to the SNP50 (Table 4:2), but incorporates a greater than 10x increase in genotypes per sample (n=726,515) ([http://www.illumina.com/products/bovinehd\\_whole-genome\\_genotyping\\_kits.ilmn](http://www.illumina.com/products/bovinehd_whole-genome_genotyping_kits.ilmn)).

### **4.2.3 CPC98 SNP genotyping**

CPC98 was genotyped by Stephen Park as part of the aurochs genome-sequencing project (Stephen Park, Personal Communication). A minimum depth of 10x was used to select SNPs for analysis.



| Code | Breed Name                   | Type | Land of Origin  | Number |
|------|------------------------------|------|-----------------|--------|
|      |                              |      |                 |        |
| ABO  | Abondance                    | EUT  | France (SE)     | 22     |
| ANG  | Angus                        | EUT  | Scotland        | 40     |
| AUB  | Aubrac                       | EUT  | France (S)      | 22     |
| BAO  | Baoule                       | WAT  | Burkina Faso    | 29     |
| BMA  | Beef Master                  | SYN  | USA             | 24     |
| BOR  | Borgou                       | WAH  | Benin           | 30     |
| BPN  | Bretonne Black Pied          | EUT  | France (NW)     | 18     |
| BRM  | Brahman                      | ZEB  | USA (India)     | 25     |
| BRU  | French Brown Swiss           | EUT  | Switzerland     | 18     |
| BSW  | Brown Swiss                  | EUT  | Switzerland     | 24     |
| CHA  | Charolais                    | EUT  | France (M)      | 20     |
| CHL  | Charolais                    | EUT  | France (M)      | 26     |
| GAS  | Gascon                       | EUT  | France (SW)     | 22     |
| GBV  | Gelbvieh                     | EUT  | Germany         | 3      |
| GIR  | Gir                          | ZEB  | India           | 24     |
| GNS  | Guernsey                     | EUT  | Channel Islands | 21     |
| HFD  | Hereford                     | EUT  | England         | 31     |
| HO2  | Holstein (US)                | EUT  | Northern Europe | 30     |
| HOL  | Holstein (FR)                | EUT  | Northern Europe | 30     |
| JE2  | Jersey (US)                  | EUT  | Channel Islands | 21     |
| JER  | Jersey (FR)                  | EUT  | Channel Islands | 28     |
| KER  | Kerry                        | EUT  | Ireland         | 23     |
| KUR  | Kuri                         | WAH  | Chad            | 30     |
| LAG  | Lagune                       | WAT  | Benin           | 30     |
| LMS  | Limousin                     | EUT  | France (MW)     | 44     |
| MAN  | Maine-Anjou (Rouge des Près) | EUT  | France (NW)     | 16     |



|       |                          |     |             |      |
|-------|--------------------------|-----|-------------|------|
| MAR   | Maraichine (Parthenaise) | EUT | France (NW) | 19   |
| MON   | Montbeliarde             | EUT | France (E)  | 30   |
| ND1   | N'Dama                   | WAT | West Africa | 25   |
| ND2   | N'Dama                   | WAT | West Africa | 14   |
| ND3   | N'Dama                   | WAT | West Africa | 17   |
| NEL   | Nelore                   | ZEB | India       | 21   |
| NOR   | Normande                 | EUT | France (NW) | 30   |
| NRC   | Norwegian Red Cattle     | EUT | Norway      | 21   |
| OUL   | Oulmès Zaer              | NAT | Morocco     | 26   |
| PMT   | Piedmontese              | EUT | Italy (N)   | 24   |
| PRP   | French Red Pied Lowland  | EUT | France (NW) | 22   |
| RGU   | Red Angus                | EUT | Scotland    | 15   |
| RMG   | Romagnola                | EUT | Italy       | 24   |
| SAL   | Salers                   | EUT | France (S)  | 22   |
| SGT   | Santa Gertrudis          | SYN | USA         | 24   |
| SHK   | Sheko                    | EAT | East Africa | 20   |
| SIM   | Simmental                | EUT | France      | 3    |
| SOM   | Somba                    | WAT | Togo        | 30   |
| TAR   | Tarine                   | EUT | France (SE) | 18   |
| VOS   | Vosgienne                | EUT | France (NE) | 19   |
| ZBO   | Zebu Bororo              | WAZ | West-Africa | 23   |
| ZFU   | Zebu Fulani              | WAZ | West-Africa | 30   |
| ZMA   | Zebu from Madagascar     | ZEB | Madagascar  | 30   |
|       |                          |     |             |      |
| Total |                          |     |             | 1158 |

**Table 4:1 Geographic location and numbers of samples studied 50K. EAT = East African Taurus, EUT = European taurine, NAT = North African taurine, SYN = synthetics, WAH = West African Hybrids, WAT = West African taurine, WAZ = West African Zebu and ZEB = zebu/indicine.**

| Code  | Breed               | Type | Land of Origin | Number |
|-------|---------------------|------|----------------|--------|
|       |                     |      |                |        |
| SOM   | Somba               | WAT  | Togo           | 11     |
| MAR   | Maremmana           | EUT  | Italy          | 4      |
| ANG   | Angus               | EUT  | Scotland       | 21     |
| ANB   | Anatolian Black     | NEH  | Turkey         | 13     |
| ALE   | Alentejana          | EUT  | Portugal       | 5      |
| BBL   | Belgium Blue        | EUT  | Belgium        | 21     |
| CHL   | Charolais           | EUT  | French         | 21     |
| EAR   | East Anatolian Red  | NEH  | Turkey         | 12     |
| HAR   | Hariana             | ZEB  | India          | 11     |
| HER   | Hereford            | EUT  | England        | 21     |
| HIG   | Highland            | EUT  | Scotland       | 3      |
| KER   | Kerry               | EUT  | Ireland        | 5      |
| LMS   | Limousin            | EUT  | France         | 21     |
| MUR   | Muturu              | WAT  | Cameroon       | 11     |
| SIS   | Sistani             | NEH  | Iran           | 5      |
| SAR   | South Anatolian Red | NEH  | Turkey         | 9      |
| SIM   | Simmental           | EUT  | Switzerland    | 21     |
| SIK   | Sikia               | EUT  | Greece         | 5      |
| TAL   | Taleshi             | NEH  | Iran           | 4      |
|       |                     |      |                |        |
| YAK   | Yak                 | YAK  | Bhutan         | 1      |
| GAR   | Gaur                | GAR  | ?              | 1      |
|       |                     |      |                |        |
| Total |                     |      |                | 226    |

**Table 4:2: Geographic location and numbers of samples studied 770K. EUT = European taurine, NEH = Near Eastern hybrid, WAT = West African taurine, ZEB=zebu/indicine.**



#### 4.2.4 PLINK analysis

All SNP data management, removal of loci and animals was completed using PLINK v1.07 (<http://pngu.mgh.harvard.edu/~purcell/plink>) with the --cow flag applied. This option allows PLINK to recognise the extra chromosomes present in the bovine genome as compared to humans

#### 4.2.5 Eigensoft Analysis

Principal components for plotting were produced using the SMARTPCA package of EIGENSOFT 4.2 (PATTERSON *et al.* 2006). Following Skoglund *et al.* (2012) all pairs of SNPs with an  $r^2 > 0.2$  had one of the pair removed. The parameter file of SMARTPCA was also modified to recognise the 29 pairs of autosomes found in cattle. All other parameters were left at their default settings.

#### 4.2.6 Structure analysis

STRUCTURE is a widely implemented program for the detection of population structure (PRITCHARD *et al.* 2000); in this analysis version 2.3.3 of the program was used. Optional parameters were left at their defaults and a burn in of 10,000 and a further 100,000 repetitions of the algorithm were used to generate the data sets for this analysis. A range of assumed populations  $K$  from 3 to 7 were explored (PRITCHARD *et al.* 2000). Genotype input files for STRUCTURE were produced from the SNP data via PLINK using the --recode-structure flag. Bar charts of the structure output were then drawn using *distruct*, selecting standard settings (<http://www.stanford.edu/group/rosenberglab/distruct.html>).



#### 4.2.7 Phylogenetic tree production

The --cluster and --distance-matrix commands also of PLINK (v1.06) were used to produce an identity by decent (IBD) matrix for phylogenetic tree drawing. The output was then converted to a PHYLIP (version 3.69) (FELSENSTEIN 1989) matrix file with a custom Python script. Subsequently neighbour joining trees were produced using the NEIGHBOUR module of PHYLIP and visualized and annotated using FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>).

#### 4.2.8 Green et al 2010 *D* Statistic

The *D*-statistic of Green *et al* (2010) (SOM15) [1] was implemented in the Python script D\_Test.py (Appendix:2). This test was preformed in its basic version as shown in the equation below.

$$D(H_1, H_2, H_3, OutGroup) = \frac{\sum_{i=1}^n [C_{ABBA}(i) - C_{BABA}(i)]}{\sum_{i=1}^n [C_{ABBA}(i) + C_{BABA}(i)]} \quad [1]$$

For each SNP position in the 770K dataset (chosen for its greater SNP sample size over the 50K) for which we have genotype information for all 4 populations ( $H_1$  = test population 1,  $H_2$  = test population 2,  $H_3$  = CPC98 and OutGroup = Yak), one allele is selected at random from each population's genotype. The outgroups allele is then designated "A" and the analysis restricted to sites where  $H_1$  and  $H_2$  vary and the alternative allele "B" is called

in CPC98 (Figure 4:1)(GREEN *et al.* 2010).  $C_{ABBA}(i)$  and  $C_{BABA}(i)$  are used as indicator variables and can be 0 or 1 if an ABBA or BABA pattern is seen at base  $i$  (GREEN *et al.* 2010). The difference in the count of ABBA ( $C_{ABBA}(i)$ ) and BABA ( $C_{BABA}(i)$ ) sites (Figure 4:1) across the  $n$  base pairs analysed is then normalised by the total number of observations to give the final score  $D$  (GREEN *et al.* 2010). 10 representative breeds were chosen for this analysis 8 from Europe (Hereford, Angus, Kerry, Maremmana, Sikia, Alentejana, Simmental and Charolais) and 2 from Africa (Somba and Murutu). In all runs of the algorithm  $H_3$  was set as CPC98 and the outgroup used was Yak. The Yak was chosen rather than an indicine breed for two reasons. Firstly, choosing an indicine outgroup could lead to taurine breeds with low levels of indicine introgression gaining an unintentional increase in ancestral alleles due to their recent shared ancestry. Secondly the outgroup chosen for the initial analysis of Green *et al.* (2010) was the chimpanzee for a comparison of human populations and Neanderthals. The estimated yak/cattle divergence of 4.9 million years (QIU *et al.* 2012) is more similar to that of human/chimpanzee 6–7 million years (THE CHIMPANZEE GENOME SEQUENCING AND ANALYSIS CONSORTIUM 2005; GREEN *et al.* 2010), than that of *Bos taurus* and *Bos indicus* (200,000 years).  $H_1$  and  $H_2$  were rotated by choosing 1 animal at random from the 10 representative breeds. Each analysis was repeated 100 times to try to counteract the biases introduced by the random sampling of alleles from each genotype and the average score of all 100 runs calculated.

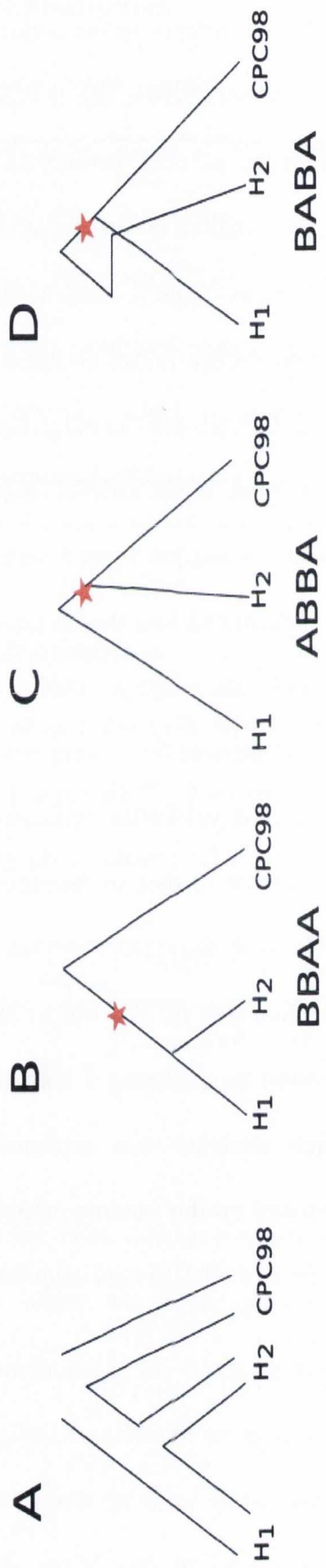


Figure 4:1: Diagram adapted from Green *et al* (2010) (SOM 15), *D*-statistic to test the null hypothesis that H1 and H2 form a clade, with CPC98 more distantly related. [A] Phylogenetic tree for cattle H1 = Test population 1, H2 = Test population 2, CPC98. [B-D] Genealogical trees "A" = ancestral allele, "B" = alternative allele, BBAA sites [B] are expected to occur more frequently, as in this configuration the genealogy mirrors the phylogeny. ABBA [C] and BABA [D] sites are then expected to occur at equal frequencies, if lineages H1 and H2 go back to the time of divergence from CPC98 without coalescing.



## **4.3 Results**

### **4.3.1 SNPs genotyped SNP50**

In total the 50K dataset consisted of 33,077 SNPs, typed in 1,158 animals, however, after filtering for a read depth of 10x in CPC98, 17,443 SNPs remained for further analysis. The total genotyping rate in the final dataset was 99.7%.

### **4.3.2 50K Eigensoft analysis worlds populations and CPC98**

The result of the Eigensoft analysis of the combined modern and CPC98 50K data set is presented in Figure 4:2 (n=1,159). After SMARTPCAs filters for informative markers and a between pair  $r^2$  value of less than 0.2 were applied, 12,418 SNPs were retained for further analysis. The modern cattle populations split as shown previously into three main groups (MACHUGH *et al.* 1997; BOVINE HAPMAP CONSORTIUM *et al.* 2009; GAUTIER *et al.* 2010). Firstly, the European taurine animals cluster on the left hand side of the plot, with the African animals forming the cline previously described, depending on the contribution of indicine genetics into the breed. The purebred indicine/zebu (ZEB) animals cluster to the right of the plot. The one North African breed in this analysis Oulmès Zaer is also located roughly halfway between the African taurine (Somba, N'Dama and Lagune) and the European animals. The Italian breeds also show some gross structure as suggested by mtDNA analyses (BONFIGLIO *et al.* 2010) with the Piedmontese breed falling within the

European taurine major cluster where as the Romagnola (“R” Figure 4:2) sits outside. CPC98 falls outside of the taurine and indicine groupings but closer to the taurine than the indicine, repeating the results of the mtDNA analysis (TROY *et al.* 2001; EDWARDS *et al.* 2010). However unlike the previously published mtDNA analyses, CPC98 falls closer to the Europe taurine than African taurine animals, with its closest neighbours the synthetic breeds and the Romagnola of Italy.

#### **4.3.3 50K STRUCTURE Analysis world populations**

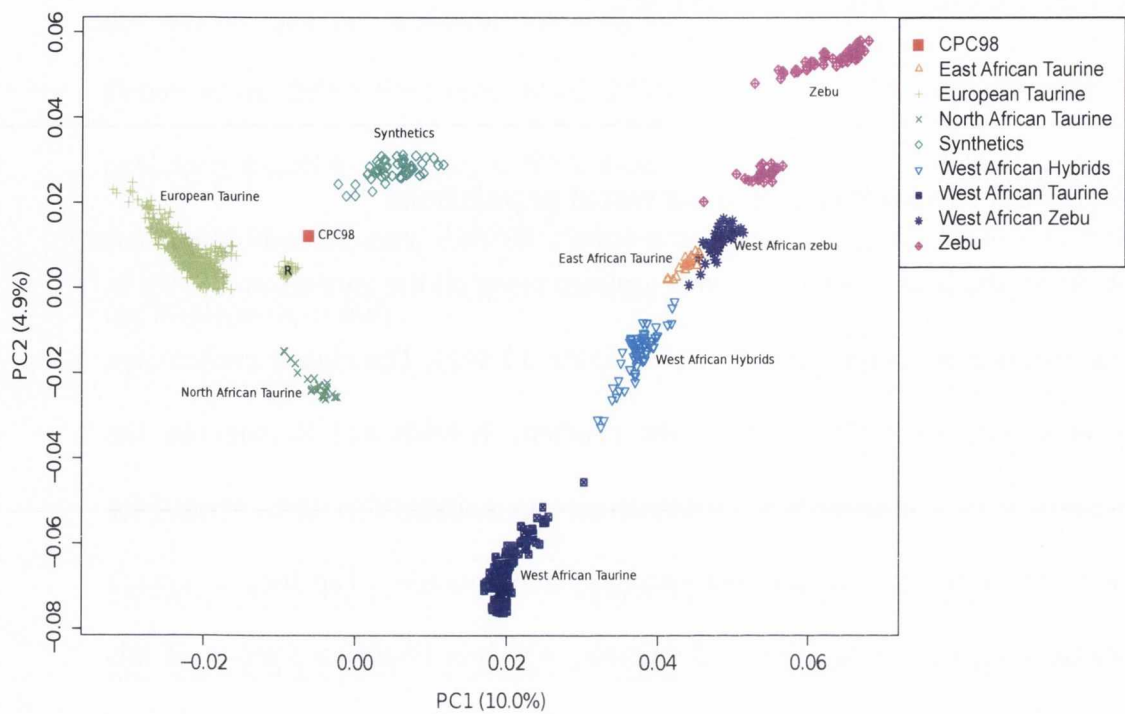
To complement the PCA a STRUCTURE analysis was also preformed on the full SNP data set (SNPs=17,443) with the assumed number of populations,  $K$ , set from 3 to 7. The results of the analysis for  $K=3$  (Figure 4:3) follows the previously described structure of cattle populations (BOVINE HAPMAP CONSORTIUM *et al.* 2009). With the three-population model distinguishing European taurine (red), African taurine (green) and indicine animals (purple). With  $K=3$  CPC98 (AUR) shows most similarity to the Romagnola animals, with a predominance of a taurine contributions with limited effects from indicine. This is an interesting result and may reflect the aurochs geographical range that was closer to the Near East than the Indus Valley. As the number of assumed populations increases, however, these populations diverge (Figure 4:3  $K=4-7$ ) Interestingly the  $K=7$  analyses sees the emergence of blocks that may be suggestive of beef and dairy breeds. The first (orange) is seen heavily in Angus, and other beef breeds, such as the Hereford but is almost absent from Holstein Friesians (Figure 4:3). The second (red) is the predominant part of the Holstein Friesian bars and may be a marker for the selective crossbreeding of these animals to increase dairy production traits in multiple modern day

European taurine populations (Figure 4:3). At higher values of  $K$  CPC98 emerges as a mixture of parts from all the major cattle populations, and is distinct from any of the breeds analysed.

#### **4.3.4 50K Phylogenetic analysis world populations**

A phylogenetic analysis was also completed using all the animals and SNPs in the 50K data set (Figure 4:4,  $n=1,159$ , SNPs=17,443). The closest modern day populations to CPC98 are again the synthetic hybrids and Romagnola, the separation of European and African taurine animals is also seen, along with that of African hybrids and purebred indicine animals. The longest branch lengths are seen in the Hereford animals, which is likely an artefact of this breed being used in SNP ascertainment. This leads to a greater number of heterozygous SNPs being called in this breed and longer branch lengths.





**Figure 4:2: SMARTPCA analysis of cattle and CPC98, bovine 50K data set (12,418 SNPs) Romagnola distribution shown by “R”.**

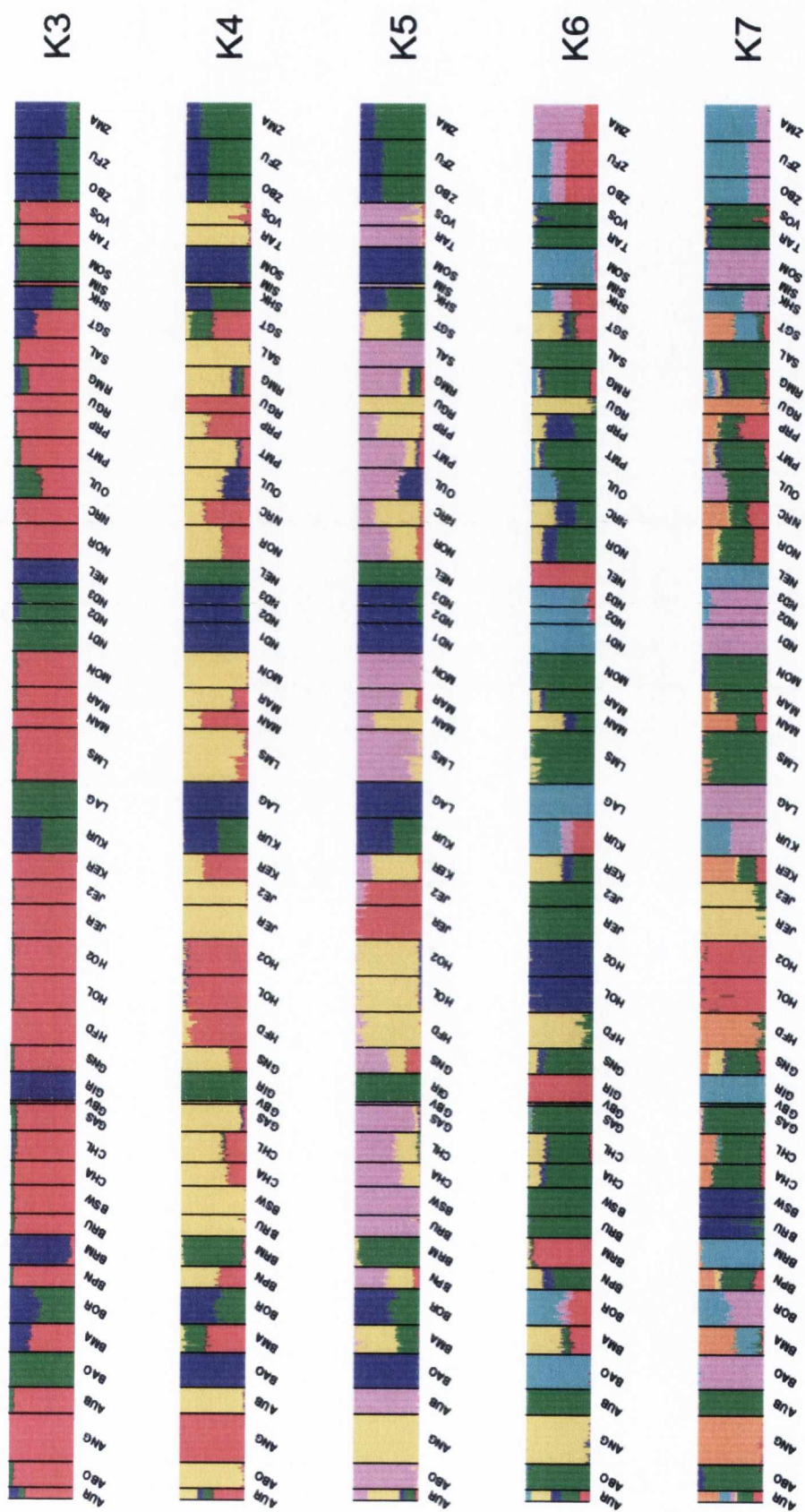


Figure 4:3 STRUCTURE analysis of CPC98 (AUR) and modern cattle populations (17,443 SNPs,  $K=3-7$ ). Note each line in the bar graph denotes 1 individual apart from CPC98 (AUR), that was placed 10 times post analysis for ease of viewing. Bar graphs drawn with *distrupt*.

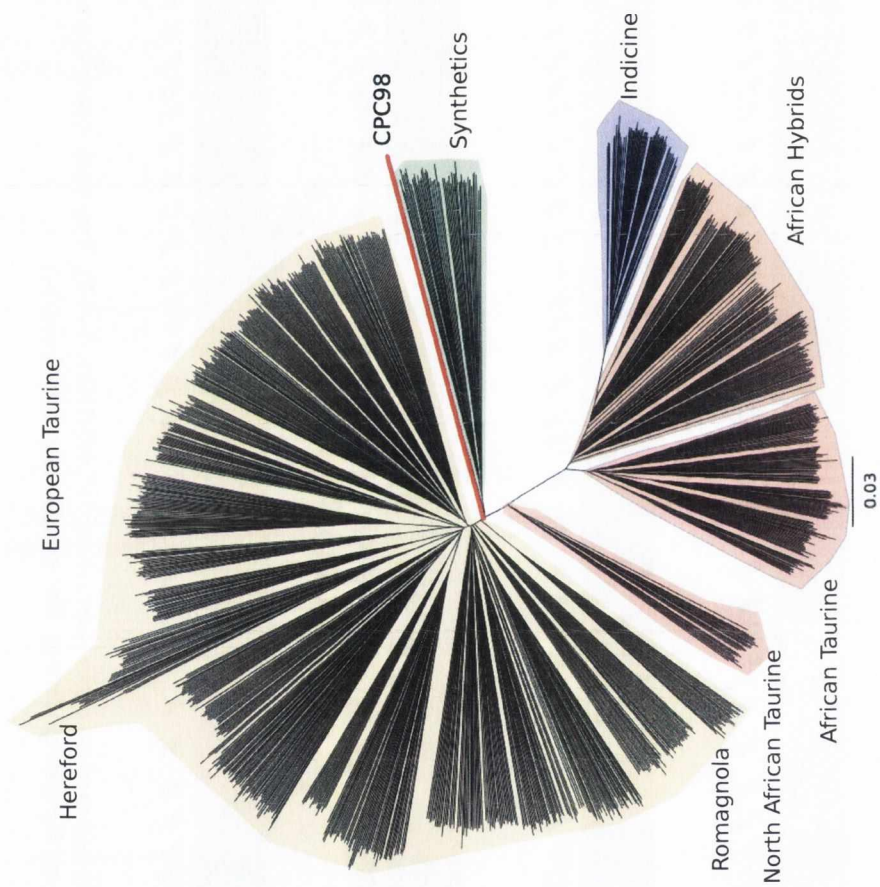
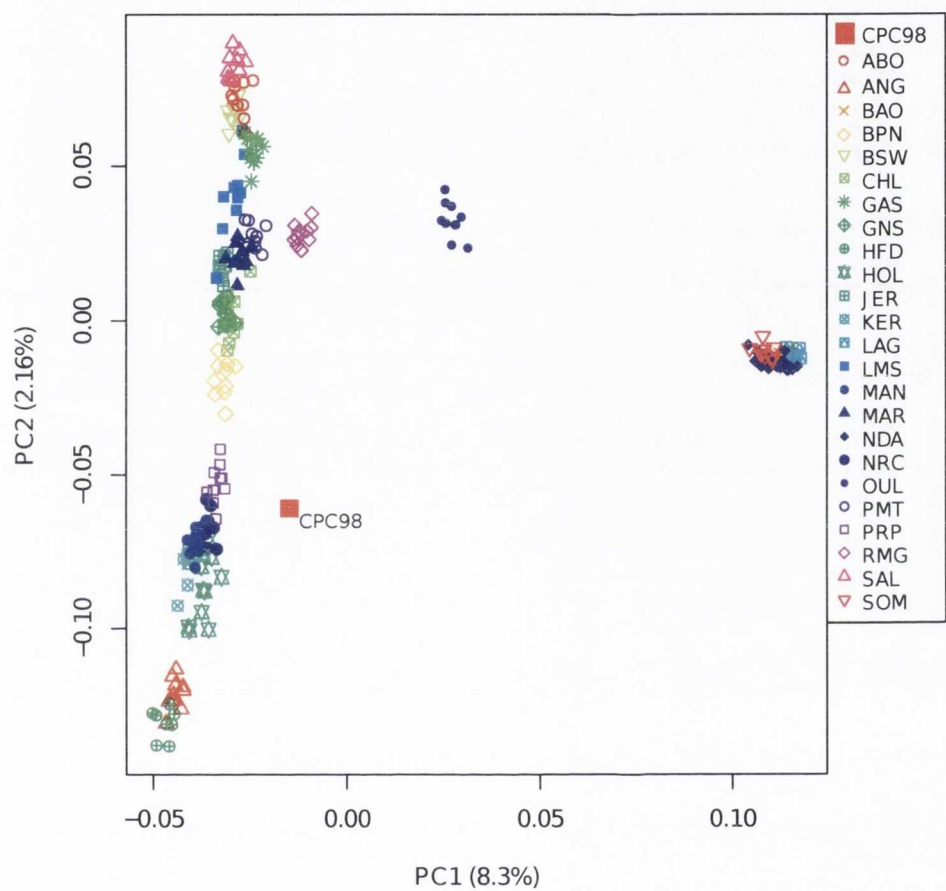


Figure 4:4: Neighbour joining tree produced from PLINK IBD analysis of 1,159 animals and 17,443 SNPs.





**Figure 4:5: SMARTPCA analysis of taurine breeds from the 50K dataset. Encompassing 262 animals from 25 breeds and CPC98 (12,418 SNPs).**

#### **4.3.5 50K Eigensoft analysis of taurine Populations and European aurochs**

Given CPC98s position close to the European taurine populations in the worldwide PCA analysis (Figure 4:2), the SMARTPCA analysis was repeated using a reduced data set of 262 (10 per breed) taurine animals. This analysis aimed to visualise the nature of CPC98s relationship to the taurine population. As can be seen in the plot above the analysis mirrors the separation of the European and African breeds shown in the previous PCA (Figure 4:5). CPC98 is again positioned outside of the modern day populations. However, it no longer clusters with the Romagnola and is closer to the northern European animals.

#### **4.3.6 50K European taurine and European aurochs**

To further investigate CPC98s relationship to the European taurine animals described above, the 50K dataset was filtered again to include only animals of European origin (n=231). The results of the initial PCA analysis seemed to suggest a large effect on both PC1 and PC2 from the Hereford animals (Blue crosses in panel A of Figure 4:6). This effect is likely caused by the Hereford breeds central role in both the bovine genome-sequencing project and SNP ascertainment. As you would expect the Hereford breed to group with other northern European animals and not out on its own. To counter this bias these animals were subsequently removed and the analysis repeated. The results of the second PCA analysis without the Hereford animals, places CPC98 closest to the breeds from northern European, specifically the French Red Pied Lowland (PRP) (Figure 4:6, panel B). This breed is itself a cross of 3 further

northern European breeds Armorican, Maine-Anjou and German Red Pied  
(FELIUS 1995).



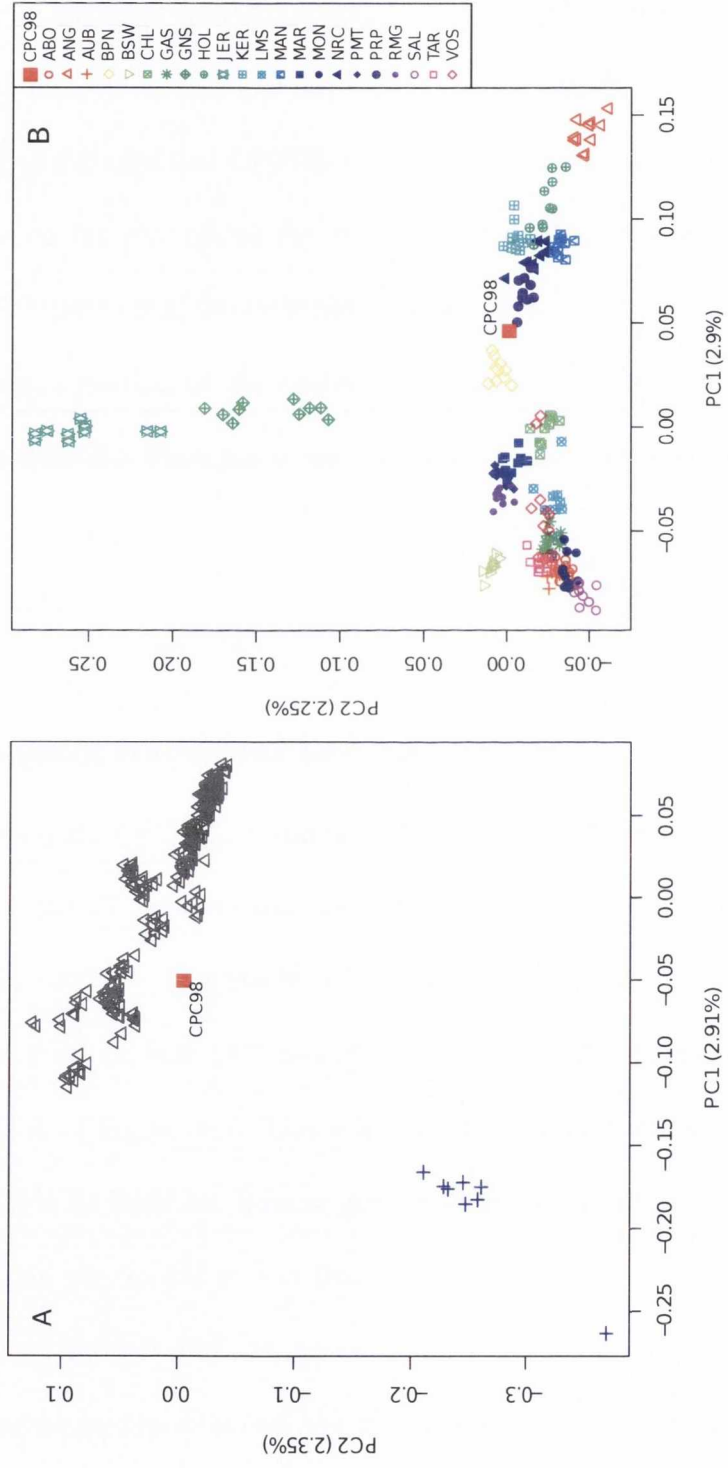


Figure 4:6 European taurine only analysis. [A] PCA showing CPC98 (red square), Hereford (blue crosses) and the rest of European Taurine animals (grey triangles). This analysis highlights probable SNP ascertainment effects, which intern likely cause an artefactual positioning of CPC98. [B] Analysis repeated with the Hereford animals removed (12,418 SNPs).

#### 4.3.7 SNP 770K analysis

During the course of this project the Illumina BovineHD BeadChip chip was released, allowing for the analysis of CPC98's affinities at a much greater resolution (~770K). However, the genotyping experiments for a large number of breeds are still in the early stages of collection and subsequently this analysis uses a more limited number of animals and breeds. Overall 186,198 SNPs from the 770K dataset could be called at 10x or greater in CPC98 and so were retained for further analysis. The final dataset composed the 226 bovines shown in table 4:2 plus CPC98, with a total genotyping rate of 97.9%. This reduced genotyping percentage when compared to the 50K analysis is likely caused by the inclusion of the outgroups yak and gaur in this analysis.

A SMARTPCA analysis was also completed for the 770K SNP dataset without outgroups. After SMARTPCAs filters for informative markers and  $r^2$  were applied, the 770K SNP dataset was reduced to 161,898 genotypes. The results of the 770K analysis mirrors that of the 50K, though with a reduced number of breeds (Figure 4:7). CPC98 again clusters with the European taurine animals and we see the three major cattle populations split (European taurine, African taurine and indicine). As mentioned above this dataset lacks the diversity of the 50K and there are a number of outliers. Three animals from the Muturu breed fall further to the right of this plot than the rest of the samples studied from this population, which likely represents some level of introgression from indicine populations into this breed. One Iranian animal also falls away from the rest of breed suggestive of a greater taurine influence.

This genotyping analysis also presents a novel 770K SNP data set from a number of Near Eastern/Anatolian breeds, specifically chosen as they originate from locations close to the centre of domestication (Anatolian Black, East Anatolian Red, South Anatolian Red, Sistani and Taleshi). These animals form a cluster in the centre of the plot approximately halfway between the European taurine and indicine animals, repeating a result previously reported that these breeds have experienced a certain level of indicine introgression (LOFTUS *et al.* 1999). Two Anatolian Black individuals and one South Anatolian Red can also be seen to cluster more closely to European taurine animals than the rest of their breed, suggestive of a lesser indicine influence in these animals.

#### **4.3.8 Green *et al* 2010, *D*-Statistic.**

The results from the *D*-statistic (Figure 4:8) also supports those of the PCAs shown above by suggesting there is a measurable amount of gene flow from the European aurochs represented in this analysis by CPC98 into all European taurine populations when compared to African (Figure 4:8). Suggesting gene flow into all European breeds. The greatest *D* score in the African/European taurine analysis is that of the comparison between Hereford and Somba animals (-0.149). The lowest (-0.02) when Maremmana are compared to Muturu, negative values indicating gene flow from CPC98 into the test population 1 (H1). There is also a suggestion of a difference in the level of gene flow between the aurochs and taurine populations of northern and southern Europe. With all northern European taurine animals showing higher levels of *D*, in both the European only and European/African taurine analyses.

The Hereford animal consistently shows the highest levels of introgression in all pairwise analyses. These patterns are repeated when other animals from these populations are chosen.

#### **4.3.9 Structure Analysis 770K**

As with the PCA results the 770K STRUCTURE analysis (Figure 4:9) mirrors that of the 50K, with CPC98 shown to be a mixture of all the population blocks. This is likely a reflection of CPC98 being a relic of a diverse ancestral population whose genetics makes up the present day cattle populations. This leads to artefacts in the STRUCTURE analysis, which assumes no knowledge about the population and so it has distinguished that these animals share some genetics.  $K = 3$  nicely highlights the admixture inherent in the Anatolian breeds and the major outliers from the PCA can also be seen as more admixed individuals (Muturu, Sistani). From  $K=5$  onwards there is differentiation of individual breeds with both the Hereford and Angus showing breed specific blocks.



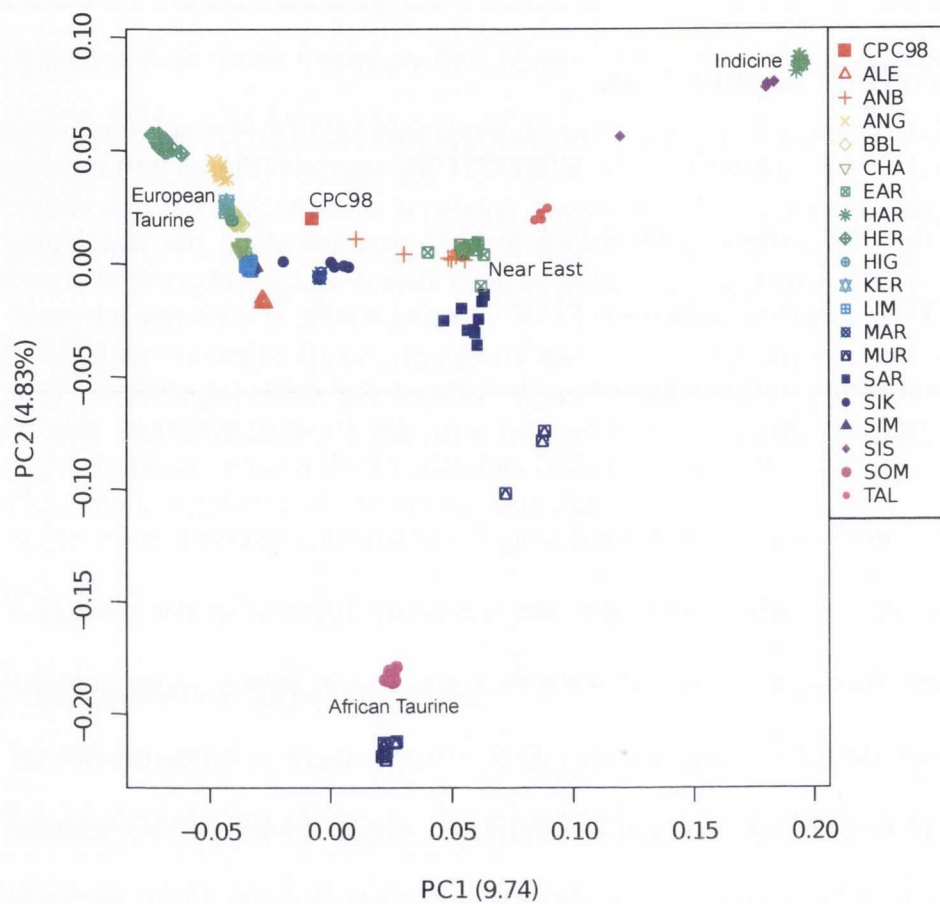
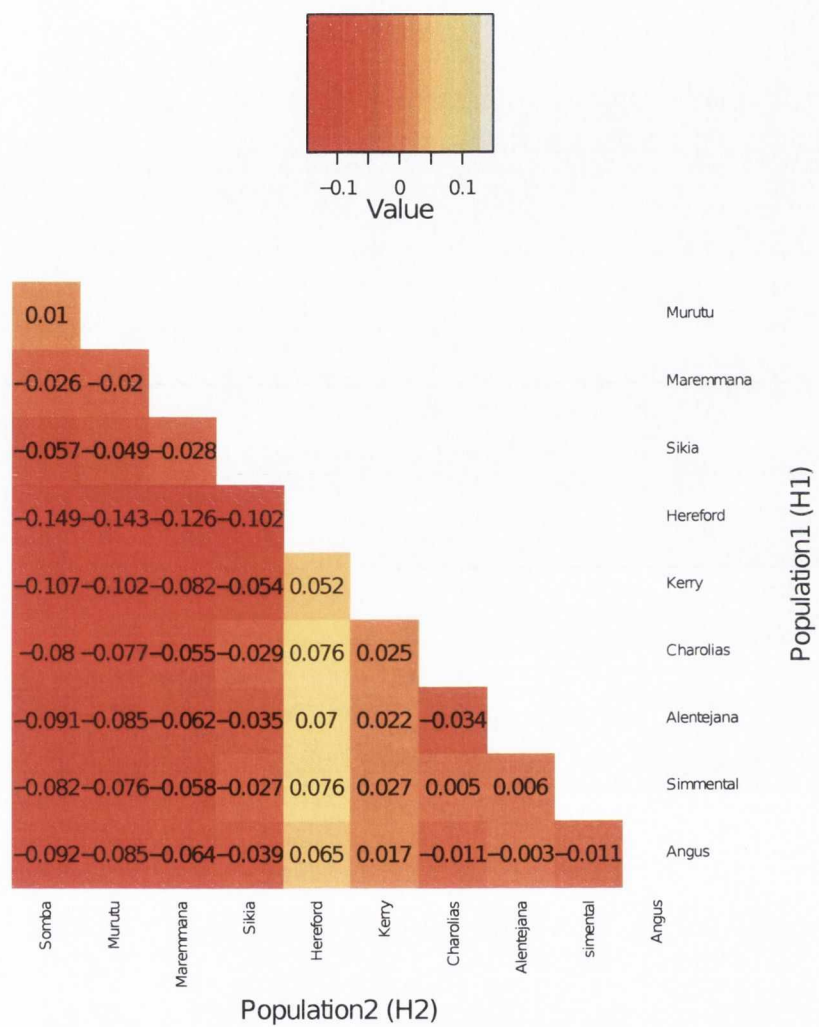


Figure 4:7 770K SMARTPCA analyses of cattle and CPC98 (161,898 SNPs).



**Figure 4:8: Pairwise *D*-statistic analysis of 770K genotype data. Negative values indicate that CPC98 is closer to population1 (H1), positive scores indicate that the converse is true (186,198 SNPs).**

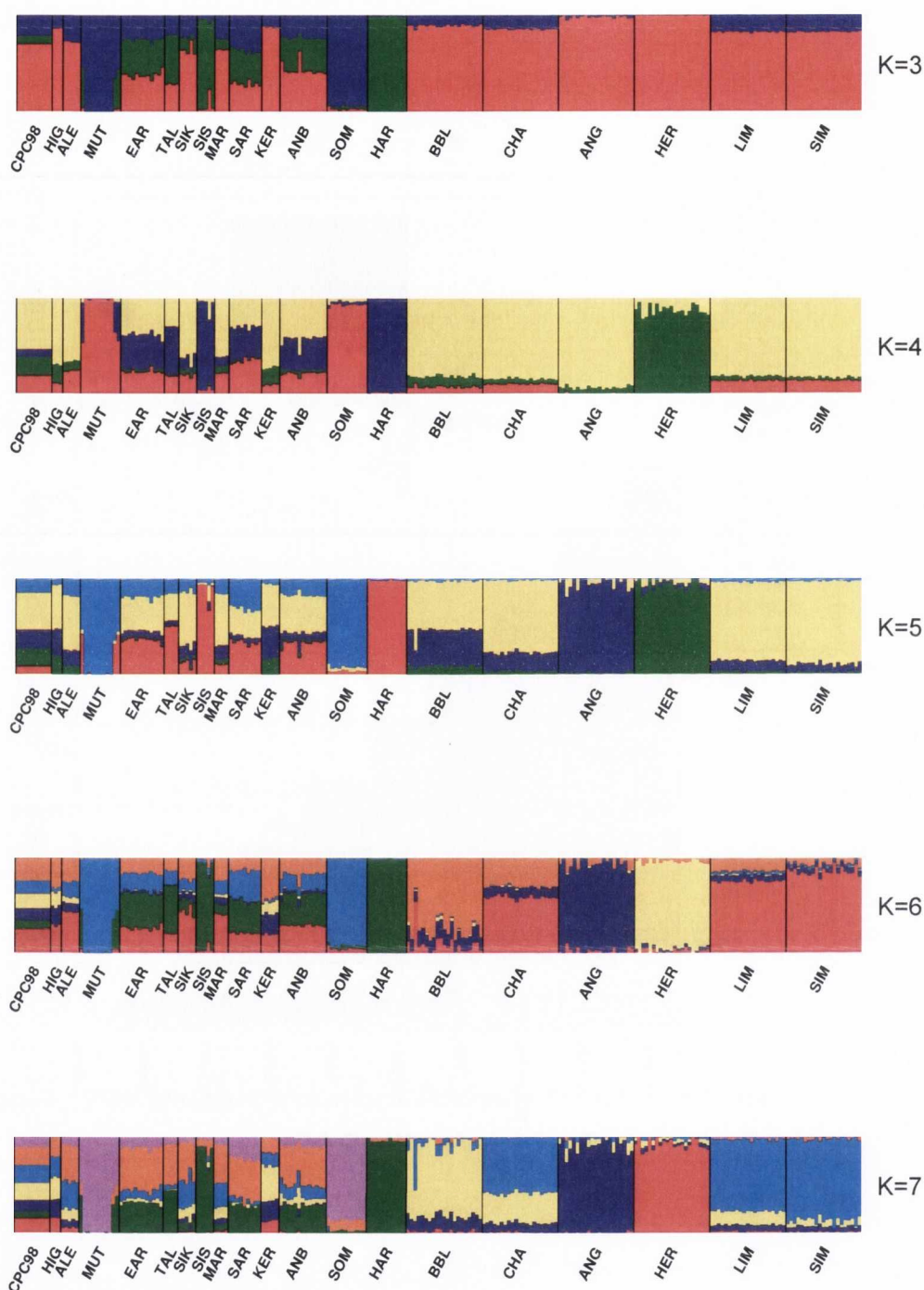


Figure 4:9 770K structure analysis. Note all bars indicate 1 animal apart from CPC98 which is placed 10 times post analysis for ease of viewing (186,198 SNPs).

## **4.4 Discussion**

The population genetic analysis of ancient DNA is seeing an increase of general interest given the new possibilities and huge increase in data that next generation sequencing provides from archaeological remains (GREEN *et al.* 2010; RASMUSSEN *et al.* 2010; SKOGLUND and JAKOBSSON 2011; KELLER *et al.* 2012; SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012). Outlying samples such as CPC98, which have a relatively large endogenous DNA content, are providing the first markers in a range of species, for the genetic analysis of ancient populations (SANGER *et al.* 1977; GREEN *et al.* 2010; RASMUSSEN *et al.* 2010; REICH *et al.* 2010; RASMUSSEN *et al.* 2011; MEYER *et al.* 2012). However, to my knowledge no further P haplotype populations are being currently studied a paucity that should be addressed to corroborate the results of this analysis

### **4.4.1 PCA Analyses**

PCA analyses, specifically those implementing the methods of EIGENSOFT are becoming a widely used technique for the comparison of modern and ancient specimens, using both high coverage and 1x data sequencing data (LANDER *et al.* 2001; VENTER 2001; GREEN *et al.* 2010; REICH *et al.* 2010; KELLER *et al.* 2012; SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012; MEYER *et al.* 2012). Interestingly in this principal component analysis of a single British aurochs specimen (CPC98), it was continually placed closest to the modern day European taurine cattle populations, than either the African taurine or indicine breeds in both the 50K and 770K analysis. Given the close level of relatedness between the African and European taurine mito-genomes



and the consistent placing of CPC98s mitochondrial P haplotype outside of the taurine T clade, this may be a tentative indication of introgression between northern European Aurochs and European *Bos taurus* animals. This conclusion that is further supported by phylogenetic analysis, which finds the African and European taurine populations to split and CPC98 to be placed in the middle, though closer to Europe.

The close proximity of CPC98 to the synthetic breeds highlighted by the 50K analysis is likely the result of two different population histories producing similar contributions of indicine and taurine genetics. In CPC98 this is likely the result of sampling an individual from a diverse ancestral population, where as the synthetic breeds reflect the effect of recent hybridisations (FELIUS 1995; METZKER 2010).

#### **4.4.2 Structure analysis**

The STRUCTURE analysis of the 50K data at  $K=3$  shows the widely observed result that modern domestic cattle split into three main populations, European Taurine, African Taurine and Indicine (KAHVEJIAN *et al.* 2008; BOVINE HAPMAP CONSORTIUM *et al.* 2009). CPC98 is seen to most closely resemble the Romagnola cattle of Italy at this value of  $K$ . However, it then diverges from this population at larger values of  $K$ , which again might be indicative of two different population histories giving similar results at this low resolution. The suggestion of beef and milk signatures in the STRUTURE analysis at  $K=7$  is interesting and likely represents the selective crossbreeding of highly predictive breeds to improve others (METZKER 2010; FELIUS *et al.* 2011). The

structure analysis at 770K SNPs supports the major conclusions of the 50K analysis though with a reduced number of breeds. The appearance of breed specific blocks at lower values of  $K$  likely represents the increased resolution in the data set.

#### **4.4.3 Utility of current SNP data in the analysis of ancient bovids**

Though much improved to the initial bovine HapMap genotyping chip (Affymetrix GeneChip® Bovine Genome Array) the BovineSNP50 (50K) chip still includes biases towards European taurine animals (MATUKUMALLI *et al.* 2009; GAUTIER *et al.* 2010; 1000 GENOMES PROJECT CONSORTIUM 2010). This is especially true in the Hereford breed and these animals can be seen to clump on the outskirts of plots in the European taurine 50K analysis (Figure 4:6). This result as stated previously is likely caused by the Herefords dual role in both the subject of the bovine genome sequencing effort and as a reference breed for SNP ascertainment (GAUTIER *et al.* 2009; BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM *et al.* 2009; MATUKUMALLI *et al.* 2009), tending to produce artefactual allele frequency bias in this breed. This effect is lessened somewhat when the global cattle populations are considering and large numbers of animals are used (Figure 4:2). The greater number of markers included in the 770K chip makes this an attractive source of data to compare shotgun sequenced ancient samples too, however, this platform is also not immune to European and Hereford biases, as is seen in the 770K PCA analysis (Figure 4:7). Though this effect can be weakened by restricting the analysis to a data set with an  $r^2$  value between SNPs of less than 0.2 (Figure 4:7).

With the reduction in cost of genome sequencing and the introduction of new population genetic methods for the analysis of aDNA, for example the enhanced *D*-statistic of Meyer *et al* (2012). Low coverage next-generation sequencing of ancient and modern samples may reduce the current reliance on modern SNP datasets for the population genetic analysis of ancient samples. Allowing for biases introduced by these methods to be addressed and a greater accuracy to be achieved (SKOGLUND and JAKOBSSON 2011; MEYER *et al*. 2012).

#### **4.4.4 *D*-statistic**

The *D*-statistic of Green *et al* (2010) further suggests a measurable level of gene flow from northern European Aurochs into the genomes of taurine cattle. There is also a hint of a cline in relative affinity to CPC98, as you move from Southern to northern Europe. Though this effect may be shown to be artifactual when a greater sample of Southern European breeds is analysed. Moreover, as cautioned by Skoglund and Jakobsson (2011) an alternative explanation could be that these scores are enhanced by biases in the modern data, which are outlined below.

Firstly as CPC98 was aligned to the BosTau6 genome build, which itself is based on a Hereford animal and both the alignment and SNP calling algorithms are designed to tolerate few miss matches to the reference genome. True heterozygotes in CPC98 may drop out of the SNP calling analysis, if the alternative allele is not called at sufficient coverage, leading to an inflation in the number shared variants between CPC98 and Hereford animals. We tried to

control for this effect by only considering SNPs with coverage of 10x or greater, however, this may not be sufficient to control for all positions.

Secondly the modern SNP dataset is based on SNP sites that have been selected to vary using Herefords and other European taurines as a reference. By chance it is likely that CPC98 will contain more alternative alleles with these animals, again causing an artificial increase in the  $D$ -statistic.

The release of further full genomes and 770K genotype data from more African taurine and indicine breeds, would allow for the re-alignment and re-analysis of CPC98 to check for these biases. Moreover, this analysis is based on only one sample; the analysis of further aurochs samples from Europe, Africa and Asia will allow for a more refined picture to emerge.

## ***4.5 Conclusion***

The results of this chapter using multiple population genetic methods and both the bovine SNP 50K and 770K datasets suggest a closer relationship of a northern European aurochs to European taurine animals than those of Africa. This result suggests an introgression of wild European aurochs into the European domestic cattle gene pool. Given the limited evidence for aurochs matrilineal survival to the present day, this introgression may have been driven by males either through unintentional interbreeding or encompassing deliberate capture from the wild. This result could have profound implications when understanding the genetic history of domestic cattle, as introgression from their



wild ancestors can't be discounted. This influx of genetics from animals adapted to the conditions the newly domesticated animals were being brought into could have influenced regions target for selection in the present European bovine genomes. However, overall the results of this analysis can be thought of as a first draft. As millennia of genetic drift and population movement has occurred since CPC98 died, the cattle populations we see today in a particular geographic location are unlikely to be representative of those in that location over history, even relatively recently.

This analysis also highlights some fundamental biases in the SNP datasets when used to deduce the placement of ancient samples within the modern populations. These issues are likely to cause even grater problems when the animals studied have poorer genome coverage than that of CPC98 or are more closely related to the present day populations. This chapter therefore provides a useful test case for the analysis of aDNA samples within these datasets, namely the removal of the Hereford breed used in SNP ascertainment.

Further analyses of the domestic bovine genome, encompassing the last 10,000 years will likely reveal a picture that is as complex as the one beginning to emerge for humans (GREEN *et al.* 2010; REICH *et al.* 2010; SKOGLUND and JAKOBSSON 2011; KELLER *et al.* 2012; SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012). This work is needed if we want to understand the evolutionary forces that have shaped the genomes of the cattle populations of today, whose welfare is of great importance to global food production.

Finally this case study shows that the same tools which haven been successfully applied to ancient hominids (GREEN *et al.* 2010; RASMUSSEN *et al.* 2010; MEYER *et al.* 2012) can also be used with domesticates. As described above they have indicated that interactions may have occurred between the nascent *Bos taurus* populations and the aurochs and subject of long genetic and archaeological debate.

## **5. Evaluating exon capture for use with ancient genomes**

### **5.1 Introduction**

#### **5.1.1 Selection and the bovine genome**

The domestication of plants and animals is one of the major turning points in human evolution, for the first time these early pastoralists could directly control their access to and supply of food (BRUFORD *et al.* 2003; ZEDER *et al.* 2006; ZEDER 2008; VIGNE 2011). The subsequent 10,000 years has seen farming grow to an industrial scale, which now places huge demands on domestic species to support an ever-increasing human population (GODDARD and HAYES 2009). Almost one and a half billion cows were bred worldwide during 2010 (<http://faostat3.fao.org>); Irish farmers produced approximately 560 million tonnes of meat and milk during the same year, at a value to the economy over 3.5 billion Euros (<http://faostat3.fao.org>).

The cow has long been the subject of extensive genetic study for the purposes of breed improvement (GODDARD and HAYES 2009); a milestone was passed in 2009 with the publication of the draft bovine genome (BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM *et al.* 2009), the first from a domestic livestock species. Before the advent of genome sequencing the genetic analysis of cattle concentrated on the use of genomic markers to highlight large regions of the bovine genome hypothesised to be under selection (<http://genomes.sapac.edu.au/bovineqtl>). The introduction of low

cost, high density SNP genotyping resources for cattle, as well as cost effective genome sequencing has allowed for these regions to be explored and the causative genes identified (BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM *et al.* 2009; GODDARD and HAYES 2009; FLORI *et al.* 2009; LARKIN *et al.* 2012).

For the last 50 years intense artificial selection has been a mainstay of domestic cattle breeding, the effects of this drive to produce highly productive animals can now be seen in modern cattle populations (GODDARD and HAYES 2009). For example, the French Holstein Friesian herd of 2 million has been found to have an effective population size of roughly 50 animals (TABERLET *et al.* 2008). This heavy selection and decrease in effective population size, likely masks some older genetic markers from population genetic analysis. Ancient DNA therefore provides a useful tool for analysing bovine genomic variation through time, enabling comparisons before and after the introduction of industrial farming. Moreover, it provides the opportunity to study the effects of selection in bovids that likely experience little or no human influence.

### **5.1.2 Nuclear sequence capture and ancient DNA.**

Hybridisation capture as highlighted in chapter 3 is an emerging technique to interrogate archaeological samples. Although exon capture technologies have been used successfully in a range of modern disease studies (NG *et al.* 2009; LANDOURÉ *et al.* 2012), their use in ancient DNA is limited at present to just two (BURBANO *et al.* 2010; ÁVILA-ARCOS *et al.* 2011). This dearth of analyses is likely due to the high initial cost of these techniques, making routine analysis



of aDNA uneconomical, though prices are now falling. Furthermore, the varying preservation state of archaeological samples means that the regions targeted for capture may no longer be retrievable. A significant amount of pre-screening is therefore required before an exon targeting analysis can be undertaken. However, as methods for the production of ancient DNA sequencing libraries improve, nuclear aDNA capture is likely to become more commonplace.

| Sample  | Dating                                 |
|---------|----------------------------------------|
|         |                                        |
| CPC98   | 6,738 YBP<br>(Direct carbon date bone) |
|         |                                        |
| Ard1    | 8,630 YBP<br>(Direct carbon date bone) |
| Ard2    | 8,640 YBP<br>(Direct carbon date bone) |
| Turk11  | 8,000 YBP<br>(Dated by context)        |
| Turk13  | 8,000 YBP<br>(Dated by context)        |
| UKR08   | 7,000 YBP<br>(Dated by context)        |
| Bercy13 | 6,000 YBP<br>(Dated by context)        |

Table 5.1: Samples chosen for exon capture analysis

5.1.3 Objectives for this case study

The main objective of this case study was to assess the utility exon hybridisation capture to the field of aDNA as no studies have as yet been published using this method. This technology was evalutated by using samples chosen to represent both pre and post-domestic populations and differing bone

preservation states. Mutations discovered in the genes targeted and in off-target sequences are discussed using their significance to modern populations.

## **5.2 Methods**

### **5.2.1 Samples chosen for analysis**

Six archaeological bovid samples were chosen for a pilot exon capture analysis using Agilent Technologies “SureSelect” target enrichment protocol. This protocol as described in chapter 2 uses biotinylated RNA probes to capture sequences of interest (METZKER 2010). All samples chosen for this analysis were either directly dated or from securely dated contexts (Table 5:1). The samples selected had also performed well in previous NGS experiments, either shotgun sequencing (CPC98) or mitochondrial capture (Ardeche2, Turk11, Turk13, UKR08 and Bercy13).

The exon capture analysis was split into 3 main parts. Firstly, as CPC98 was undergoing whole genome sequencing at the time of this analysis, libraries produced for this purpose (without barcodes) were captured in two independent reactions. Ardeche2 as the best performing sample from the mtDNA analysis (Chapter 3) was also captured and sequenced individually. The remaining 4 samples (Turk11, Turk13, UKR08 and Bercy13) were pooled, captured and sequenced together, to try to evaluate the amount of DNA that could be retrieved from combined samples of differing preservation states.

### 5.2.2 Genes targeted for sequencing

In total 121 genes were included for capture, from which the exons, UTRs and 2kb of upstream sequence from the start of transcription were selected for analysis (Figure 5:1). A full list of the genes targeted can be found in appendix1:B, those selected fall into three main functional categories productivity (n=77), immunity (n=35) and pigmentation (n=9). Genes were selected for inclusion into the RNA capture after consulting relevant literature and previous studies within the lab. Those with functions in immunity, pigmentation and milking were of specific interest, as these pathways are hypothesized to be under selection early in domestication (BEJA-PEREIRA *et al.* 2003; VIGNE and HELMER 2007; SVENSSON *et al.* 2007; LUDWIG *et al.* 2009).

### 5.2.3 Bait design

RNA baits targeting these regions were designed as follows. Firstly, the RefSeq positions from the bosTau4 (Oct 2007) assembly of the bovine genome were downloaded from the UCSC genome browser (<http://genome.ucsc.edu/>). A custom Python script was then used to parse this list for genes of interest, adding 2kb of upstream sequence to search for mutations in regulatory sequences and extending exon locations 20bp into the introns to cover splice sites (Figure 5:1). These new custom positions were then uploaded to the eArray website (<http://earray.chem.agilent.com/earray>), to produce the final RNA bait sequences.

One of the design issues that has to be considered carefully when developing an exon capture experiment, is to avoid targeting repeat regions that can reduce

the overall efficiency of sequence capture, by enabling the non-specific binding of capture probes. For this analysis eArray's default settings were modified to their most conservative settings allowing for only 1bp of overlap of probes into repeat regions. For the genes selected for inclusion whose RefSeq positions were unknown, either the Bovine EST or human tblastn positions from the UCSC genome browser were used to provide the loci for analysis.

The designed probes were then visually examined, by downloading the associated BED file for each gene targeted from the eArray web service and uploading it as a custom track to the UCSC genome browser (Figure 5:1). This analysis allowed the capture design to be manually modified to accommodate problem positions. The final capture contained 57,750 probes and targeted approximately 3.30Mb of genomic sequence.

#### **5.2.4 Sequencing library preparation**

The NGS libraries used in this analysis were produced by Frauke Stock (Stock F, PhD Thesis, Trinity College Dublin Library, 2011), David Magee and Ceiridwen Edwards (EDWARDS *et al.* 2010), using a modified version of the standard Illumina protocol to account for the aDNA substrate. All samples were indexed using the Craig *et al* (2008) method of barcoding (Chapter 2). Before the capture assay, library concentrations were assessed using a Qubit<sup>TM</sup> fluorometer (Invitrogen) and samples to be combined (Turk11, Turk13, UKR08 and Bercy13) were pooled in equimolar ratios.



### **5.2.5 RNA capture assay**

As with the previous hybridisation capture experiments (Chapter 3) all manufacturers instructions were followed, except half reaction volumes were used. Genomic libraries from Ardeche2 and the pooled samples were vacuum concentrated prior to pooling and capture using a speed-vac concentrator (Savant DNA 110) to attain the required starting concentration of 250ng/ $\mu$ l. Due to the limited amount of sample available, CPC98 was captured from two aliquots with a starting concentration of roughly 145ng each. The genomic and RNA capture libraries were then incubated at 65°C for 24 hours. The captured genomic sequences were subsequently separated from the rest of the bone meta-genome using Dynabeads (280-Streptavidin, Invitrogen), prepared for RNA work following the manufactures instructions.

### **5.2.6 PCR enrichment of RNA captured library**

Three enrichment 20 $\mu$ l PCR amplifications were preformed post capture for each captured library (CPC98-1, CPC98-2, Ardeche2 and pooled) comprising 3  $\mu$ l of captured DNA, 10 $\mu$ l Phusion® high-fidelity PCR master mix with HF buffer (NEB, catalogue no. M0530S) 6.2 $\mu$ l ddH<sub>2</sub>O and 0.4 $\mu$ l each of both the forward and reverse primers (Illumina, catalogue no. FC-102-1003). Amplification reactions per library consisted of an initial denaturation step of 98°C for 30s, then 20 cycles (30 cycles for each CPC98 library) of 98°C for 10s, 65°C for 30s and 72°C for 30 s, followed by a final extension step of 72°C for 5 min.

### **5.2.7 Sequencing and assembly and SNP calling**

The 4 target enriched libraries (CPC98-1, CPC98-2, Ardeche2, Pooled Samples) were each sequenced on individual lanes of an Illumina Genome Analyzer II at the Conway Institute, University College Dublin.

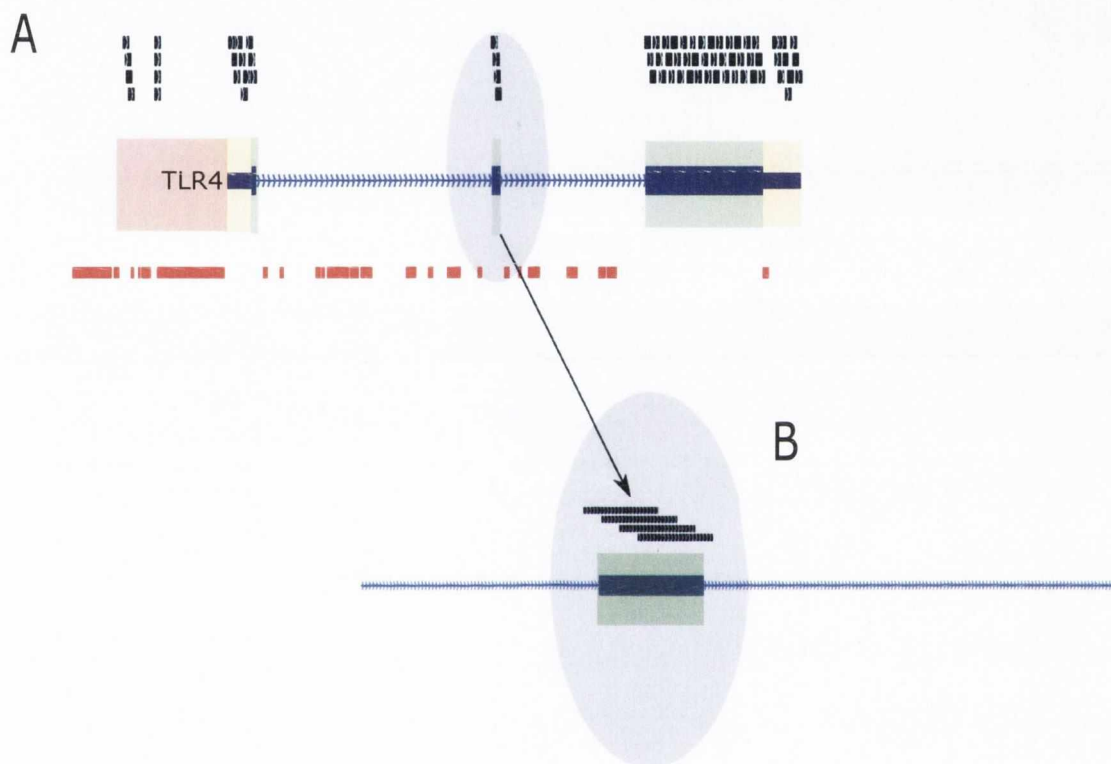
All CPC98 bosTau6 alignments were performed by Stephen Park as part of the aurochs genome-sequencing project the following describes the alignment analysis of the Ardeche2 and pooled samples. BAM files produced from the CPC98 analysis re-enter the pipeline at the variant calling and filtering stage.

All other samples were aligned to the bosTau6 reference genome following the pipeline outlined in the general methods section. Additionally SNPs were called using the SAMTOOLS mpileup command with default settings and reported in the Variant Calling Format (VCF) for ease of use in downstream analysis.

### **5.2.8 Variant Filtering**

Given the large amount of genomic sequence data targeted by exon capture studies, a high proportion of the variants identified will either be of poor coverage, or synonymous mutations with limited functional impact. It is therefore necessary to filter the raw SNP calls to identify candidate mutations, which can then be further manually curated. In this analysis the ANNOVAR (<http://www.openbioinformatics.org/annovar>) software package was used for

this purpose. The two main advantages of ANNOVAR for this study were firstly, its speed, with a typical run taking under two minutes and secondly, its ability to handle non-human reference genomes. The program was first used to identify the genomic locations of variants (Intronic, Exonic, Intragenic and UTR), and secondly to classify the effect of the discovered mutation as synonymous or non-synonymous. ANNOVAR takes as input a VCF file, which was produced from the SAMTOOLS mpileup analysis described above. The VCF file from each of the captured samples was first converted to an ANNOVAR input file using the `convert2annovar.pl` script from the ANNOVAR package. The SNP data in ANNOVAR format was then compared with the `annotate_variation.pl` script of ANNOVAR to a database of RefSeq gene positions and mRNA sequences. This database was produced from the bosTau6 (Downloaded 6-5-2012) build of the bovine genome, using the `retrieve_seq_from_fasta.pl` script of ANNOVAR. The annotation files produced from these analyses were then further manually curated and variants of interest identified for further analysis.



**Figure 5:1 RNA capture design overview.** A: Baits produced for TLR4 (Exons, UTRs, 2kb upstream) shown by black bars above the exon diagram; repeat regions marked by red bars weren't covered by baits to try to increase unique alignments. B: Zoomed view of TLR4 exon 2 a justified bait tiling design was selected to try to improve the coverage of exons and flanking sequence. Figure represents a modified UCSC genome browser output (<http://genome.ucsc.edu/>). Red highlight = 2kb upstream sequence, yellow = UTRs and green = Exons.



### **5.2.9 SMART analysis**

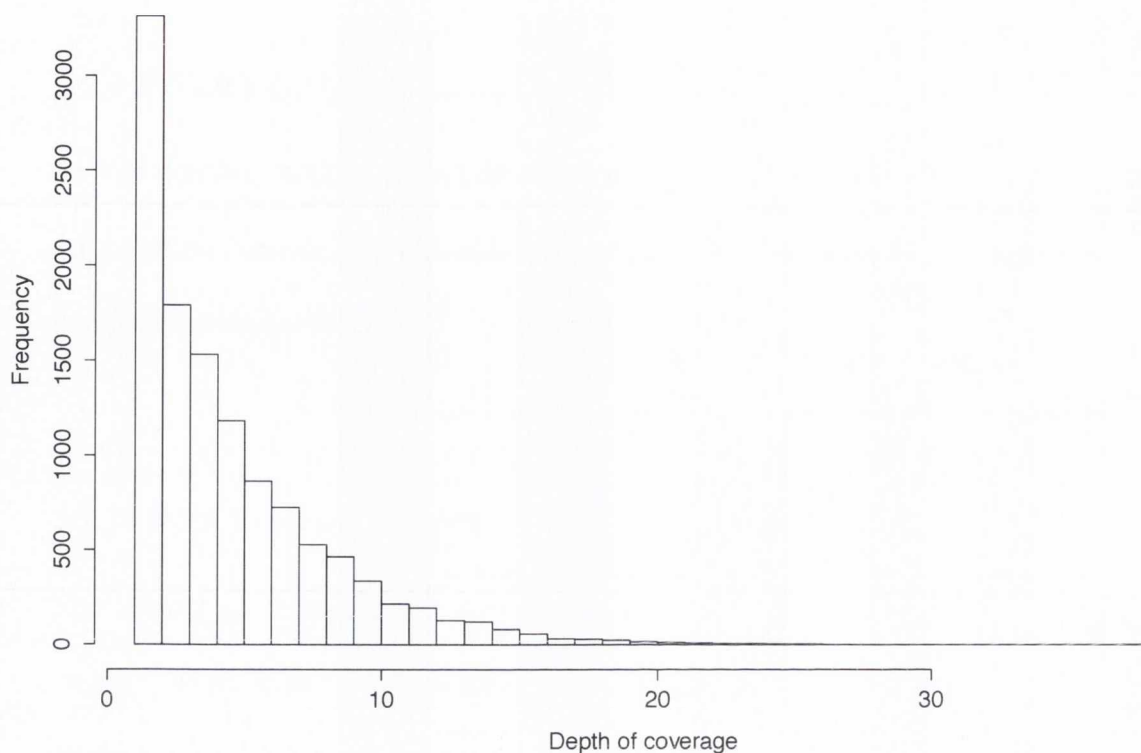
Protein domain prediction was completed using the SMART (Simple Modular Architecture Research Tool) web resource (LETUNIC *et al.* 2011), with all prediction matrixes turned on.

### **5.2.10 RNA function analysis**

The impact of mutations discovered in the 5' and 3' UTRs of the genes targeted were assessed via RegRNA (<http://regrna.mbc.nctu.edu.tw/>), a web resources for RNA function predication, using default settings (HUANG *et al.* 2006).

| Sample   | Total reads | Aligned reads | Aligned reads duplicates removed | Duplicate aligned reads (%) | Final aligned reads (%) | Regions targeted covered $\geq 1x$ (%) | Mean depth of coverage | Range | On target (%) | Adapter dimer (%) | Total sample duplicate rate (%) |
|----------|-------------|---------------|----------------------------------|-----------------------------|-------------------------|----------------------------------------|------------------------|-------|---------------|-------------------|---------------------------------|
| Ardeche2 | 14,464,431  | 243,690       | 136,187                          | 44.11                       | 0.94                    | 8.39                                   | 1x                     | 1-4   | 1.08          | 7.1               | 92%                             |
| Turk11   | 3,479,052   | 51,786        | 28,573                           | 44.82                       | 0.82                    | 0.55                                   | 1x                     | 1-2   | 1.47          | 22.1              | 91.6                            |
| Turk13   | 4,849,040   | 82,647        | 9,684                            | 88.28                       | 0.20                    | 1.94                                   | 1x                     | 1-3   | 1.00          | 31.1              | 96.1                            |
| UKR08    | 2,585,855   | 28,309        | 13,696                           | 51.62                       | 0.53                    | 1.51                                   | 1x                     | 1-3   | 0.56          | 31.4              | 84.3                            |
| Bercy 13 | 3,341,536   | 13,638        | 5,678                            | 58.37                       | 0.17                    | 0.26                                   | 1x                     | 1-1   | 0.38          | 57.6              | 93                              |
| CPC98    | N/A         | N/A           | 1,925,952                        | N/A                         | N/A                     | 75.68                                  | 3x                     | 1-53  | 2.36          | N/A               | N/A                             |

Table 5:2 : Raw sequencing alignment results all samples



**Figure 5:2: Depth of coverage histogram Aurochs (CPC98) capture (non-zero positions).**

### **5.3 Results**

The results of the exon capture analysis are presented in two parts: the first of which are the results from the analysis of CPC98, the second presents the results of Ardeche2 and the pooled samples (Turk11, Turk13, UKR08 and Bercy13).

### 5.3.1 CPC Results

Both lanes of sequencing from CPC98 were merged and analysed together to produce a final data set of 1,925,952 (C1=892,121, C2=1,033,831) aligned reads. 76% of all the regions targeted were covered to a depth of at least 1x, with a mean coverage of 3.6x (median = 2x) and a range from 1 to 53x (Figure 5:2). This allowed for the calling of 11,672 variants of which 5,041 were called to a depth of  $\geq 5x$  (1,278  $\geq 10x$ ). The high quality variants ( $>5x$  coverage, mapping quality  $\geq 30$ ) were then filtered to select only those that were located within the exons of genes, which produced a reduced set of 62 SNPs of which 18 were non-synonymous (Table 5:3) . All heterozygous variants had to have at least 3 reads from alternative alleles to be considered. These 18 mutations were then checked for concordance with the complete CPC98 genome produced from shotgun reads without capture, to assess any contamination that may have been introduced by the RNA capture. A similar analysis was also performed to look at the number of mutations discovered in the untranslated regions of the genes targeted, which produced a list of 68 variants in these regions (Appendix:1C). The read on-target percentage, defined here, as the total number of bases sequenced within the targeted regions out of the total of alignable bases was 2%. Mutations are described irrespectively of whether they were located in genes target or in off-target sequence.

### 5.3.2 Ardeche2 and Pooled samples

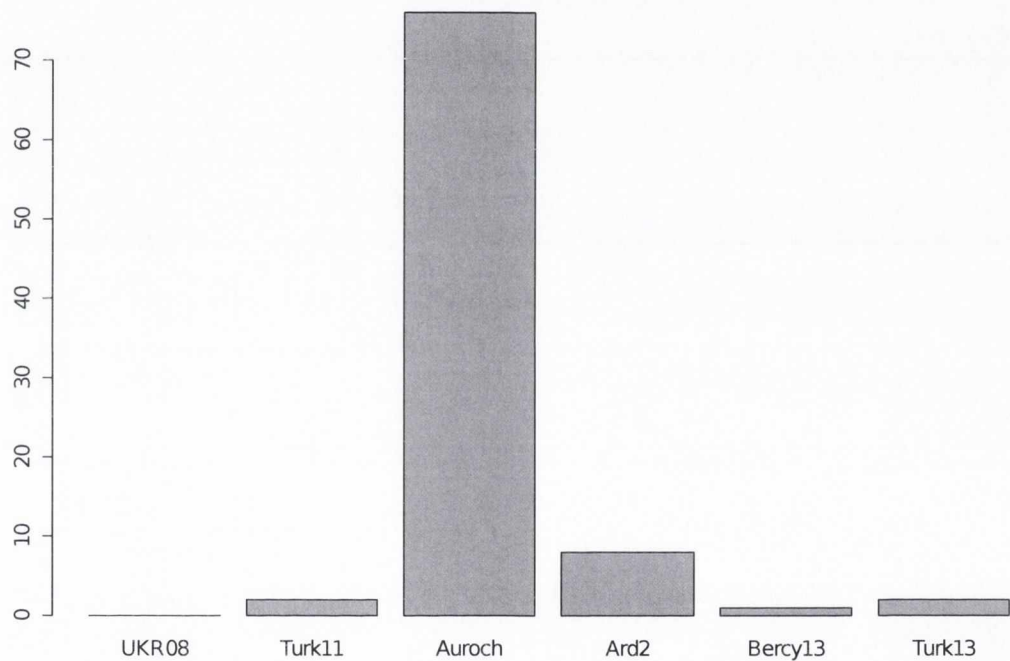
In total, sequencing of Ardeche2 produced 136,187 (0.94%) uniquely aligning reads after duplicate removal. 8% of targeted bases were covered to at least 1x in this analysis with a mean coverage of 1x and a range of 1 to 4x. From this



data 2,873 variants could be successfully typed, these raw variants were also filtered for exonic variants to produce a final set of 154 SNPs. The exonic variants were then further reduced to 12 (Table 5:4), using a less stringent quality filter than that of CPC98, given the poorer mapping percentage and sequencing quality of Ardeche2 (Mapping quality  $\geq 25$ , read depth  $\geq 2$ ). The on target percentage was almost half of that of CPC98 with only 1% of the alignable bases sequenced being inside the targeted regions. The results from the pooled samples vary from 0.8% alignment of Turk11 to a 0.2% alignment of Turk13 (Table 5:2). No non-CPC98 capture produced more than a 1% alignment. Turk13 performed the worst of all the samples tested with a high number of reads sequenced (4,849,040) but only 9,684 aligning to the bovine genome, this sample was likely hampered by its high duplicate rate of 88.28%. Turk11 was the best of the pooled sample with 28,573 (0.82%) reads aligning to the bovine genome, also likely assisted by its relatively low duplication rate of 44.82%. However, this trend was reversed in the final SNP calls with Turk13 producing 6 SNPs of greater than 1x with Turk11 only 1 (Mapping quality  $\geq 25$ , read depth  $\geq 2$  as Ardeche2). However, both of these numbers are extremely low when compared to CPC98.

Although steps were taken prior to capture to combine the pooled samples in equal-molar ratios using a Qubit fluorometer, there are clearly biases in the number of reads produced for each of the samples. UKR08 has almost a million fewer reads than the other samples with Turk13 over a million more. A high duplication rate is also seen in all samples ranging from 44% to 88%. There is also a relatively high percentage of adapter reads in all pooled samples, which are dimmers formed by the self-ligation of sequencing adapters

without an insert. Bercy13 contains the most of all the pooled samples at 57.6% with Turk11 the least at 22.1%. The final variants produced by the analysis of the pooled samples were also screened using ANNOVAR to produce a final gene list of exonic variants table (Table 5:5).



**Figure 5:3: Percentage 1x coverage of targeted bases in all samples tested.**

| Variant description | Gene: Exon: Mutation                         | Chr   | Start    | End      | WG     | Het/Hom | Coverage(X) | MQ |
|---------------------|----------------------------------------------|-------|----------|----------|--------|---------|-------------|----|
| nonsynonymous SNV   | TLR2:NM_174197:exon2:c.A631G:p.I211V,        | chr17 | 3952556  | 3952556  | Y(het) | het     | 23          | 37 |
| nonsynonymous SNV   | TLR2:NM_174197:exon2:c.G602A:p.S201N,        | chr17 | 3952585  | 3952585  | Y(het) | het     | 23          | 37 |
| nonsynonymous SNV   | TSEN15:NM_001076312:exon4:c.T395C:p.L132S,   | chr16 | 66609042 | 66609042 | Y(het) | het     | 12          | 37 |
| nonsynonymous SNV   | ZNF804B:NM_001191503:exon4:c.C1136T:p.A379V  | chr4  | 73894130 | 73894130 | Y(hom) | hom     | 12          | 37 |
| nonsynonymous SNV   | HIPK3:NM_001193007:exon8:c.G1981A:p.V661M,   | chr15 | 64752653 | 64752653 | Y(hom) | hom     | 11          | 37 |
| stopgain SNV        | BUB3:NM_001076177:exon7:c.C802T:p.R268X,     | chr26 | 43286046 | 43286046 | Y(het) | het     | 11          | 37 |
| nonsynonymous SNV   | CSN2:NM_181008:exon7:c.C245A:p.P82H,         | chr6  | 87181619 | 87181619 | Y(het) | het     | 11          | 37 |
| nonsynonymous SNV   | PCDHA13:NM_001102506:exon1:c.C1028T:p.T343I, | chr7  | 53664924 | 53664924 | Y(hom) | hom     | 10          | 37 |
| nonsynonymous SNV   | RAB3C:NM_001046606:exon4:c.A463G:p.T155A,    | chr20 | 20482430 | 20482430 | Y(het) | het     | 9           | 37 |
| nonsynonymous SNV   | TOPORS:NM_001192578:exon3:c.G709A:p.V237I,   | chr8  | 11398105 | 11398105 | Y(het) | het     | 9           | 37 |
| nonsynonymous SNV   | DNA2:NM_001192802:exon20:c.A3082G:p.R1028G,  | chr28 | 24967427 | 24967427 | Y(hom) | hom     | 8           | 37 |
| nonsynonymous SNV   | SLC9B2:NM_001193108:exon5:c.C634T:p.L212F,   | chr6  | 23148401 | 23148401 | Y(hom) | hom     | 8           | 37 |
| nonsynonymous SNV   | ATRN:NM_173995:exon15:c.G2375A:p.R792K,      | chr13 | 52084988 | 52084988 | Y(het) | het     | 7           | 37 |
| nonsynonymous SNV   | GHR:NM_176608:exon10:c.C1688T:p.T563I,       | chr20 | 31891025 | 31891025 | Y(hom) | hom     | 6           | 37 |
| nonsynonymous SNV   | MC4R:NM_174110:exon1:c.C856G:p.L286V,        | chr24 | 59670860 | 59670860 | Y(hom) | hom     | 6           | 37 |
| nonsynonymous SNV   | ACRV1:NM_001075926:exon2:c.T158C:p.V53A,     | chr29 | 29433038 | 29433038 | Y(hom) | hom     | 6           | 37 |
| nonsynonymous SNV   | CD2:NM_001011676:exon3:c.C467T:p.T156M,      | chr3  | 26593559 | 26593559 | Y(hom) | hom     | 6           | 33 |
| nonsynonymous SNV   | TICAM1:NM_001030301:exon1:c.A1028C:p.Q343P,  | chr7  | 20548991 | 20548991 | Y(hom) | hom     | 6           | 37 |

Table 5:3: Non-synonymous variants discovered during the CPC98 exon capture analysis. WG = mutation checked in CPC98 whole genome sequence with captured reads removed. Het/hom = SNP is heterozygous or homozygous as assessed by ANNOVAR. MQ = mapping quality of variant encoded as a phred quality score as assessed by ANNOVAR.



| Variant           | Gene: Exon: Mutation                         | Chr            | Start    | End      | Het/Hom | Coverage (X) | MQ |
|-------------------|----------------------------------------------|----------------|----------|----------|---------|--------------|----|
| synonymous SNV    | LEP:NM_173928:exon3:c.G240A:p.A80A,          | chr4           | 93263980 | 93263980 | het     | 3            | 37 |
| nonsynonymous SNV | GTPBP5:NM_001014874:exon8:c.T1343G:p.L448R,  | chr13          | 55531471 | 55531471 | hom     | 2            | 37 |
| nonsynonymous SNV | GNB1:NM_175777:exon6:c.G310A:p.A104T,        | chr16          | 52172898 | 52172898 | het     | 2            | 37 |
| nonsynonymous SNV | G0S2:NM_001192147:exon2:c.C283A:p.L95M,      | chr16          | 75539871 | 75539871 | hom     | 2            | 17 |
| synonymous SNV    | TBX5:NM_001192749:exon1:c.A12G:p.A4A,        | chr17          | 62532629 | 62532629 | hom     | 2            | 25 |
| nonsynonymous SNV | PHOSPHO1:NM_001193119:exon4:c.C640A:p.L214M, | chr19          | 37984200 | 37984200 | het     | 2            | 26 |
| synonymous SNV    | STAT3:NM_001012671:exon9:c.A909G:p.P303P,    | chr19          | 43074278 | 43074278 | hom     | 2            | 37 |
| nonsynonymous SNV | MUTED:NM_001098875:exon5:c.T529C:p.Y177H,    | chr23          | 47268144 | 47268144 | het     | 2            | 32 |
| synonymous SNV    | IGSF9:NM_001205532:exon7:c.T879C:p.P293P,    | chr3           | 9869511  | 9869511  | hom     | 2            | 25 |
| synonymous SNV    | RAB3A:NM_174446:exon5:c.T513C:p.N171N,       | chr7           | 4949197  | 4949197  | het     | 2            | 26 |
| nonsynonymous SNV | PCYOX1L:NM_001075673:exon1:c.T41G:p.L14R,    | chr7           | 62752534 | 62752534 | het     | 2            | 26 |
| nonsynonymous SNV | RNH1:NM_001035319:exon2:c.G199C:p.G67R,      | chrUn_GJ060032 | 7685     | 7685     | het     | 2            | 26 |

Table 5:4: Results of the Ardeche2 exon capture analysis, exonic variants retrieved ≥ 1x. Het/hom= SNP is heterozygous or homozygous as assessed by ANNOVAR. MQ = mapping quality of variant encoded as a phred quality score as assessed by ANNOVAR.



| Variant | Gene: Exon: Mutation                         | Chr   | Start     | End       | Het/Hom | Coverage (X) | MQ |
|---------|----------------------------------------------|-------|-----------|-----------|---------|--------------|----|
| Turk13  | ZER1:NM_001205808:exon4:c.T554C:p.M185T,     | chr11 | 99270799  | 99270799  | het     | 2            | 32 |
|         | TMEM55A:NM_001040594:exon1:c.T28A:p.S10T,    | chr14 | 75708393  | 75708393  | het     | 2            | 26 |
|         | KLHL25:NM_001192686:exon1:c.C1227G:p.V409V,  | chr21 | 16959193  | 16959193  | het     | 2            | 32 |
|         | ATP9B:NM_001080255:exon1:c.G35T:p.S12I,      | chr24 | 1265068   | 1265068   | het     | 2            | 32 |
|         | MARVELD1:NM_001101262:exon1:c.C324G:p.L108L, | chr26 | 18733319  | 18733319  | het     | 2            | 27 |
|         | CDK13:NM_001205431:exon1:c.A790G:p,          | chr4  | 81806761  | 81806761  | het     | 2            | 37 |
| Turk11  | CYP27B1:NM_001192284:exon2:c.A359T:p.Q120L,  | chr5  | 56025233  | 56025233  | het     | 2            | 26 |
|         | CIH3orf58:NM_001205651:exon1:c.C296G:p.P99R, | chr1  | 125884672 | 125884672 | hom     | 2            | 37 |

Table 5.5: Results of the Pooled exon capture analysis, exonic variants retrieved  $\geq 1x$ . Bercy13 provided no exonic variant. Het/hom= SNP is heterozygous or homozygous as assessed by ANNOVAR. MQ = mapping quality of variant encoded as a phred quality score as assessed by ANNOVAR.

## **5.4 Discussion**

### **5.4.1 Aurochs capture analysis**

The exon capture of CPC98 was a success, with over three quarters of the regions selected for targeting successfully retrieved in this analysis. The number of variants discovered was also large over 5,000 with a depth of 5x or greater. It should be noted that as with most SNP discovery experiments, the vast majority of these variants were synonymous and likely of little functional impact. CPC98 is comfortably dated to the British Mesolithic and is therefore a representative of a European aurochs population that likely experienced little human intervention, apart from predation via hunting. The results of the CPC98 exon capture analysis therefore provides a number of interesting target mutations that may offer clues to bovid evolution pre and post domestication.

### **5.4.2 CPC98 TOLL like receptors**

A number of genes in the TOLL like receptor pathway were discovered to harbour either heterozygous or homozygous non-synonymous mutations (TLR2,TLR10,TICAM1) in this analysis. The heterozygous I211V mutation (ss104796312) discovered in the *TLR2* gene of CPC98, has also been found previously in both domestic taurine and indicine cattle and confidently predicts a low complexity protein domain in SMART analyses (SEABURY and WOMACK 2008). Low complexity regions (LCR) are hypothesised to play a variety of roles in protein structure and interactions, with LCR containing

proteins represented strongly in hub nodes of biological networks (COLETTA *et al.* 2010). That this mutation was found to be conserved from *Bos taurus* to *Bos indicus* but missing from four other mammalian species (Human, Mouse, Rat and Dog) sequenced by Sebury and Womack (2008), suggests that this region may have a role in the bovine immune response. That this variant is missing from the bovine reference sequence for TLR2 (NC\_007315) is interesting, given its general conservation. This anomaly may be indicative of a *De novo* mutation in this animal or limited on-going selection at this locus given its heterozygous state in other cattle populations. However, it could also be the result of poor sequence quality at this location in the genome build.

The second high confidence heterozygous non-synonymous polymorphism identified in the aurochs *TLR2* gene was S201N (rs110491977). SMART analysis of the bovine TLR2 protein sequence including this polymorphism, increases the predicted likelihood of a Leucine-rich repeats (LRR) region in the protein (0 to 4.30e02). This region was also found to be bovine specific by Sebury and Womack (2008), but is again missing from the bovine reference sequence. LRR are proposed to have roles in antigen binding and therefore this mutation may have a positive effect on pathogen recognition. Moreover, the S201N mutation has been found to be fixed in 3 British breeds (Highlands, Jersey and South Devon) tested by Mariotti *et al* (2009).



### 5.4.2 CPC98 other genes and pathways highlighted

Melanocortin receptor 4 (*MC4R*) is a well known as a marker for hereditary obesity, with roles in reproductive pathways also hypothesised (VAN DER PLOEG *et al.* 2002). The homozygous L286V mutation (rs108968214) discovered in the *MC4R* gene of CPC98 was previously shown to vary in Korean Hanwoo cattle (SEONG *et al.* 2011). The homozygous non-reference state is thought to be responsible for a significant increase in the muscle marbling score in these animals (SEONG *et al.* 2011). Sequencing of further ancient *Bos primigenius* and domestic *Bos taurus* animals would elucidate if this mutation varied in ancient populations, or whether the variation we see at this locus today is young.

The attractin (*ATRN*) gene of CPC98 was also found to contain a novel heterozygous non-synonymous mutation in exon 15 (R792K). *ATRN* is hypothesized to play an important role in coat colour (SEO *et al.* 2007) and has been identified in a quantitative trait loci for the UV-protective eye area pigmentation (PAUSCH *et al.* 2012). Pigmentation is suggested to be an early target of positive selection in horses and cattle (SVENSSON *et al.* 2007; LUDWIG *et al.* 2009) making variation in this protein an interesting target for further study.

Finally a diagnostic mutation (P82H) in the beta casein gene (*CSN2*) was also discovered in a heterozygous state in CPC98. This SNP (rs43703011) can distinguish between the natural variants of beta casein A1, B, C, F and G from



A2 (KAMIŃSKI *et al.* 2007), suggesting that CPC98 carried the two major types. Both taurine and indicine breeds have been found to differ in the parentage of beta casein variants their milk contains, with some variants almost fixed in certain populations (KAMIŃSKI *et al.* 2007). Diversity within this locus in both *Bos taurus* and *Bos indicus* animals, suggests that these mutations are of long standing in all cattle populations and occurred prior to domestication. An increase in the diversity at the casein locus in cattle from European countries with high incidence of lactase persistence, has been proposed as a marker for gene culture evolution between humans and cattle (BEJA-PEREIRA *et al.* 2003). Further studies in a range of animals from aurochs to early and late domestics would elucidate if this increase in variation can be dated in a similar way to that of the reciprocal lactase persistence mutations in humans (BURGER *et al.* 2007; ITAN *et al.* 2009).

#### **5.4.3 CPC98 UTR analysis**

As described previously ANNOVAR analysis of the UTRs sequenced in the exon capture identified 68 high quality SNPs ( $\geq 5\times$  coverage, mapping quality  $\geq 30$ ) (Appendix1:C). The 3' UTR of genes are well known to contain regulatory sequences with important roles in protein synthesis and transcript localisation (KUERSTEN and GOODWIN 2003). A number of genes had multiple mutations identified in their 3' UTRs (Appendix1:C), one of which was leptin (*LEP*) (n=5), variants of which are thought to be causative for obesity in humans (STROBEL *et al.* 1998). Given leptins important dietary role in Humans, it has been widely studied in cattle and is thought to effect body fat deposition, feed and energy use (LUSK 2007; WYLIE 2010; WORONUK *et al.* 2012; AKHTER *et al.* 2012). Modifications in the 3'UTR of this protein have been described in

humans, some of which are significantly associated with hypertension (AKHTER *et al.* 2012). However the analysis of the 3' UTR of the leptin gene in bovids is at present unpublished. *In silico* analysis of the 3' UTR of leptin including these mutations however failed to find any functional variants. Other production genes identified with multiple 3'UTR mutations include myostatin (*MSTN*) and two casein genes (*CSNIS2* and *CSN3*).

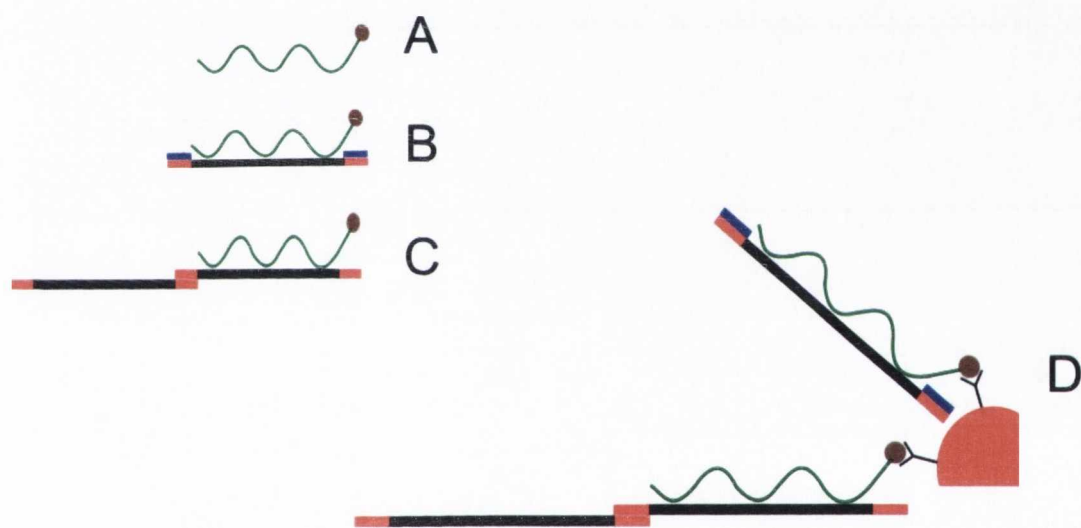
Two genes found to have high quality variants in their 5' UTRs were *HTR2A* a serotonin receptor and alpha-S2-casein *CSNIS2*. The heterozygous 5'UTR T>C mutation of CPC98's *CSNIS2* 5' UTR at position 87264294 on chromosome 6 has recently been described in Norwegian Red cattle (SODELAND *et al.* 2011). The SNP -5bp from the start of transcription is located within the Kozak sequence and is therefore proposed to effect the efficiency of *CSNIS2* transcription (SODELAND *et al.* 2011). Moreover, the associated haplotype in Norwegian Red cattle is found to correlated with milk protein yield (SODELAND *et al.* 2011). The 5'UTR sequences for all four casein genes are highly conserved both within cattle and with other ruminants (BEVILACQUA *et al.* 2006; SODELAND *et al.* 2011). The T>C variant was found by Sodeland *et al* (2011) to more closely resemble the Kozak consensus sequence for all 4 casein genes and is therefore likely to be the ancestral state. That this mutation is found in a heterozygous state in both the aurochs and domestic cattle and on a haplotype under selection in Norwegian Red cattle (SODELAND *et al.* 2011). Suggests that this SNP may be an interesting marker for artificial selection of an ancestral allele, if its causative effects can be biochemically proven.

The second 5UTR mutation of interest was a novel homozygous mutation in the 5'UTR of *HTR2A* a serotonin receptor, however, *In silico* analysis again proposed no functional role for this mutation. Given the widely reported differences in aggression between aurochs and domestic cattle, genes with proposed roles in docility may be a fruitful line of aDNA enquiry.

#### 5.4.4 Off target sequence

CPC98s low on-target coverage of roughly 2% is a concern and likely reflects the composition of aDNA sequencing libraries, which are known to consist of short insert fragments of roughly 40-80bp (EDWARDS *et al.* 2010). This short fragment length is likely to impede the efficient hybridisation of the 120bp RNA capture probes, which are designed for modern DNA sequencing libraries with inserts in the range of 200bp. This problem may have also been exacerbated by the inability to adequately block the ligated sequencing adapters during capture. These un-blocked sequences will have provided areas of homology between a single library molecule and any other non-blocked molecule in the library as a whole (Figure 5:4, C and D). Hybridisation (Daisy chaining) of these adapter sequences, could lead to unintentional capture of unwanted sequence (Figure 5:4, D). Moreover, this daisy chaining of unblocked sequence could also be one way that adapter dimers are retained post capture. Anecdotal evidence suggests that the Craig *et al* (2008) method of barcoding samples may also reduce capture efficiency; by introducing non-homologous sequences either side of the genomic insert. This is likely to be of greater concern with an ancient DNA library of 40-80bp insert plus 12bp (6bp each side) of flanking barcodes, than it would be in a modern DNA library with a 200bp insert. Finally the large number of PCR cycles (30x) post-capture

could also compound this problem, with off targeted sequences that might have persisted throughout the capture at low concentrations amplified to sequenceable levels. However, the 2% on-target percentage can still be considered an enrichment of roughly 200x, when coverage is assessed at the same number of random genomic locations the on target percentage is 0.01%.



**Figure 5:4:** Cartoon of RNA capture with daisy chained baits. Green line = RNA sequence, red line = NGS adapter sequence, brown circle = biotin, blue line = DNA adapter blocks, red oval streptavidin coated bead. [A] RNA probe, [B] Adapter sequences (red) correctly blocked (blue) no daisy chaining. [C] Baits incorrectly blocked with daisy chaining. [D] Bead capture of both the correctly and incorrectly blocked DNA/RNA molecules.

### 5.4.5 Contamination

Exon capture is hypothesised to increase exogenous contamination when compared to pure shotgun sequencing, as the RNA probes have the ability to target non-endogenous sequences that are homologous to the regions selected



for capture. To assess if the high quality non-synonymous SNPs (depth  $\geq 5$ , mapping quality  $\geq 30$ ) called in this analysis (especially the heterozygous positions) could be attributed to exogenous contamination, all positions were independently corroborated using shotgun reads from the aurochs genome project (Chapter 4). This analysis failed to identify any of the variants as being RNA capture specific; the authenticity of variants below these stringent cut offs and in samples of poorer sequence coverage is likely to be less discernable. Some of the SNPs identified were also known mutations with variance in indicine breeds suggesting that they are real and a feature of the bovine genome.

#### **5.4.6 Ardeche2**

The results of the exon capture analysis for Ardeche2 were markedly less successful than that of the CPC98, which is not surprising given the difference in preservation states of these samples. However, Ardeche2 did perform extremely well in the previous mtDNA capture experiment and it was hoped that this would continue to nuclear capture. Subsequent shotgun analysis of Ardeche2 has shown this sample to contain only a moderate amount of nuclear DNA (Chapter 3) compared to the other samples tested in this analysis. This sample was further hampered by an initially low library concentration post capture, which required 20 rounds of PCR to generate a concentration high enough to sequence, this in turn led to a high duplication rate of 44.11% (Table 5:2). Only 8% of the regions targeted in Ardeche2 were retrieved and 12 SNPs called at greater than 2X. At this level of coverage it is hard to distinguish true sequence polymorphisms from nucleotide changes produced by PCR and sequencing error, as well as aDNA degradation and exogenous

contamination (MEYER *et al.* 2012). The only SNP called at greater than 2x in exonic regions was a synonymous mutation in the leptin (*LPT*) gene. Ardeche2 also suffers with a very low on target percentage of 1.08%. No mutations in either of the UTR regions were discovered at greater than 1x and are not reported.

Given the low percentage of retrieved sequence, it was hypothesised that sequencing adapter dimer could be affecting the amount of data produced in this analysis in a similar fashion to that of the mtDNA captures (Chapter 3). Dimer percentage was initially estimated using CUTADAPT at only 7.1%. However, subsequent metagenomic analysis of both aligned and unaligned reads found 91% of Ardeche2's reads to have no match in the NCBI nucleotide database. Suggesting that experimentally introduce adapter and PCR artefacts could account for a large percentage of the unmapped reads. Ardeche2 also has an excessively high total duplication rate of 92% (including bacterial contamination), reflecting the low starting concentration of the sample and the need for multiple rounds of PCR.

#### **5.4.7 Pooled samples**

The results of the combined analysis were also disappointing with relatively few positions and SNPs called in each library, none of which were greater than 2x. Subsequent shotgun analysis of three of these samples UKR08, Turk11 and Turk13 found two Turk11 and Turk13 to contain significant amounts of endogenous DNA (Chapter 6). These combined results highlight one of the main issues concerning hybridisation capture and ancient DNA; that many of

the regions targeted may no longer be retrievable and a high initial DNA concentration is a must. Newer library preparation protocols that take into account the limitations of ancient DNA, may improve these results and allow more of the targeted regions to be retrieved (MEYER *et al.* 2012). The low on target coverage for these samples is also disappointing although Turk11 outperforms Ardeche2 with 1.47% of reads on target. This reflects the high endogenous DNA content in this sample found in later analyses. Again no UTR variants for any of the pooled samples were found at sufficient depth or quality to be reported. Unlike Ardeche2 these samples were found to contain a high number of adapter dimer in their sequencing reads with CUTADAPT in region of 22.1 to 57.6%. FastQC also estimated a high total duplication rate (including bacterial contamination) for all these samples from 84 to 96%.

Ardeche2 didn't significantly outperform the pooled samples in terms of percentage alignment even though it was sequenced individually. However, it did have a larger percentage of the genome covered with 8% as opposed to the pooled samples none of which none had a coverage greater than 2% (Figure 5:3). However, it is likely that these percentages would improve if these samples were also sequenced independently.

## 5.5 Conclusions

Exon capture has been is as an emerging technique that could have massive implications for the successful interrogation of archaeological samples. This promise has to be qualified by the specific challenges that aDNA presents, both in the low starting concentration of DNA in archaeological samples and the



fragile nature of the aDNA molecule. As shown by the discrepancy between the results of the samples selected for analysis in this case study. These deficiencies are unlikely to be address by the alternative RNA based capture technologies available at present for example MYbaits from Mycroarray ([www.mycroarray.com](http://www.mycroarray.com)) which uses the same RNA probe capture method as this chapter. Newer DNA based capture methods designed for aDNA using custom biotinylated PCR products are also intractable at present as they target too few base pairs to be considered for exon capture (MARICIC *et al.* 2010). These problems along with the initial high cost of this technique have stopped nuclear capture from becoming a regular part of the aDNA tool kit, in the way that similar techniques for mtDNA have (MARICIC *et al.* 2010) become (BRIGGS *et al.* 2009; SCHUENEMANN *et al.* 2011). Improvements to extraction methods and library preparation protocols are likely to be the only effective ways of improving these types of analyses long term. That this method has been used extensively for the capture of mtDNA genomes (SCHUENEMANN *et al.* 2011; KNAPP *et al.* 2012; FU *et al.* 2013) (chapter 3) gives hope that this will be possible in the future. These problems are beginning to be addressed by novel new extraction and library preparation methods (MEYER *et al.* 2012).

The aim of this case study was to assess the power of nuclear hybridisation capture for study of aDNA samples. This chapter shows this technique to be a qualified success with the analysis of CPC98 proving largely successful and identifying multiple known and novel variants that warrant further study. The use of surrounding un-translated regions in the capture analysis allowed for the identification of non-coding variants, one of which was subsequently identified be under selection in modern cattle populations. These results show the power



of this technique to be used in aDNA to enable genes important to production traits and immunity in modern day cattle to be followed over time. However, CPC98 is an especially well-preserved sample and the results from other samples were less impressive. Out of the 5 samples tested only one had a variant called at greater than 3X. This result may reflect the limits of the protocols used at the time of this analysis, which have subsequently been improved to allow shotgun sequencing of Turk11 and Turk13 to retrieve 1-4% endogenous DNA (Chapter 6). Moreover these results may highlight the limit of this technique when applied to aDNA, with all samples, regardless of the their preservation state, showing only 1 to 2% on target sequence; albeit that this still classes as enrichment of over 100x.

In summary exon capture has the potential to follow alleles through archaeological time and for selection to be observed in temporal transects, without the need to sequence full genomes. At this time however, given the results of this case study it is likely that these analyses will be reserved for the best preserved of the samples subjected to ancient DNA analysis.

## **6. Genomic affinities of two 8,000-year-old *Bos taurus* animals from Anatolia**

### **6.1 Introduction**

#### **6.1.1 The domestic cattle of Anatolia**

Archaeological evidence suggests that domestic *Bos taurus* cattle were present in Anatolia from approximately 8,500BP (ZEDER 2008). It is likely that these populations were tightly managed, with evidence of complex animal husbandry practices for example dairying, present in the region from the seventh millennium BC (VIGNE and HELMER 2007; EVERSLED *et al.* 2008). However, genetic studies have shown the current Anatolian cattle populations to be hybrids of both taurine and indicine animals (LOFTUS *et al.* 1999; EDWARDS *et al.* 2007b) and indicate that this admixture is likely to be ancient (FREEMAN *et al.* 2006b). Archaeological data also provides further evidence for this assertion by suggesting Zebu animals were present at least in Jordan from the fourth millennium BC onwards (CLASON 1978; EDWARDS *et al.* 2007a).

#### **6.1.2 Shotgun sequencing of ancient DNA**

Next generation shotgun sequencing of aDNA extracts has become the *de facto* standard for the genome scale analysis of archaeological specimens, with four ancient genomes already published using these techniques (MILLER *et al.* 2008; GREEN *et al.* 2010; RASMUSSEN *et al.* 2010; REICH *et al.* 2010). However, as

discussed in previous chapters most archaeological samples contain much less DNA than would be required to make whole genome sequencing an economic possibility (KNAPP and HOFREITER 2010). Other methods for the genomic analysis of aDNA are therefore required to adequately return the investment (both economic and in the destructive sampling of artefacts), which is needed for the production of aDNA sequencing libraries.

One-way to accomplish a gain in sequence coverage is through the use of hybridisation capture as described in chapters 3 and 5, though these methods have been shown to be successful for mtDNA and bacterial genomes (BRIGGS *et al.* 2009; SCHUENEMANN *et al.* 2011; BOS *et al.* 2011) (Chapter3); their use for the analysis of ancient nuclear genomes may be limited at present to well preserved samples (BURBANO *et al.* 2010; ÁVILA-ARCOS *et al.* 2011)(Chapter 5). Moreover, only a small percentage of a genome can ever be successfully sequenced using current DNA capture technologies, with this effect likely to be exacerbated in archaeological samples where the regions targeted may no longer be recoverable or available in such small concentrations that the resultant sequencing library has to be heavily amplified, leading to a reduction in overall coverage.

### **6.1.3 Objectives for this case study**

The analysis of haploid SNP data from low coverage shotgun sequencing is an emerging technique for the study of archaeological samples, these methods have already enabled the modern European populations who most closely resemble their Mesolithic and Neolithic counterparts to be elucidated

(SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012). In this chapter they are applied to two 8,000-year-old *Bos taurus* samples from western Anatolia.

**6.2 Methods**

**6.2.1 Samples analysed in this study.**

Two samples from a single archaeological site in western Turkey (Menteşe) dated to approximately 8,000YBP were used in this analysis (Table 6:1).

| Sample | Sample location | Age of sample (YBP)          |
|--------|-----------------|------------------------------|
|        |                 |                              |
| Turk11 | Turkey          | 8,000 YBP (Dated by context) |
| Turk13 | Turkey          | 8,000 YBP (Dated by context) |

**Table 6:1: Turkish Samples chosen for shotgun analysis**

**6.2.2 Ancient DNA extraction and library preparation**

All aDNA extractions and library preparations were completed following the protocols described in the general methods and sequenced together on a single lane of an Illumina HiSeq2000 at the Beijing Genomics Institute.

**6.2.5 BWA sequence Alignment**

The raw reads were then trimmed split and aligned to the *Bos taurus* reference genome (bosTau6, downloaded 19-4-2012) following the standard alignment



protocol also in the general methods section of this thesis (chapter 2)(LI and DURBIN 2009), the mitochondrial genome of bosTau6 was also replaced with the Anderson (V00654) reference sequence for ease of comparison (ANDERSON *et al.* 1982).

### **6.2.7 Alignment to human genome**

All sequencing reads were further aligned to the human genome (hg19) using BWA with identical parameters to the sheep (OAR2) alignments above. Any reads that aligned to the human genome were then removed from subsequent analysis irrespective of mapping quality.

### **6.2.8 SNP calling, Eigensoft and Procrustes analysis Analysis**

SNP calling and Eigensoft/Procrustes analysis was completed following established protocols for 1x aDNA NGS data (SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012) as general methods (chapter 2) with the following modifications. Plink and Eigensoft were passed the flags required ‘–cow’ and ‘numchrom = 29’ respectively to recognise the extra autosomes present as compared to humans in bovine genotype data.

### **6.2.7 770K bovine SNP dataset**

The 770K SNP dataset of locally held samples as described in chapter 4 was also used for the analysis of the Turkish samples described above. A/T and G/C SNPs were removed when combining data from multiple datasets and genotypes where required were flipped using PLINK to be in the bosTau6

orientation, SNPs with a MAF  $\leq 0.5\%$  were excluded leaving a final dataset of 719,758 genotypes.

|        | Raw reads  | Aligned reads<br>mapping<br>quality $\geq 30$ | Aligned reads<br>mapping<br>quality $\geq 30$<br>(% of total<br>reads) | Reads<br>remaining<br>after<br>duplicate<br>removal<br>(SAMtools<br>rmdup) | Reads<br>remaining<br>after<br>duplicate<br>removal (% of<br>total reads)<br>(SAMtools<br>rmdup) | Reads<br>remaining after<br>filters were<br>applied<br>(Human reads<br>removed<br>XT,X1 tags) | Reads<br>remaining<br>after filters<br>were<br>applied<br>(%) | SNPs<br>called in<br>770K<br>dataset |
|--------|------------|-----------------------------------------------|------------------------------------------------------------------------|----------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|---------------------------------------------------------------|--------------------------------------|
| Turk11 | 16,561,501 | 326,702                                       | 1.97                                                                   | 312,794                                                                    | 1.89                                                                                             | 305,242                                                                                       | 1.84                                                          | 2,296                                |
| Turk13 | 16,849,508 | 771,387                                       | 4.58                                                                   | 740,731                                                                    | 4.40                                                                                             | 723,535                                                                                       | 4.29                                                          | 5,975                                |

Table 6:2: Sequencing results for the 2 bovine samples studied.

## **6.3 Results**

### **6.3.1 Sequencing Results**

In total 33,411,009 (Turk11=16,561,501, Turk13=16,849,508) 49bp reads were generated by this study once adapter reads were removed, both samples were found to contain significant amounts of bovine DNA, 1.84% for Turk11 and 4.29% for Turk13 (mapping quality  $\geq 30$ ). These alignment rates correlate well with the results of a previous mtDNA RNA capture experiment (Chapter 3), which produced completed mtDNA genomes for each sample. These reads covered 12,851,982 and 30,390,084 positions of the bovine genome (bosTau6) in Turk11 and Turk13 respectively, which corresponds to a retrieval of 0.48% of the cattle genome for Turk11 and 1.15% for Turk13.

### **6.3.2 Mitochondrial genomes**

The sequencing results from both Turk13 and Turk11 provided sufficient mitochondrial genome coverage to allow for comparisons with the consensus sequence produced by the capture methods described above (Turk11 = 87% and Turk13 = 92%). For Turk11 we were able to confidently corroborate 75% (6/8) of the SNPs and identify no further mutations; for the two remaining positions the 1x base called at these locations in this study also matched that of the capture consensus. For Turk13 we could reproduce 58.6% (17/29) of the SNPs and again identify no further mutations, 4 captured genotypes could also be identified at 1X raising the final percentage assessed to 72% (21/29). These



SNPs also allowed for an independent haplotype assignments of both samples, which repeated the (*Bos taurus*) T3 and Q haplotypes assigned to Turk11 and Turk13 by the capture analysis (Chapter 3).

The SNPs called in the mtDNA analysis above also allowed for a rough estimation of the contamination rate in these samples to be calculated (both ancient and modern). To do this heterogeneities at SNP locations in both samples that could be contamination, sequencing error, hetroplasmy or damage were analyzed (SÁNCHEZ-QUINTO *et al.* 2012). At these positions the consensus base was found in 100% (13/13) of sequences for Turk11 and 98.6% (68/69) of sequences for Turk13. Giving estimated contamination rates of 0% and 1.4% for Turk11 and Turk13 respectively. These figures are markedly lower than a study by Sanchez-Quinto *et al* (2012) of a complete ancient human mitochondria, which estimated exogenous contamination at 8%. These figures compare to 1.2% (254/257) and 1.4% (582/590) when the contamination rates for the captured genomes are considered for Turk11 and Turk13 (mean coverage Turk11 44x and Turk13 22x), here the higher sequence coverage of Turk11 produced a more realistic contamination rate of 1.2%, interestingly, the captured estimate for Turk13 (1.4%) matches that of the shotgun sequencing.

### **6.3.3 MAPDAMAGE analysis**

As a further test of the authenticity of both samples all sequences retrieved from Turk11 and Turk13 were submitted to a MAPDAMGE (version 0.3.3.pl)

analysis, which indicated DNA fragmentation patterns suggestive of ancient DNA (Appendix1:D).

| Position | Anderson genotype (V00654) | Consensus base capture | Consensus base shotgun | Haplotype distinction |
|----------|----------------------------|------------------------|------------------------|-----------------------|
|          |                            |                        |                        |                       |
| 169      | A                          | G                      | G                      | T3b → T3              |
| 363      | C                          | G                      | (1G)                   | ?                     |
| 587      | -                          | +C                     | +C                     | BRS → T3b             |
| 1594     | -                          | -A                     | -A                     | ?                     |
| 2536     | C                          | A                      | A                      | BRS → T3b             |
| 9682     | G                          | C                      | (1C)                   | BRS → T3b             |
| 13310    | A                          | C                      | C                      | BRS → T3b             |
| 14043    | G                          | A                      | A                      | ?                     |

Table 6:3: Turk11 SNPs capture and whole sample shotgun, haplotype distinctions as Achilli *et al* (2009), local mutations marked with (?) . For positions with insufficient coverage to call SNPs 1X genotype data is shown in brackets.

| Position | Anderson<br>genotype<br>(V00654) | Consensus<br>base capture | Consensus<br>base shotgun | Haplotype<br>distinction |
|----------|----------------------------------|---------------------------|---------------------------|--------------------------|
| 169      | A                                | G                         | -                         | T3b → T3                 |
| 587      | -                                | +C                        | (1+C)                     | BRS→T3b                  |
| 1457     | G                                | A                         | A                         | ?                        |
| 2536     | C                                | A                         | A                         | BRS→T3b                  |
| 2558     | G                                | A                         | A                         | T → QT                   |
| 3238     | T                                | C                         | (1T,1C)                   | QT → Q                   |
| 3415     | A                                | T                         | T                         | QT → Q                   |
| 4319     | T                                | C                         | C                         | ?                        |
| 5501     | C                                | T                         | T                         | T → QT                   |
| 7918     | C                                | T                         | T                         | QT → Q                   |
| 8318     | T                                | C                         | -                         | QT → Q                   |
| 8370     | A                                | C                         | (1C)                      | T → QT                   |
| 9682     | G                                | C                         | (1C)                      | BRS → T3b                |
| 10717    | T                                | A                         | -                         | QT → Q                   |
| 10927    | T                                | C                         | C                         | QT → Q                   |
| 11000    | C                                | T                         | (1T)                      | T → QT                   |
| 11089    | T                                | C                         | C                         | QT → Q                   |
| 12433    | T                                | C                         | -                         | QT → Q                   |
| 12468    | T                                | C                         | C                         | T → QT                   |
| 12675    | C                                | T                         | T                         | T → QT                   |
| 12750    | T                                | C                         | C                         | T → QT                   |
| 13005    | A                                | G                         | G                         | T1'2'3 → T               |
| 13310    | A                                | C                         | -                         | BRS → T3b                |
| 14036    | G                                | A                         | A                         | T → QT                   |
| 14108    | T                                | C                         | C                         | QT → Q                   |
| 15134    | C                                | T                         | T                         | T → QT                   |
| 15627    | A                                | G                         | G                         | Q1a                      |
| 15953    | C                                | G                         | G                         | T → QT                   |
| 16255    | T                                | C                         | C                         | T3 → T1'2'3              |

**Table 6:4: Turk13 SNPs captured and whole sample shotgun, haplotype assignments as Achilli *et al* (2009), local mutation marked with (?). Dashed line (-) marks SNP that aren't covered at all in this dataset. For positions with insufficient coverage to confidently call SNPs genotypes are shown in brackets.**



#### 6.3.4 PCA analysis

2,296 and 5,975 SNPs could be called from the bovine 770K panel for Turk11 and Turk13 respectively, after SMARTPCAs filters for informative markers and  $r^2$  were applied, 1,776 (Turk11) and 4,633 (Turk13) SNPs remained for study.

The initial PCA analysis centred on placing both ancient samples within the global genetic map of cattle using a diverse collection of samples covering all the major cattle populations (European taurine, African taurine and Indicine) as well as a number of Near Eastern hybrid breeds (Figure 6:1). As can be seen on the resultant plots both Turk11 and Turk13 fall squarely within the European taurine distribution, away from the current Near Eastern/Anatolian populations. Turk11 is shown to have a greater affinity to a cluster of Limousin (LIM), Simmental (SIM) and Alentejana (ALE) animals, whereas as Turk13 places outside of this cluster and next to a single Alentejana individual. When the analyses are combined using the Procrustes method of Skoglund *et al* (2012), this trend is repeated with Turk11 located inside the Limousin/Simmental cluster, in this case also joined by all the Alentejana animals; Turk13 is positioned by itself outside of this grouping (Figure 6:2).

To further investigate the relationship between the ancient and modern day European and Anatolian animals the analysis was repeated using only breeds with a European/Near Eastern Origin (n=165, minus Herford because of the biases outlined in chapter 4). These plots describe a similar picture to the ones drawn from the global cattle populations (Figure 6:3) in that both Turkish



samples share more affinity to the European taurine animals than they do with Near Eastern/Anatolian samples. In this reduced dataset Turk11's affinity to the Limousin/Simmental breeds is reduced and it now falls next to a single Alentejana individual. Turk13 sits further outside the Alentejana grouping roughly halfway between it and the two cattle populations from Greece and Italy (Maranamana and Sikia). When the analyses are combined (again using the Procrustes method), both animals are now shown to fall next to a cluster of Alentejana animals (Figure 6:4).

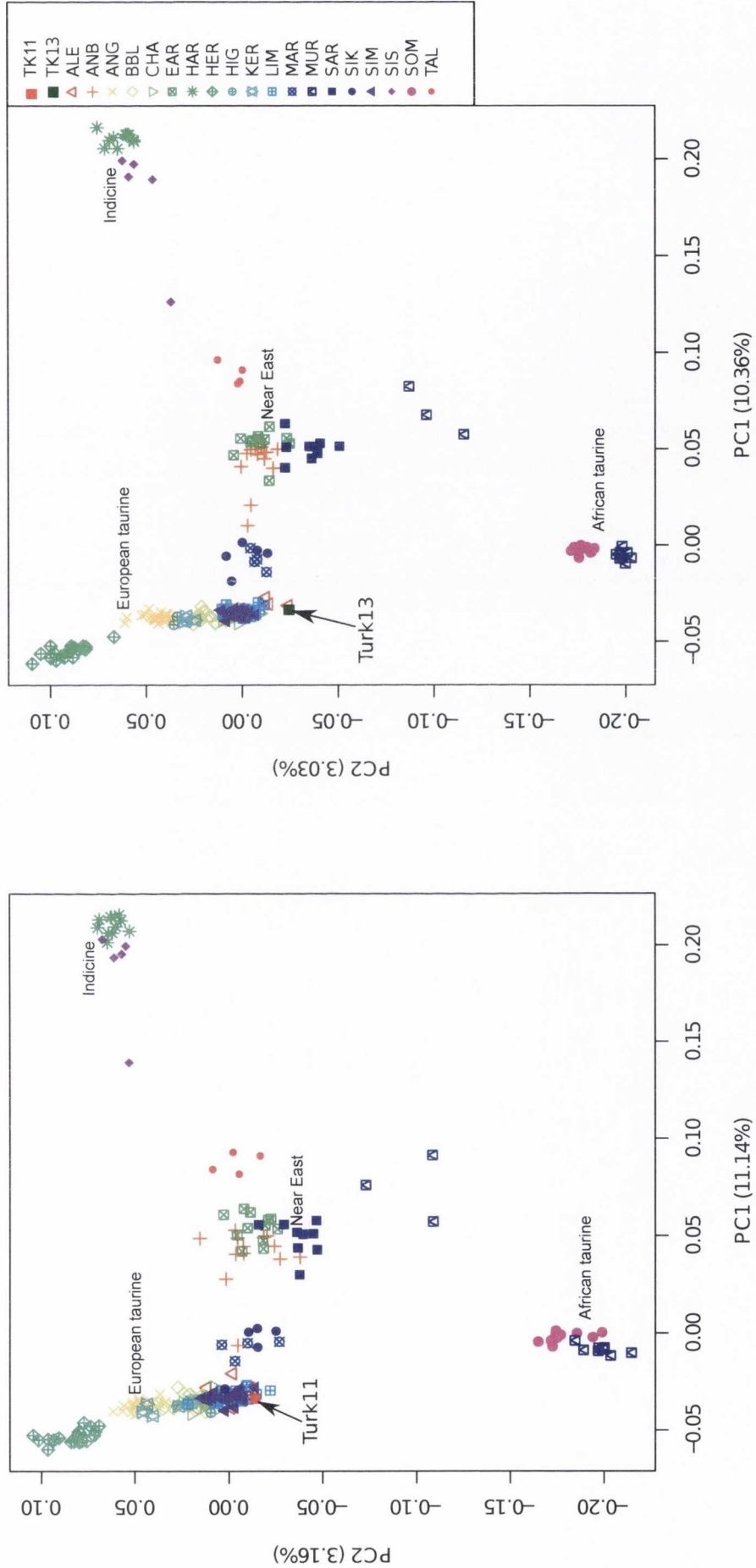
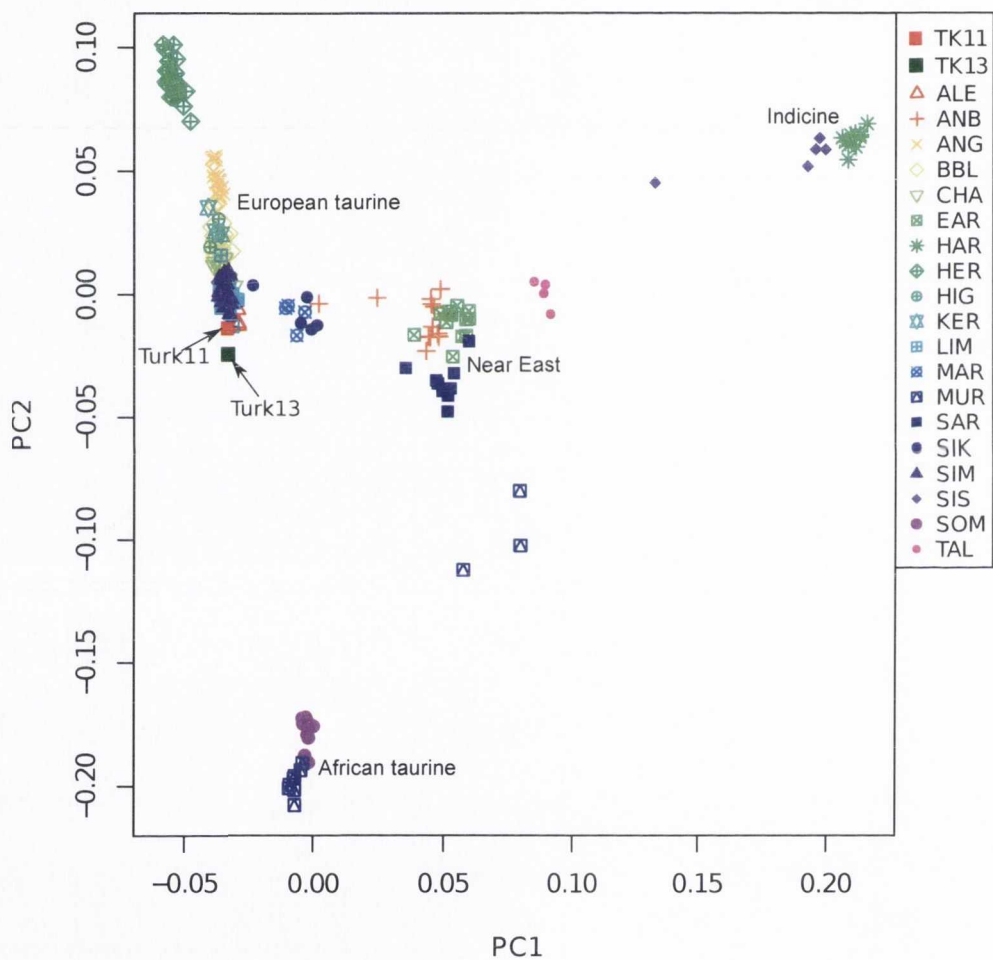


Figure 6:1: 770K SMARTPCA analysis Turk11, left (TK11, 1,776 SNPs) and TURK13, right (TK13, 4,633 SNPs). Full 770K dataset.



**Figure 6:2: Merge of Turk11 and Turk13 770K SMARTPCA analyses, full 770K dataset.**

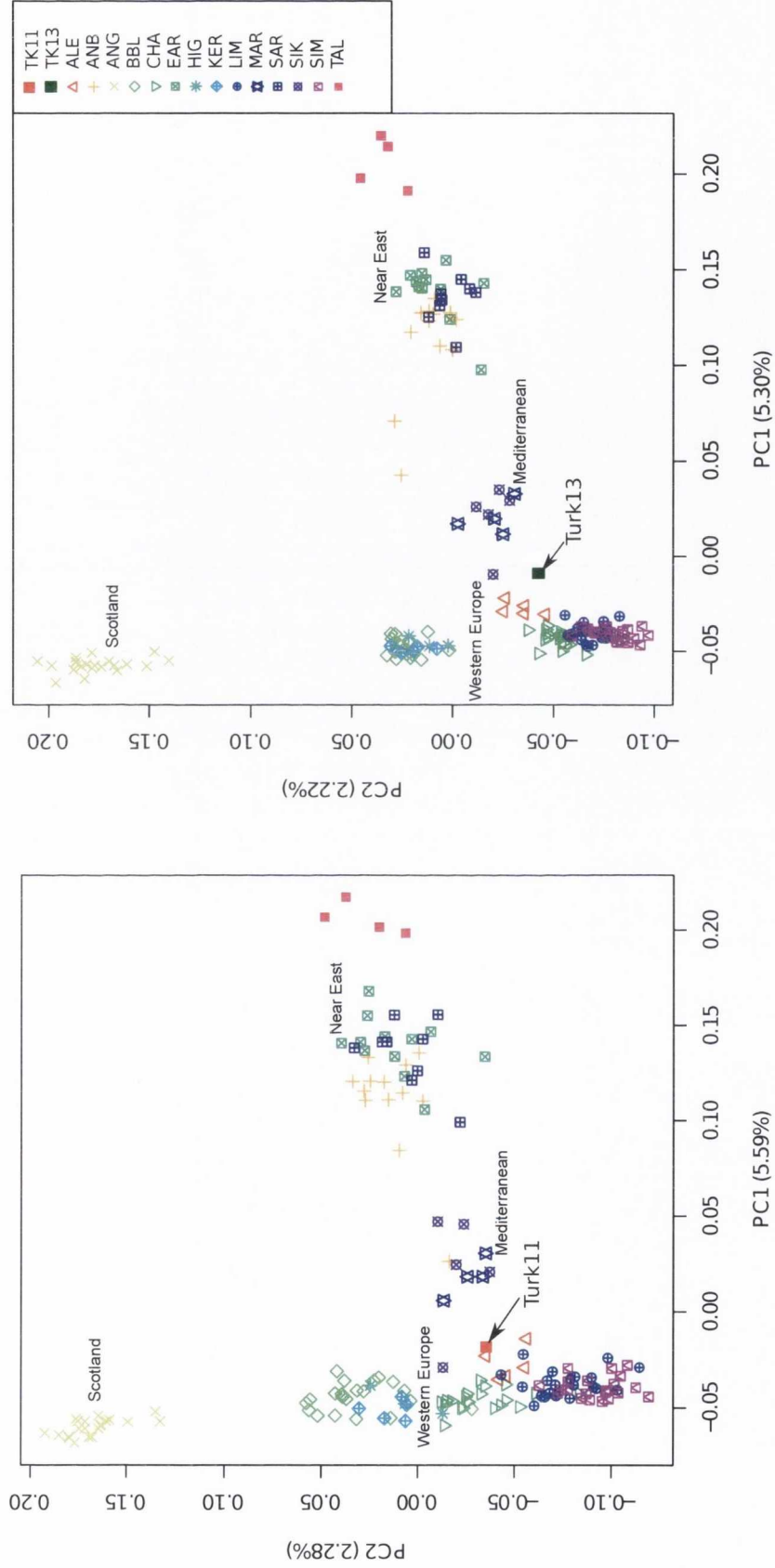
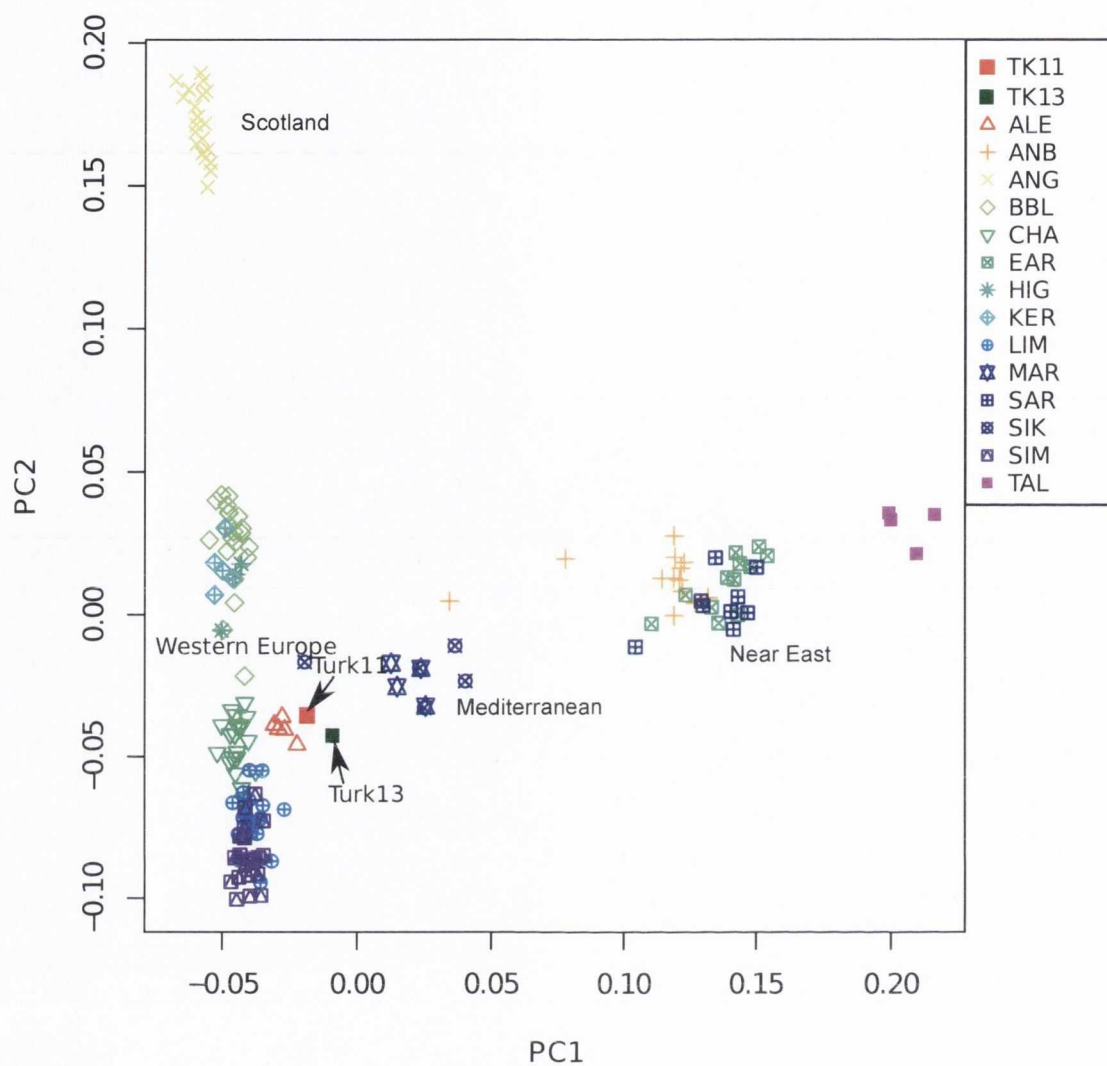


Figure 6:3: 770K SMARTPCA analysis Turk11 (TK11, 1,776 SNPs) and TURK13 (TK13, 4,633 SNPs) European and Near Eastern samples only.





**Figure 6:4:** Merge of Turk11 and Turk13 770K SMARTPCA analyses, European and Near Eastern dataset.

## 6.4 Discussion

The hybrid nature of modern Anatolia cattle populations has been well described at the genetic level (LOFTUS *et al.* 1999; FREEMAN *et al.* 2006b; EDWARDS *et al.* 2007b). Though this admixture is predicted to be ancient (FREEMAN *et al.* 2006b) the first domestic cattle of Anatolia were likely to have been newly domesticated unadmixed *Bos taurus*. In this chapter we presented the results of a population genetic study of two bovid samples from the site of Menteşe in western Anatolia, dated by context to roughly 8,000 BP. Both samples were incredibly well preserved with high quality unique alignment percentages (mapping quality  $\geq 30$ ) of 1.84% and 4.29% for Turk11 and Turk13 respectively (genome coverage 0.4% Turk11, 1.15% for Turk13). This result is surprising given that these samples were discovered in Turkey; a location that, although of huge archaeological significance is not one normally predicted to be good for the survival of aDNA (HOFREITER *et al.* 2001). Given Turkey's lower latitude and high daytime temperatures, which are not thought to be suited to DNA retrieval (ALLENTOFT *et al.* 2012; OTTONI *et al.* 2012).

### 6.4.1 Test of authenticity

To assess the authenticity of the results presented for Turk11 and Turk13 two tests were performed, the first a MAPDAMAGE analysis provided fragmentation profiles for both samples that are suggestive of ancient DNA. The second, an estimation of the within-species contamination rate was also encouraging, with values of 0% for Turk11 and 1.4% for Turk13. This compares favourably to a similar analysis completed for a human mitochondrial genome with an estimated rate of 8% (SÁNCHEZ-QUINTO *et al.*

2012). However the lower sequence coverage of Turk13 and 11 (mean depth of coverage Turk11 = 2x and Turk13 = 2x) compared to Sanchez-Quinto *et al* (2012) (28x) is likely a mitigating factor, as well as the analysis of faunal remains compared to human. The predicted contamination rate for Turk11 was also found to increase to a more believable 1.2% when the RNA captured genomes are considered; interestingly the rate for Turk13 was maintained at 1.4%. These concordant results suggest that capture isn't adding significantly to the mtDNA contamination rate.

#### **6.4.2 PCA Analysis**

In the global bovine 770K PCA analyses (Figure 6:1, Figure 6:2) both animals were found to be more closely related to western European taurine animals than either African taurine, Indicine or Near Eastern/Anatolian animals. The consistent placing of these ancient western Turkish samples next to Alentejana animals in the European/Near Eastern analysis is interesting given that this population is identified as an old breed of centuries long importance for draft in the hills of Alentjo Portugal (FELIUS 1995). However, this result is likely to be largely artifactual given the likely hood that the breed itself is only hundreds of years old and has likely experience admixture with multiple European taurine populations.

It is also worth noting that in the global PCA analysis PC2 distinguishes between taurine populations and on this axis the Turkish samples fall into the middle, with a marked discontinuity seen, at least in this analysis between the African taurines and a clump of all European taurines / Near Eastern animals

(Figure 6:4). When resolved on this axis the Turkish samples are found to fall with the Mediterranean and Near Eastern animals.

The analysis unsurprisingly suggest that the first cattle populations of western Anatolia were unadmixed *Bos taurus* animals and thus they resemble the present day European populations more than they do the local modern hybrid cattle (FREEMAN *et al.* 2006b; EDWARDS *et al.* 2007b). However, this assertion requires the reasonable assumption of geographic isolation of cattle populations between the nascent and ancestral taurine and incine populations.

The more divergent positioning of Turk13 to Turk11 in both these analyses may be a reflection of its Q mitochondrial haplotype, which as discussed in chapter 3 is found to place outside of the predominant T grouping of European taurine animals, to which the Turk11 mtDNA chromosome belongs. Suggesting some level of population differentiation in early domestic taurine animals. However as the Q haplotype is closely related to T and only discovered through whole genome sequencing it is likely that the nuclear differentiation is equally minor.

Interestingly, neither ancient western Anatolian sample placed near to the African taurine animals, which are consistent outliers among *Bos taurus* across multiple analyses (BRADLEY *et al.* 1996; MACHUGH *et al.* 1997; TROY *et al.* 2001; BOVINE HAPMAP CONSORTIUM *et al.* 2009; GAUTIER *et al.* 2010). Whereas the affinity of contemporary European taurine cattle to these early domestic cattle (~8,000YBP) seems consistent with an ancestry stretching back



to Anatolia the distinction between the Mentese samples and the Africans suggests a different relationship. Multiple theories could explain this divergence, including admixture of African aurochs into the African taurine populations in a similar manor to that proposed for indicine populations (BRIGGS *et al.* 2009; CHEN *et al.* 2009; MURRAY *et al.* 2010) or perhaps population structure in the original domestic *Bos taurus* animals with separate Near Eastern contributions to the two derivative continents. Ascertainment biases in the modern 770K SNP dataset at the resolutions explored in this study may also be a factor in this pattern. These biases are caused by the ascertainment of SNPs in European animals this ascertainment process gives these animals a greater level of heterozygosity, where as the more divergent indicine animals are fixed at these taurine variable positions. Given that unlike in chapter 4 where full diploid data was examined this chapter was considering pseudo diploid data, based on haploid calls from ancient samples. The ancient samples may by chance look more European if a low frequency heterozygous allele is sequenced at these sites in these samples.

Further study of cattle samples of a range of ages and locations in Anatolia and the Near East would allow for full affinities of these ancient populations to be assessed. It will also be possible for the time depth of the profound indicine admixture in modern breeds of the region to be estimated, providing insights into the human population movements that likely generated these first hybrids. Further analyses of the modern genomes of Anatolian cattle may also allow for the taurine-specific regions of current genomes to be deduced and compared separately to the ancient population.

## 6.5 Conclusion

The analysis of ancient haploid data has already provided insights into the population history of modern European populations (BURBANO *et al.* 2010; ÁVILA-ARCOS *et al.* 2011; SCHUENEMANN *et al.* 2011; SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012). The results of this case study suggest that these methods can also be applied to the study of domesticated genomes; illustrated by the placing of two roughly 8,000-year-old Anatolian bovid samples within the variation of domestic cattle populations. Moreover this chapter highlights the power of these techniques to answer important archaeological and biological questions when applied to domestic species. In this chapter they were used to answer the question of where the genetic affinities of the early domestic populations of Anatolia lie. This analysis suggested that these animals more closely resemble modern day European cattle and that the introgression seen with zebu animals in present day Anatolian populations happened post domestication. This supports the notion of two independent domestications of cattle, with an admixture of these two domestic sub species accruing during the next 10,000 years

This analysis also provided enough sequencing coverage to corroborate the majority of the mitochondrial SNPs produced for these animals in a RNA hybridisation capture experiment (Chapter 3). An estimated contamination rate was also deduced which was comparable to other ancient DNA samples; moreover this rate seems to be unaffected by the RNA capture process.

This chapter supports the notion that further study of both modern and ancient Anatolian cattle populations should be able to describe the affinities of these populations and help deduce the genomic ancestry of European cattle. Moreover, that the new methods of 1x SNP typing in ancient samples will be useful in completing this analysis.

## **7. Evaluating parchment as a reservoir of ancient DNA for next generation sequencing**

### ***7.1 Introduction***

#### **7.1.1 Parchment and the historical record**

Parchment provides an invaluable historical archive, not only as a source of written knowledge, but also as a dated store of biological information (BOWER *et al.* 2010). Parchment first appears in the archaeological record as the writing material for the Dead Sea Scrolls dated from 250BC to AD68 (LARSON 2002). However, further evidence from Egypt and the Middle East may place parchments origins even earlier to the 3<sup>rd</sup> century BC (LARSON 2002).

Parchment is a durable writing material, which has enabled its use in a multitude of documents prior to the introduction of mass-produced paper in the early 19<sup>th</sup> century (BOWER *et al.* 2010; CAMPANA *et al.* 2010). This wide adoption and the fact that parchment can be made from cattle, sheep and goat skin, makes it an intriguing substrate for the study of domestic *ungulate* genetics.

#### **7.1.2 Manufacturing parchment**

Parchment differs to leather in that its manufacture rarely requires treatment with tanning materials, making it an attractive source of aDNA (LARSON 2002;



VUISSOZ *et al.* 2007). Descriptions of the parchment making process survive from as early as the 8<sup>th</sup> century.

*“How parchment is to be prepared: place [the skin] in lime water and leave it there for a few days. Then extend it on a frame and scrape it on both sides with a sharp knife and leave it to dry.”*

(<http://www.sca.org.au/scribe/articles/parchment.htm>, REED 1975).

Generally parchments production requires the animal's skin to be first washed clean, then soaked in multiple lime baths to soften its hairs. These loosened hairs are then scraped from the skin with a blade and the de-haired skin is further stretched scraped and treated with chalk to produce a flat writing surface (LARSON 2002; CAMPANA *et al.* 2010).

### **7.1.3 Parchment and aDNA**

Though next generation sequencing (NGS) has reinvigorated the analysis of ancient bone specimens (POINAR *et al.* 2006; EDWARDS *et al.* 2010; GREEN *et al.* 2010; REICH *et al.* 2010; ORLANDO *et al.* 2011; SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012), the routine analysis of ancient nuclear genomes is limited by the availability of well-preserved historic and archaeological samples (KNAPP and HOFREITER 2010; KIRSANOW and BURGER 2012). Most bone samples are found to contain high levels of bacterial contamination and low levels of endogenous DNA (~0.1-5%) (POINAR *et al.* 2006; GREEN *et al.* 2010; ORLANDO *et al.* 2011; SKOGLUND *et al.* 2012). With some notable high

profile exceptions (REICH *et al.* 2010; MEYER *et al.* 2012). Hair and eggshell appear to be two promising alternative substrates due to their greater resilience to decontamination techniques (GILBERT *et al.* 2007; 2008; OSKAM *et al.* 2010; RASMUSSEN *et al.* 2010; 2011; BENGTSSON *et al.* 2012). A third candidate for NGS could be parchment given its ubiquity in the historical record. However, parchment has seemed an enigmatic aDNA source material, with publications describing both successful single-species amplifications (BURGER *et al.* 2001; POULAKAKIS *et al.* 2007; PANGALLO *et al.* 2010) and a “DNA soup” of multiple species (CAMPANA *et al.* 2010). This heterogeneous environment, is likely the result of the parchment making processes, which may introduce DNA from multiple species and animals (CAMPANA *et al.* 2010). To date however, no NGS analysis of parchment has been published.

#### **7.1.4 Sheep hapmap and breed formation**

The sheep diversity project collected DNA from 74 diverse sheep breeds from Asia, Africa, the Middle East, the Caribbean, North and South America, Europe, and Australasia (<http://www.sheephapmap.org/>). These animals were genotyped at roughly 50,000 evenly spaced loci, to provide a HapMap dataset that covers a large sample of domestic sheep diversity. SNPs were discovered using both the Illumina and Roche 454 platforms and show some ascertainment bias towards European breeds. However, this was found to have a limited effect on the SNP panels ability to distinguish genetic distances between the breeds (KIJAS *et al.* 2012). Analyses of the 50K sheep SNP dataset have provided insights into the origin and relationship between sheep breeds and indicate that the initial sheep domestication involved the incorporations of a

larger number of animals into the initial domestic population than that of the two other deeply studied domesticates to date: Dog and Cow (LINDBLAD-TOH *et al.* 2005; BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM *et al.* 2009; KIJAS *et al.* 2012). This data set provides an excellent genetic map to which ancient sheep samples can be placed.

#### **7.1.4 Objectives of this case study**

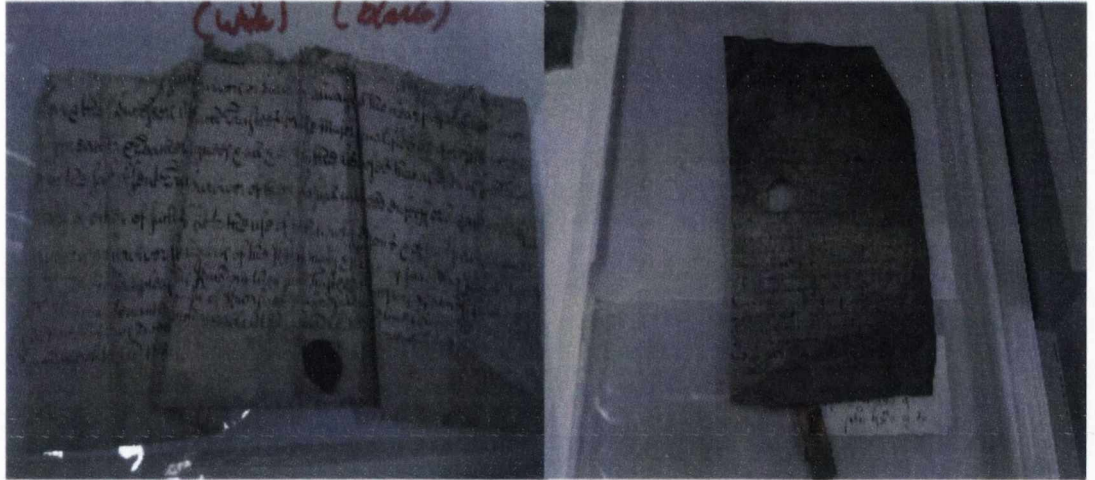
Given parchments high value and ubiquitous use from roughly the 13<sup>th</sup> to the 18<sup>th</sup> century, it provides an enticing source for studies of domesticate aDNA. However previously PCR based work has raised issues with the authenticity of data generated from the substrait (CAMPANA *et al.* 2010). This chapter presents the results of a pilot aDNA analysis using parchment as a source for NGS analysis. It aims to evaluate the use of parchment for aDNA studies, addresses multiple issues raised during previous PCR studies of the material and evaluates it as a source of aDNA.

## **7.2 Methods**

### **7.2.1 Parchment samples chosen for aDNA analysis**

Two parchment samples (PA1 and PA2) were chosen for a preliminary aDNA analysis from those housed at the Borthwick Institute for Archives, The University of York. These parchments were dated by handwriting analysis to

the 18<sup>th</sup>(PA1) and 17<sup>th</sup>(PA2) century respectively (Christopher Webb, personal communication).



**Figure 7.1: Parchment samples chosen for aDNA analysis, left PA1 18<sup>th</sup> century, right PA2 17<sup>th</sup> century.**

### **7.2.2 Ancient DNA extraction**

All aDNA extractions were performed in a dedicated ancient DNA laboratory at Trinity College Dublin. Given the pilot nature of this analysis the maximum amount of starting material available was used for each extraction, with DNA retrieved from 2x2cm<sup>2</sup> pieces cut from both parchments. To lightly clean the parchments surface all pieces were washed in a 12% sodium hypochlorite solution for ten seconds followed by two, two-minute washes in UV irradiated molecular grade water prior to DNA extraction. DNA was prepared using a small-scale version of the modified Yang *et al.* (1998) procedure, previously described by our group (EDWARDS *et al.* 2004; 2010). Briefly parchment samples were incubated in extraction buffer (200µM proteinase K, 0.5M EDTA) for 48 hours (55°C for 24 hours and then 37°C for 24 hours).



Concentrated to 100 $\mu$ l on a 30kD Amicon filter with two 3ml washes of 10mM Tris EDTA and cleaned of any salt contaminants using a Quiagen min-elute column, following manufacturers instructions.

### **7.2.3 Illumina sequencing library preparation**

Illumina single end sequencing libraries were produced for each of the two parchment samples via PCR enrichment of end repaired adaptor-ligated DNA templates using the standard protocol outline in chapter 2.

The final PCR products from both samples (2 per parchment) were pooled and visualized on an Agilent 2100 Bioanalyzer using a DNA 1000 chip. Both samples were then sequenced together on a single lane of an Illumina HiSeq2000 at the Beijing Genomics Institute.

### **7.2.4 Initial parchment read QC**

Initial quality control of sequencing reads were preformed using the protocol described in chapter 2, reads were then separated by their barcodes also following the protocols described in chapter 2.

### **7.2.5 FastQ Screen**

*FastQ Screen* (version 0.3.1) is a PERL rapper script for the alignment algorithm BOWTIE (version 0.12.7) (LANGMEAD *et al.* 2009), which allows for the same sequencing library to be easily aligned to multiple reference genomes ([http://www.bioinformatics.babraham.ac.uk/projects/fastq\\_screen](http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen)).

The percentage of raw parchment reads that aligned uniquely to a single genome, to multiple places in the same genome, uniquely in multiple genomes and to multiple places in multiple genomes can then be assessed. *FastQ Screen* was used in this analysis to align the CUTADAPT trimmed reads to 3 genomes, cow (bosTau6, Downloaded 14-8-12), human (hg19, Downloaded 14-8-12) and sheep (OAR2, Downloaded 14-8-12). BOWTIE and *FastQ Screen* alignment settings were modified to allow the number of mismatches between the total read length and the reference genome to vary between 0 and 3.

#### **7.2.6 BWA Sequence Alignment**

The raw trimmed reads were also aligned to the Sheep reference genome (OAR2) using BWA (version 0.6.1) (LI and DURBIN 2009). Standard alignment settings for BWA were used (Chapter 2).

#### **7.2.7 Alignment to human genome**

All sequencing reads were further aligned to the human genome (hg19) using BWA with identical parameters to the sheep (OAR2) alignments above. Any reads that aligned to the human genome were then removed from subsequent analysis irrespective of mapping quality.

#### **7.2.8 SNP calling**

SNPs were called using established protocols for 1x aDNA NGS data (SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012) as methods chapter.

### **7.2.9 Whole Mitochondrial genomes**

Full mitochondrial genomes were produced from alignments of the parchment reads to a modern mitochondrial reference genome HM236176 with BWA. Redundant reads were then removed using the SAMTOOLS “rmdup” command and genotypes called using the combination of 3 algorithms as described in chapter 3.

### **7.2.10 Sequence alignment and tree drawing**

Modern sheep mitochondrial reference genomes were downloaded from GenBank (n=24) to allow placement of the parchment samples within this dataset. Multiple sequence alignments of both the modern and parchment mtDNA genomes were completed using version 3.8.31 of the Muscle alignment algorithm implemented in Seaview Version 4.3.4. Neighbour joining trees (1,000 bootstrap replicates) were then produced from the alignment data in Seaview using default setting (BioNJ model) with no partitioning and visualised using Figtree.

### **7.2.11 Principal Component Analysis**

The ovine HapMap dataset (KIJAS *et al.* 2012) provided a modern sheep SNP reference for use in this study, SNPs with  $MAF \leq 0.5\%$  were excluded leaving 48,596 SNPs for further analysis. The modern and ancient datasets were then merged using PLINK, removing A/T and G/C SNPs when multiple datasets were combined. PCA analysis was completed using the standard protocols as described in chapter 2 with the following modifications Plink and Eigensoft

were pass the following flags to recognize the differing number of autosomes in the sheep genome compared to humans ‘-sheep’ and ‘numchrom = 26’ respectively.



| Sample | Total reads | Aligned reads | Aligned reads (%) | Aligned reads duplicates removed (SAMtools rmdup) | Aligned reads duplicates removed (%) (SAMtools rmdup) | Mapping quality $\geq 30$ | Mapping quality $\geq 30$ (%) | Final aligned reads, post filters | Final aligned reads, post filters (%) | SNPs called ovine HapMap |
|--------|-------------|---------------|-------------------|---------------------------------------------------|-------------------------------------------------------|---------------------------|-------------------------------|-----------------------------------|---------------------------------------|--------------------------|
| PA1    | 27,955,895  | 21,271,013    | 76.1              | 16,924,142                                        | 60.5                                                  | 5,692,171                 | 20.4                          | 5,508,263                         | 19.7                                  | 1,834                    |
|        |             |               |                   |                                                   |                                                       |                           |                               |                                   |                                       |                          |
| PA2    | 15,881,027  | 13,660,034    | 86.0              | 11,476,494                                        | 72.3                                                  | 6,174,808                 | 38.9                          | 6,041,816                         | 38.0                                  | 2,515                    |

Table 7:1: Sequencing results PA1 and PA2

## **7.3 Results**

### **7.3.1 Species identification, ZooMS**

Species identification of both parchments was initially completed at The University of York using the proteomics method ZooMS (Zoo Archaeology by mass spectroscopy) (BUCKLEY *et al.* 2009). ZooMS is used routinely for the taxonomic assignment of both archaeological and modern bone specimens (BUCKLEY *et al.* 2009; DOORN *et al.* 2011; RICHTER *et al.* 2011) and works by sequencing the collagen fragments preserved in the remains using mass spectrometry. The results of this protein-based analysis suggested that both parchments were derived from sheepskin, with no spectra from other domesticates found (van Doorn, N. Personal communication). This preliminary species identification was used to inform the subsequent aDNA alignments.

### **7.3.2 Species identification, FASTQ\_SCREEN**

The NGS sequencing of the parchment samples produced 43,836,922 raw 49bp reads, after adapter dimer removal (PA1=27,955,895, PA2=15,881,027) (Table 7:1).

To corroborate the species identification provided by ZooMS above *FastQ Screen* was used to align the parchment sequences to three mammalian genomes; firstly, sheep the source species identified by ZooMS, then human and cow as two likely contaminants. The results of this analysis supported that

of ZooMS by indicating the predominant source of aDNA in the parchment dataset was sheep (Figure 2, full data Appendix:1F). 26% of the raw reads from PA1 and 37% of the raw reads from PA2 aligned solely (Red and orange bars) to the sheep genome when the number of mismatches allowed between the read and genome ( $n$ ) for a successful alignment was set to 0. This compares to just 1% for both PA1 and PA2 when *Bos taurus* reads are considered at the same stringency, and alignment percentages of only 0.06% (PA1) and 0.03% (PA2) to the human genome. When reads that also found alignments to other genomes are considered (Dark and light blue bars), the percentage alignment to the sheep genome rises to 59% and 68% for PA1 and PA2 respectively (same *FastQ Screen* settings). These reads which cannot be attributed to a unique species are likely representative of the homology between the genomes, particularly that of sheep and cow. This assertion was tested by aligning one million simulated reads generated from the ovine genome to the bovine genome, which found roughly 40% of reads to align to the bovine genome none uniquely by chance.

That relatively few reads aligned solely to either the human (hg19) or the bovine (bosTau6) genome at all values of  $n$  tested (Figure 7:2) suggests that both PA1 and PA2 may suffer from a very small degree of exogenous contamination. However the relative completeness of the other two genomes may also be a factor. Unsurprisingly, the percentage of reads that aligned to multiple regions in the same genome and to both single and multiple positions in two or more genomes increase when a greater number of mismatches ( $n=1-3$ ) are allowed by *FastQ Screen*.

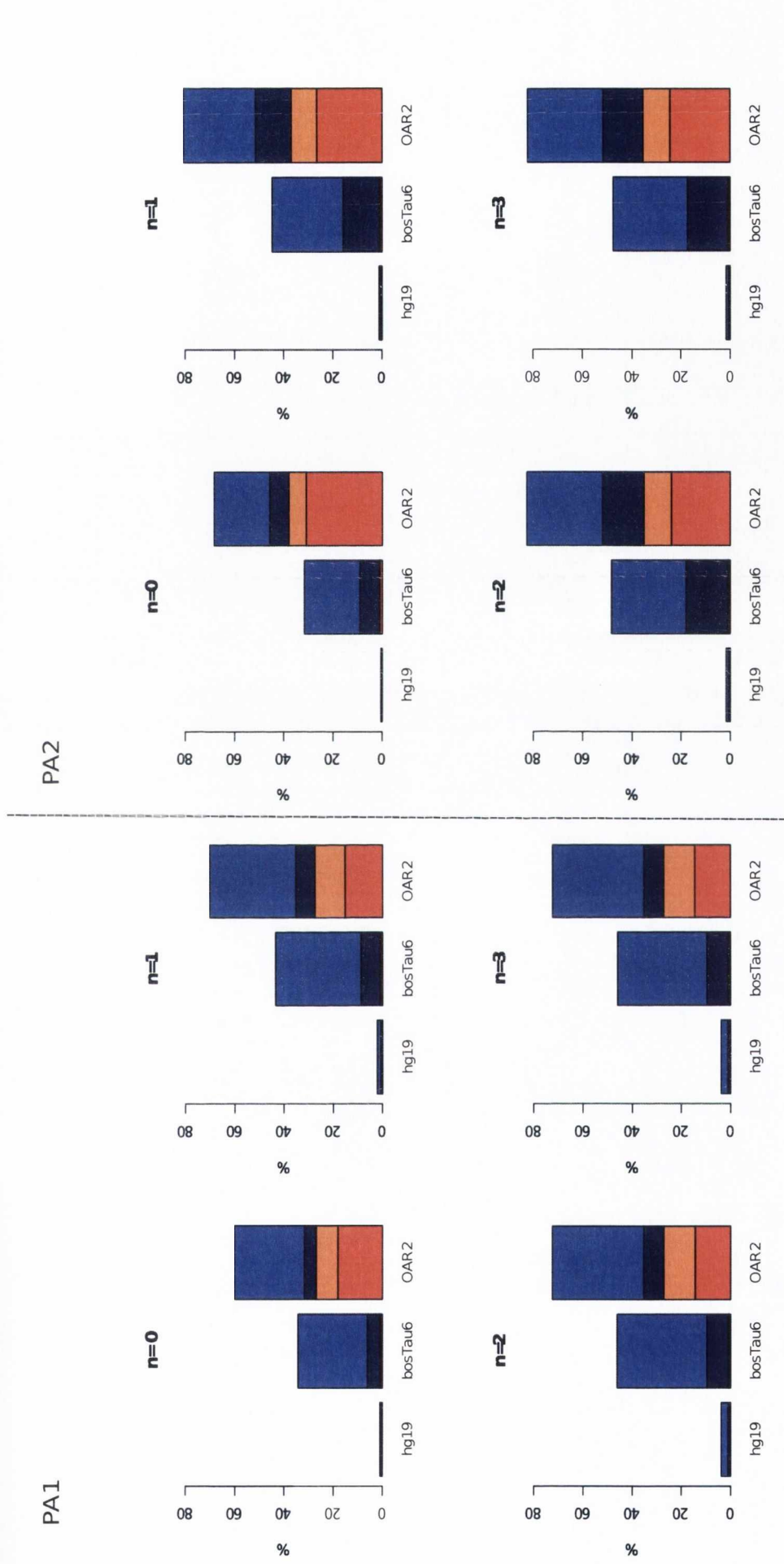


Figure 7.2: *FastQ Screen* analysis of the two parchments (PA1 and PA2), showing percentage of reads that had a unique alignment to a single genome (red bar) non-unique alignment to a single genome (orange bar), a unique alignment in multiple genomes (dark blue bar), non-unique alignment to multiple genomes (light blue bar). n= number of mismatches allowed. Note that at all levels of stringency most single species alignments are found to the sheep (OAR2) genome.



### 7.3.3 Sequencing Results

Given the identification of sheep as the likely source material for both parchments using both the proteomic and ancient DNA screening methods described above. The parchment samples were then re-aligned to the Ovine genome (OAR2) using BWA; in total 28,400,636 (64.7%) reads were retained post PCR duplicate removal (PA1=16,924,142 and PA2=11,476,494). After filtering for a mapping quality of greater than or equal to 30, 5,692,171 (20.4%) reads remained for PA1 and 6,174,808 (38.9%) for PA2. These reads were then further filtered by the removal of any reads that aligned to the human genome (irrespective of PCR duplication or mapping quality) and reads identified in the *FastQ Screen* analysis above that aligned solely to the bovine genome with no mismatches. These datasets were then further filtered for uniquely aligning ovine reads, leaving a final set of 5,508,263 (19.7%) reads for PA1 and 6,041,816 (38.0%) for PA2, covering 202,396,766 OAR2 positions in PA1 and 233,806,579 in PA2, which equates to a retrieval of 8.4% and 9.7% of the sheep genome in PA1 and PA2 respectively.

### 7.3.4 Mitochondrial Genomes

In total 27,274 reads aligned to the ovine mitochondrial genome after duplicate removal (PA1 = 15,584, PA2 = 11,690), which allowed for the production of whole mitochondrial genomes from both samples with an average read depth of 35x and 28x for PA1 and PA2 respectively (Figure 7:4). These genomes were then compared to a modern reference dataset (n=24) including 18

domestic sheep and 6 other wild ovid samples. Both parchments were found to locate within haplogroup B (HB) (Red clade , Figure 7:3), which is found in large numbers in both the modern and ancient sheep populations of Europe (HIENDLEDER *et al.* 1998; OLIVIERI *et al.* 2012).

The high level of mitochondria genome coverage achieved in the parchment sequencing, allows for a rough estimation of the contamination rate in these samples to be calculated (both ancient and modern). To do this heterogeneities at SNP locations in both samples that could be contamination, sequencing error, hetroplasmmy or damage were analyzed (SÁNCHEZ-QUINTO *et al.* 2012). At these positions we found the sample consensus base in 94% of sequences (450/475) for PA1 and 96% (287/296) of sequences for PA2, giving estimated contamination rates of 6% for PA1 and 4% for PA2. These figures are slightly less than a study by Sanchez-Quinto *et al* (2012) of a complete ancient human mitochondria, which estimated exogenous contamination at 8%.



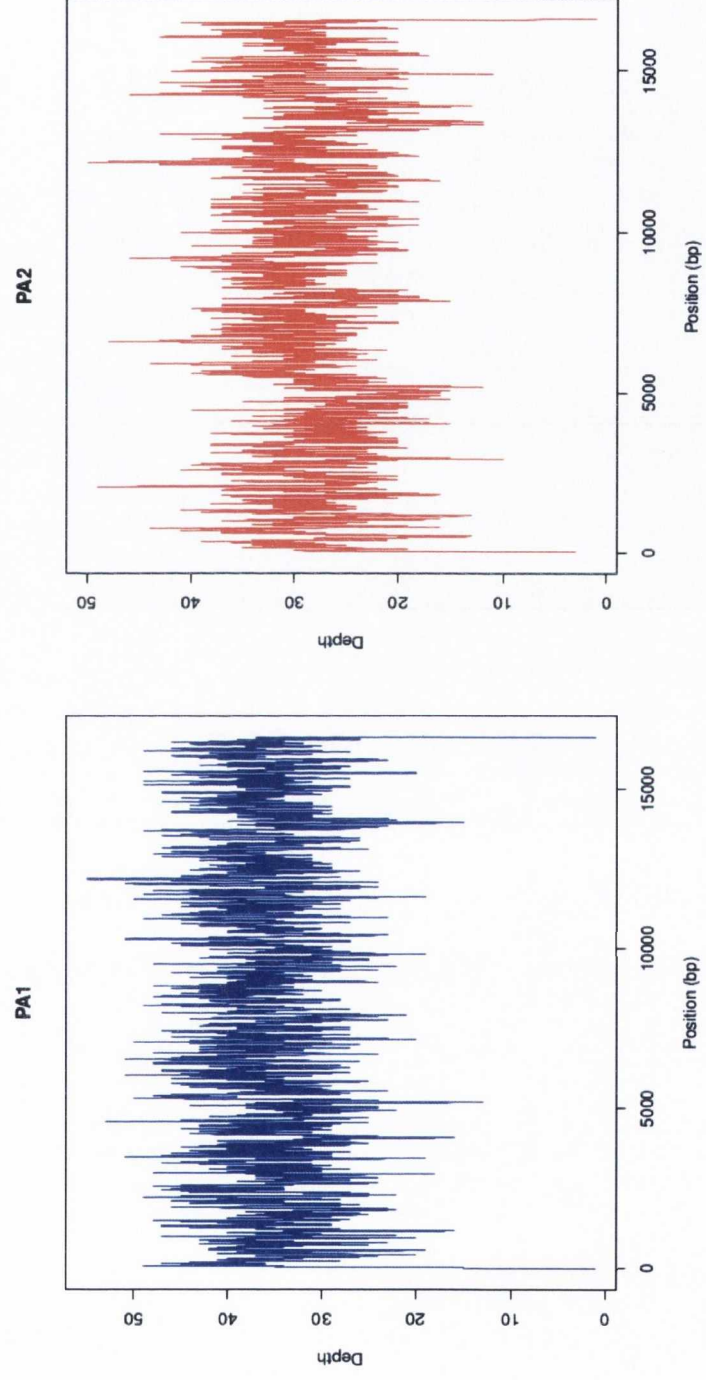


Figure 7:4: Coverage graphs of PA1 and PA2 mitogenomes.



### 7.3.5 SNP analysis

Both parchment samples were genotyped following established protocols for 1x aDNA data (SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012). 1,985 SNPs were called for PA1 and 2,702 for PA2. PCA analyses were completed for both samples using EIGENSOFT 4.2 and the Ovine HapMap was used as a modern reference population (KIJAS *et al.* 2012). After SMARTPCAs filters were applied, 1,834 and 2,515 SNPs were retained for analysis in PA1 and PA2 respectively. In the global PCA analyses both parchment samples were found to cluster within the distribution of British sheep as seen in Figure 7:5 (Panel A = PA1, panel B = PA2 ), this result is to be expected as both samples were likely produced from 17<sup>th</sup> and 18<sup>th</sup> century Yorkshire sheep populations. It is interesting to note that PA2 seems to show slightly more affinity to northern European breeds given its greater age than PA1, which clusters more toward the centre of the British distribution. To further investigate the relationship between the two parchment samples and the British and northern European breeds the analysis was repeated with a reduced dataset (10 animals per breed where available) containing only these populations (Figure 7:6, Appendix1:E, n=189). As can be seen in the plot (Figure 7:6) both PA1 and PA2 are located in a similar position in this analysis, with PA2s affinity to the northern European breeds reduced. PA1 is shown to cluster with the Romney (ROM) and Blackheaded Mountain sheep (BHM) populations (Figure 7:6, panel A), where as PA2 sits outside of this grouping showing a closer relation to two Australian Coopworth animals (CPW, a relatively recent cross of two old British breeds the Romney and Border Leicester) (Figure 7:6, panel B). To aid in the visualisation of these two parchment samples on a single plot, the Procrustes analysis of Skoglund *et al.* (2012) was implemented and the samples

re-drawn together (Figure 7:7). The result of this analysis highlights the trends shown in the previous two plots with PA1 showing more resemblance to the Romney and the Blackheaded Mountain sheep with PA2 placing outside this grouping closer to the two Coopworth animals.

The limited number of SNPs used in these PCA analyses (1,985 SNPs for PA1 and 2,702 for PA2) were found to not overly bias the results of both PCAs (Full and Europe), as plots produced from PCAs of the complete dataset (48,596 SNPs) showing similar placement of the modern sheep populations to those observed when only the parchment genotyped SNPs are used (Appendix1:G). However, the clustering of animals to their breed is more pronounced and less outliers are observed.

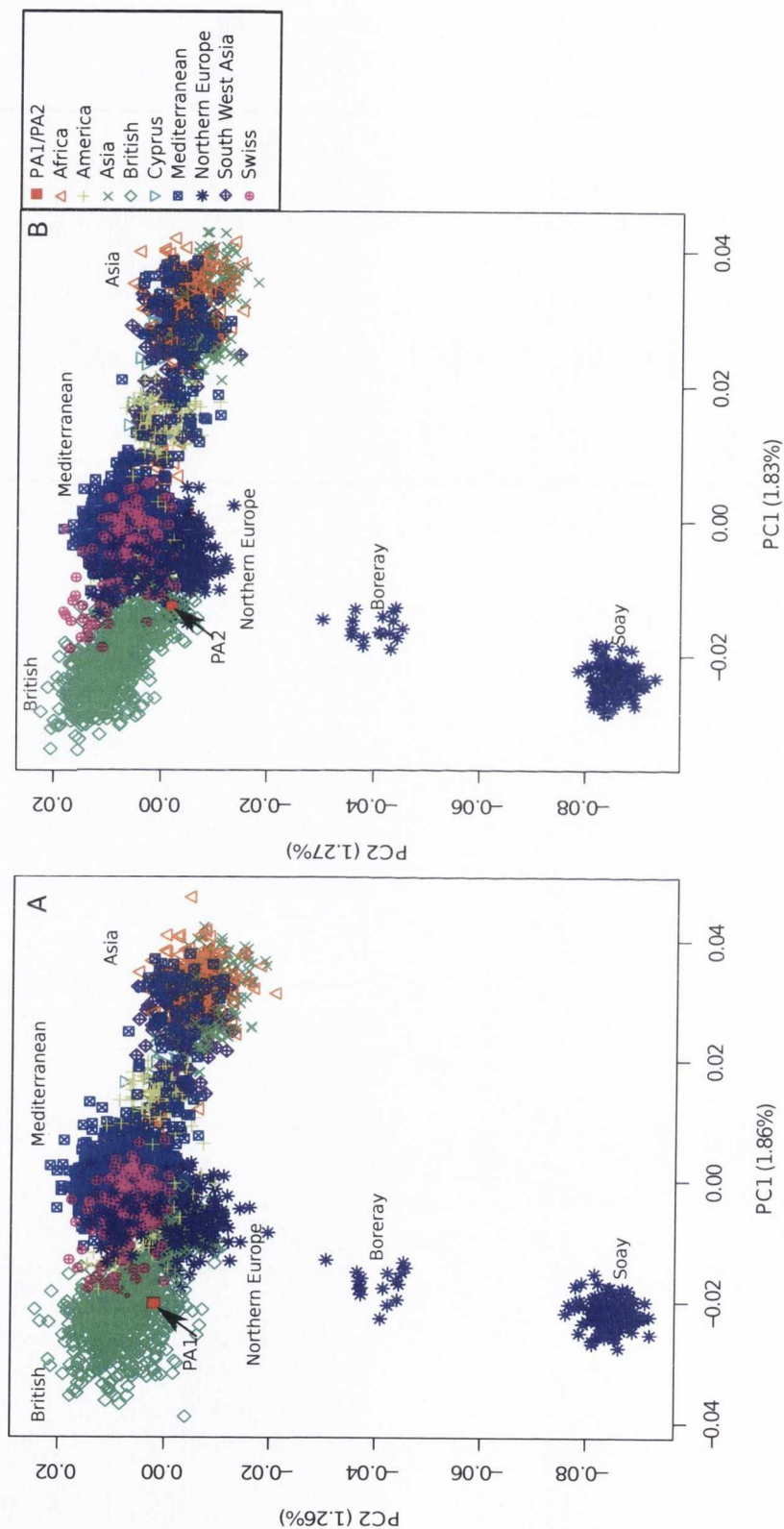


Figure 7:5: SMARTPCA analysis of PA1 (1,834 SNPs) [A] and PA2 (2,515 SNPs) [B]. Note: the PCA axes describe a small amount of variation.

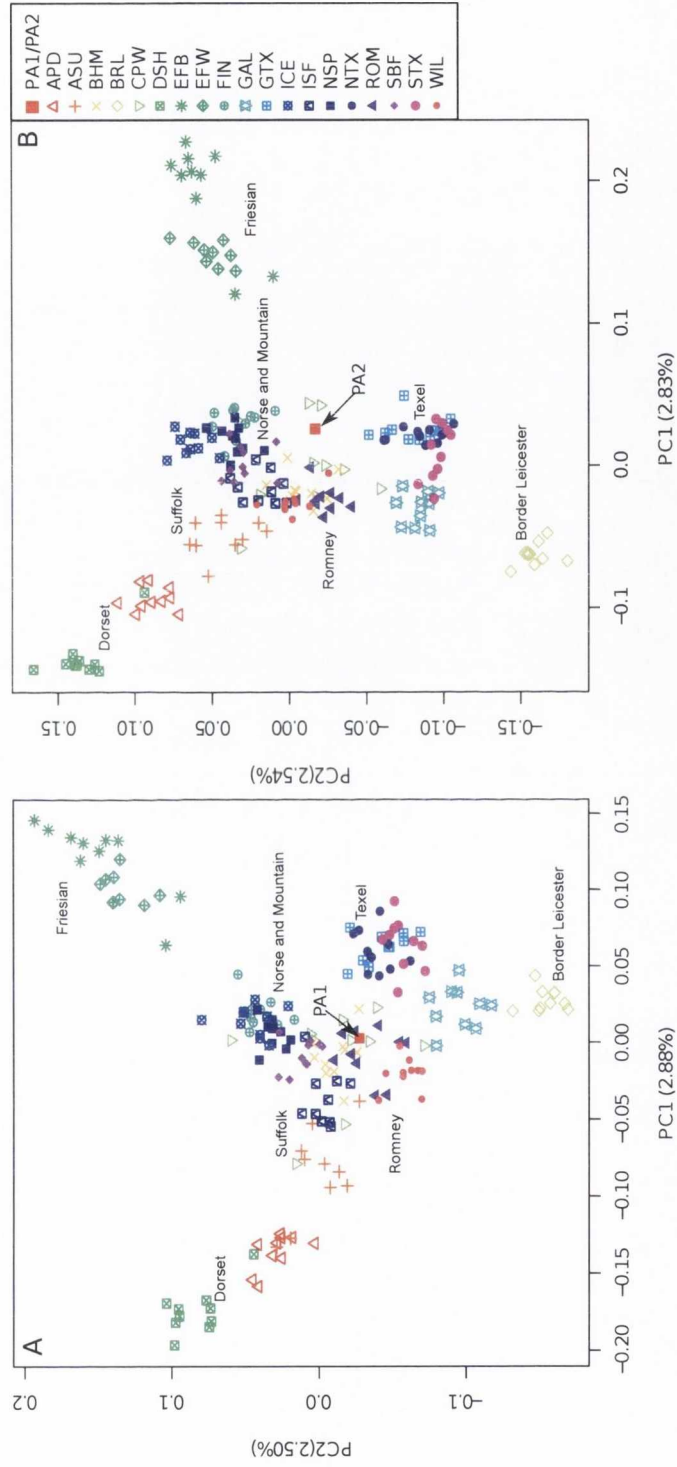
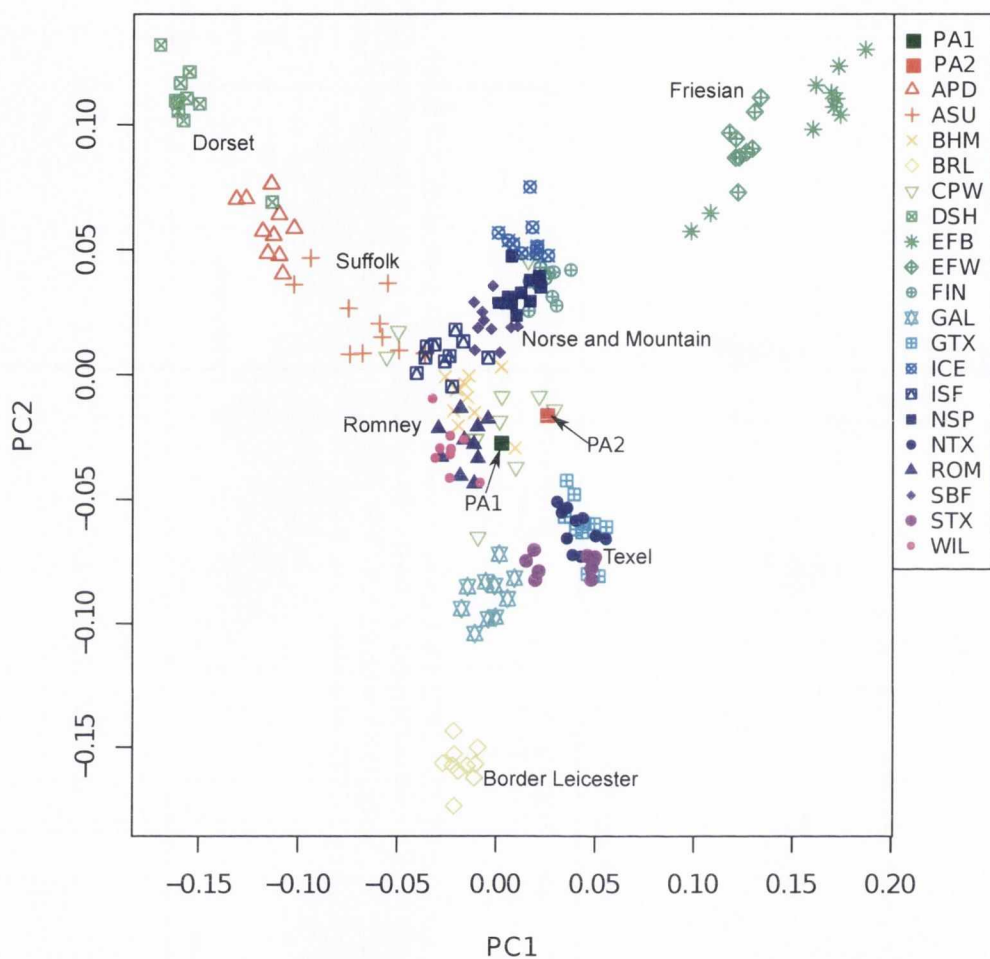


Figure 7:6: SMARTPCA analysis of PA1 [A] (1,834 SNPS ) and PA2 [B] (2,515 SNPS ), European breeds only. APD=Australian Poll Dorset, ASU = Australian Suffolk, BHM = Black Headed Mountain, BRL = Border Leicester, CPW = Australian Coopworth, DSH = Dorset Horn, EFB = East Friesian Brown, EFW = East Friesian White, FIN = Finn sheep, GAL = Galway, GTX = German Texel, ICE = Icelandic, ISF = Irish Suffolk, NSP = Old Norwegianpaelsau, NTX = New Zealand Texel, PA1 = Parchment 1, ROM = New Zealand Romney, SBF = Scottish Black face = STX = Scottish Texel and WIL = Wiltshire.





**Figure 7:7 Procrustes SMARTPCA analysis of PA1(green square) and PA2(red square).** APD=Australian Poll Dorset, ASU = Australian Suffolk, BHM = Black Headed Mountain, BRL = Border Leicester, CPW = Australian Coopworth, DSH = Dorset Horn, EFB = East Friesian Brown, EFW = East Friesian White, FIN = Finn sheep, GAL = Galway, GTX = German Texel, ICE = Icelandic, ISF = Irish Suffolk, NSP = Old Norwegianspaelsau, NTX = New Zealand Texel, PA1 = Parchment 1, ROM = New Zealand Romney, SBF = Scottish Black face = STX = Scottish Texel and WIL = Wiltshire.

## **7.4 Discussion**

The routine analysis of ancient nuclear genomes has been hampered by the availability of suitable substrates for genomic scale analyses. Given that the average bone sample has in the region of 0.1-5% endogenous DNA (ORLANDO *et al.* 2011; KIRSANOW and BURGER 2012), novel substrates are required to allow for a more complete analysis of ancient and historical populations. Two other sources of aDNA (hair and egg shell) have been shown to provide for successful aDNA retrieval (OSKAM *et al.* 2010; RASMUSSEN *et al.* 2010). With hair samples in particular having had major successes, allowing for the sequencing of two whole nuclear genomes, one from a Palaeo-Eskimo and one from an Aboriginal Australian (RASMUSSEN *et al.* 2010; 2011). Given that parchment is ubiquitous in the historical record and produced from a variety of *ungulate* species, it is an attractive source for aDNA analysis if the challenges in its sampling and analysis can be overcome.

### **7.4.1 Suitability of parchment for aDNA analysis**

As described previously parchment has long been a understudied source of ancient DNA, possibly due to the widely reported difficulties in extracting aDNA from animal hide material (BURGER *et al.* 2001; VUISOZ *et al.* 2007; CAMPANA *et al.* 2010). Moreover, the destructive sampling methods necessary for an aDNA study are much less routinely performed in the analysis of documents as they are for bone. The published reports of parchment aDNA investigations also indicate conflicting results, reporting both unique PCR amplification products from a single species (BURGER *et al.* 2001; POULAKAKIS

*et al.* 2007; PANGALLO *et al.* 2010) and multiple sequences from multiple species (BOWER *et al.* 2010; CAMPANA *et al.* 2010).

#### **7.4.2 Single species identification**

The first part of this pilot parchment study was concerned with assessing whether a single species could be identified as the source material for each of the two samples. This analysis was completed using both proteomic and aDNA approaches. ZooMS failed to retrieve any spectra that indicated the presence of domesticate collagen molecules other than sheep. Also the *FastQ Screen* analysis of the raw sequencing data identified the sheep genome as the most likely source for the majority of the sequences. This analysis also identified relatively few molecules that aligned uniquely to either the bovine or human genome 2% and 0.1% of all reads respectively. This small background contamination rate may be one reason why PCR based analyses of parchment have identified DNA from multiple species (CAMPANA *et al.* 2010). The greater resolution of the aDNA extract provide by next generation sequencing, allows for these contaminating sequences to be identified and filtered bioinformatically post sequencing. These percentages may also be inflated by the differences in the completeness of genome builds between the species (BAI *et al.* 2012), with ovine genome the least complete.

The results of the aDNA and protein studies taken together confidently predict sheepskin as the origin for both of the parchments. However, in the re-alignment of all reads to the sheep genome stringent filters were applied to

remove both putative human and bovine contaminants, to try to provide as pure a sheep dataset as possible.

### 7.4.3 Sheep aDNA genomics

Both parchment samples analysed in this study showed a greater than 60% alignment to the Ovine genome once clonally amplified reads are removed. These numbers are reduced somewhat to 19.7% (PA1) and 38.0% (PA2) when filters for mapping quality ( $q = 30$ ) are applied. This result is a feature of most whole genome shotgun sequencing efforts where the sequencing of repetitive DNA reduces mapping quality (in a recent bovine genome re-sequencing study 29.8% of reads were found to have a mapping quality  $\geq 30$ , Connell S, personal communication). These values however, still equate to a 8.4% and 9.7% retrieval of the sheep genome in PA1 and PA2 respectively and the mapping percentages are high in comparison to the amount of material retrieved from bone extracts, typically in the range of 0.1 to 5%, with some notable exceptions mentioned previously (REICH *et al.* 2010). This large percentage of endogenous DNA recovery is likely helped by the parchments age, material analysed (skin instead of bone) and the conditions in which they were stored. The two samples analysed in this study came from The Borthwick Institute for Archives where they were stored at a regulated temperature and humidity, likely aiding aDNA survival. Moreover, the fact that the parchment stored within these achieves contained valuable information suggests that they would have also been carefully looked after during their owners' lifetime. If this percentage DNA retrieval can be repeated in other parchments, the production of historic domesticated genomes over a range of time periods could be accomplished, providing wide-ranging insights into domestic animal breeding.



### 7.4.3 Mitochondrial genomes

Both parchment samples allowed for the production of whole sheep mitochondrial genomes at high levels of coverage (35x PA1 and 28x PA2) with their mtDNA sequences falling within haplogroup B, which has a wide distribution in both modern and ancient European populations.

The high coverage mitochondrial genomes also allowed for the estimation of a within species contamination rate for both samples which were found to be lower than that previously reported for human aDNA samples at 8% (6% for PA1 and 4% for PA2). Moreover, the contamination rate for PA1 falls to 2% when only SNPs outside the difficult to align 75bp repeat regions of the ovine MT genome are considered, suggesting that alignment errors may be contributing to the higher contamination rate in this sample compared to PA2. Furthermore, metagenomic analyses of all reads suggest a bacterial contamination rate of roughly 3% in both samples, which is also noticeably lower than bone samples (POINAR *et al.* 2006).

### 7.4.4 SNP analysis

In the 50K HapMap SNP analysis both parchments were found to sit within the distribution of modern British sheep, next to a cluster of Romney, Blackheaded Mountain and Coopworth animals (Figure 7:6 and Figure 7:7). All these breeds likely descend from the medieval long wool sheep, which were widely distributed throughout Britain and are hypothesised to be the ancestors of

numerous modern breeds (RYDER 1964a). The suggestion that the sheep used to produce the two parchments are long woolled is not surprising given the origin of the parchments in the North of England where these breeds predominate. That we were able to locate the two parchments within this genetic landscape using 1x-sequencing data opens up the possibility of repeating this analysis with a wider range of parchments from a greater time depth. These results would allow for the changes introduced by the onset of industrial farming to be deduced and yield insights into sheep population movements from approximately the 13<sup>th</sup> century onwards. Moreover, by sampling parchment genomes in depth from similar time periods it may be possible to answer important ethnographic questions, for example whether all the parchments contained within the archbishop of Yorks register were made locally, or commandeered from wider population sources.

## ***7.6 Conclusions***

The findings of this pilot next generation sequencing study of parchment have shown it to be an amenable substrate for large-scale aDNA analysis. This result could have wide implications for domestic research given the vast quantities of parchments from various location and time points available for analysis. We were able to retrieve 8% and 9% of the ovine genome from PA1 and PA2 respectively at high quality (mapping quality  $\geq 30$ ) using half a lane each of an Illumina HiSeq 2000. This result suggests that the production of whole historic domestic genomes or targeted exon sequencing from parchment is a realistic possibility. These analyses could provide a fundamental understanding of the effects of the industrial breeding practices of

the 17<sup>th</sup> and 18<sup>th</sup> century (RYDER 1964a; b) had on domesticated genomes, as well as allowing a local domesticated populations to be studied over historical time. However, these ancient sheep populations may still look very different to the modern populations we see today.

We were able to provide a unique species assignment for both pieces using both proteomic and aDNA methods and estimate an external contamination rate comparable to other aDNA sources. This analysis contradicts claims made previously using PCR based methods which suggests parchment was a poor substrate for aDNA analysis (CAMPANA *et al.* 2010) and indicates that parchment should be viewed as an exciting source of aDNA. Whole mitochondrial genomes produced from both samples allowed their placement within the most populous sheep haplogroup and 1x SNP data allowed the modern day sheep breeds that most closely resemble the ancient samples to be identified. As stated previously, further sequencing of parchment samples from a variety of times periods and locations should allow for genetic maps from a variety of domestic species to be built, providing important insights into the relatively recent history of animal breeding.

## **8. Molecular archaeological analysis of a Croatian Avar with Pott's disease.**

### ***8.1 Introduction***

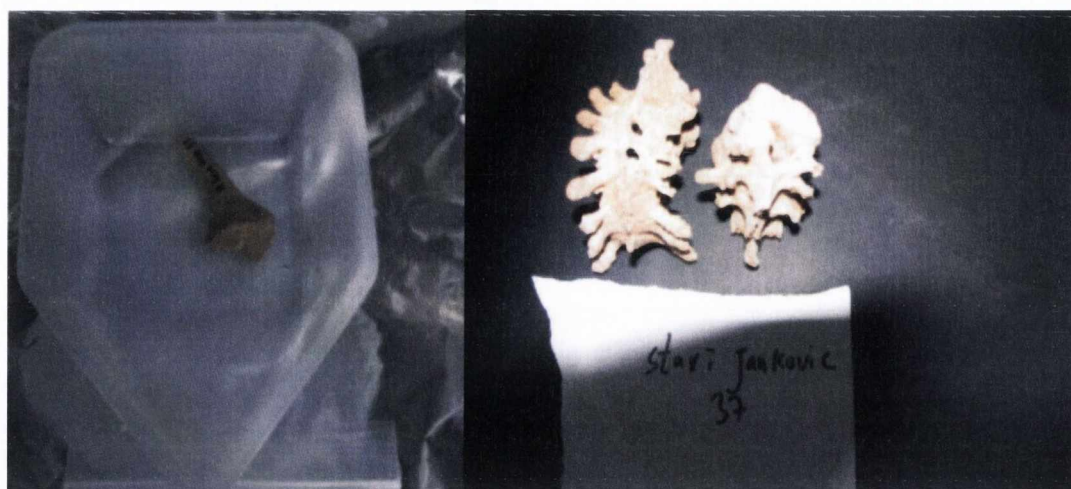
#### **8.1.1 The Avars**

The Avars were a major Eurasian horse riding population dating from the 6<sup>th</sup> to the 9<sup>th</sup> century AD (CURTA 2006; 2008) although little is known of their origins they are thought to have emerged as a group of elite nomadic horsemen circa 570AD from a homeland located in central Asia (CURTA 2008). The Avars had a rich material culture, which has allowed for the tracking of their interactions with local populations as they grew larger and became more sedentary (SINOR 1990; CURTA 2008); this culture therefore provides a useful reference from which to study medieval population movements.

#### **8.1.2 CTB6; an Avar individual with Pott's disease**

The individual analysed in this pilot study CTB6 was sampled from skeletal remains recovered from the Stari Jankavci cemetery (grave 37) dating from the 7<sup>th</sup> to 8<sup>th</sup> century AD; this site has been previously described as having excellent osteological preservation (SLAUS 1994). Anthropological studies on the full skeletal remains also indicate that this individual was suffering from tuberculosis arthritis of the intervertebral joints or Pott's disease (Ron Pinhasi, personal communication), suggestive of a systemic *Mycobacterium tuberculosis* infection (Figure 8:1).





**Figure 8:1** CTB6 metatarsal sample submitted for aDNA right and vertebra showing signs of Pott's disease left.

### 8.1.3 Ancient DNA and pathogens

The study of ancient bacterial pathogens using PCR based methods has been plagued by controversy concerning the authenticity of results, with the technical difficulties of teasing apart true pathogen sequences from homologous environmental contaminants widely discussed in the literature (RAOULT *et al.* 2000; GILBERT *et al.* 2004; BARNES and THOMAS 2006). The introduction of NGS and hybridisation capture techniques have however, reinvigorated the field with the retrieval of the first whole ancient pathogen genome (BOS *et al.* 2011). NGS methods, as discussed in previous chapters, also provide the added benefit of allowing for a greater depth of analysis to be undertaken into the metagenome of archaeological samples, which may allow for a more accurate taxonomic assignment of the discovered sequences.

8.1.4 Objectives of this chapter

This chapter presents a case study of a single human individual with Pott’s disease, an osteological marker for a systemic *Mycobacterium* infection. It aims to assess the authenticity of the *Mycobracterium* sequences found in the ancient DNA extract. Moreover to use 1x sequencing of SNPs to locate the Avar individual within the modern human populations of the Human genome diversity project.

8.2 Materials and methods

8.2.1 Archaeological samples

DNA was extracted from a single Avar individual CTB6 with morphological features suggestive of Pott’s disease.

| Sample | Site                             | Bone sample | Age                                                               |
|--------|----------------------------------|-------------|-------------------------------------------------------------------|
|        |                                  |             |                                                                   |
| CTB6   | Croatia, Stari Jankavci grave 37 | Metatarsal  | 7 <sup>th</sup> -8 <sup>th</sup> century AD<br>(Dated by context) |

Table 8:1: Archaeological sample CTB6 used in this study.

8.2.2 Ancient DNA extraction

All aDNA extractions were preformed in a dedicated ancient DNA laboratory at Trinity College Dublin following the standard protocol outline in the methods chapter (chapter 2).

### **8.2.3 Illumina sequencing library preparation**

Illumina single-end sequencing libraries were produced from CTB6 via the PCR enrichment of end repaired adaptor-ligated DNA templates again following the protocol outline in the methods chapter. CTB6 was sequenced twice on a single lane of an Illumina HiSeq2000 at the Beijing Genomics Institute as part of a multiplex with other samples (data not shown).

### **8.2.4 Initial sequencing QC**

Initial quality control of sequencing reads were preformed using the standard methods outlined in chapter 2

### **8.2.5 BWA sequence Alignment**

The raw trimmed reads were then aligned to the human hg18 reference genome (downloaded 24-8-12) using BWA (version 0.6.1) (LI and DURBIN 2009). Standard alignment settings for BWA were used (Chapter 2).

### **8.2.6 Bowtie sequence Alignment**

All reads were further aligned to the *M. tuberculosis* H37Rv (Downloaded 24-8-12) genome using BOWTIE. BOWTIE was chosen for this analysis as this program allowed a greater level of control over the alignment. BOWTIEs options were modified to allow for only 1 mismatch between the sequencing read and the *M. tuberculosis* genome. The resultant reads were then further filtered for clonal PCR amplified products and a mapping quality of 30 using the SAMTOOLS rmdup and view commands.

### **8.2.7 SNP calling and analysis**

SNPs were analysed following established protocols for 1X aDNA NGS data (SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012) and analysed using EIGENSOFT following the standard method of chapter 2 with no modifications.

## **8.3 Results**

### **8.3.1 CTB6 sequencing results**

In total CTB6 produced 39,053,538 reads of which 386,346 (0.98%) aligned to the human genome (hg18) with a minimum mapping quality of 30. Once duplicate clonal amplified reads were removed this value fell to 358,735 (0.91%).

### **8.3.2 PCA analysis**

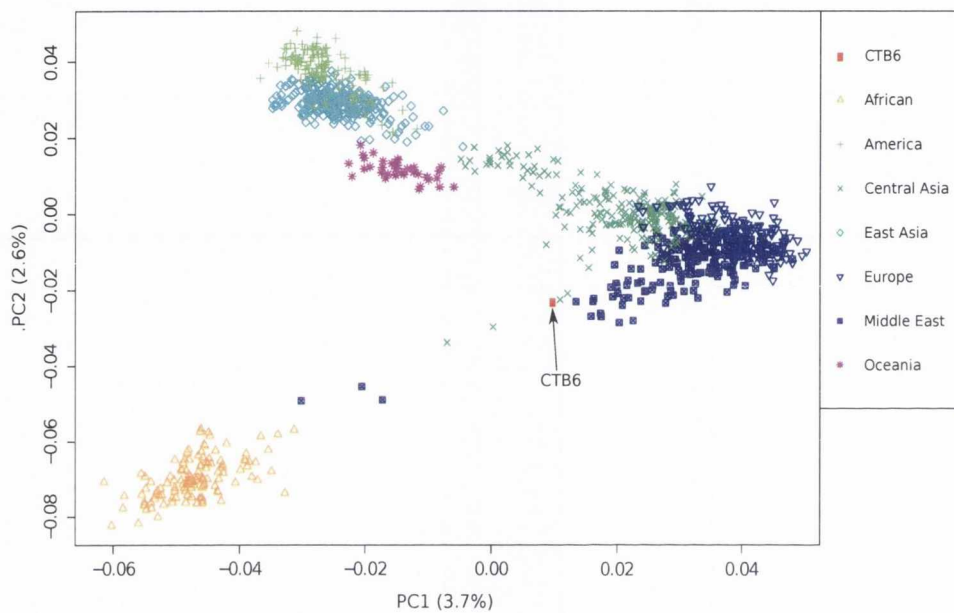
As described in previous chapters, low coverage sequencing of ancient genomes allows for the calling of haploid SNP genotypes, which can then be compared to high-density SNP typed modern reference datasets. In this analysis CTB6 was placed within a worldwide human SNP dataset the HGDP panel. After EIGENSOFTs filters were applied 2,244 SNPs were retained for further study. The first PCA analysis centred on placing CTB6 within the worldwide dataset as seen in figure 8:2, in this analysis we see CTB6 falling



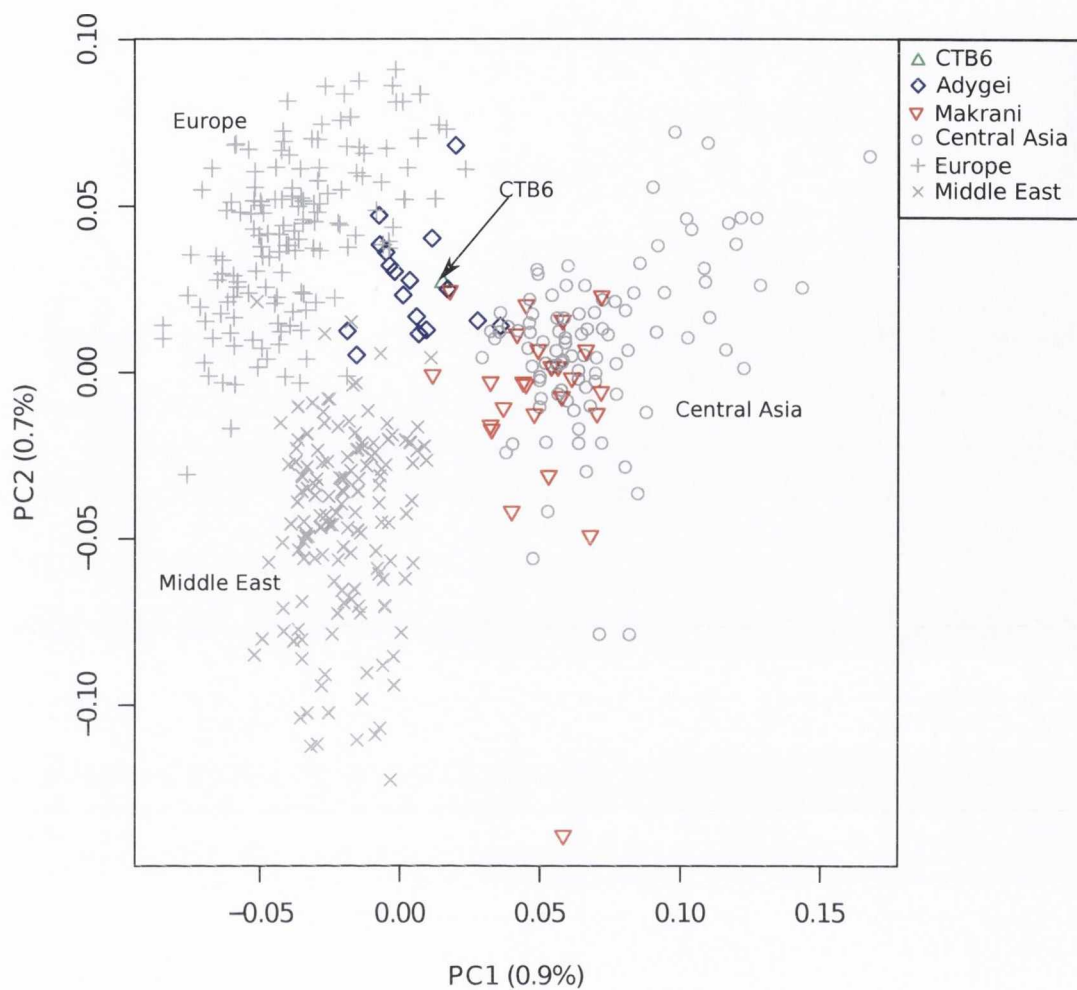
outside of the major European grouping next to an individual from central Asia. To further investigate this relationship the analysis was repeated using only populations from Europe, central Asia and the Middle East. In this analysis CTB6 falls between two central Asian/Caucasus populations; the Adeygi and the Makrani.

|      | Raw reads  | Aligned reads<br>mapping<br>quality $\geq 30$ | Aligned reads<br>mapping<br>quality $\geq 30$<br>(%) | Duplicates<br>removed | Duplicates<br>removed (%) | Filters applied<br>(XT,X1 tags) | Filters<br>applied<br>(%) | SNPs called<br>HGDp |
|------|------------|-----------------------------------------------|------------------------------------------------------|-----------------------|---------------------------|---------------------------------|---------------------------|---------------------|
| CTB6 | 39,053,538 | 386,346                                       | 0.98                                                 | 358,735               | 0.91                      | 358,735                         | 0.91                      | 2,245               |

Table 8:2 CTB6 shotgun sequencing results.



**Figure 8:2: SMARTPCA analysis, CTB6 and worldwide populations (2,244SNPs)**



**Figure 8:3 PCA analysis CTB and European Central Asian and Middle Eastern populations (2,244SNPs)**



### 8.3.3 Pathogen aDNA analysis.

Given that CTB6's skeleton showed signs of Pott's disease, the possibility that *M. tuberculosis* sequences could be recovered from CTB6's sequencing reads was investigated. To do this all reads were re-aligned to the *M.tuberculosis* H37Rv genome using BOWTIE and strict alignment settings of only one mismatch between the length of the read and the reference genome were implemented, to try to counteract the effects of homology between bacterial genomes. This analysis produced very few hits to the bacterium's genome 10,612 (0.027%) in total. However, this result might be expected as we are analysing a metatarsal bone from the skeleton, which, is unlikely to contain high concentrations of the *M.tuberculosis* bacterium. To further investigate the authenticity of the proposed *M.tuberculosis* reads recovered, these sequences were blasted against a custom database containing only mycobacterium sequences from a phylogenetic analysis by Dai *et al* (2011). The results of this analysis suggest that only 10% (1,113) of CTB6 TB aligned reads could be unambiguously assigned to the *M tuberculosis* in this manner. Moreover as shown in figure 8:5 read stacking is seen around non-coding RNA genes in the *M.tuberculosis* genome when the alignments are visualised suggesting that many of the sequences retrieved may be from homologous bacterial genomes.

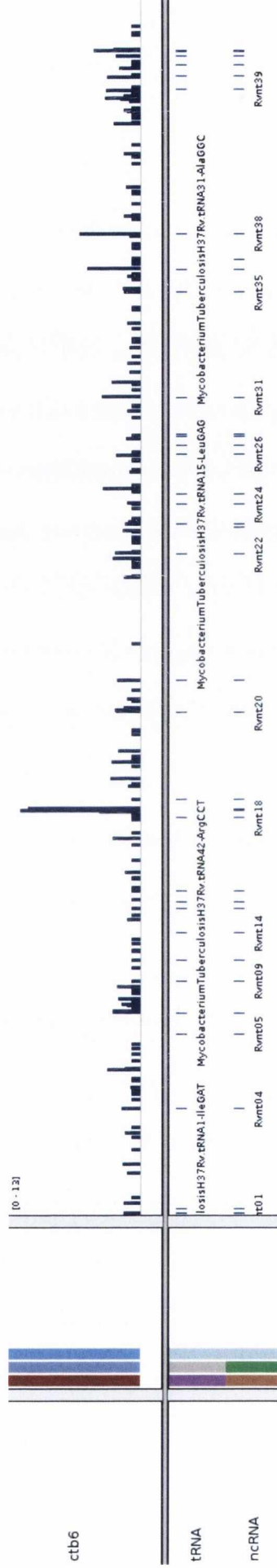


Figure 8:4 Screenshot from IGV viewer showing stacking of CTB6 reads around ncRNA genes in the *M. tuberculosis* genome. The top line of the plot depicts the depth of coverage over complete *M. tuberculosis* genome the next two lines positions of tRNA and all ncRNA genes.

## **8.4 Discussion.**

Human population movements are a constant theme in archaeology from Palaeolithic to relatively modern times (PINHASI *et al.* 2012). In this chapter the results of a pilot population genetic analysis of a single Avar individual from Croatia are presented. As discussed previously the Avar's had a rich material culture (CURTA 2008), which allows for the accurate tracking of their dispersals making them a useful test population for the study of medieval Europe.

### **8.4.1. PCA analyses.**

The results of the CTB6/HGDP PCA analysis suggest that CTB6 shares more affinity to some outlying European and central Asia populations than that of central Europeans as would be hypothesized from a sample retrieved from Croatia. The affinity of CTB6 to the Adygei and the Makrani populations is interesting given that one of these groups the Adygei are located around the Caspian and Black Sea coasts of Russian a propose Avar homeland and the other the border of Iran and Pakistan. Further analyses incorporating CTB into SNP dataset which include more European populations like that produced by the POPRES study (NELSON *et al.* 2008) and an addition of a HGDP consistent dataset with increase caucous populations would allow the genetic affinities of CTB to be asses in greater detail via the use of PCA and structure analyses. These analyses would compliment the analysis presented in this chapter, which is limited, by the relatively small number of European populations in the

HGDP dataset. However unlike the HGDP dataset the POPRES data is not widely publically available and access was not granted in time for use in this study.

#### **8.4.2. Molecular investigation for *M tuberculosis***

CTB6 was originally chosen for study because of skeleton signs of Pott's disease, making it a good candidate for the study of ancient pathogen DNA. However, the results of the analysis presented above suggests that if *M. tuberculosis* sequences exist in these data then they do so at extremely low concentrations. Moreover, of the sequences that could be retrieved and aligned to *M. tuberculosis* only 10% could be confidently assigned to *M. tuberculosis* in this analysis.

The bone sample chosen for this analysis however, is unlikely to provide high concentrations of pathogens. In this choice however a balance was sought between a sample whose integrity was sufficiently maintained to allow DNA retrieval and one that contained sufficient levels of pathogens for detection. The results of the alignments to the human genome suggest that this sample does provide good endogenous aDNA recovery rates.

### **8.5 Conclusion**

In this study of a single Avar individual sufficient amounts of ancient DNA were retrieved to provide a first assessment of this individual's affinity in a



modern high density SNP dataset, this analysis suggests that this individual shares some relation with either Caucasus or central Asian populations. This result adds to the growing literature that suggests 1x genotyping allows the placement of aDNA samples within modern day populations. However it is likely that these affinities may not be a true result due to the decades of inbreeding and migration between these populations. In the case of the molecular assignment of a *M. tuberculosis* infection, insufficient data was retrieved to make an accurate assessment although some authentic sequence matches were proposed. As the Avars represent well-defined cultural group further studies of this population including more samples of varying locations and a population based haploid PCA approach should provide insights into the population dynamics of early medieval Europe.

## 9. General conclusions

### *9.1 Summary of results and their significance*

The introduction of next generation DNA sequencing technologies into field of aDNA has seen a revolution in the amount of data generated. However as mentioned previously methodological improvements are require for the power of these techniques to be fully utilised. This thesis presents the results from multiple pilot studies applying new analysis methods to samples from multiple species and substrates. The following chapter provides a brief overview of the conclusions from each of these studies and a discussion of the future of genomic scale aDNA analyses.

#### **9.1.1 Hybridisation capture, mtDNA and exon**

Chapters 3 and 5 present the results of two hybridisation capture experiments applied to ancient DNA. The first of these chapter 3 presents the results of a mtDNA capture study. This study was a success producing 11 high-coverage genomes (3x coverage greater than 85% of the genome), 10 of which were then subsequently used in a Bayesian phylogenetic analyses (BEAST) to try to provide a greater resolution of the domestication process in two species, sheep and cow. This analysis supported the two domestications predicted for domestic cattle (LOFTUS *et al.* 1994) but suggested that a single domestication could be possible for sheep something that contradicts previous reports (MEADOWS *et al.* 2010).

In the bovine analysis estimates of the divergence times between all domestic cattle clades fall into a time period that significantly predates that of the domestication process. Moreover, the result suggested that the genetic variation within the major domestic taurine cattle T clade cannot all be explained by post domestication mutation, suggesting at least some genetic diversity incorporated from an earlier population expansion. This population expansion was likely a product of the range expansion of these species from their refugia populations after the last glacial maximum (EDWARDS *et al.* 2007b). Further simulation data has suggested that this initial domestic maternal population is likely to have been small encompassing only roughly 80 animals (BOLLONGINO *et al.* 2012), providing further evidence that this ancestral population may have been diverse as a small number of animals has capture a range of haplogroups.

The ovine analysis however, presents a more ambiguous interpretation, which is likely due to the limited number of both modern and ancient samples used in this analysis and their less secure dating. In two BEAST analyses we found sheep haplotype divergence times before or after the domestication horizon depending on the ancient samples used. These widely varying dates produced by this analysis call for further studies to be undertaken to provide a clearer picture of the timing of the divergences between these groups. The results of these two experiments taken together however, reiterate the power of hybridisation capture in the study of ancient mtDNA, which is likely helped by the high copy number of mtDNA genomes in cells making them amenable to the capture of aDNA.

The results of a nuclear exon capture study were less conclusive with the analysis working well for CPC98 highlighting multiple known and novel variants. However, this bone sample has provided good aDNA results for almost 20 years and is the subject of a whole genome sequence (Chapter 4). The results from 5 less well-preserved samples are more modest, suggesting there may be a limit to the preservation state of samples that can be used in these analyses, although methodological problems could also not be discounted in the poorer quality of these alignments.

In sum the use of hybridisation capture for the analysis of both nuclear and mtDNA aDNA is still in its infancy and further improvements in both aDNA capture and extraction methods should provide further insights into the evolution of multiple extant and extinct species.

### **9.1.2 Population genetic analysis of a Mesolithic aurochs CPC98**

As mentioned above the British Mesolithic aurochs specimen CPC98 has provided consistent aDNA results for almost two decades and has recently had a draft whole genome sequencing analysis completed. High coverage SNP genotypes produced by this study allowed for multiple population genetic analyses to be undertaken (Chapter 4) and provided a first analysis of CPC98's position in relation to the global domestic cattle populations using nuclear SNP data. The tentative results from this analysis suggest that CPC98 has a closer relationship to European taurine animals than African implying that the European aurochs may have introgressed into the *Bos taurus* population post domestication. However, this analysis centres on a single individual and further



study of more aurochs specimens and domestic cattle from a range of ages is needed for a complete picture to be produced.

### 9.1.3 Haploid ancient DNA data

Recent studies have shown the power of haploid SNP data for the analysis of ancient DNA samples (GREEN *et al.* 2010; REICH *et al.* 2010; SKOGLUND *et al.* 2012; SÁNCHEZ-QUINTO *et al.* 2012) we applied these methods to two sets of bone specimens, two 8,000-year-old cattle samples (Chapter 6) and one from a Avar individual from Croatia (Chapter 8).

The analysis of two samples from western Anatolia suggested unsurprisingly that they share a greater genetic affinity towards the domestic cattle of Europe than that of present-day indicine admixed Anatolian populations and that the first domestic cattle populations of Anatolia were likely unadmixed *Bos taurus*. Interestingly, even at this early stage in the history of domestic cattle neither ancient western Anatolian sample placed near to the African taurine animals, which are consistent outliers among multiple analyses including *Bos taurus*. Multiple theories could explain this divergence, including admixture of African aurochs or perhaps population structure in the original domestic *Bos taurus* animals with separate Near Eastern contributions to the two derivative continents.

In the second analysis of a Croatia Avar sample we found this individual to share a greater affinity to Caucas and central Asian populations than that of

Western Europe, suggesting that this person may have an ancestry indicating towards the Avar homeland.

#### **9.1.4 Parchment as a source of aDNA**

Given the power of next generation sequencing to allow the analysis of ancient genomes, new and more efficient substrates for analysis are required. In chapter 7 the results of a pilot analysis using two pieces of parchment dated to the 18<sup>th</sup> and 17<sup>th</sup> century was described. Parchment was found to be amenable for large-scale aDNA analysis with over 8% of the ovine genome recovered in each sample using half a lane of an Illumina HiSeq. These sequencing results allowed for the placing of these parchments within the distribution of domestic sheep populations and suggested an affinity to breeds of British long wool sheep that is consistent with their proposed origin in Britain (RYDER 1964a; b).

### ***9.2 Future prospects***

#### **9.2.1 New NGS library making technologies and 3<sup>rd</sup> generation sequencing**

Next generation sequencing has provided a vast amount of data for ancient DNA studies. However, the pace of the field continues to increase with the recent publications of enhanced aDNA library production protocols (ORLANDO *et al.* 2011; MEYER *et al.* 2012; GANSAUGE and MEYER 2013) subsequently allowing for the 30x sequencing of the Denisovan genome (MEYER *et al.* 2012). The incorporation of new substrates such as parchment (chapter 7),

which are not encumbered by the limitations of bone (high exogenous DNA concentration and low total concentrations of DNA) are likely to increase this flood of genomic level data.

The introduction of third generation sequencing methodologies to aDNA is also likely to swell the amount of data generated further. One of these new techniques is proposed to allow for the sequencing of large quantities of DNA using a USB key sized device without the need for PCR amplification (<http://www.nanoporetech.com/>). This methodology may have profound implication for aDNA sampling as this device could in principle allow the sequencing of ancient samples in their host institutions, without the need to bring the sample to specialised ancient DNA facilities. This could lead to samples being assessed in situ and then only move to aDNA labs when they have been deduced to contain high levels of endogenous DNA and are therefore tractable for large scale NGS analysis. This pipeline would vastly reduce the number of samples needed to be removed from collections as well as decreasing the cost of analysis substantially. However, these new technologies will likely bring with them their own challenges and adaptations needed for their successful use in aDNA.

### **9.2.1 Further studies**

Given the vast quantities of data provided by NGS analysis of ancient DNA it can be asserted that aDNA has entered the genomics age. aDNA experiments are no longer confined to PCR analysis of small parts of the mitochondrial genome, which only offer limited resolution into the genetics of past

populations. It is also sensible to assume that studying the genomic effects of the forces that have shaped the human and major farm animal genomes since domestication will allow for major advances in the fields of medicine and agricultural science. The domestication process has brought about substantial changes to both these populations and their genomes will have adapted to the challenges. The fluid parts of the genome that can respond quickly to these challenges particularly those relating to the immune response are likely to be again the targets of adaption to new disease challenges to these populations. Both species are also likely to suffer from pathologies that were once selected benefits from this coevolution. Ancient DNA allows for this process to be tracked in real time and these regions to be identified. The paragraphs below outline two hypothetical experiments that are likely to become possible in the next 10 years and build upon the analyses and technologies outlined in this thesis and have major implications for agricultural science.

### **9.2.2 Population genetics of bovine domestication**

It has long been asserted that cattle were subjected to two independent domestications one in the Near East and one in the Indus valley (LOFTUS *et al.* 1994; BRADLEY *et al.* 1996; CHEN *et al.* 2009). This assumption has been repeated multiple times in this thesis and substantiated by genomewide data (BOVINE HAPMAP CONSORTIUM *et al.* 2009) and analyses in chapters 3,4 and 6. However, a recent publication has suggested a single domestication location for all cattle in the Near East with decedents of these animals used to seed the indicine population (LARSON and BURGER 2013). Ancient DNA provides an excellent way to test both hypotheses by allowing the sequencing of a genome of an early domesticate zebu animal. The Indus valley because of its latitude



was once not thought to be tractable for ancient DNA analysis however new methods like hybridisation capture may make aDNA analysis from some samples from this region possible in the near future.

At present only one low quality indicus genome has been published (CANAVEZ *et al.* 2012) and all modern day analyses are likely plagued by cryptic taurine introgression. An ancient genome would therefore allow for unbiased analyses based on both nuclear and mtDNA to see if at this early stage in Zebu cattle domestication, we do have two independent genomes or a hybrid *Bos taurus* genome. Two methods outlined already in this thesis would allow this analysis to be completed firstly the D-statistic of Green *et al* (2010) could be applied to even a low coverage 1x genome. If larger genome coverage was achieved then a STRUTURE analysis could be attempted as well. Both these methods should show clearly if the early zebu animals are introgressed or pure bred and therefore answer this open question as to the fundamental origins of domestic indicine cattle.

### **9.2.2 York parchments**

As outline in chapter 7 parchment provides an excellent substrate for the analysis of historic domesticate genomes. The parchment archive of York University library is an immensely important store of genetic information, containing manuscripts from the 13<sup>th</sup> to the 18<sup>th</sup> century. These artefacts provide a sample of sheep diversity in York over important historical periods that cover major social and economic changes. Specifically in the intensification of animal breeding (RYDER 1964b; a). If money was no object an analysis could be undertaken to look at sheep diversity in York over a 600-

year period. An initial program could be undertaken to sample parchments in 25-year intervals from the 13<sup>th</sup> to the 19<sup>th</sup> century these could then be genotyped using RNA capture of SNPs from the currently developed ovine 770k chip. RNA capture would allow the genotyping of many samples to a usable depth of coverage to call full genotype data and counter some of the problems with the loss of data introduced by using 1x genotypes (chapters 6,7,8). This analysis would also lower the cost of the analysis when compared to 1x genotyping where you sequence large regions of the genomes which at present contain no modern SNP information. These genotypes could be then placed into PCA analysis to look for profound genetic changes over time for example the introduction of Leicester rams into what was once locally developed breeds (RYDER 1964a; b). The loss of heterozygosity and effective population size associated with domestication could also be investigated using currently available tools in PLINK. These genotypes could also be supplemented by the genotyping of modern rare breeds from the location, which the parchment samples have been taken to increase the resolution of the analysis.

It has already been shown that the parchment quality decreases with time in the York archive (Fiddymment S, Personal communication) using a proteomic analysis. It would be interesting to compare this to a genetic analysis of the parchment to see if this reduction of quality can be explained by a shift to a different type of sheep breeding. For example away from wool to meat production which may explain the reduction in parchment quality.

The 25-year samples could then be used to identify outlying or novel samples that would be worth whole genome sequencing. These whole genome sequences would then allow pairwise comparisons between modern and historic samples to identify vast numbers of genome wide SNPs between the animals. This study would likely highlight interesting regions of the genome to study through time. It would also allow genomewide population genetic tools to be implemented for example the PSMC program of Li and Durbin (2011) already used to great effect in modern populations to look at the effective population size of a population through time (LI and DURBIN 2011).

Both of the analyses described above highlight the importance of aDNA to answer fundamental questions in animal breeding.

## 10. References

- 1000 GENOMES PROJECT CONSORTIUM, 2010 A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.
- ACHILLI A., BONFIGLIO S., OLIVIERI A., MALUSÀ A., PALA M., KASHANI B. H., PEREGO U. A., AJMONE-MARSAN P., LIOTTA L., SEMINO O., BANDELT H.-J., FERRETTI L., TORRONI A., 2009 The Multifaceted Origin of Taurine Cattle Reflected by the Mitochondrial Genome (V Macaulay, Ed.). *PLoS ONE* **4**: e5753.
- ACHILLI A., OLIVIERI A., PELLECCIA M., UBOLDI C., COLLI L., AL-ZAHERY N., ACCETTURO M., PALA M., HOOSHIAR KASHANI B., PEREGO U. A., BATTAGLIA V., FORNARINO S., KALAMATI J., HOUSHMAND M., NEGRINI R., SEMINO O., RICHARDS M., MACAULAY V., FERRETTI L., BANDELT H.-J., AJMONE-MARSAN P., TORRONI A., 2008 Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Curr. Biol.* **18**: R157–8.
- AKHTER Q., MASOOD A., ASHRAF R., MAJID S., RASOOL S., KHAN T., RASHID T., SAMEER A. S., GANAI B. A., 2012 Polymorphisms in the 3'UTR of the human leptin gene and their role in hypertension. *Mol Med Report* **5**: 1058–1062.
- ALLENTOFT M. E., COLLINS M., HARKER D., HAILE J., OSKAM C. L., HALE M. L., CAMPOS P. F., SAMANIEGO J. A., GILBERT M. T. P., WILLERSLEV E., ZHANG G., SCOFIELD R. P., HOLDAWAY R. N., BUNCE M., 2012 The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences*.
- ANDERSON S., DE BRUIJN M. H., COULSON A. R., SANGER F., YOUNG I. G., 1982 Complete sequence of bovine mitochondrial DNA. Conserved features of the mammalian mitochondrial genome. *Journal of molecular* ....
- ARBUCKLE B. S., 2013 The late adoption of cattle and pig husbandry in Neolithic Central Turkey. *Journal of Archaeological Science* **40**: 1805–1815.
- ÁVILA-ARCOS M. C., CAPPELLINI E., ROMERO-NAVARRO J. A., WALES N., MORENO-MAYAR J. V., RASMUSSEN M., FORDYCE S. L., MONTIEL R., VIELLE-CALZADA J.-P., WILLERSLEV E., GILBERT M. T. P., 2011 Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Sci. Rep.* **1**.
- BAI Y., SARTOR M., CAVALCOLI J., 2012 Current status and future perspectives for sequencing livestock genomes. *Journal of Animal Science and Biotechnology* **3**: 8.
- BAILEY J. F., RICHARDS M. B., MACAULAY V. A., COLSON I. B., JAMES I. T., BRADLEY D. G., HEDGES R. E., SYKES B. C., 1996 Ancient DNA suggests a recent expansion of European cattle from a diverse wild progenitor



species. *Proc. Biol. Sci.* **263**: 1467–1473.

BAR YOSEF O., 1998 The Natufian culture in the Levant, threshold to the origins of agriculture. *Evol. Anthropol.* **6**: 159–177.

BARNES I., THOMAS M. G., 2006 Evaluating bacterial pathogen DNA preservation in museum osteological collections. *Proceedings of the Royal Society B: Biological Sciences* **273**: 645–653.

BARNES I., MATHEUS P., SHAPIRO B., JENSEN D., COOPER A., 2002 Dynamics of Pleistocene population extinctions in Beringian brown bears. *Science* **295**: 2267–2270.

BARNES I., SHAPIRO B., LISTER A., KUZNETSOVA T., SHER A., GUTHRIE D., THOMAS M. G., 2007 Genetic Structure and Extinction of the Woolly Mammoth, *Mammuthus primigenius*. *Current Biology* **17**: 1072–1075.

BEJA-PEREIRA A., CAMELLI D., LALUEZA-FOX C., VERNESI C., FERRAND N., CASOLI A., GOYACHE F., ROYO L. J., CONTI S., LARI M., MARTINI A., OURAGH L., MAGID A., ATASH A., ZSOLNAI A., BOSCATO P., TRIANTAPHYLIDIS C., PLOUMI K., SINEO L., MALLEGNI F., TABERLET P., ERHARDT G., SAMPIETRO L., BERTRANPETIT J., BARBUJANI G., LUIKART G., BERTORELLE G., 2006 The origin of European cattle: evidence from modern and ancient DNA. *Proc Natl Acad Sci U S A* **103**: 8113–8118.

BEJA-PEREIRA A., LUIKART G., ENGLAND P. R., BRADLEY D. G., JANN O. C., BERTORELLE G., CHAMBERLAIN A. T., NUNES T. P., METODIEV S., FERRAND N., ERHARDT G., 2003 Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nature Genetics* **35**: 311–313.

BENGTTSSON C. F., OLSEN M. E., BRANDT L. Ø., BERTELSEN M. F., WILLERSLEV E., TOBIN D. J., WILSON A. S., GILBERT M. T. P., 2012 DNA from keratinous tissue. Part I: hair and nail. *Ann. Anat.* **194**: 17–25.

BENTLEY D. R., BALASUBRAMANIAN S., SWERDLOW H. P., SMITH G. P., MILTON J., BROWN C. G., HALL K. P., EVERS D. J., BARNES C. L., BIGNELL H. R., BOUTELL J. M., BRYANT J., CARTER R. J., KEIRA CHEETHAM R., COX A. J., ELLIS D. J., FLATBUSH M. R., GORMLEY N. A., HUMPHRAY S. J., IRVING L. J., KARBELASHVILI M. S., KIRK S. M., LI H., LIU X., MAISINGER K. S., MURRAY L. J., OBRADOVIC B., OST T., PARKINSON M. L., PRATT M. R., RASOLONJATOVO I. M. J., REED M. T., RIGATTI R., RODIGHIERO C., ROSS M. T., SABOT A., SANKAR S. V., SCALLY A., SCHROTH G. P., SMITH M. E., SMITH V. P., SPIRIDOU A., TORRANCE P. E., TZONEV S. S., VERMAAS E. H., WALTER K., WU X., ZHANG L., ALAM M. D., ANASTASI C., ANIEBO I. C., BAILEY D. M. D., BANCARZ I. R., BANERJEE S., BARBOUR S. G., BAYBAYAN P. A., BENOIT V. A., BENSON K. F., BEVIS C., BLACK P. J., BOODHUN A., BRENNAN J. S., BRIDGHAM J. A., BROWN R. C., BROWN A. A., BUERMANN D. H., BUNDU A. A., BURROWS J. C., CARTER N. P., CASTILLO N., CHIARA E CATENAZZI M., CHANG S., NEIL COOLEY R., CRAKE N. R., DADA O. O., DIAKOUMAKOS K. D., DOMINGUEZ-FERNANDEZ B., EARNSHAW D. J., EGBUJOR U. C., ELMORE D. W., ETCHIN S. S., EWAN M. R., FEDURCO M., FRASER L. J., FUENTES FAJARDO K. V., SCOTT FUREY W., GEORGE D.,

GIETZEN K. J., GODDARD C. P., GOLDA G. S., GRANIERI P. A., GREEN D. E., GUSTAFSON D. L., HANSEN N. F., HARNISH K., HAUDENSCHILD C. D., HEYER N. I., HIMS M. M., HO J. T., HORGAN A. M., HOSCHLER K., HURWITZ S., IVANOV D. V., JOHNSON M. Q., JAMES T., HUW JONES T. A., KANG G.-D., KERELSKA T. H., KERSEY A. D., KHRIBTUKOVA I., KINDWALL A. P., KINGSBURY Z., KOKKO-GONZALES P. I., KUMAR A., LAURENT M. A., LAWLEY C. T., LEE S. E., LEE X., LIAO A. K., LOCH J. A., LOK M., LUO S., MAMMEN R. M., MARTIN J. W., MCCAULEY P. G., MCNITT P., MEHTA P., MOON K. W., MULLENS J. W., NEWINGTON T., NING Z., LING NG B., NOVO S. M., O'NEILL M. J., OSBORNE M. A., OSNOWSKI A., OSTADAN O., PARASCHOS L. L., PICKERING L., PIKE A. C., PIKE A. C., CHRIS PINKARD D., PLISKIN D. P., PODHASKY J., QUIJANO V. J., RACZY C., RAE V. H., RAWLINGS S. R., CHIVA RODRIGUEZ A., ROE P. M., ROGERS J., ROBERT BACIGALUPO M. C., ROMANOV N., ROMIEU A., ROTH R. K., ROURKE N. J., RUEDIGER S. T., RUSMAN E., SANCHES-KUIPER R. M., SCHENKER M. R., SEOANE J. M., SHAW R. J., SHIVER M. K., SHORT S. W., SIZTO N. L., SLUIS J. P., SMITH M. A., ERNEST SOHNA SOHNA J., SPENCE E. J., STEVENS K., SUTTON N., SZAJKOWSKI L., TREGIDGO C. L., TURCATTI G., VANDEVONDELE S., VERHOVSKY Y., VIRK S. M., WAKELIN S., WALCOTT G. C., WANG J., WORSLEY G. J., YAN J., YAU L., ZUERLEIN M., ROGERS J., MULLIKIN J. C., HURLES M. E., MCCOOKE N. J., WEST J. S., OAKS F. L., LUNDBERG P. L., KLENERMAN D., DURBIN R., SMITH A. J., 2008 Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

BEVILACQUA C., HELBLING J. C., MIRANDA G., MARTIN P., 2006 Translational efficiency of casein transcripts in the mammary tissue of lactating ruminants. *Reprod. Nutr. Dev.* **46**: 567–578.

BOLLONGINO R., BURGER J., POWELL A., MASHKOUR M., VIGNE J.-D., THOMAS M. G., 2012 Modern Taurine Cattle Descended from Small Number of Near-Eastern Founders. *Mol Biol Evol.*

BOLLONGINO R., EDWARDS C. J., ALT K. W., BURGER J., BRADLEY D. G., 2006 Early history of European domestic cattle as revealed by ancient DNA. *Biology Letters* **2**: 155–159.

BOLLONGINO R., ELSNER J., VIGNE J.-D., BURGER J., 2008 Y-SNPs Do Not Indicate Hybridisation between European Aurochs and Domestic Cattle (S Shennan, Ed.). *PLoS ONE* **3**: e3418.

BONFIGLIO S., ACHILLI A., OLIVIERI A., NEGRINI R., COLLI L., LIOTTA L., AJMONE-MARSAN P., TORRONI A., FERRETTI L., 2010 The Enigmatic Origin of Bovine mtDNA Haplogroup R: Sporadic Interbreeding or an Independent Event of *Bos primigenius* Domestication in Italy? (T Kivisild, Ed.). *PLoS ONE* **5**: e15760.

BOS K. I., SCHUENEMANN V. J., GOLDING G. B., BURBANO H. A., WAGLECHNER N., COOMBES B. K., MCPHEE J. B., DEWITTE S. N., MEYER M., SCHMEDES S., WOOD J., EARN D. J. D., HERRING D. A., BAUER P., POINAR H. N., KRAUSE J., 2011 A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature*: 1–6.



BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM, ELSIK C. G.,  
 TELLAM R. L., WORLEY K. C., GIBBS R. A., MUZNY D. M., WEINSTOCK G.  
 M., ADELSON D. L., EICHLER E. E., ELNITSKI L., GUIGÓ R., HAMERNIK D.  
 L., KAPPES S. M., LEWIN H. A., LYNN D. J., NICHOLAS F. W., REYMOND  
 A., RIJNKELS M., SKOW L. C., ZDOBNOV E. M., SCHOOK L., WOMACK J.,  
 ALIOTO T., ANTONARAKIS S. E., ASTASHYN A., CHAPPLE C. E., CHEN H.-  
 C., CHRAST J., CÂMARA F., ERMOLAEVA O., HENRICHSEN C. N., HLAVINA  
 W., KAPUSTIN Y., KIRYUTIN B., KITTS P., KOKOCINSKI F., LANDRUM M.,  
 MAGLOTT D., PRUITT K., SAPOJNIKOV V., SEARLE S. M., SOLOVYEV V.,  
 SOUVOROV A., UCLA C., WYSS C., ANZOLA J. M., GERLACH D., ELHAIK E.,  
 GRAUR D., REESE J. T., EDGAR R. C., MCEWAN J. C., PAYNE G. M.,  
 RAISON J. M., JUNIER T., KRIVENTSEVA E. V., EYRAS E., PLASS M.,  
 DONTU R., LARKIN D. M., REECY J., YANG M. Q., CHEN L., CHENG Z.,  
 CHITKO-MCKOWN C. G., LIU G. E., MATUKUMALLI L. K., SONG J., ZHU  
 B., BRADLEY D. G., BRINKMAN F. S. L., LAU L. P. L., WHITESIDE M. D.,  
 WALKER A., WHEELER T. T., CASEY T., GERMAN J. B., LEMAY D. G.,  
 MAQBOOL N. J., MOLENAAR A. J., SEO S., STOTHARD P., BALDWIN C. L.,  
 BAXTER R., BRINKMEYER-LANGFORD C. L., BROWN W. C., CHILDERS C.  
 P., CONNELLEY T., ELLIS S. A., FRITZ K., GLASS E. J., HERZIG C. T. A.,  
 IIVANAINEN A., LAHMERS K. K., BENNETT A. K., DICKENS C. M., GILBERT  
 J. G. R., HAGEN D. E., SALIH H., AERTS J., CAETANO A. R., DALRYMPLE  
 B., GARCIA J. F., GILL C. A., HIENDLEDER S. G., MEMILI E., SPURLOCK D.,  
 WILLIAMS J. L., ALEXANDER L., BROWNSTEIN M. J., GUAN L., HOLT R. A.,  
 JONES S. J. M., MARRA M. A., MOORE R., MOORE S. S., ROBERTS A.,  
 TANIGUCHI M., WATERMAN R. C., CHACKO J., CHANDRABOSE M. M.,  
 CREE A., DAO M. D., DINH H. H., GABISI R. A., HINES S., HUME J.,  
 JHANGIANI S. N., JOSHI V., KOVAR C. L., LEWIS L. R., LIU Y.-S., LOPEZ J.,  
 MORGAN M. B., NGUYEN N. B., OKWUONU G. O., RUIZ S. J., SANTIBANEZ  
 J., WRIGHT R. A., BUHAY C., DING Y., DUGAN-ROCHA S., HERDANDEZ J.,  
 HOLDER M., SABO A., EGAN A., GOODELL J., WILCZEK-BONEY K.,  
 FOWLER G. R., HITCHENS M. E., LOZADO R. J., MOEN C., STEFFEN D.,  
 WARREN J. T., ZHANG J., CHIU R., SCHEIN J. E., DURBIN K. J., HAVLAK P.,  
 JIANG H., LIU Y., QIN X., REN Y., SHEN Y., SONG H., BELL S. N., DAVIS  
 C., JOHNSON A. J., LEE S., NAZARETH L. V., PATEL B. M., PU L.-L.,  
 VATTATHIL S., WILLIAMS R. L., CURRY S., HAMILTON C., SODERGREN E.,  
 WHEELER D. A., BARRIS W., BENNETT G. L., EGGEN A., GREEN R. D.,  
 HARHAY G. P., HOBBS M., JANN O., KEELE J. W., KENT M. P., LIEN S.,  
 MCKAY S. D., MCWILLIAM S., RATNAKUMAR A., SCHNABEL R. D., SMITH  
 T., SNELLING W. M., SONSTEGARD T. S., STONE R. T., SUGIMOTO Y.,  
 TAKASUGA A., TAYLOR J. F., VAN TASSELL C. P., MACNEIL M. D.,  
 ABATEPAULO A. R. R., ABBEY C. A., AHOLA V., ALMEIDA I. G., AMADIO  
 A. F., ANATRIELLO E., BAHADUE S. M., BIASE F. H., BOLDT C. R.,  
 CARROLL J. A., CARVALHO W. A., CERVELATTI E. P., CHACKO E., CHAPIN  
 J. E., CHENG Y., CHOI J., COLLEY A. J., DE CAMPOS T. A., DE DONATO M.,  
 SANTOS I. K. F. de M., DE OLIVEIRA C. J. F., DEOBALD H., DEVINOY E.,  
 DONOHUE K. E., DOVC P., EBERLEIN A., FITZSIMMONS C. J., FRANZIN A.  
 M., GARCIA G. R., GENINI S., GLADNEY C. J., GRANT J. R., GREASER M.  
 L., GREEN J. A., HADSELL D. L., HAKIMOV H. A., HALGREN R., HARROW J.  
 L., HART E. A., HASTINGS N., HERNANDEZ M., HU Z.-L., INGHAM A., ISO-  
 TOURU T., JAMIS C., JENSEN K., KAPETIS D., KERR T., KHALIL S. S.,  
 KHATIB H., KOLBEHDARI D., KUMAR C. G., KUMAR D., LEACH R., LEE J.  
 C.-M., LI C., LOGAN K. M., MALINVERNI R., MARQUES E., MARTIN W. F.,

MARTINS N. F., MARUYAMA S. R., MAZZA R., MCLEAN K. L., MEDRANO J. F., MORENO B. T., MORÉ D. D., MUNTEAN C. T., NANDAKUMAR H. P., NOGUEIRA M. F. G., OLSAKER I., PANT S. D., PANZITTA F., PASTOR R. C. P., POLI M. A., POSLUSNY N., RACHAGANI S., RANGANATHAN S., RAZPET A., RIGGS P. K., RINCON G., RODRIGUEZ-OSORIO N., RODRIGUEZ-ZAS S. L., ROMERO N. E., ROSENWALD A., SANDO L., SCHMUTZ S. M., SHEN L., SHERMAN L., SOUTHEY B. R., LUTZOW Y. S., SWEEDLER J. V., TAMMEN I., TELUGU B. P. V. L., URBANSKI J. M., UTSUNOMIYA Y. T., VERSCHOOR C. P., WAARDENBERG A. J., WANG Z., WARD R., WEIKARD R., WELSH T. H., WHITE S. N., WILMING L. G., WUNDERLICH K. R., YANG J., ZHAO F.-Q., 2009 The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**: 522–528.

BOVINE HAPMAP CONSORTIUM, GIBBS R. A., TAYLOR J. F., VAN TASSELL C. P., BARENDSE W., EVERSOLE K. A., GILL C. A., GREEN R. D., HAMERNIK D. L., KAPPES S. M., LIEN S., MATUKUMALLI L. K., MCEWAN J. C., NAZARETH L. V., SCHNABEL R. D., WEINSTOCK G. M., WHEELER D. A., AJMONE-MARSAN P., BOETTCHER P. J., CAETANO A. R., GARCIA J. F., HANOTTE O., MARIANI P., SKOW L. C., SONSTEGARD T. S., WILLIAMS J. L., DIALLO B., HAILEMARIAM L., MARTINEZ M. L., MORRIS C. A., SILVA L. O. C., SPELMAN R. J., MULATU W., ZHAO K., ABBEY C. A., AGABA M., ARAUJO F. R., BUNCH R. J., BURTON J., GORNI C., OLIVIER H., HARRISON B. E., LUFF B., MACHADO M. A., MWAKAYA J., PLASTOW G., SIM W., SMITH T., THOMAS M. B., VALENTINI A., WILLIAMS P., WOMACK J., WOOLLIAMS J. A., LIU Y., QIN X., WORLEY K. C., GAO C., JIANG H., MOORE S. S., REN Y., SONG X.-Z., BUSTAMANTE C. D., HERNANDEZ R. D., MUZNY D. M., PATIL S., SAN LUCAS A., FU Q., KENT M. P., VEGA R., MATUKUMALLI A., MCWILLIAM S., SCLEP G., BRYC K., CHOI J., GAO H., GREFFENSTETTE J. J., MURDOCH B., STELLA A., VILLA-ANGULO R., WRIGHT M., AERTS J., JANN O., NEGRINI R., GODDARD M. E., HAYES B. J., BRADLEY D. G., BARBOSA DA SILVA M., LAU L. P. L., LIU G. E., LYNN D. J., PANZITTA F., DODDS K. G., 2009 Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**: 528–532.

BOWER M. A., CAMPANA M. G., CHECKLEY-SCOTT C., KNIGHT B., HOWE C. J., 2010 The potential for extraction and exploitation of DNA from parchment: a review of the opportunities and hurdles. *Journal of the Institute of Conservation* **33**: 1–11.

BRADLEY D. G., MACHUGH D. E., CUNNINGHAM P., LOFTUS R. T., 1996 Mitochondrial diversity and the origins of African and European cattle. *Proc Natl Acad Sci U S A* **93**: 5131–5135.

BRADLEY D. G., MACHUGH D. E., LOFTUS R. T., SOW R. S., HOSTE C. H., CUNNINGHAM E. P., 1994 Zebu-aurine variation in Y chromosomal DNA: a sensitive assay for genetic introgression in west African trypanotolerant cattle populations. *Anim. Genet.* **25**: 7–12.

BRIGGS A. W., GOOD J. M., GREEN R. E., KRAUSE J., MARICIC T., STENZEL U., LALUEZA-FOX C., RUDAN P., BRAJKOVIC D., KUCAN Z., GUSIC I., SCHMITZ R., DORONICHEV V. B., GOLOVANOV L. V., LA RASILLA DE M., FORTEA J., ROSAS A., PÄÄBO S., 2009 Targeted Retrieval and Analysis of Five



Neandertal mtDNA Genomes. *Science* **325**: 318–321.

BRUFORD M. W., BRADLEY D. G., LUIKART G., 2003 DNA markers reveal the complexity of livestock domestication. *Nature Publishing Group* **4**: 900–910.

BUCKLEY M., COLLINS M., THOMAS-OATES J., WILSON J. C., 2009 Species identification by analysis of bone collagen using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **23**: 3843–3854.

BURBANO H. A., HODGES E., GREEN R. E., BRIGGS A. W., KRAUSE J., MEYER M., GOOD J. M., MARICIC T., JOHNSON P. L. F., XUAN Z., ROOKS M., BHATTACHARJEE A., BRIZUELA L., ALBERT F. W., LA RASILLA DE M., FORTEA J., ROSAS A., LACHMANN M., HANNON G. J., PÄÄBO S., 2010 Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *Science* **328**: 723–725.

BURGER J., KIRCHNER M., BRAMANTI B., HAAK W., THOMAS M. G., 2007 Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proc Natl Acad Sci U S A* **104**: 3736–3741.

BURGER J., PFEIFFER I., HUMMEL S., FUCHS R., BRENIG B., HERRMANN B., 2001 Mitochondrial and nuclear DNA from (pre) historic hide-derived material. *Ancient Biomolecules* **3**: 227–238.

CAESAR J., MACDEVITT W. A., FORD J. H., 2005 *Caesar's Commentaries: On the Gallic War and on the Civil War*. El Paso Norte Press.

CAMPANA M. G., BOWER M. A., BAILEY M. J., STOCK F., CONNELL T. C. O., EDWARDS C. J., CHECKLEY-SCOTT C., KNIGHT B., SPENCER M., HOWE C. J., 2010 A flock of sheep, goats and cattle: ancient DNA analysis reveals complexities of historical parchment manufacture. *Journal of Archaeological Science* **37**: 1317–1325.

CANAVEZ F. C., LUCHE D. D., STOTHARD P., LEITE K. R. M., SOUSA-CANAVEZ J. M., PLASTOW G., MEIDANIS J., SOUZA M. A., FEIJAO P., MOORE S. S., CAMARA-LOPES L. H., 2012 Genome Sequence and Assembly of *Bos indicus*. *Journal of Heredity* **103**: 342–348.

CAUVIN J., WATKINS T., 2000 *The Birth of the Gods and the Origins of Agriculture*. Cambridge University Press.

CHEN S., LIN B. Z., BAIG M., MITRA B., LOPES R. J., SANTOS A. M., MAGEE D. A., AZEVEDO M., TARROSO P., SASAZAKI S., OSTROWSKI S., MAHGOUB O., CHAUDHURI T. K., ZHANG Y. P., COSTA V., ROYO L. J., GOYACHE F., LUIKART G., BOIVIN N., FULLER D. Q., MANNEN H., BRADLEY D. G., BEJA-PEREIRA A., 2009 Zebu Cattle Are an Exclusive Legacy of the South Asia Neolithic. *Mol Biol Evol* **27**: 1–6.

CHESSA B., PEREIRA F., ARNAUD F., AMORIM A., GOYACHE F., MAINLAND I., KAO R. R., PEMBERTON J. M., BERARDI D., STEAR M. J., ALBERTI A., PITTAU M., IANNUZZI L., BANABAZI M. H., KAZWALA R. R., ZHANG Y.-P., ARRANZ J. J., ALI B. A., WANG Z., UZUN M., DIONE M. M., OLSAKER I.,

- HOLM L.-E., SAARMA U., AHMAD S., MARZANOV N., EYTHORSDDOTTIR E., HOLLAND M. J., AJMONE-MARSAN P., BRUFORD M. W., KANTANEN J., SPENCER T. E., PALMARINI M., 2009 Revealing the history of sheep domestication using retrovirus integrations. *Science* **324**: 532–536.
- CLASON A. T., 1978 Late Bronze Age - Iron Age zebu cattle in Jordan? *Journal of Archaeological Science* **5**: 91–93.
- COLETTA A., PINNEY J. W., SOLÍS D., MARSH J., PETTIFER S. R., ATTWOOD T. K., 2010 Low-complexity regions within protein sequences have position-dependent roles. *BMC Syst Biol* **4**: 43.
- COOPER A., LALUEZA-FOX C., ANDERSON S., RAMBAUT A., AUSTIN J., WARD R., 2001 Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* **409**: 704–707.
- CURTA F., 2006 *Southeastern Europe in the Middle Ages, 500-1250*. Cambridge University Press.
- CURTA F., 2008 *The Other Europe in the Middle Ages: Avars, Bulgars, Khazars, and Cumans*.
- DOORN N. L., HOLLUND H., COLLINS M. J., 2011 A novel and non-destructive approach for ZooMS analysis: ammonium bicarbonate buffer extraction. *Archaeol Anthropol Sci* **3**: 281–289.
- DRUMMOND A. J., SUCHARD M. A., XIE D., RAMBAUT A., 2012 Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**: 1969–1973.
- EDGAR R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.
- EDWARDS C. J., BAIRD J. F., MACHUGH D. E., 2007a Taurine and zebu admixture in Near Eastern cattle: a comparison of mitochondrial, autosomal and Y-chromosomal data. *Anim. Genet.* **38**: 520–524.
- EDWARDS C. J., BOLLONGINO R., SCHEU A., CHAMBERLAIN A., TRESSET A., VIGNE J. D., BAIRD J. F., LARSON G., HO S. Y. W., HEUPINK T. H., SHAPIRO B., FREEMAN A. R., THOMAS M. G., ARBOGAST R. M., ARNDT B., BARTOSIEWICZ L., BENECKE N., BUDJA M., CHAIX L., CHOYKE A. M., COQUEUGNIOT E., DOHLE H. J., GOLDNER H., HARTZ S., HELMER D., HERZIG B., HONGO H., MASHKOUR M., OZDOGAN M., PUCHER E., ROTH G., SCHADE-LINDIG S., SCHMOLCKE U., SCHULTING R. J., STEPHAN E., UERPMANN H. P., VOROS I., VOYTEK B., BRADLEY D. G., BURGER J., 2007b Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. *Proceedings of the Royal Society B: Biological Sciences* **274**: 1377–1385.
- EDWARDS C. J., MACHUGH D. E., DOBNEY K. M., MARTIN L., RUSSELL N., HORWITZ L. K., MCINTOSH S. K., MACDONALD K. C., HELMER D., TRESSET A., VIGNE J.-D., BRADLEY D. G., 2004 Ancient DNA analysis of 101 cattle remains: limits and prospects. *Journal of Archaeological Science*



- EDWARDS C. J., MAGEE D. A., PARK S. D. E., MCGETTIGAN P. A., LOHAN A. J., MURPHY A., FINLAY E. K., SHAPIRO B., CHAMBERLAIN A. T., RICHARDS M. B., BRADLEY D. G., LOFTUS B. J., MACHUGH D. E., 2010 A Complete Mitochondrial Genome Sequence from a Mesolithic Wild Aurochs (*Bos primigenius*) (M Knapp, Ed.). PLoS ONE 5: e9255.
- EVERSHED R. P., PAYNE S., SHERRATT A. G., COPLEY M. S., COOLIDGE J., UREM-KOTSU D., KOTSAKIS K., ÖZDOĞAN M., ÖZDOĞAN A. E., NIEUWENHUYSE O., AKKERMANS P. M. M. G., BAILEY D., ANDEESCU R.-R., CAMPBELL S., FARID S., HODDER I., YALMAN N., ÖZBAŞARAN M., BİÇAKCI E., GARFINKEL Y., LEVY T., BURTON M. M., 2008 Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. Nature 455: 528–531.
- FELIUS M., 1995 *Cattle Breeds - an Encyclopedia*. Misset uitgeverij bv.
- FELIUS M., KOOLMEES P. A., THEUNISSEN B., EUROPEAN CATTLE GENETIC DIVERSITY CONSORTIUM, LENSTRA J. A., 2011 On the Breeds of Cattle—Historic and Current Classifications. Diversity 3: 660–692.
- FELSENSTEIN J., 1989 PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.
- FERNÁNDEZ H., HUGHES S., VIGNE J.-D., HELMER D., HODGINS G., MIQUEL C., HÄNNI C., LUIKART G., TABERLET P., 2006 Divergent mtDNA lineages of goats in an Early Neolithic site, far from the initial domestication areas. Proc Natl Acad Sci U S A 103: 15375–15379.
- FLICEK P., BIRNEY E., 2009 Sense from sequence reads: methods for alignment and assembly. Nat Meth 6: S6–S12.
- FLORI L., FRITZ S., JAFFRÉZIC F., BOUSSAHA M., GUT I., HEATH S., FOULLEY J.-L., GAUTIER M., 2009 The Genome Response to Artificial Selection: A Case Study in Dairy Cattle (H Harpending, Ed.). PLoS ONE 4: e6595.
- FREEMAN A. R., BRADLEY D. G., NAGDA S., GIBSON J. P., HANOTTE O., 2006a Combination of multiple microsatellite data sets to investigate genetic diversity and admixture of domestic cattle. Anim. Genet. 37: 1–9.
- FREEMAN A. R., HOGGART C. J., HANOTTE O., BRADLEY D. G., 2006b Assessing the relative ages of admixture in the bovine hybrid zones of Africa and the Near East using X chromosome haplotype mosaicism. Genetics 173: 1503–1510.
- FU Q., MEYER M., GAO X., STENZEL U., BURBANO H. A., KELSO J., PÄÄBO S., 2013 DNA analysis of an early modern human from Tianyuan Cave, China. PNAS: 1–14.
- FULLER D. Q., 2006 Agricultural Origins and Frontiers in South Asia: A Working Synthesis. J World Prehist 20: 1–86.
- GANSAUGE M.-T., MEYER M., 2013 Single-stranded DNA library preparation

for the sequencing of ancient or damaged DNA. *Nat Protoc* **8**: 737–748.

GAUTIER M., FLORI L., RIEBLER A., JAFFRÉZIC F., LALOÉ D., GUT I., MOAZAMI-GOUDARZI K., FOULLEY J.-L., 2009 A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genomics* **10**: 550.

GAUTIER M., LALOÉ D., MOAZAMI-GOUDARZI K., 2010 Insights into the Genetic History of French Cattle from Dense SNP Data on 47 Worldwide Breeds (A Brandstaetter, Ed.). *PLoS ONE* **5**: e13038.

GILBERT M. T. P., CUCCUI J., WHITE W., LYNNERUP N., TITBALL R. W., COOPER A., PRENTICE M. B., 2004 Absence of *Yersinia pestis*-specific DNA in human teeth from five European excavations of putative plague victims. *Microbiology (Reading, Engl.)* **150**: 341–354.

GILBERT M. T. P., KIVISILD T., GRONNOW B., ANDERSEN P. K., METSPALU E., REIDLA M., TAMM E., AXELSSON E., GÖTHERSTRÖM A., CAMPOS P. F., RASMUSSEN M., METSPALU M., HIGHAM T. F. G., SCHWENNINGER J. L., NATHAN R., DE HOOG C. J., KOCH A., MOLLER L. N., ANDREASEN C., MELDGAARD M., VILLEMS R., BENDIXEN C., WILLERSLEV E., 2008 Paleo-Eskimo mtDNA Genome Reveals Matrilineal Discontinuity in Greenland. *Science* **320**: 1787–1789.

GILBERT M. T. P., TOMSHO L. P., RENDULIC S., PACKARD M., DRAUTZ D. I., SHER A., TIKHONOV A., DALÉN L., KUZNETSOVA T., KOSINTSEV P., CAMPOS P. F., HIGHAM T., COLLINS M. J., WILSON A. S., SHIDLOVSKIY F., BUIGUES B., ERICSON P. G. P., GERMONPRÉ M., GÖTHERSTRÖM A., IACUMIN P., NIKOLAEV V., NOWAK-KEMP M., WILLERSLEV E., KNIGHT J. R., IRZYK G. P., PERBOST C. S., FREDRIKSON K. M., HARKINS T. T., SHERIDAN S., MILLER W., SCHUSTER S. C., 2007 Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science* **317**: 1927–1930.

GILBERT M. T. P., WILLERSLEV E., HANSEN A. J., BARNES I., RUDBECK L., LYNNERUP N., COOPER A., 2003 Distribution patterns of postmortem damage in human mitochondrial DNA. *Am. J. Hum. Genet.* **72**: 32–47.

GILLIS R., CHAIX L., VIGNE J.-D., 2011 An assessment of morphological criteria for discriminating sheep and goat mandibles on a large prehistoric archaeological assemblage (Kerma, Sudan). *Journal of Archaeological Science*: 1–16.

GINOLHAC A., RASMUSSEN M., GILBERT M. T. P., WILLERSLEV E., ORLANDO L., 2011 mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27**: 2153–2155.

GODDARD M. E., HAYES B. J., 2009 Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Publishing Group* **10**: 381–391.

GOURICHON L., HELMER D., 2008 *Life and Death in a Prehistoric Settlement in Northwest Anatolia. The Ilipinar Excavations, Volume III* (J Roodenberg and S Alpaslan-Roodenburg, Eds.). Nederlands Instituut voor het Nabije



Oosten.

- GOUY M., GUINDON S., GASCUEL O., 2010 SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Mol Biol Evol* **27**: 221–224.
- GÖTHERSTRÖM A., ANDERUNG C., HELLBORG L., ELBURG R., SMITH C., BRADLEY D. G., ELLEGREN H., 2005 Cattle domestication in the Near East was followed by hybridization with aurochs bulls in Europe. *Proc. Biol. Sci.* **272**: 2345–2350.
- GREEN R. E., KRAUSE J., BRIGGS A. W., MARICIC T., STENZEL U., KIRCHER M., PATTERSON N., LI H., ZHAI W., FRITZ M. H. Y., HANSEN N. F., DURAND E. Y., MALASPINAS A. S., JENSEN J. D., MARQUES-BONET T., ALKAN C., PRUFER K., MEYER M., BURBANO H. A., GOOD J. M., SCHULTZ R., AXIMU-PETRI A., BUTTHOF A., HOBER B., HOFFNER B., SIEGEMUND M., WEIHMANN A., NUSBAUM C., LANDER E. S., RUSS C., NOVOD N., AFFOURTIT J., EGHOLM M., VERNA C., RUDAN P., BRAJKOVIC D., KUCAN Z., GUSIC I., DORONICHEV V. B., GOLOVANOV L. V., LALUEZA-FOX C., LA RASILLA DE M., FORTEA J., ROSAS A., SCHMITZ R. W., JOHNSON P. L. F., EICHLER E. E., FALUSH D., BIRNEY E., MULLIKIN J. C., SLATKIN M., NIELSEN R., KELSO J., LACHMANN M., REICH D., PÄÄBO S., 2010 A Draft Sequence of the Neandertal Genome. *Science* **328**: 710–722.
- HADDRATH O., BAKER A. J., 2001 Complete mitochondrial DNA genome sequences of extinct birds: ratite phylogenetics and the vicariance biogeography hypothesis. *Proceedings of the Royal Society B: Biological Sciences* **268**: 939–945.
- HELMER D., GOURICHON L., H M., PETERS J., M S. A. S., 2005 Identifying early domestic cattle from Pre-Pottery Neolithic sites on the Middle Euphrates using sexual dimorphism. In: Vigne JD, Helmer D, Peters J (Eds.), *First steps of animal domestication. New archaeozoological approaches*, Oxbow Books, London, pp. 86–95.
- HIENDLER S., MAINZ K., PLANTE Y., LEWALSKI H., 1998 Analysis of mitochondrial DNA indicates that domestic sheep are derived from two different ancestral maternal sources: no evidence for contributions from urial and argali sheep. *J. Hered.* **89**: 113–120.
- HIGUCHI R., BOWMAN B., FREIBERGER M., RYDER O. A., WILSON A. C., 1984 DNA sequences from the quagga, an extinct member of the horse family. *Nature* **312**: 282–284.
- HO S. Y. W., GILBERT M. T. P., 2010 Ancient mitogenomics. *Mitochondrion* **10**: 1–11.
- HO S. Y. W., LARSON G., 2006 Molecular clocks: when times are a-changin'. *Trends Genet.* **22**: 79–83.
- HO S. Y. W., LARSON G., EDWARDS C. J., HEUPINK T. H., LAKIN K. E., HOLLAND P. W. H., SHAPIRO B., 2008 Correlating Bayesian date estimates with climatic events and domestication using a bovine case study. *Biology Letters* **4**: 370–374.

- HOFREITER M., SERRE D., POINAR H. N., KUCH M., PÄÄBO S., 2001 Ancient DNA. *Nat. Rev. Genet.* **2**: 353–359.
- HUANG H. Y., CHIEN C. H., JEN K. H., HUANG H. D., 2006 RegRNA: an integrated web server for identifying regulatory RNA motifs and elements. *Nucleic Acids Research* **34**: W429–W434.
- HUSON D. H., MITRA S., RUSCHEWEYH H. J., WEBER N., SCHUSTER S. C., 2011 Integrative analysis of environmental sequences using MEGAN4. *Genome Research* **21**: 1552–1560.
- ITAN Y., POWELL A., BEAUMONT M. A., BURGER J., THOMAS M. G., 2009 The Origins of Lactase Persistence in Europe (MM Tanaka, Ed.). *PLoS Comp Biol* **5**: e1000491.
- KAHVEJIAN A., QUACKENBUSH J., THOMPSON J. F., 2008 What would you do if you could sequence everything? *Nat. Biotechnol.* **26**: 1125–1133.
- KAMIŃSKI S., CIESLIŃSKA A., KOSTYRA E., 2007 Polymorphism of bovine beta-casein and its potential effect on human health. *J Appl Genetics* **48**: 189–198.
- KELLER A., GRAEFEN A., BALL M., MATZAS M., BOISGUERIN V., MAIXNER F., LEIDINGER P., BACKES C., KHAIRAT R., FORSTER M., STADE B. O. R., FRANKE A., MAYER J., SPANGLER J., MCLAUGHLIN S., SHAH M., LEE C., HARKINS T. T., SARTORI A., MORENO-ESTRADA A., HENN B., SIKORA M., SEMINO O., CHIARONI J., ROOTSI S., MYRES N. M., CABRERA V. M., UNDERHILL P. A., BUSTAMANTE C. D., VIGL E. E., SAMADELLI M., CIPOLLINI G., HAAS J., KATUS H., O'CONNOR B. D., CARLSON M. R. J., MEDER B., BLIN N., MEESE E., PUSCH C. M., ZINK A., 2012 New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Communications* **3**: 698–9.
- KIJAS J. W., LENSTRA J. A., HAYES B., BOITARD S., PORTO NETO L. R., SAN CRISTOBAL M., SERVIN B., MCCULLOCH R., WHAN V., GIETZEN K., PAIVA S., BARENDSE W., CIANI E., RAADSMA H., MCEWAN J., DALRYMPLE B., INTERNATIONAL SHEEP GENOMICS CONSORTIUM MEMBERS, 2012 Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *Plos Biol* **10**: e1001258.
- KIRSANOW K., BURGER J., 2012 Ancient human DNA. *Annals of Anatomy* **194**: 121–132.
- KNAPP M., HOFREITER M., 2010 Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives. *Genes* **1**: 227–243.
- KNAPP M., HORSBURGH K. A., PROST S., STANTON J.-A., BUCKLEY H. R., WALTER R. K., MATISOO-SMITH E. A., 2012 Complete mitochondrial DNA genome sequences from the first New Zealanders. *PNAS* **109**: 18350–18354.
- KRAUSE J., DEAR P. H., POLLACK J. L., SLATKIN M., SPRIGGS H., BARNES I., LISTER A. M., EBERSBERGER I., PÄÄBO S., HOFREITER M., 2005 Multiplex amplification of the mammoth mitochondrial genome and the



evolution of Elephantidae. *Nature* **439**: 724–727.

KUERSTEN S., GOODWIN E. B., 2003 The power of the 3' UTR: translational control and development. *Nat. Rev. Genet.* **4**: 626–637.

LALUEZA-FOX C., GILBERT M. T. P., 2011 Paleogenomics of archaic hominins. *Curr. Biol.* **21**: R1002–9.

LANDER E. S., LINTON L. M., BIRREN B., NUSBAUM C., ZODY M. C., BALDWIN J., DEVON K., DEWAR K., DOYLE M., FITZHUGH W., FUNKE R., GAGE D., HARRIS K., HEAFORD A., HOWLAND J., KANN L., LEHOCZKY J., LEVINE R., MCEWAN P., MCKERNAN K., MELDRIM J., MESIROV J. P., MIRANDA C., MORRIS W., NAYLOR J., RAYMOND C., ROSETTI M., SANTOS R., SHERIDAN A., SOUGNEZ C., STANGE-THOMANN N., STOJANOVIC N., SUBRAMANIAN A., WYMAN D., ROGERS J., SULSTON J., AINSCOUGH R., BECK S., BENTLEY D., BURTON J., CLEE C., CARTER N., COULSON A., DEADMAN R., DELOUKAS P., DUNHAM A., DUNHAM I., DURBIN R., FRENCH L., GRAHAM D., GREGORY S., HUBBARD T., HUMPHRAY S., HUNT A., JONES M., LLOYD C., MCMURRAY A., MATTHEWS L., MERCER S., MILNE S., MULLIKIN J. C., MUNGALL A., PLUMB R., ROSS M., SHOWNKEEN R., SIMS S., WATERSTON R. H., WILSON R. K., HILLIER L. W., MCPHERSON J. D., MARRA M. A., MARDIS E. R., FULTON L. A., CHINWALLA A. T., PEPIN K. H., GISH W. R., CHISSOE S. L., WENDL M. C., DELEHAUNTY K. D., MINER T. L., DELEHAUNTY A., KRAMER J. B., COOK L. L., FULTON R. S., JOHNSON D. L., MINX P. J., CLIFTON S. W., HAWKINS T., BRANSCOMB E., PREDKI P., RICHARDSON P., WENNING S., SLEZAK T., DOGGETT N., CHENG J. F., OLSEN A., LUCAS S., ELKIN C., UBERBACHER E., FRAZIER M., GIBBS R. A., MUZNY D. M., SCHERER S. E., BOUCK J. B., SODERGREN E. J., WORLEY K. C., RIVES C. M., GORRELL J. H., METZKER M. L., NAYLOR S. L., KUCHERLAPATI R. S., NELSON D. L., WEINSTOCK G. M., SAKAKI Y., FUJIYAMA A., HATTORI M., YADA T., TOYODA A., ITOH T., KAWAGOE C., WATANABE H., TOTOKI Y., TAYLOR T., WEISSENBAACH J., HEILIG R., SAURIN W., ARTIGUENAVE F., BROTTIER P., BRULS T., PELLETIER E., ROBERT C., WINCKER P., SMITH D. R., DOUCETTE-STAMM L., RUBENFIELD M., WEINSTOCK K., LEE H. M., DUBOIS J., ROSENTHAL A., PLATZER M., NYAKATURA G., TAUDIEN S., RUMP A., YANG H., YU J., WANG J., HUANG G., GU J., HOOD L., ROWEN L., MADAN A., QIN S., DAVIS R. W., FEDERSPIEL N. A., ABOLA A. P., PROCTOR M. J., MYERS R. M., SCHMUTZ J., DICKSON M., GRIMWOOD J., COX D. R., OLSON M. V., KAUL R., RAYMOND C., SHIMIZU N., KAWASAKI K., MINOSHIMA S., EVANS G. A., ATHANASIOU M., SCHULTZ R., ROE B. A., CHEN F., PAN H., RAMSER J., LEHRACH H., REINHARDT R., MCCOMBIE W. R., LA BASTIDE DE M., DEDHIA N., BLÖCKER H., HORNISCHER K., NORDSIEK G., AGARWALA R., ARAVIND L., BAILEY J. A., BATEMAN A., BATZOGLOU S., BIRNEY E., BORK P., BROWN D. G., BURGE C. B., CERUTTI L., CHEN H. C., CHURCH D., CLAMP M., COPLEY R. R., DOERKS T., EDDY S. R., EICHLER E. E., FUREY T. S., GALAGAN J., GILBERT J. G., HARMON C., HAYASHIZAKI Y., HAUSSLER D., HERMJAKOB H., HOKAMP K., JANG W., JOHNSON L. S., JONES T. A., KASIF S., KASPRYZK A., KENNEDY S., KENT W. J., KITTS P., KOONIN E. V., KORF I., KULP D., LANCET D., LOWE T. M., MCLYSAGHT A., MIKKELSEN T., MORAN J. V., MULDER N., POLLARA V. J., PONTING C. P., SCHULER G., SCHULTZ J., SLATER G., SMIT A. F., STUPKA E., SZUSTAKOWSKI J., THIERRY-MIEG D., THIERRY-MIEG J., WAGNER L.,

- WALLIS J., WHEELER R., WILLIAMS A., WOLF Y. I., WOLFE K. H., YANG S. P., YEH R. F., COLLINS F., GUYER M. S., PETERSON J., FELSENFELD A., WETTERSTRAND K. A., PATRINOS A., MORGAN M. J., DE JONG P., CATANESE J. J., OSOEGAWA K., SHIZUYA H., CHOI S., CHEN Y. J., SZUSTAKOWKI J., INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM, 2001 Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- LANDOURÉ G., SULLIVAN J. M., JOHNSON J. O., MUNNS C. H., SHI Y., DIALLO O., GIBBS J. R., GAUDET R., LUDLOW C. L., FISCHBECK K. H., TRAYNOR B. J., BURNETT B. G., SUMNER C. J., 2012 Exome sequencing identifies a novel TRPV4 mutation in a CMT2C family. *Neurology*.
- LANGMEAD B., TRAPNELL C., POP M., SALZBERG S. L., 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- LARI M., RIZZI E., MONA S., CORTI G., CATALANO G., CHEN K., VERNESI C., LARSON G., BOSCATO P., DE BELLIS G., COOPER A., CARAMELLI D., BERTORELLE G., 2011 The Complete Mitochondrial Genome of an 11,450-year-old Aurochs (Bos primigenius) from Central Italy. *BMC Evolutionary Biology* **11**: 32.
- LARKIN D. M., DAETWYLER H. D., HERNANDEZ A. G., WRIGHT C. L., HETRICK L. A., BOUCEK L., BACHMAN S. L., BAND M. R., AKRAIKO T. V., COHEN-ZINDER M., THIMMAPURAM J., MACLEOD I. M., HARKINS T. T., MCCAGUE J. E., GODDARD M. E., HAYES B. J., LEWIN H. A., 2012 Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *Proc Natl Acad Sci U S A*.
- LARSON G., BURGER J., 2013 A population genetics view of animal domestication. *Trends Genet.*: 1–9.
- LARSON G., ALBARELLA U., DOBNEY K., ROWLEY-CONWY P., SCHIBLER J., TRESSET A., VIGNE J.-D., EDWARDS C. J., SCHLUMBAUM A., DINU A., BALAŞESCU A., DOLMAN G., TAGLIACCOZZO A., MANASERYAN N., MIRACLE P., VAN WIJNGAARDEN-BAKKER L., MASSETI M., BRADLEY D. G., COOPER A., 2007 Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proc Natl Acad Sci U S A* **104**: 15276–15281.
- LARSON R. (Ed.), 2002 *Microanalysis of Parchment*. Archetype Publications, London.
- LETUNIC I., DOERKS T., BORK P., 2011 SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Research* **40**: D302–D305.
- LI H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987–2993.
- LI H., DURBIN R., 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.



- LI H., DURBIN R., 2011 Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493–496.
- LI H., HANDSAKER B., WYSOKER A., FENNELL T., RUAN J., HOMER N., MARTH G., ABECASIS G., DURBIN R., 1000 GENOME PROJECT DATA PROCESSING SUBGROUP, 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
- LINDBLAD-TOH K., WADE C. M., MIKKELSEN T. S., KARLSSON E. K., JAFFE D. B., KAMAL M., CLAMP M., CHANG J. L., KULBOKAS E. J., ZODY M. C., MAUCELI E., XIE X., BREEN M., WAYNE R. K., OSTRANDER E. A., PONTING C. P., GALIBERT F., SMITH D. R., DEJONG P. J., KIRKNESS E., ALVAREZ P., BIAGI T., BROCKMAN W., BUTLER J., CHIN C.-W., COOK A., CUFF J., DALY M. J., DECAPRIO D., GNERRE S., GRABHERR M., KELLIS M., KLEBER M., BARDELEBEN C., GOODSTADT L., HEGER A., HITTE C., KIM L., KOEPFLI K.-P., PARKER H. G., POLLINGER J. P., SEARLE S. M. J., SUTTER N. B., THOMAS R., WEBBER C., BALDWIN J., ABEBE A., ABOUELLEIL A., AFTUCK L., AIT-ZAHRA M., ALDREDGE T., ALLEN N., AN P., ANDERSON S., ANTOINE C., ARACHCHI H., ASLAM A., AYOTTE L., BACHANTSANG P., BARRY A., BAYUL T., BENAMARA M., BERLIN A., BESSETTE D., BLITSHTEYN B., BLOOM T., BLYE J., BOGUSLAVSKIY L., BONNET C., BOUKHGALTER B., BROWN A., CAHILL P., CALIXTE N., CAMARATA J., CHESHATSANG Y., CHU J., CITROEN M., COLLYMORE A., COOKE P., DAWOE T., DAZA R., DECKTOR K., DEGRAY S., DHARGAY N., DOOLEY K., DOOLEY K., DORJE P., DORJEE K., DORRIS L., DUFFEY N., DUPES A., EGBIREMOLEN O., ELONG R., FALK J., FARINA A., FARO S., FERGUSON D., FERREIRA P., FISHER S., FITZGERALD M., FOLEY K., FOLEY C., FRANKE A., FRIEDRICH D., GAGE D., GARBER M., GEARIN G., GIANNOUKOS G., GOODE T., GOYETTE A., GRAHAM J., GRANDBOIS E., GYALTSSEN K., HAFEZ N., HAGOPIAN D., HAGOS B., HALL J., HEALY C., HEGARTY R., HONAN T., HORN A., HOUDE N., HUGHES L., HUNNICUTT L., HUSBY M., JESTER B., JONES C., KAMAT A., KANGA B., KELLS C., KHAZANOVICH D., KIEU A. C., KISNER P., KUMAR M., LANCE K., LANDERS T., LARA M., LEE W., LEGER J.-P., LENNON N., LEUPER L., LEVINE S., LIU J., LIU X., LOKYITSANG Y., LOKYITSANG T., LUI A., MACDONALD J., MAJOR J., MARABELLA R., MARU K., MATTHEWS C., McDONOUGH S., MEHTA T., MELDRIM J., MELNIKOV A., MENEUS L., MIHALEV A., MIHOVA T., MILLER K., MITTELMAN R., MLENGA V., MULRAIN L., MUNSON G., NAVIDI A., NAYLOR J., NGUYEN T., NGUYEN N., NGUYEN C., NGUYEN T., NICOL R., NORBU N., NORBU C., NOVOD N., NYIMA T., OLANDT P., O'NEILL B., O'NEILL K., OSMAN S., OYONO L., PATTI C., PERRIN D., PHUNKHANG P., PIERRE F., PRIEST M., RACHUPKA A., RAGHURAMAN S., RAMEAU R., RAY V., RAYMOND C., REGE F., RISE C., ROGERS J., ROGOV P., SAHALIE J., SETTIPALLI S., SHARPE T., SHEA T., SHEEHAN M., SHERPA N., SHI J., SHIH D., SLOAN J., SMITH C., SPARROW T., STALKER J., STANGE-THOMANN N., STAVROPOULOS S., STONE C., STONE S., SYKES S., TCHUINGA P., TENZING P., TESFAYE S., THOULUTSANG D., THOULUTSANG Y., TOPHAM K., TOPPING I., TSAMLA T., VASSILIEV H., VENKATARAMAN V., VO A., WANGCHUK T., WANGDI T., WEIAND M., WILKINSON J., WILSON A., YADAV S., YANG S., YANG X., YOUNG G., YU Q., ZAINOUN J., ZEMBEK L., ZIMMER A., LANDER E. S., 2005 Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**: 803–819.

- LOFTUS R. T., ERTUGRUL O., HARBA A. H., EL-BARODY M. A., MACHUGH D. E., PARK S. D., BRADLEY D. G., 1999 A microsatellite survey of cattle from a centre of origin: the Near East. *Molecular Ecology* **8**: 2015–2022.
- LOFTUS R. T., MACHUGH D. E., BRADLEY D. G., SHARP P. M., CUNNINGHAM P., 1994 Evidence for two independent domestications of cattle. *Proc Natl Acad Sci U S A* **91**: 2757–2761.
- LUDWIG A., PRUVOST M., REISSMANN M., BENECKE N., BROCKMANN G. A., CASTANOS P., CIESLAK M., LIPPOLD S., LLORENTE L., MALASPINAS A. S., SLATKIN M., HOFREITER M., 2009 Coat Color Variation at the Beginning of Horse Domestication. *Science* **324**: 485–485.
- LUSK J. L., 2007 Association of single nucleotide polymorphisms in the leptin gene with body weight and backfat growth curve parameters for beef cattle. *J. Anim. Sci.* **85**: 1865–1872.
- MACHUGH D. E., SHRIVER M. D., LOFTUS R. T., CUNNINGHAM P., BRADLEY D. G., 1997 Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics* **146**: 1071–1086.
- MALMSTRÖM H., 2005 Extensive Human DNA Contamination in Extracts from Ancient Dog Bones and Teeth. *Mol Biol Evol* **22**: 2040–2047.
- MARICIC T., WHITTEN M., PÄÄBO S., 2010 Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products (RC Fleischer, Ed.). *PLoS ONE* **5**: e14004.
- MARTIN M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**: pp. 10–12.
- MATUKUMALLI L. K., LAWLEY C. T., SCHNABEL R. D., TAYLOR J. F., ALLAN M. F., HEATON M. P., O'CONNELL J., MOORE S. S., SMITH T. P. L., SONSTEGARD T. S., VAN TASSELL C. P., 2009 Development and Characterization of a High Density SNP Genotyping Assay for Cattle (AE Toland, Ed.). *PLoS ONE* **4**: e5350.
- MCKENNA A., HANNA M., BANKS E., SIVACHENKO A., CIBULSKIS K., KERNYTSKY A., GARIMELLA K., ALTSHULER D., GABRIEL S., DALY M., DEPRISTO M. A., 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.
- MEADOW R. H., 1989 Prehistoric wild sheep and sheep domestication on the Eastern margin of the Middle East (C Pam, C Douglas, and K Ryan, Eds.). *MASCA Research Papers in Science and Archaeology* **6**: 24–36.
- MEADOW R. H., 1991 Harappa excavations 1986--1990: A multi-disciplinary approach to 3rd millennium urbanism.
- MEADOW R. H., 1996 The origins and spread of agriculture and pastoralism in northwestern South Asia. In: Haris DR (Ed.), *The Origins and Spread of Agriculture and Pastoralism in Eurasia*, Smithsonian Inst Pr, p. 594.



- MEADOWS J. R. S., HIENDLEDER S., KIJAS J. W., 2010 Haplogroup relationships between domestic and wild sheep resolved using a mitogenome panel. *Heredity* **106**: 700–706.
- METZKER M. L., 2010 Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**: 31–46.
- MEYER M., KIRCHER M., GANSAUGE M.-T., LI H., RACIMO F., MALICK S., SCHRAIBER J. G., JAY F., PRÜFER K., DE FILIPPO C., SUDMANT P. H., ALKAN C., FU Q., DO R., ROHLAND N., TANDON A., SIEBAUER M., GREEN R. E., BRYC K., BRIGGS A. W., STENZEL U., DABNEY J., SHENDURE J., KITZMAN J., HAMMER M. F., SHUNKOV M. V., DEREVIANKO A. P., PATTERSON N., ANDRÉS A. M., EICHLER E. E., SLATKIN M., REICH D., KELSO J., PÄÄBO S., 2012 A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*.
- MILLER F. J., ROSENFELDT F. L., ZHANG C., LINNANE A. W., NAGLEY P., 2003 Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age. *Nucleic Acids Research* **31**: e61.
- MILLER W., DRAUTZ D. I., RATAN A., PUSEY B., QI J., LESK A. M., TOMSHO L. P., PACKARD M. D., ZHAO F., SHER A., TIKHONOV A., RANEY B., PATTERSON N., LINDBLAD-TOH K., LANDER E. S., KNIGHT J. R., IRZYK G. P., FREDRIKSON K. M., HARKINS T. T., SHERIDAN S., PRINGLE T., SCHUSTER S. C., 2008 Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456**: 387–390.
- MILLER W., SCHUSTER S. C., WELCH A. J., RATAN A., BEDOYA-REINA O. C., ZHAO F., KIM H. L., BURHANS R. C., DRAUTZ D. I., WITTEKINDT N. E., TOMSHO L. P., IBARRA-LACLETTE E., HERRERA-ESTRELLA L., PEACOCK E., FARLEY S., SAGE G. K., RODE K., OBBARD M., MONTIEL R., BACHMANN L., INGÓLFSSON O., AARS J., MAILUND T., WIIG O., TALBOT S. L., LINDQVIST C., 2012 Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *PNAS* **109**: E2382–90.
- MURRAY C., HUERTA-SANCHEZ E., CASEY F., BRADLEY D. G., 2010 Cattle demographic history modelled from autosomal sequence variation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**: 2531–2539.
- NELSON M. R., BRYC K., KING K. S., INDAP A., BOYKO A. R., NOVEMBRE J., BRILEY L. P., MARUYAMA Y., WATERWORTH D. M., WAEBER G., VOLLENWEIDER P., OKSENBERG J. R., HAUSER S. L., STIRNADEL H. A., KOONER J. S., CHAMBERS J. C., JONES B., MOOSER V., BUSTAMANTE C. D., ROSES A. D., BURNS D. K., EHM M. G., LAI E. H., 2008 The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. *The American Journal of Human Genetics* **83**: 347–358.
- NG S. B., TURNER E. H., ROBERTSON P. D., FLYGARE S. D., BIGHAM A. W., LEE C., SHAFFER T., WONG M., BHATTACHARJEE A., EICHLER E. E.,

- BAMSHAD M., NICKERSON D. A., SHENDURE J., 2009 Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**: 272–276.
- OLIVIERI C., ERMINI L., RIZZI E., CORTI G., LUCIANI S., MAROTA I., DE BELLIS G., ROLLO F., 2012 Phylogenetic Position of a Copper Age Sheep (*Ovis aries*) Mitochondrial DNA (C Lalueza-Fox, Ed.). *PLoS ONE* **7**: e33792.
- ORLANDO L., GINOLHAC A., RAGHAVAN M., VILSTRUP J., RASMUSSEN M., MAGNUSSEN K., STEINMANN K. E., KAPRANOV P., THOMPSON J. F., ZAZULA G., FROESE D., MOLTKE I., SHAPIRO B., HOFREITER M., AL-RASHEID K. A. S., GILBERT M. T. P., WILLERSLEV E., 2011 True single-molecule DNA sequencing of a pleistocene horse bone. *Genome Research* **21**: 1705–1719.
- OSKAM C. L., HAILE J., MCLAY E., RIGBY P., ALLENTOLT M. E., OLSEN M. E., BENGTSSON C., MILLER G. H., SCHWENNINGER J. L., JACOMB C., WALTER R., BAYNES A., DORTCH J., PARKER-PEARSON M., GILBERT M. T. P., HOLDAWAY R. N., WILLERSLEV E., BUNCE M., 2010 Fossil avian eggshell preserves ancient DNA. *Proceedings of the Royal Society B: Biological Sciences* **277**: 1991–2000.
- OTTONI C., GIRDLAND FLINK L., EVIN A., GEORG C., DE CUPERE B., VAN NEER W., BARTOSIEWICZ L., LINDERHOLM A., BARNETT R., PETERS J., DECORTE R., WAELEKENS M., VANDERHEYDEN N., RICAUT F. X., CAKIRLAR C., CEVIK O., HOELZEL A. R., MASHKOUR M., MOHASEB KARIMLU A. F., SHEIKHI SENO S., DAUJAT J., BROCK F., PINHASI R., HONGO H., PEREZ-ENCISO M., RASMUSSEN M., FRANTZ L., MEGENS H. J., CROOIJMANS R., GROENEN M., ARBUCKLE B., BENECKE N., STRAND VIDARSDOTTIR U., BURGER J., CUCCHI T., DOBNEY K., LARSON G., 2012 Pig Domestication and Human-Mediated Dispersal in Western Eurasia Revealed through Ancient DNA and Geometric Morphometrics. *Mol Biol Evol.*
- PAIJMANS J. L. A., GILBERT M. T. P., HOFREITER M., 2012 Mitogenomic analyses from ancient DNA. *MOLECULAR PHYLOGENETICS AND EVOLUTION*: 1–12.
- PANGALLO D., CHOVANOVA K., MAKOVA A., 2010 Identification of animal skin of historical parchments by polymerase chain reaction (PCR)-based methods. *Journal of Archaeological Science* **37**: 1202–1206.
- PATTERSON N., PRICE A. L., REICH D., 2006 Population Structure and Eigenanalysis. *PLoS Genet* **2**: e190.
- PAUSCH H., WANG X., JUNG S., KROGMEIER D., EDEL C., EMMERLING R., GÖTZ K.-U., FRIES R., 2012 Identification of QTL for UV-Protective Eye Area Pigmentation in Cattle by Progeny Phenotyping and Genome-Wide Association Analysis (S Moore, Ed.). *PLoS ONE* **7**: e36346.
- PÄÄBO S., 1985 Molecular cloning of Ancient Egyptian mummy DNA. *Nature* **314**: 644–645.
- PÄÄBO S., POINAR H., SERRE D., JAENICKE-DESPRES V., HEBLER J., ROHLAND N., KUCH M., KRAUSE J., VIGILANT L., HOFREITER M., 2004 Genetic



- analyses from ancient DNA. *Annu. Rev. Genet.* **38**: 645–679.
- PINHASI R., THOMAS M. G., HOFREITER M., CURRAT M., BURGER J., 2012  
The genetic history of Europeans. *Trends Genet.* **28**: 496–505.
- POINAR H. N., SCHWARZ C., QI J., SHAPIRO B., MACPHEE R. D. E., BUIGUES B.,  
TIKHONOV A., HUSON D. H., TOMSHO L. P., AUCH A., RAMPP M., MILLER  
W., SCHUSTER S. C., 2006 Metagenomics to paleogenomics: large-scale  
sequencing of mammoth DNA. *Science* **311**: 392–394.
- POULAKAKIS N., TSELIKAS A., BITSAKIS I., MYLONAS M., LYMBERAKIS P.,  
2007 Ancient DNA and the genetic signature of ancient Greek  
manuscripts. *Journal of Archaeological Science* **34**: 675–680.
- PRITCHARD J. K., STEPHENS M., DONNELLY P., 2000 Inference of population  
structure using multilocus genotype data. *Genetics* **155**: 945–959.
- PRUVOST M., SCHWARZ R., CORREIA V. B., CHAMPLIT S., BRAGUIER S.,  
MOREL N., FERNANDEZ-JALVO Y., GRANGE T., GEIGL E.-M., 2007  
Freshly excavated fossil bones are best for amplification of ancient DNA.  
*Proc Natl Acad Sci U S A* **104**: 739–744.
- PURCELL S., NEALE B., TODD-BROWN K., THOMAS L., FERREIRA M. A. R.,  
BENDER D., MALLER J., SKLAR P., DE BAKKER P. I. W., DALY M. J., SHAM  
P. C., 2007 PLINK: A Tool Set for Whole-Genome Association and  
Population-Based Linkage Analyses. *The American Journal of Human  
Genetics* **81**: 559–575.
- QIU Q., ZHANG G., MA T., QIAN W., WANG J., YE Z., CAO C., HU Q., KIM J.,  
LARKIN D. M., AUVIL L., CAPITANU B., MA J., LEWIN H. A., QIAN X.,  
LANG Y., ZHOU R., WANG L., WANG K., XIA J., LIAO S., PAN S., LU X.,  
HOU H., WANG Y., ZANG X., YIN Y., MA H., ZHANG J., WANG Z., ZHANG  
Y., ZHANG D., YONEZAWA T., HASEGAWA M., ZHONG Y., LIU W., ZHANG  
Y., HUANG Z., ZHANG S., LONG R., YANG H., WANG J., LENSTRA J. A.,  
COOPER D. N., WU Y., WANG J., SHI P., WANG J., LIU J., 2012 The yak  
genome and adaptation to life at high altitude. *Nature Publishing Group*  
**44**: 946–949.
- RAOULT D., ABOUDHARAM G., CRUBÉZY E., LARROUY G., LUDÉS B.,  
DRANCOURT M., 2000 Molecular identification by “suicide PCR” of  
*Yersinia pestis* as the agent of medieval black death. *Proc Natl Acad Sci U  
S A* **97**: 12800–12803.
- RASMUSSEN M., GUO X., WANG Y., LOHMEUELLER K. E., RASMUSSEN S.,  
ALBRECHTSEN A., SKOTTE L., LINDGREEN S., METSPALU M., JOMBART T.,  
KIVISILD T., ZHAI W., ERIKSSON A., MANICA A., ORLANDO L., LA VEGA  
DE F. M., TRIDICO S., METSPALU E., NIELSEN K., AVILA-ARCOS M. C.,  
MORENO-MAYAR J. V., MULLER C., DORTCH J., GILBERT M. T. P., LUND  
O., WESOŁOWSKA A., KARMIN M., WEINERT L. A., WANG B., LI J., TAI S.,  
XIAO F., HANIHARA T., VAN DRIEM G., JHA A. R., RICAUT F. X., DE KNIJFF  
P., MIGLIANO A. B., GALLEGU ROMERO I., KRISTIANSEN K., LAMBERT D.  
M., BRUNAK S., FORSTER P., BRINKMANN B., NEHLICH O., BUNCE M.,  
RICHARDS M., GUPTA R., BUSTAMANTE C. D., KROGH A., FOLEY R. A.,

- LAHR M. M., BALLOUX F., SICHERITZ-PONTEN T., VILLEMS R., NIELSEN R., WANG J., WILLERSLEV E., 2011 An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. *Science* **334**: 94–98.
- RASMUSSEN M., LI Y., LINDGREEN S., PEDERSEN J. S., ALBRECHTSEN A., MOLTKE I., METSPALU M., METSPALU E., KIVISILD T., GUPTA R., BERTALAN M., NIELSEN K., GILBERT M. T. P., WANG Y., RAGHAVAN M., CAMPOS P. F., KAMP H. M., WILSON A. S., GLEDHILL A., TRIDICO S., BUNCE M., LORENZEN E. D., BINLADEN J., GUO X., ZHAO J., ZHANG X., ZHANG H., LI Z., CHEN M., ORLANDO L., KRISTIANSEN K., BAK M., TOMMERUP N., BENDIXEN C., PIERRE T. L., GRØNNOW B., MELDGAARD M., ANDREASEN C., FEDOROVA S. A., OSIPOVA L. P., HIGHAM T. F. G., RAMSEY C. B., O HANSEN VON T., NIELSEN F. C., CRAWFORD M. H., BRUNAK S., SICHERITZ-PONTÉN T., VILLEMS R., NIELSEN R., KROGH A., WANG J., WILLERSLEV E., 2010 Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**: 757–762.
- REICH D., GREEN R. E., KIRCHER M., KRAUSE J., PATTERSON N., DURAND E. Y., VIOLA B., BRIGGS A. W., STENZEL U., JOHNSON P. L. F., MARICIC T., GOOD J. M., MARQUES-BONET T., ALKAN C., FU Q., MALLICK S., LI H., MEYER M., EICHLER E. E., STONEKING M., RICHARDS M., TALAMO S., SHUNKOV M. V., DEREVIANKO A. P., HUBLIN J.-J., KELSO J., SLATKIN M., PÄÄBO S., 2010 Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**: 1053–1060.
- RICHTER K. K., WILSON J., JONES A. K. G., BUCKLEY M., VAN DOORN N., COLLINS M. J., 2011 Fish 'n chips: ZooMS peptide mass fingerprinting in a 96 well plate format to identify fish bone fragments. *Journal of Archaeological Science* **38**: 1502–1510.
- ROGAEV E. I., MOLIKA Y. K., MALYARCHUK B. A., KONDRASHOV F. A., DERENKO M. V., CHUMAKOV I., GRIGORENKO A. P., 2006 Complete Mitochondrial Genome and Phylogeny of Pleistocene Mammoth *Mammuthus primigenius*. *Plos Biol* **4**: e73.
- RYDER M. L., 1964a The Histort of Sheep Breeds in Britain (Cont). The Agricultural History Review.
- RYDER M. L., 1964b The History of Sheep Breeds in Britain. The Agricultural History Review **12**: 1–12.
- SANGER F., NICKLEN S., COULSON A. R., 1977 DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**: 5463–5467.
- SÁNCHEZ-QUINTO F., SCHROEDER H., RAMÍREZ O., ÁVILA-ARCOS M. C., PYBUS M., OLALDE I., VELAZQUEZ A. M. V., MARCOS M. E. P., ENCINAS J. M. V., BERTRANPETIT J., ORLANDO L., GILBERT M. T. P., LALUEZA-FOX C., 2012 Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers. *Current Biology*: 1–6.
- SCHUENEMANN V. J., BOS K., DEWITTE S., SCHMEDES S., JAMIESON J., MITTNIK A., FORREST S., COOMBS B. K., WOOD J. W., EARN D. J. D., WHITE W., KRAUSE J., POINAR H. N., 2011 Targeted enrichment of



- ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death. *Proc Natl Acad Sci U S A* **108**: E746–52.
- SEABURY C. M., WOMACK J. E., 2008 Analysis of sequence variability and protein domain architectures for bovine peptidoglycan recognition protein 1 and Toll-like receptors 2 and 6. *Genomics* **92**: 235–245.
- SEO K., MOHANTY T. R., CHOI T., HWANG I., 2007 Biology of epidermal and hair pigmentation in cattle: a mini-review. *Vet Dermatol* **18**: 392–400.
- SEONG J., SUH D. S., PARK K. D., LEE H. K., KONG H. S., 2011 Identification and analysis of MC4R polymorphisms and their association with economic traits of Korean cattle (Hanwoo). *Mol Biol Rep* **39**: 3597–3601.
- SHAPIRO B., DRUMMOND A. J., RAMBAUT A., WILSON M. C., MATHEUS P. E., SHER A. V., PYBUS O. G., GILBERT M. T. P., BARNES I., BINLADEN J., WILLERSLEV E., HANSEN A. J., BARYSHNIKOV G. F., BURNS J. A., DAVYDOV S., DRIVER J. C., FROESE D. G., HARRINGTON C. R., KEDDIE G., KOSINTSEV P., KUNZ M. L., MARTIN L. D., STEPHENSON R. O., STORER J., TEDFORD R., ZIMOV S., COOPER A., 2004 Rise and fall of the Beringian steppe bison. *Science* **306**: 1561–1565.
- SINOR D., 1990 *The Cambridge history of early inner Asia*. Cambridge University Press.
- SKOGLUND P., JAKOBSSON M., 2011 Archaic human ancestry in East Asia. *PNAS* **108**: 18301–18306.
- SKOGLUND P., MALMSTRÖM H., RAGHAVAN M., STORÅ J., HALL P., WILLERSLEV E., GILBERT M. T. P., GÖTHERSTRÖM A., JAKOBSSON M., 2012 Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336**: 466–469.
- SLAUS M., 1994 Cranial Variation and Mircoevolution in two early middle age sites from Croatia: Privlaka and Stari Jankovci. *Opvsc archaeol*: 273–307.
- SODELAND M., GROVE H., KENT M., TAYLOR S., SVENDSEN M., HAYES B. J., LIEN S., 2011 Molecular characterization of a long range haplotype affecting protein yield and mastitis susceptibility in Norwegian Red cattle. *BMC Genet* **12**: 70.
- STOCK F., EDWARDS C. J., BOLLONGINO R., FINLAY E. K., BURGER J., BRADLEY D. G., 2009 Cytochrome b sequences of ancient cattle and wild ox support phylogenetic complexity in the ancient and modern bovine populations. *Anim. Genet.* **40**: 694–700.
- STONEKING M., KRAUSE J., 2011 Learning about human population history from ancient and modern genomes. *Nature Publishing Group* **12**: 603–614.
- STROBEL A., ISSAD T., CAMOIN L., OZATA M., STROSBERG A. D., 1998 A leptin missense mutation associated with hypogonadism and morbid obesity. *Nature Genetics* **18**: 213–215.

- SVENSSON E. M., ANDERUNG C., BAUBLIENE J., PERSSON P., MALMSTRÖM H., SMITH C., VRETEMARK M., DAUGNORA L., GÖTHERSTRÖM A., 2007 Tracing genetic change over time using nuclear SNPs in ancient and modern cattle. *Anim. Genet.* **38**: 378–383.
- TABERLET P., VALENTINI A., REZAEI H. R., NADERI S., POMPANON F., NEGRINI R., AJMONE-MARSAN P., 2008 Are cattle, sheep, and goats endangered species? *Molecular Ecology* **17**: 275–284.
- THE CHIMPANZEE GENOME SEQUENCING AND ANALYSIS CONSORTIUM, 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.
- THORVALDSDOTTIR H., ROBINSON J. T., MESIROV J. P., 2012 Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*.
- TRESSET A., VIGNE J.-D., 2011 Last hunter-gatherers and first farmers of Europe. *Comptes rendus - Biologies* **334**: 182–189.
- TRESSET A., BOLLONGINO R., EDWARDS C. J., 2009 Early diffusion of domestic bovids in Europe. In: dErrico F, Hombert JM (Eds.), *Becoming Eloquent: Advances in the Emergence of Language, Human Cognition, and Modern Cultures*, Amsterdam, pp. 69–93.
- TROY C. S., MACHUGH D. E., BAILEY J. F., MAGEE D. A., LOFTUS R. T., CUNNINGHAM P., CHAMBERLAIN A. T., SYKES B. C., BRADLEY D. G., 2001 Genetic evidence for Near-Eastern origins of European cattle. *Nature* **410**: 1088–1091.
- VAN DER PLOEG L. H. T., MARTIN W. J., HOWARD A. D., NARGUND R. P., AUSTIN C. P., GUAN X., DRISKO J., CASHEN D., SEBHAT I., PATCHETT A. A., FIGUEROA D. J., DiLELLA A. G., CONNOLLY B. M., WEINBERG D. H., TAN C. P., PALYHA O. C., PONG S.-S., MACNEIL T., ROSENBLUM C., VONGS A., TANG R., YU H., SAILER A. W., FONG T. M., HUANG C., TOTA M. R., CHANG R. S., STEARNS R., TAMVAKOPOULOS C., CHRIST G., DRAZEN D. L., SPAR B. D., NELSON R. J., MACINTYRE D. E., 2002 A role for the melanocortin 4 receptor in sexual function. *Proc Natl Acad Sci U S A* **99**: 11381–11386.
- VAN VUURE C., 2005 *Retracing the Aurochs*.
- VENTER J. C., 2001 The Sequence of the Human Genome. *Science* **291**: 1304–1351.
- VIGNE J. D., HELMER D., 2007 Was milk a “secondary product” in the Old World Neolithisation process? Its role in the domestication of cattle, sheep and goats. *Anthropozoologica*.
- VIGNE J.-D., 2011 The origins of animal domestication and husbandry: A major change in the history of humanity and the biosphere. *Comptes rendus - Biologies* **334**: 171–181.
- VIGNE J.-D., BRIOIS F., ZAZZO A., WILLCOX G., CUCCHI T., THIÉBAULT S.,



- CARRÈRE I., FRANEL Y., TOUQUET R., MARTIN C., MOREAU C., COMBY C.,  
 GUILAINE J., 2012 First wave of cultivators spread to Cyprus at least  
 10,600 y ago. *Proc Natl Acad Sci U S A* **109**: 8445–8449.
- VUISSOZ A., WOROBAY M., ODEGAARD N., BUNCE M., MACHADO C. A.,  
 LYNNERUP N., PEACOCK E. E., GILBERT M. T. P., 2007 The survival of  
 PCR-amplifiable DNA in cow leather. *Journal of Archaeological Science*  
**34**: 823–829.
- WILLERSLEV E., COOPER A., 2005 Review Paper. Ancient DNA. *Proceedings  
 of the Royal Society B: Biological Sciences* **272**: 3–16.
- WORONUK G. N., MARQUESS F. L., JAMES S. T., PALMER J., BERRYERE T.,  
 DEOBALD H., HOWIE S., KONONOFF P. J., 2012 Association of leptin  
 genotypes with beef cattle characteristics. *Anim. Genet.*
- WYLIE A. R. G., 2010 Leptin in farm animals: where are we and where can  
 we go? *Animal* **5**: 246–267.
- YANG D. Y., ENG B., WAYE J. S., DUDAR J. C., SAUNDERS S. R., 1998  
 Technical note: improved DNA extraction from ancient bones using silica-  
 based spin columns. *Am. J. Phys. Anthropol.* **105**: 539–543.
- ZEDER M. A., 2008 Domestication and early agriculture in the Mediterranean  
 Basin: Origins, diffusion, and impact. *Proc Natl Acad Sci U S A* **105**:  
 11597–11604.
- ZEDER M. A., 2011 The Origins of Agriculture in the Near East. *Current  
 Anthropology* **52**: S221–S235.
- ZEDER M. A., EMSHWILLER E., SMITH B. D., BRADLEY D. G., 2006  
 Documenting domestication: the intersection of genetics and archaeology.  
*Trends in Genetics* **22**: 139–155.

# 11. APPENDIX:1 – Extra tables and figures

Appendix:1A: Sheep sequences used for tree drawing and analysis in chapter 3 and 7

| Accession number | Haplogroup |
|------------------|------------|
| HM236175         | A          |
| HM236174         | A          |
| AY858379         | A          |
| NC_001941        | B          |
| HM236177         | B          |
| HM236176         | B          |
| EF490456         | B          |
| EF490454         | B          |
| EF490452         | B          |
| EF490455         | B          |
| EF490453         | B          |
| EF490451         | B          |
| HM236179         | C          |
| HM236178         | C          |
| HM236181         | D          |
| HM236180         | D          |
| HM236183         | E          |
| HM236182         | E          |
| HM236186         | O. ammon   |
| HM236185         | O. musimon |
| HM236184         | O. musimon |
| HM236187         | O.vignei   |
| HM236188         | O.vignei   |
| HM236189         | O.vignei   |

**Appendix 1:B Full list of genes used in capture analysis chapter**

**5**

| Full name                                                                           | Protein code |
|-------------------------------------------------------------------------------------|--------------|
| ATP-binding cassette sub-family G member 2                                          | ABCG2        |
| AGRP agouti related protein homolog (mouse)                                         | AGRP         |
| Fetuin                                                                              | AHSG         |
| Serum albumin                                                                       | ALB          |
| Apolipoprotein A-I                                                                  | APOA1        |
| Apolipoprotein E                                                                    | APOE         |
| Rho GDP-dissociation inhibitor 1                                                    | ARHGDI A     |
| ADP-ribosylation factor-like protein 2                                              | ARL2         |
| ADP-ribosyltransferase 4                                                            | ART4         |
| Agouti-signaling protein                                                            | ASIP         |
| Attractin                                                                           | Atrn         |
| calpain 1, (mu/I) large subunit                                                     | CAPN1        |
| calpain 2, (m/II) large subunit                                                     | CAPN2        |
| calpain 3, (p94)                                                                    | CAPN3        |
| calpastatin                                                                         | CAST         |
| Monocyte differentiation antigen (CD14)                                             | CD14         |
| T-cell surface antigen CD2                                                          | CD2          |
| Tumor necrosis factor receptor superfamily member 5                                 | CD40         |
| CD46 molecule                                                                       | CD46         |
| Alpha-S1-casein                                                                     | CSN1S1       |
| Alpha-S2-casein                                                                     | CSN1S2       |
| Beta-casein                                                                         | CSN2         |
| Kappa-casein                                                                        | CSN3         |
| DBH dopamine beta-hydroxylase (dopamine beta-monooxygenase)                         | DBH          |
| DCT dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2) | DCT          |
| Diacylglycerol O-acyltransferase 1                                                  | DGAT1        |
| DLK1 delta-like 1 homolog (Drosophila)                                              | DLK1         |
| MHC class II antigen DRB3                                                           | DRB3         |
| DRD1 dopamine receptor D1                                                           | DRD1         |
| DRD2 dopamine receptor D2                                                           | DRD2         |
| Dopamine receptor D4                                                                | DRD4         |
| Endothelin B receptor                                                               | EDNRB        |
| Fatty acid synthase                                                                 | FASN         |
| Fez family zinc finger proteinA                                                     | FEZF2        |
| GH1 growth hormone                                                                  | GH1          |
| GHR growth hormone receptor                                                         | GHR          |
| GNAS complex locus                                                                  | GNAS         |
| GTP binding protein G I G S G T beta subunit 1                                      | GNB1         |



|                                                                              |             |
|------------------------------------------------------------------------------|-------------|
| Grb10 growth factor receptor-bound protein 10                                | GRB10       |
| Gluthathione-S-transferase theta 2(Human theta 2 maps to theta 1 in cattle)  | GSTT1       |
| Gluthathione-S-transferase theta 2(Human theta 2 maps to theta 3 in cattle)  | GSTT3       |
| Histatherin                                                                  | HSTN        |
| HTR1A 5-hydroxytryptamine (serotonin) receptor 1A                            | HTR1A       |
| HTR2A 5-hydroxytryptamine (serotonin) receptor 2A                            | HTR2A       |
| IFNG interferon, gamma                                                       | IFNG        |
| IGF1 insulin-like growth factor 1 (somatomedin C)                            | IGF1        |
| IGF1R insulin-like growth factor 1 receptor                                  | IGF1R       |
| IGF2R insulin-like growth factor 2 receptor                                  | IGF2R       |
| Insulin-like growth factor II                                                | IGF2        |
| Interleukin-8                                                                | IL-8        |
| Interleukin-13                                                               | IL13        |
| Interleukin-1 beta                                                           | IL1B        |
| Interleukin-2                                                                | IL2         |
| Interleukin-6                                                                | IL6         |
| INS insulin                                                                  | INS         |
| INSR insulin receptor                                                        | INSR        |
| Janus kinase and microtubule-interacting protein 1                           | JAKMIP<br>1 |
| KIT v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog            | KIT         |
| Kit ligand                                                                   | KITLG       |
| Alpha-lactalbumin                                                            | LALBA       |
| Leptin                                                                       | LEP         |
| Beta-lactoglobulin                                                           | LGB         |
| Lactotransferrin(Lactoferrin)                                                | LTF         |
| MAOA monoamine oxidase A                                                     | MAOA        |
| MC1R melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor) | MC1R        |
| Mahogunin                                                                    | Mgrn1       |
| Microphthalmia associated transcription factor isoform M                     | MITF        |
| 39S ribosomal protein L30, mitochondrial                                     | MRPL30      |
| 28S ribosomal protein S14, mitochondrial                                     | MRPS14      |
| Macrophage scavenger receptor types I and II                                 | MSR1        |
| myostatin                                                                    | MSTN        |
| Myocilin                                                                     | MYOC        |
| NFKB1 nuclear factor of kappa light polypeptide gene enhancer in B-cells 1   | NFKB1       |
| Neuropeptide Y                                                               | NPY         |
| Pregnancy-associated endometrial alpha-2 globulin                            | PAEP        |
| PCYOX1 protein                                                               | PCYOX1      |
| PEG10 paternally expressed 10                                                | PEG10       |
| PEG3 paternally expressed 3                                                  | PEG3        |
| Peptidoglycan recognition protein 1                                          | PGLYRP<br>1 |
| Polymeric immunoglobulin receptor                                            | PIGR        |



|                                                                                        |         |
|----------------------------------------------------------------------------------------|---------|
| Pro-melanin concentrating hormone                                                      | PMCH    |
| Pro-opiomelanocortin                                                                   | POMC    |
| Major prion protein                                                                    | PRNP    |
| Pentraxin-3                                                                            | PTX3    |
| R3H domain-containing protein 1                                                        | R3HDM1  |
| Rab 11B                                                                                | RAB11B  |
| Rab 18                                                                                 | RAB18   |
| Rab 3A                                                                                 | RAB3A   |
| Rab 3C                                                                                 | RAB3C   |
| Rap 1b                                                                                 | RAP1B   |
| RASGRF1 (Bos taurus)                                                                   | RASGRF1 |
| RTL1 retrotransposon-like 1                                                            | RTL1    |
| S100 calcium-binding protein A12                                                       | S100A2  |
| GTP binding protein Sara                                                               | SAR1A   |
| SAR1B protein                                                                          | SAR1B   |
| SERGEF secretion regulating guanine nucleotide exchange factor                         | SERGEF  |
| Melanocyte protein Pmel 17                                                             | SILV    |
| SLAMF1                                                                                 | SLAMF1  |
| SLAMF6                                                                                 | SLAMF6  |
| SLAMF7                                                                                 | SLAMF7  |
| Solute carrier family 24 member 5                                                      | Slc24a5 |
| SLC6A4 solute carrier family 6 (neurotransmitter transporter, serotonin), member 4     | SLC6A4  |
| STAT1 signal transducer and activator of transcription 1, 91kDa                        | STAT1   |
| STAT3 signal transducer and activator of transcription 3 (acute-phase response factor) | STAT3   |
| STAT5A signal transducer and activator of transcription 5A                             | STAT5A  |
| Statherin                                                                              | STATH   |
| TBC1 domain family member 1                                                            | TBC1D1  |
| TIR domain-containing adapter molecule 1                                               | TICAM1  |
| TLR1 toll-like receptor 1                                                              | TLR1    |
| Toll-like receptor 10                                                                  | TLR10   |
| TLR2 toll-like receptor 2                                                              | TLR2    |
| TLR4 toll-like receptor 4                                                              | TLR4    |
| Toll-like receptor 6                                                                   | TLR6    |
| Toll-like receptor 9                                                                   | TLR9    |
| Tumor necrosis factor                                                                  | TNFA    |
| TYR tyrosinase                                                                         | TYR     |
| TYRO protein tyrosine kinase-binding protein                                           | TYROBP  |
| TYRP1 tyrosinase-related protein 1                                                     | TYRP1   |
| Tyrosinase-related protein 2                                                           | TYRP2   |
| 14-3-3 protein beta alpha                                                              | YWHAB   |
| ZIM2 zinc finger, imprinted 2                                                          | ZIM2    |



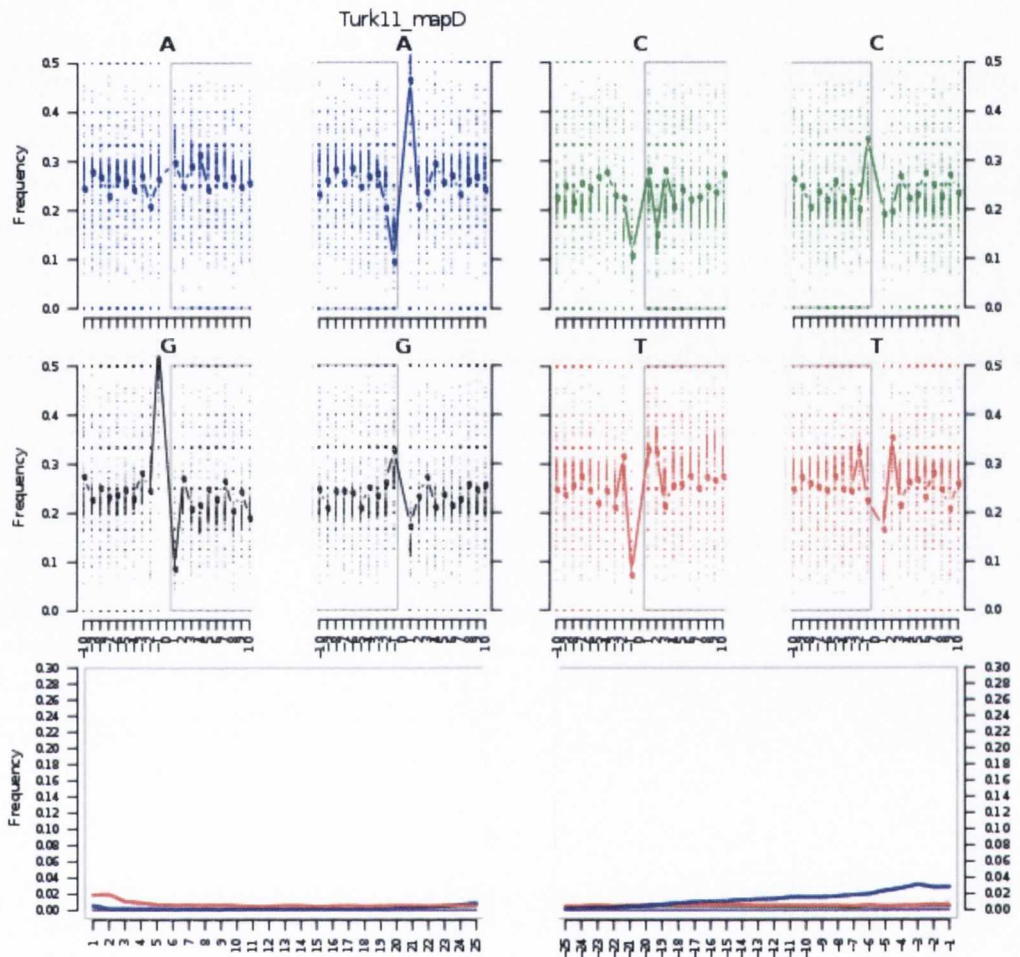
**Appendix 1:C Full list of high coverage 3' UTR mutations  
discovered in CPC98**

| Location | Gene       | Chromosome | Position  |
|----------|------------|------------|-----------|
| UTR3     | ZIM2       | chr18      | 64180100  |
| UTR3     | CNOT7      | chr27      | 19067353  |
| UTR3     | TAT        | chr18      | 39624096  |
| UTR3     | DEM1       | chr3       | 106235104 |
| UTR3     | CSN3       | chr6       | 87392665  |
| UTR3     | ADAMTS1    | chr1       | 8963404   |
| UTR3     | LEP        | chr4       | 93265006  |
| UTR3     | TOX4       | chr10      | 25732248  |
| UTR3     | RRS1       | chr14      | 32700717  |
| UTR5     | RAP1B      | chr5       | 45364944  |
| UTR3     | PDLIM2     | chr8       | 70401417  |
| UTR3     | LTA        | chr23      | 27537950  |
| UTR3     | LIX1L      | chr3       | 21549827  |
| UTR3     | ASCC3      | chr9       | 50102947  |
| UTR3     | ADNP       | chr13      | 79546003  |
| UTR3     | MYOC       | chr16      | 39957568  |
| UTR3     | TAT        | chr18      | 39624048  |
| UTR3     | ZIM2       | chr18      | 64178797  |
| UTR3     | SMARCE1    | chr19      | 41488860  |
| UTR3     | OXCT1      | chr20      | 32848666  |
| UTR3     | CSRP2      | chr5       | 6266261   |
| UTR5     | CSN1S2     | chr6       | 87264294  |
| UTR3     | PRNP       | chr13      | 47416714  |
| UTR3     | SLC7A6     | chr18      | 35793633  |
| UTR3     | ZIM2       | chr18      | 64182363  |
| UTR3     | SLC1A3     | chr20      | 37596667  |
| UTR3     | GABRB3     | chr21      | 4001747   |
| UTR3     | ZSCAN12    | chr23      | 30091124  |
| UTR3     | ACSL5      | chr26      | 33234819  |
| UTR3     | LEP        | chr4       | 93264965  |
| UTR3     | LEP        | chr4       | 93265183  |
| UTR3     | GALNT4     | chr5       | 19458651  |
| UTR3     | LALBA      | chr5       | 31349786  |
| UTR3     | CSN3       | chr6       | 87392591  |
| UTR3     | CSN3       | chr6       | 87392592  |
| UTR3     | CBR4       | chr8       | 943817    |
| UTR3     | C9H6orf120 | chr9       | 105074182 |
| UTR3     | PTTG1IP    | chr1       | 145059132 |
| UTR5     | HTR2A      | chr12      | 16888315  |

|                 |                 |       |           |
|-----------------|-----------------|-------|-----------|
| UTR3            | ZIM2            | chr18 | 64178108  |
| UTR3            | ABHD6           | chr22 | 43569849  |
| UTR3            | SEPHS2          | chr25 | 26867622  |
| UTR3            | ZDHHC4          | chr25 | 38961935  |
| UTR3            | MSR1            | chr27 | 20017481  |
| UTR3            | PCP4L1          | chr3  | 8250165   |
| UTR3            | TMEM175         | chr6  | 109056666 |
| UTR3            | YIPF2           | chr7  | 16588162  |
| UTR3            | CAST            | chr7  | 98579663  |
| UTR3            | EXOC5           | chr10 | 69862924  |
| UTR3            | ZIM2            | chr18 | 64179212  |
| exonic;splicing | TUBGCP5;TUBGCP5 | chr2  | 1164256   |
| UTR3            | CEP76           | chr24 | 43500292  |
| UTR3            | PTBP2           | chr3  | 46842537  |
| UTR5            | HSTN            | chr6  | 87189842  |
| UTR3            | GRAMD3          | chr7  | 28657444  |
| UTR3            | CAST            | chr7  | 98580401  |
| UTR3            | TMEM50B         | chr1  | 1359951   |
| UTR3            | MRPL30          | chr11 | 4321288   |
| UTR3            | YWHAB           | chr13 | 74040186  |
| UTR5            | SERTAD4         | chr16 | 74999809  |
| UTR3            | ZNF404          | chr18 | 52400195  |
| UTR3            | ZIM2            | chr18 | 64178009  |
| UTR3            | STAT3           | chr19 | 43058528  |
| UTR3            | MSTN            | chr2  | 6218869   |
| UTR3            | LEP             | chr4  | 93265322  |
| UTR3            | LEP             | chr4  | 93265455  |
| UTR3            | CSN1S2          | chr6  | 87280796  |
| UTR3            | TRIM7           | chr7  | 41718041  |

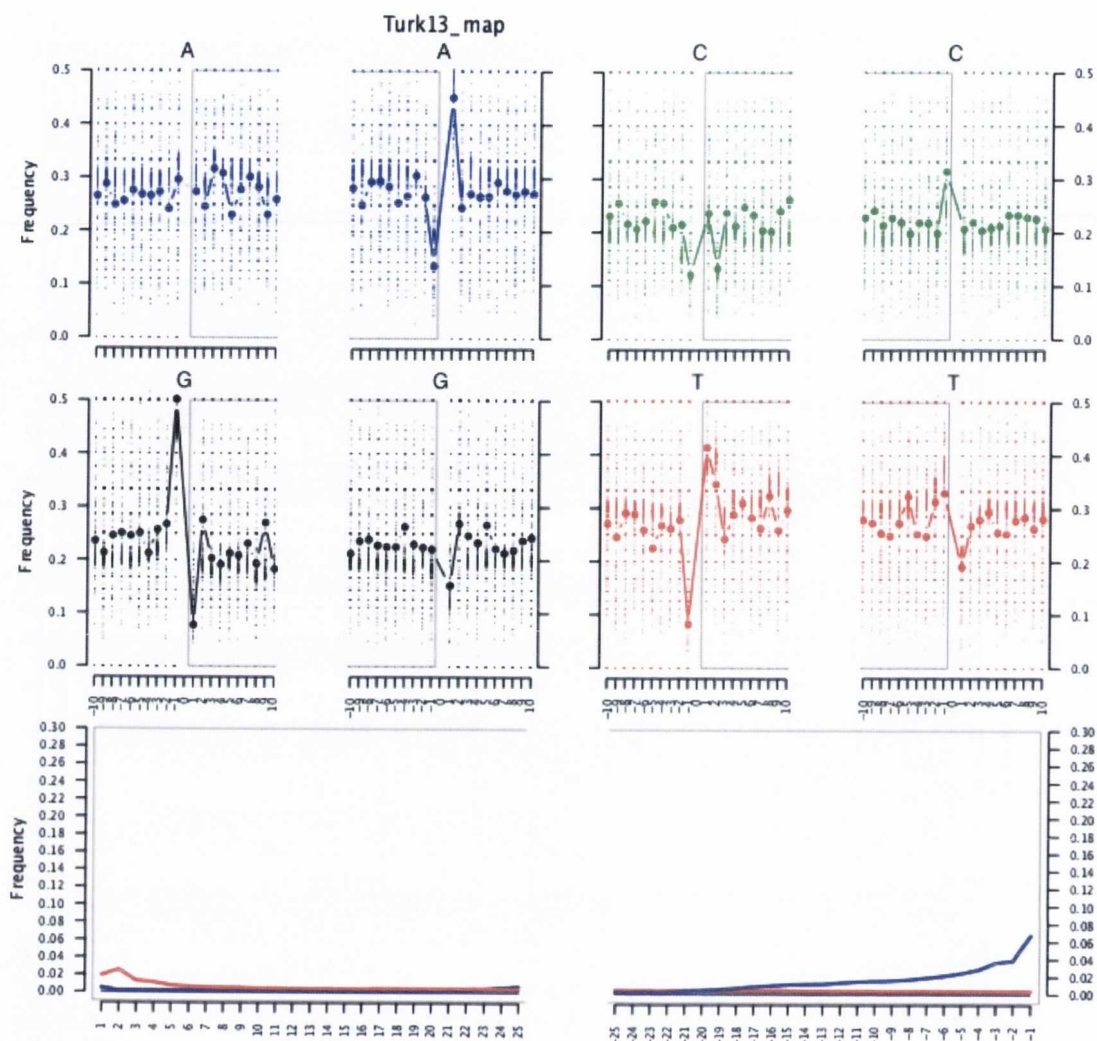


## Appendix 1:D: MAPDAMAGE results for Turk11 and Turk13



Map damage result for Turk11 showing patterns suggestive of aDNA when compared to a Phusion polymerase reference analysis of a Saqqaq individual (<http://geogenetics.ku.dk/publications/mapdamage/>).





Map damage result for Turk13 showing patterns charatertic of aDNA when compaired to a Phusion polymerase reference analysis of a Saqqaq individual (<http://geogenetics.ku.dk/publications/mapdamage/>).

## Appendix1:E sheep hapmap breeds used in this study

| Number | Breed                    |
|--------|--------------------------|
| 24     | AfecAssaf                |
| 21     | AfricanDorper            |
| 6      | AfricanWhiteDorper       |
| 37     | Afshari                  |
| 24     | Altamurana               |
| 34     | Arawapa                  |
| 3      | Arawapa                  |
| 19     | AustralianCoopworth      |
| 88     | AustralianIndustryMerino |
| 50     | AustralianMerino         |
| 108    | AustralianPollDorset     |
| 98     | AustralianPollMerino     |
| 109    | AustralianSuffolk        |
| 24     | BangladeshiBGE           |
| 24     | BangladeshiGarole        |
| 24     | BarbadosBlackBelly       |
| 24     | BlackHeadedMountain      |
| 48     | BorderLeicester          |
| 17     | Boreray                  |
| 23     | BrazilianCreole          |
| 24     | BundnerOberlanderSheep   |
| 23     | Castellana               |
| 29     | Changthangi              |
| 23     | ChineseMerino            |
| 23     | Chios                    |
| 120    | Churra                   |
| 24     | Comisana                 |
| 30     | CyprusFatTail            |
| 24     | Deccani                  |
| 21     | DorsetHorn               |
| 39     | EastFriesianBrown        |
| 9      | EastFriesianWhite        |
| 13     | EgyptianBarki            |
| 24     | EngadineRedSheep         |
| 34     | EthiopianMenz            |
| 99     | Finnsheep                |
| 49     | Galway                   |
| 22     | Garut                    |
| 46     | GermanTexel              |

|     |                               |
|-----|-------------------------------|
| 94  | GulfCoastNative               |
| 16  | Icelandic                     |
| 23  | ImprovedAwassi                |
| 26  | IndianGarole                  |
| 55  | IrishSuffolk                  |
| 18  | Karakas                       |
| 24  | Leccese                       |
| 24  | LocalAwassi                   |
| 10  | MacarthurMerino               |
| 78  | MeatLacaune                   |
| 24  | Merinolandschaf               |
| 103 | MilkLacaune                   |
| 34  | Moghani                       |
| 22  | MoradaNova                    |
| 12  | NamaquaAfrikaner              |
| 24  | NewZealandRomney              |
| 24  | NewZealandTexel               |
| 20  | Norduz                        |
| 24  | Ojalada                       |
| 15  | OldNorwegianspaelsau          |
| 35  | Qezel                         |
| 102 | Rambouillet                   |
| 22  | Rasaaragonesa                 |
| 45  | RedMaasai                     |
| 17  | RonderibAfrikaner             |
| 22  | Sakiz                         |
| 47  | SantaInes                     |
| 20  | SardinianAncestralBlack       |
| 56  | ScottishBlackface             |
| 80  | ScottishTexel                 |
| 110 | Soay                          |
| 3   | Spael-coloured                |
| 32  | Spael-white                   |
| 1   | SriLankan                     |
| 10  | StElizabeth                   |
| 24  | Sumatra                       |
| 24  | SwissBlack-BrownMountainSheep |
| 24  | SwissMirrorSheep              |
| 24  | SwissWhiteAlpineSheep         |
| 37  | Tibetan                       |
| 24  | ValaisBlacknoseSheep          |
| 24  | ValaisRedSheep                |
| 23  | Wiltshire                     |



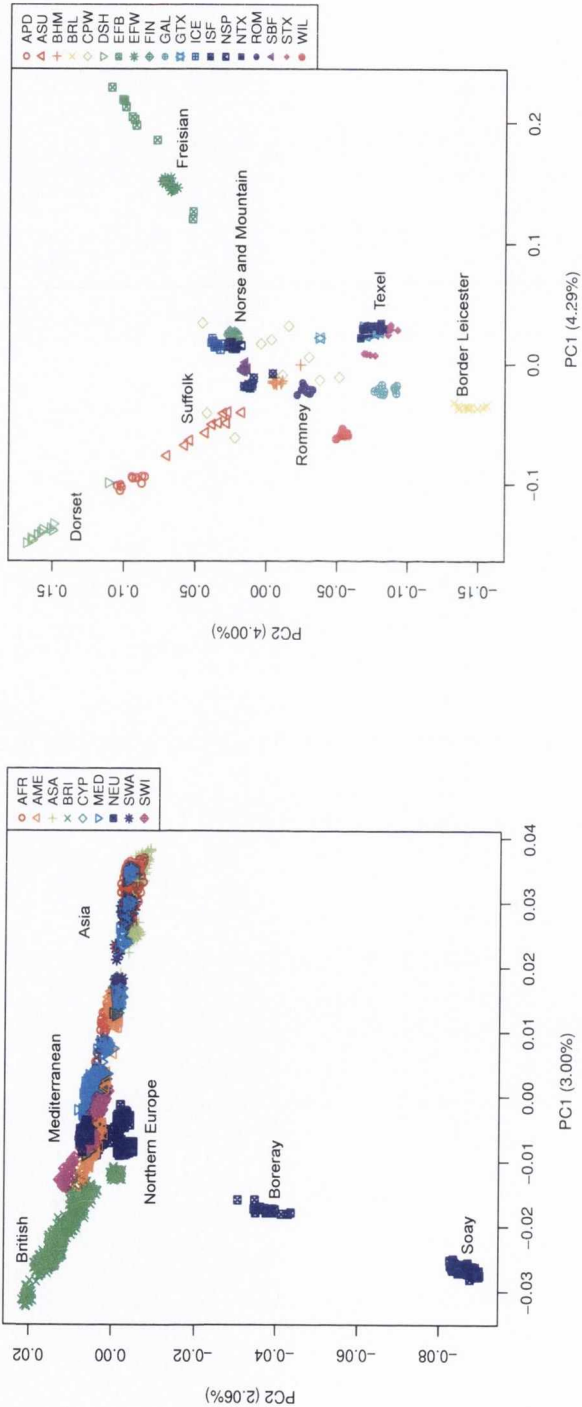
Appendix1F: Raw parchment *FastQ\_Screen* results

Table below presents the raw data used to produce (Figure 7:2) 1H1L = 1 hit 1 library/genome, MH1L = multiple hits 1 library, 1HML = multiple hits multiple libraries, MHML = multiple hits multiple libraries.

|                 | 1H1<br>L<br>(%) | MH1<br>L<br>(%) | 1HML<br>L<br>(%) | MH<br>ML<br>(%) | Tota<br>l<br>(%) | 1H1<br>L<br>(%) | MH1<br>L<br>(%) | 1HML<br>L<br>(%) | MH<br>ML<br>(%) | Tota<br>l<br>(%) | 1H1<br>L<br>(%) | MH1<br>L<br>(%) | 1HML<br>L<br>(%) | MH<br>ML<br>(%) | Tota<br>l<br>(%) | 1H1<br>L<br>(%) | MH1<br>L<br>(%) | 1HML<br>L<br>(%) | MH<br>ML<br>(%) | Tota<br>l<br>(%) |
|-----------------|-----------------|-----------------|------------------|-----------------|------------------|-----------------|-----------------|------------------|-----------------|------------------|-----------------|-----------------|------------------|-----------------|------------------|-----------------|-----------------|------------------|-----------------|------------------|
| P hg19          | 0.04            | 0.02            | 0.27             | 0.63            | 0.96             | 0.07            | 0.02            | 0.66             | 1.49            | 2.24             | 0.09            | 0.03            | 1.07             | 2.66            | 3.85             | 0.09            | 0.03            | 1.08             | 2.66            | 3.86             |
| A bosTa<br>1 u6 | 0.81            | 0.36            | 4.95             | 28.06           | 34.18            | 0.66            | 0.29            | 7.85             | 34.44           | 43.24            | 0.66            | 0.28            | 8.75             | 36.44           | 46.13            | 0.66            | 0.28            | 8.63             | 36.38           | 45.95            |
| OAR2            | 18.09           | 8.77            | 4.78             | 28.26           | 59.9             | 15.2            | 12.21           | 7.61             | 34.78           | 69.8             | 14.4            | 12.65           | 8.42             | 36.96           | 72.43            | 14.52           | 12.54           | 8.3              | 36.9            | 72.26            |
|                 |                 |                 |                  |                 |                  |                 |                 |                  |                 |                  |                 |                 |                  |                 |                  |                 |                 |                  |                 |                  |
| P hg19          | 0.02            | 0.01            | 0.25             | 0.27            | 0.55             | 0.02            | 0.01            | 0.53             | 0.65            | 1.21             | 0.03            | 0.01            | 0.71             | 1.01            | 1.76             | 0.03            | 0.01            | 0.7              | 1               | 1.74             |
| A BosTa<br>2 u6 | 0.95            | 0.38            | 7.86             | 22.45           | 31.64            | 0.78            | 0.33            | 14.62            | 28.79           | 44.52            | 0.77            | 0.33            | 16.78            | 30.49           | 48.37            | 0.77            | 0.33            | 16.26            | 30.32           | 47.68            |
| OAR2            | 30.95           | 7               | 7.73             | 22.6            | 68.28            | 26.38           | 10.4            | 14.46            | 29              | 80.24            | 23.87           | 11.36           | 16.6             | 30.75           | 82.58            | 24.47           | 11.11           | 16.07            | 30.58           | 82.23            |



Appendix 1:G PCA analysis full sheep HapMap dataset



SMARTPCA Sheep HapMap full dataset worldwide right and European left (48,596 SNPs). Note more clustering to breeds seen at this resolution.

## **12. APPENDIX:2 - Code**

A digital appendix “teasdale\_m\_code.tar.gz” including code used in this thesis can be found at.

<http://www.gen.tcd.ie/molpopgen/resources/>