Museum of London Report on the DNA Analyses of Four Roman Individuals

Supplementary Information

Katherine Eaton[1], Ana T. Duggan[1], Alison Devault[2], and Hendrik Poinar[1]

[1]McMaster Ancient DNA Centre, Department of Anthropology, McMaster University, Hamilton, ON, Canada
[2]MYcroarray, Ann Arbor, MI, USA

November 27, 2015

Contents

1 Sample Information

The four individuals examined in this study are permanently stored at the Museum of London. Osteobiographical data and excavation context for these individuals are summarized in Table 1.

Table 1: Summary of Skeletal Characteristics and Excavation Context

Individual ID	Location	Time Period	Skeletal Sex	Age
HR79 SK311 T13	Harper Road, Southwark	50-70 AD	Female	>45 Years
MNL88 SK37 T16	65-73 Mansell Street	180-400 AD	Male	>45 Years
LTU03 SK385 T17	Lant Street, Southwark	300-400 AD	Female	14 Years
LOW88 SK695.5 T18	London Wall	125-200 AD	Male	36-45 Years

Subsampling and all subsequent labwork was conducted at the McMaster Ancient DNA Centre. All pre-PCR work, including subsampling, extraction, and library prep, was conducted in dedicated ancient DNA clean rooms that have not been exposed to high-copy modern DNA or high copy PCR products. The remaining labwork (indexing amplification and targeted enrichment) was conducted in the post-PCR modern laboratories.

2 Pre-PCR Labwork

2.1 Subsampling, Demineralization, Protein Digestion

A piece of root, including the pulp cavity, was sectioned from each individual with a diamond tipped dremel wheel (Figure 1). Individuals MNL88 and LTU03 had roots sufficiently large to take two subsamples of one root. The sectioned piece was then crushed into fine pieces and weighed so that the final mass of subsampled bone used was between 50 and 100 mg. The root pieces were demineralized in 0.5 M EDTA (pH 8) in a thermomixer, set at 1000 rpm and 22°C, for 24 hours. One negative extraction blank consisting of only reagents and no DNA accompanied the four samples through all subsequent procedures. The demineralized root pieces were pelleted, and the supernatant removed and stored at -20°C. The root pellet was then digested in 0.5 mL of digestion buffer containing as final concentrations: 250 µg/ml Proteinase K, 10 mM Tris-CL, 0.5% Sarcosyl, 5 mM $CaCl_2$, 50 mM DTT, 1% Polyvinylpyrrolidone, and 2.5 mM N-phenacylthiazolium bromide. Digestion occurred over 24 hours at 25°C and under agitation in a thermomixer at 1000 rpm. The digested root pieces were pelleted, and the supernatant removed and stored at -20°C. The demineralization and digestion steps were repeated once more to produce 1 mL each of demineralization-digestion supernatant per round.

Figure 1: Molar of individual MNL88 with root subsampled.

2.2 Extraction, Concentration, Purification

Each demineralization-digestion supernatant mixture was extracted separately for each round. DNA was extracted using a modified phenol-chloroform method as previously described (Schwarz et al. 2009). The extraction products were concentrated with ultrafiltration columns (10K Amicon Ultra 0.5mL; Millipore, MA) and purified over silica columns with an additional wash step (MinElute PCR Purification Kit; Qiagen, Hilden, Germany).

3 Library Preparation and Indexing

Preparation of Illumina sequencing libraries from the purified DNA extracts was performed according to a previously published protocol (Meyer and Kircher 2010) with recommended modifications for double-indexing (Kircher et al. 2012). QIAquick MinElute PCR purification was used instead of SPRI beads for intermediate clean-up steps. After the final step of adapter fill-in, heat inactivation at 80°C for 20 minutes was substituted instead of additional purification.

Prepared libraries were double-indexed in a 40 µl reaction containing KAPA SYBR® FAST polymerase (Kircher et al. 2012). Reaction concentrations and cycling conditions were modified to accommodate the KAPA polymerase mastermix. Indexing proceeded until the first sample began to plateau at cycle 9. After indexing amplification, 40 µl of volume was purified over MinElute column with a final elution volume of 15 µl.

4 In-Solution Bait Design

Custom capture probes for targeted sequencing were designed and produced as in-solution MYbaits kits (MYcroarray, Ann Arbor, MI, USA).

4.1 Human Mitochondrial Baits

A bait set for targeted enrichment of the human mitochondrial genome had previously been designed using GenBank Accession number J01415.2. The mitochondrial bait set consists of 80mer biotinylated RNA probes with 21 bp tiling for a total of 785 baits.

4.2 Hair and Eye Color

Probes used for the identification of hair and eye color were designed using the HIrisPlex SNP assay (Table 2) (Walsh, Liu, et al. 2013). SNP information including gene name and the major and minor alleles were obtained from the original HIrisPlex assay publication (Walsh, Liu, et al. 2013). Chromosome position for each SNP with reference to hg38 was obtained from dbSNP build 142

(dbSNP Build ID: 142 2014). 80 mer biotinylated RNA probes were tiled every 20 bp so that each major and minor allele had 5X bait coverage, for a total of 10X bait coverage per locus.

Table 2: HIrisPlex SNPs

	SNP	Gene	Major Allele	Minor Allele	Position (hg38)
1	rs312262906	MC1R	-	A	Chr16: 89919342
2	rs11547464	MC1R	G	A	Chr16: 89919683
3	rs885479	MC1R	C	T	Chr16: 89919746
4	rs1805008	MC1R	C	T	Chr16: 89919736
5	rs1805005	MC1R	G	T	Chr16: 89919436
6	rs1805006	MC1R	C	A	Chr16: 89919510
7	rs1805007	MC1R	C	T	Chr16: 89919709
8	rs1805009	MC1R	G	C	Chr16: 89920138
9	rs201326893	MC1R	C	A	Chr16: 89919714
10	rs2228479	MC1R	G	A	Chr16: 89919532
11	rs1110400	MC1R	T	C	Chr16: 89919722
12	rs28777	SLC45A2	A	C	Chr5: 33958854
13	rs16891982	SLC45A2	G	C	Chr5: 33951588
14	rs12821256	KITLG	A	G	Chr12: 88934558
15	rs4959270	EXOC2	C	A	Chr6: 457748
16	rs12203592	IRF4	C	T	Chr6: 396321
17	rs1042602	TYR	G	T	Chr11: 89178528
18	rs1800407	OCA2	G	A	Chr15: 27985172
19	rs2402130	SLC24A4	A	G	Chr14: 92334859
20	rs12913832	HERC2	C	T	Chr15: 28120472
21	rs2378249	ASIP/PIGU	T	C	Chr20: 34630286
22	rs12896399	SLC24A4	T	G	Chr14: 92307319
23	rs1393350	TYR	C	T	Chr11: 89277878
24	rs683	TYRP1	T	G	Chr9: 12709305

4.3 Sex Identification

Baits used for sex chromosome typing were designed to target three Y chromosome-specific genes (AMELY, SRY, TSPY) and two X chromosome-specific genes (AMELX, SOX3) (Table 3). The amelogenin locus, which encodes a protein associated with developing tooth enamel, is a frequently used target for sex identification (Nakahori, Hamano, et al. 1991). AMELX and AMELY share a large degree of homology with approximately 89% sequence similarity between them (Nakahori, Takenaka, et al. 1991). However, length polymorphisms between the X and Y homologues allow for differentiation and detection of the X and Y chromosomes. The sex-determining region Y (SRY) gene located on the Y gene is strongly associated with the development of male reproductive organs (Gubbay et al. 1990). SRY exhibits sequence similarity to the SOX3 gene, located on the X chromosome (Katoh K 1999). The inclusion of single-copy paired genes (AMELX/AMELY and SRY/SOX3) provides the opportunity to confirm the expected 1:1 ratio of the X and Y chromosome

in males. The Y-encoded testis-specific protein 1 (TSPY1) is part of the TSPY heterogenous gene family that ranges in copy number from 11 to 76 copies on the Y chromosome in humans (Krausz et al. 2010). The high copy number of TSPY provides increased sensitivity for detecting the presence of a Y chromosome in highly degraded or low concentration samples (Yasmin et al. 2015; Kamodyová et al. 2013; Benoit et al. 2013; Jacot et al. 2013; Campos et al. 2014). The autosomal and X chromosome homologues of TSPY are single-copy genes and thus are of limited use to compare sex chromosome ratios with TSPY.

Gene sequences and NCBI genbank annotations for the five 5 sex-linked genes were downloaded from RefSeq (AMELX: NC_000023.11, AMELY: NC_000024.10, SRY: NC_000024.10, SOX3: NC_000023.11, TSPY1 NC_000024.10). Targeted regions for sex-estimation using AMELX, AMELY, SRY, and TSPY were obtained from previously published loci (Butler and R. Li 2014, Table 2, Table 5, Table 6). SOX3 regions were chosen based on areas where sequence similarity was less than 85% when compared to SRY. To further refine specificity in bait design, a series of multiple alignments to detect sequence homology were conducted using the implementation of MUSCLE (Edgar 2004) in Geneious version 7.1.9 (Kearse et al. 2012). Potential homologues of the sex chromosome-linked genes used in this study were identified through a recent review on genetic markers for sex identification (Butler and R. Li 2014) and the HomoloGene tool for detecting Homologues in the NCBI RefSeq Database (Pruitt et al. 2014) (Table 3). Any regions within the sex-linked gene containing high levels of sequence similarity (>85%) with non-target regions were removed from the bait design.

Table 3: Sex Chromosome Gene Targets

Gene	**Chromosome**	**Copy Number**	**X Homologues**	**Y Homologues**	**Autosomal Homologues**
AMELX	X	1	None	AMELY	None
AMELY	Y	1	AMELX	None	None
SOX3	X	1	None	SRY	SOX family [1]
SRY	Y	1	SOX3	None	SOX family [1]
TSPY	Y	11-76 [3]	TSPYL2	None	TSPYL Family [2]

[1] SOX gene homologues described in Table 4
[2] TSPY gene homologues described in Table 5
[3] Krausz et al. 2010

Table 4: SOX Family Homologues

Gene	Chromosome	Coordinates	Accession
SRY	Y	2786855..2787741 (complement)	NC_000024.10
SOX1	13	112067599..112071706	NC_000013.11
SOX2	3	181711924..181714436	NC_000003.12
SOX3	X	140502987..140505060	NC_000023.11
SOX4	6	21593741..21598619	NC_000006.12
SOX5	12	23529495..24562669	NC_000012.12
SOX6	11	15966449..16476388 (complement)	NC_000011.10
SOX7	8	10723768..10730574 (complement)	NC_000008.11
SOX8	16	981808..986979	NC_000016.10
SOX9	17	72121020..72126420	NC_000017.11
SOX10	22	37972312..37984532 (complement)	NC_000022.11
SOX11	2	5692667..5701385	NC_000002.12
SOX12	20	325571..330228	NC_000020.11
SOX13	1	204073109..204127743	NC_000001.11
SOX14	3	137764292..137766334	NC_000003.12
SOX15	17	7588180..7590170 (complement)	NC_000017.11
SOX17	8	54457935..54460896	NC_000008.11
SOX18	20	64047726..64049626 (complement)	NC_000020.11
SOX21	20	94709622..94712543 (complement)	NC_000020.11
SOX30	5	157624796..157672756 (complement)	NC_000005.10

Table 5: TSPY Family Homologues

Gene	Chromosome	Coordinates	Accession
TSPYL1	6	116274859..116280117 (complement)	NC_000006.12
TSPYL2	X	53082360..53088546	NC_000023.11
TSPYL4	6	116249964..116254098 (complement)	NC_000006.12
TSPYL5	8	97273486..97277948 (complement)	NC_000008.11
TSPYL6	2	54253178..54256272 (complement)	NC_000002.12

Candidate sequences were then repeat-masked using RepeatMasker (version open-4.0.5) using the search engine NCBI/RMBLAST (Smit et al. 2013-2015). The resulting sequences were sent to MYcroarray (MYcroarray, Ann Arbor, MI, USA) where they were decomposed into 80-mer probes. Each probe was then blasted against the Human hg38 genome (Altschul et al. 1990) and a hybridization temperature was predicted for each hit. Baits that did not pass the hybridization stringency filter were excluded. A final round of blasting excluding human was conducted on the final probes to ensure no hits against non-target organisms such as soil bacteria. The relative contribution of each gene within the baitset is detailed in Table 6.

Table 6: Composition of the Sex Estimation Bait Set

Gene	**Nucleotides in Bait Set (bp)**	**Proportion**
AMELX	1556	0.28
AMELY	1466	0.27
SOX3	1479	0.27
SRY	266	0.05
TSPY	760	0.14
Total	5527	1

5 Targeted Enrichment

5.1 Human Mitochondria, HIrisPlex, Sex

Two rounds of targeted capture were performed on one indexed library from each individual and the extraction blank. An enrichment blank was also included with input library replaced with water. Several modifications to the original MYbaits Protocol (version 2.3.1) were made to improve capture sensitivity and to reduce reagent waste. The Hybridization Master Mix and Capture Baits Master Mix were combined according to updated manufacturer recommendations. The input library volume was increased to 9 µl and Block #3 was replaced with custom blocking oligos complementary to the P5 and P7 adapters and flow-cell binding sequences. To accommodate the additional input library, volumes of Block #1, Block #2, and Block #3, were reduced to 1.94 µl, 1.94 µl, and 0.40 µl respectively. Volumes of HYB #1. HYB #2, HYB #3, and HYB #4 were reduced to 8.23 µl, 0.33 µl, 3.30 µl, and 0.33 µl. RNase Block (Superase) was increased to 1.2 µl and Mitochondrial and HIrisSex probes were combined, with inputs of 50 ng and 100 ng respectively, for a total of 2.32 µl of probes (1.16 µl each). The library-bait hybridization was performed at 55°C for 16 hours to improve sensitivity and to increase target complexity. The volume of streptavidin-coated magnetic beads was reduced to 20 µl and bait-bead hybridization was performed at 55°C with rotation. Wash Buffer #1 was eliminated and the concentration of Wash Buffer #2 was reduced to a five-fold dilution containing 0.08% SDS to increase wash stringency. Post-washed bait-bead complexes were resuspended in 20 µl and taken directly into post-enrichment re-amplification. Re-amplification was performed in a 40 µl reaction volume using KAPA SYBR ® FAST qPCR Master Mix (2X) at 1X, 150 µM of forward primer, 150 µM of reverse primer, and 18.80 µl of bait-bead complex. Cycling

conditions were as follows: initialization for 5 minutes at 95°C, 14 cycles of amplification including denaturation at 95°C for 30 seconds and combined annealing/extension at 60°C for 45 seconds, followed by a final extension at 60°C for 3 minutes. The entire enrichment protocol was repeated for a second round to improve target specificity.

6 Sequencing

Enriched libraries were sequenced on the Illumina HiSeq 1500 platform at the Farncombe Family Digestive Health Research Institute of McMaster University. Libraries were pooled in equimolar ratios to occupy 12.5% of one lane. We used 75 bp paired-end read chemistry but the total read length was extended to 90 bp through additional sequencing reagents. Adapter sequences were trimmed followed by overlap-merging of paired end reads using leeHom (Renaud et al. 2014) with default parameters for ancient dna. Reference sequences for analysis in this study include the mitochondrial reference sequence, the revised Cambridge Reference Sequence (rCRS) (NC_012920) which was chosen to ensure compatibility with the haplotyping tool HaploGrep2 (Kloss-Brandstätter et al. 2011). Reference sequences for the targeted nuclear loci consisted of the HIrisPlex and sex chromosome regions. The merged or single-end reads were all mapped as single-end reads using BWA aln (H. Li and Durbin 2009), with seeding disabled for greater sensitivity, the maximum number of gap opens increased to 2, and fraction of missing alignments set to 0.01. PCR duplicate removal was conducted using bam-rmdup (Available from https://github.com/udo-stenzel/biohazard). All reads shorter than 24 base pairs or with a mapping quality less than 30 were discarded. General sequencing metrics for the individuals are described in Table 7.

Table 7: Sequencing Metrics

Sample ID	**Total Reads**	**Merged or Single-end Reads**
MNL88 SK37	7,107,326	3,616,903
HR79 SK311	9,161,302	4,711,391
LTU03 SK385	6,587,932	3,348,324
LOW88 SK695.5	7,120,572	3,624,657
Extraction Blank	14	8

6.1 Mitochondrial Haplotype

High-coverage mitochondrial genomes were successfully retrieved from all four individuals (Table 8). Local realignment and quality score recalibration of mapped reads was performed using GATK v3.4-46 (McKenna et al. 2010) and coverage depth estimated with the DepthOfCoverage tool. Coverage plots were generated with BRIG (Alikhan et al. 2011) to visualize coverage across the mitochondrial genome (Figure 2). The authenticity of ancient DNA was explored using fragment length distributions plots and nucleotide misincorporation patterns with mapDamage2.0 (Jónsson et al. 2013) (Figure 3, Figure 4, Figure 5, Figure 6). All four individuals demonstrated typical ancient DNA fragment lengths with modes ranging from 42 bp to 44 bp and the mean ranging

from 52 bp to 62 bp. Deamination profiles also reflected high levels of damage with 21% to 31% of cytosines deaminated on the terminal ends. The nucleotide composition at the genomic coordinate directly preceding sequenced reads showed purine enrichment as described in previous ancient DNA studies (Briggs et al. 2007; Krause et al. 2010; Seguin-Orlando et al. 2013). Taken together, these measures demonstrate a strong authentic ancient DNA signature in the sequenced mitochondrial reads.

SNP and INDEL discovery was conducted using the assembly-based HaplotypeCaller of GATK and performed separately on each individual due to small cohort size. Standard hard filtering parameters were applied to process raw variants according to GATK Best Practices recommendations (Online Documentation: https://www.broadinstitute.org/gatk/guide/best-practices). The filtered variants were then imported into HaploGrep2 (Available Online: https://haplogrep.uibk.ac.at/) (Kloss-Brandstätter et al. 2011; van Oven and Kayser 2009) in VCF format. Haplogroup assignment was selected using the predicted haplogroup with the highest quality (Table 9). The quality measure is based on 1) the presence of expected derived mutations from the reference, 2) additional private mutations, and 3) private mutations which are diagnostic for haplogroups other than the one assigned. As HR79 SK311 demonstrates no observed polymorphisms compared to the revised Cambridge Reference Sequence, HaploGrep2 has no input data for this individual and thus returns a low quality measure. Since this individual's mitochondrial genome is identical to the revised Cambridge Refernce Sequence, HR79 SK311 was therefore classified as the same haplotype H2a2a1.

Table 8: Mitochondria Mapping and Coverage Metrics

Sample ID	**Total Mapped**	**Filtered Reads**	**Average Coverage Depth**
MNL88 SK37	3,031,307	607,155	2315.81
HR79 SK311	915,634	166,464	539.16
LTU03 SK385	1,667,516	239,852	907.07
LOW88 SK695.5	2,387,340	311,938	1225.21
Extraction Blank	0	0	0

Table 9: HaploGrep2 Mitochondrial Haplotypes

Sample ID	**Super**	**Sub**	**Quality**	**HG Quality**	**Sample Quality**
MNL88 SK37	N(R)	V16	0.934	1	0.826
HR79 SK311	N(R0)	H2a2a1	0.5	1	0
LTU03 SK385	N(R0)	HV6	0.95	1	0.899
LOW88 SK695.5	N(R*)	J1b1a1	0.963	1	0.927

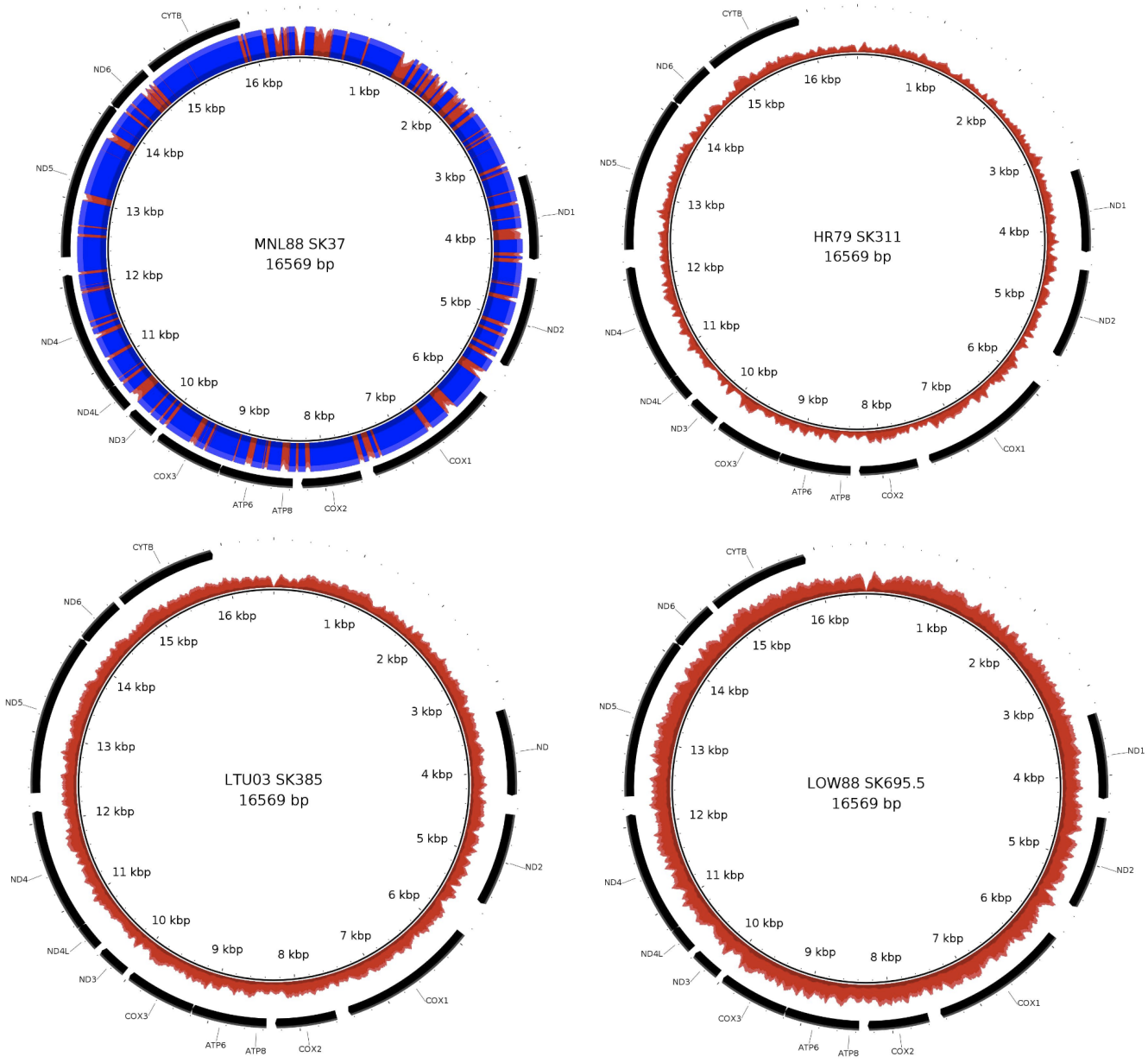

Figure 2: Coverage plots for the mitochondrial genome generated using BRIG (Alikhan et al. 2011). Read coverage in red with regions exceeding the maximum coverage depth (2000) depicted in blue.

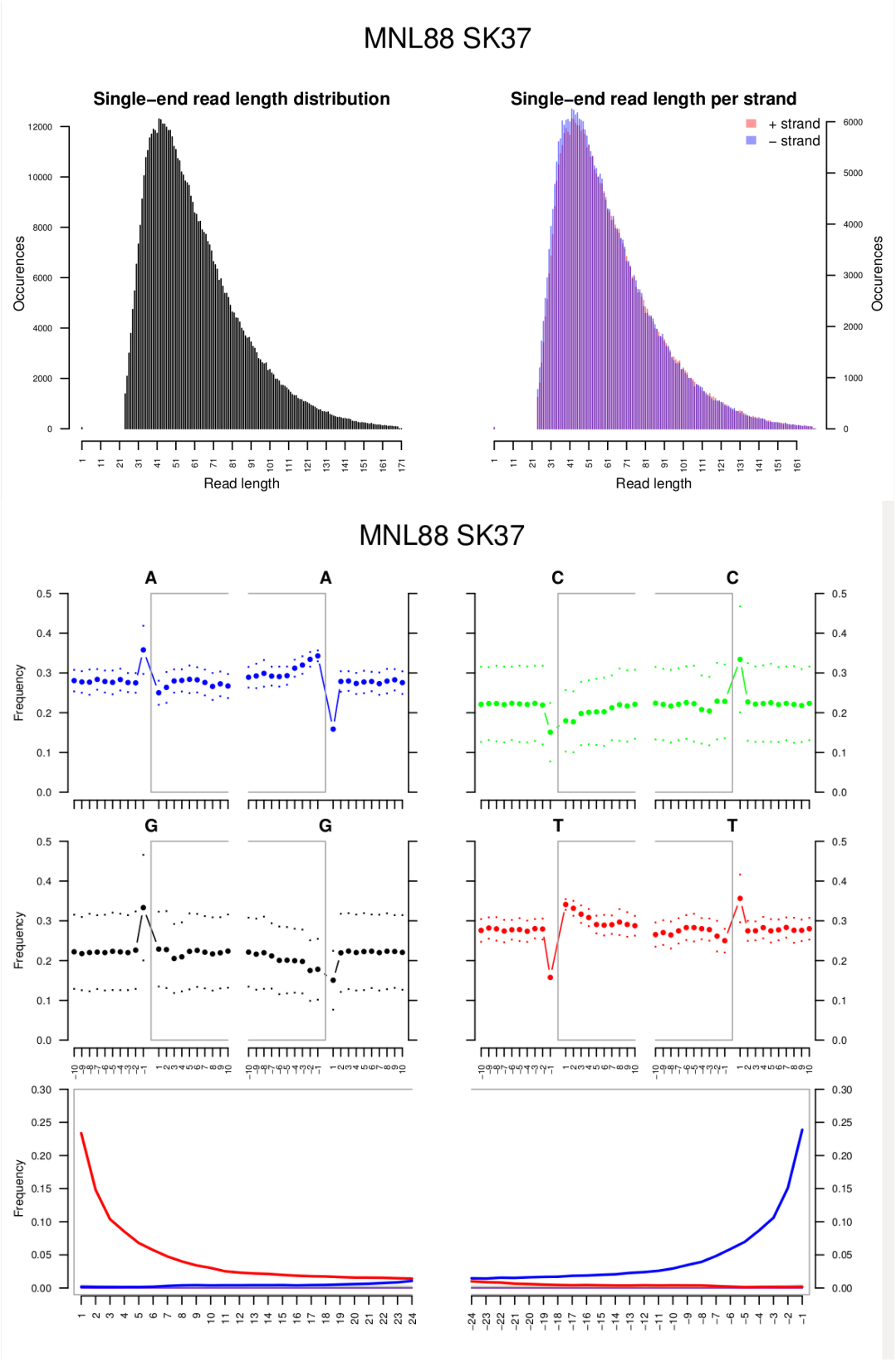

Figure 3: Fragment length distributions and nucleotide misincorporation plots of reads mapping to the mitochondrial genome for MNL88 SK37. Deamination profile: Red: C→T transitions, Blue: G→A transitions, Purple: Others

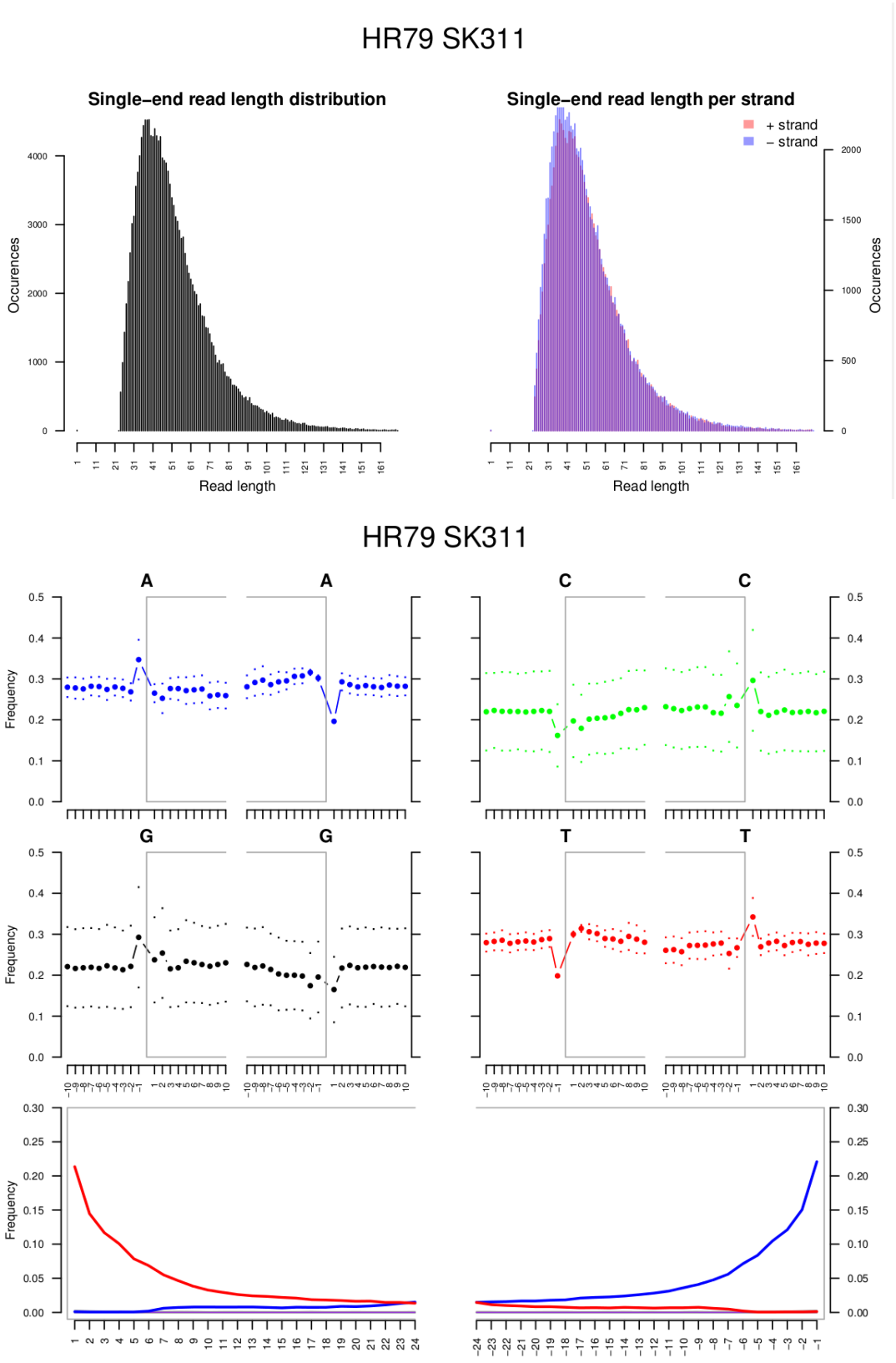

Figure 4: Fragment length distributions and nucleotide misincorporation plots of reads mapping to the mitochondrial genome for HR79 SK311. Deamination profile: Red: C→T transitions, Blue: G→A transitions, Purple: Others

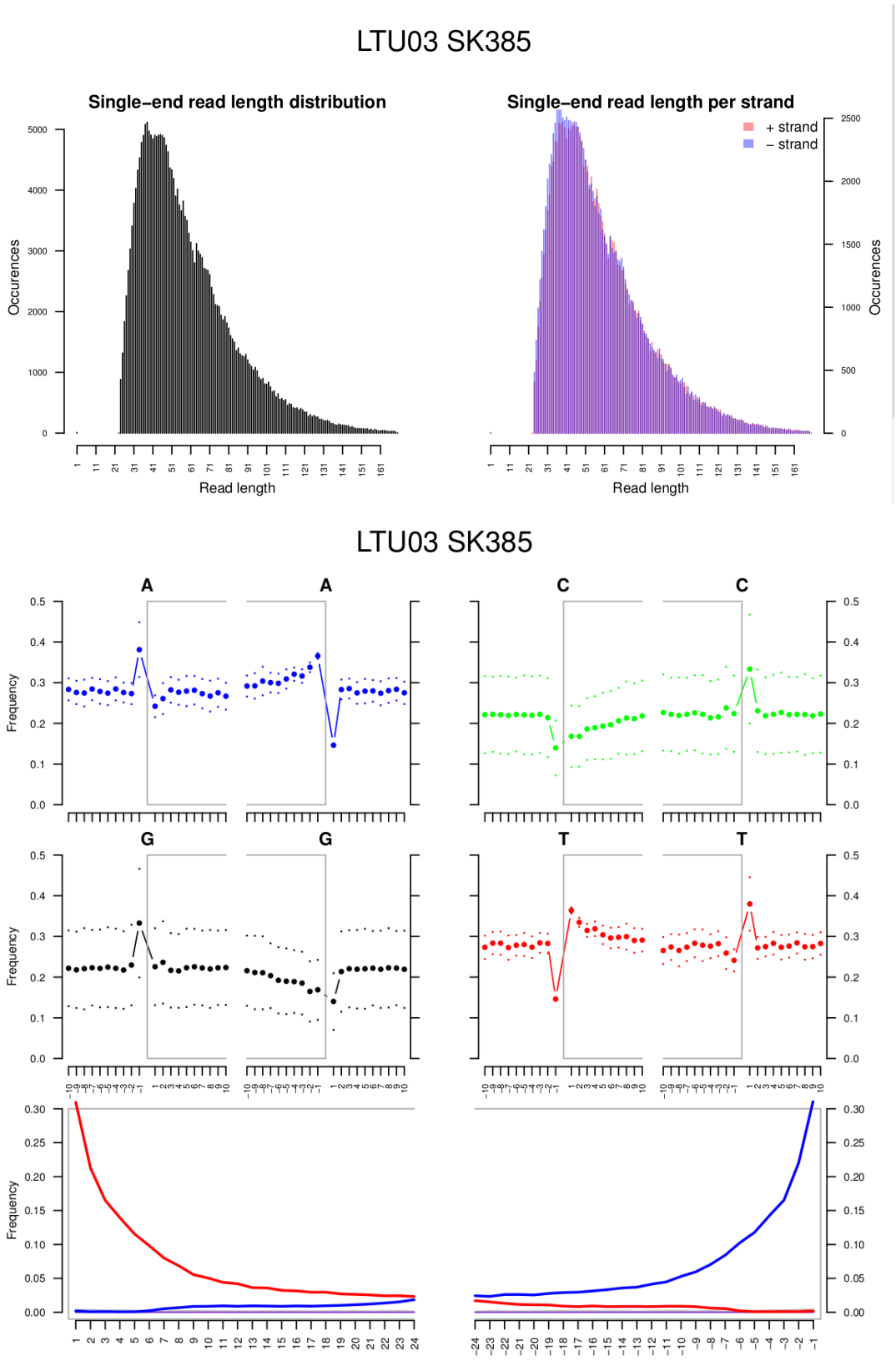

Figure 5: Fragment length distributions and nucleotide misincorporation plots of reads mapping to the mitochondrial genome for LTU03 SK385. Deamination profile: Red: C→T transitions, Blue: G→A transitions, Purple: Others

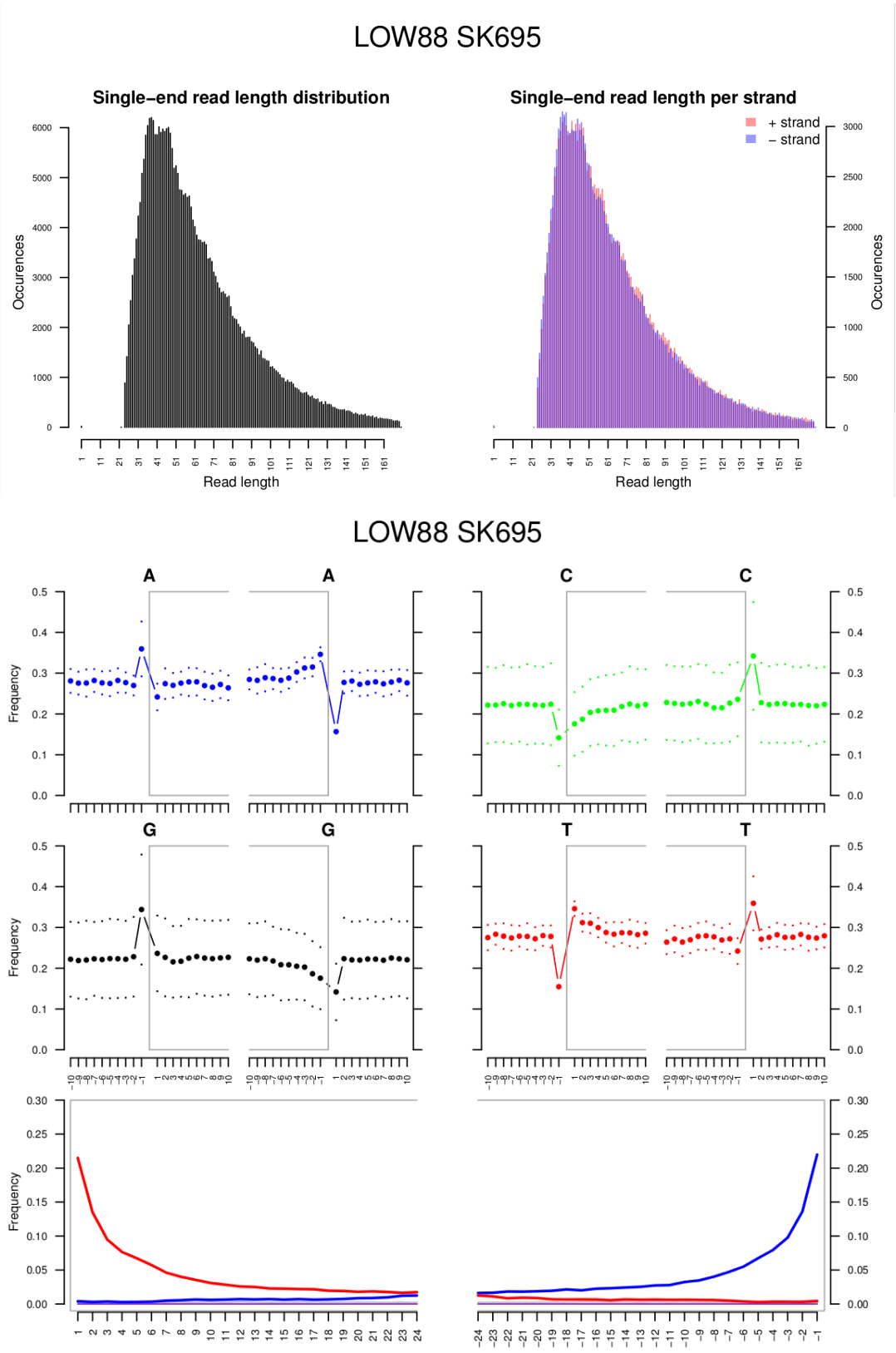

Figure 6: Fragment length distributions and nucleotide misincorporation plots of reads mapping to the mitochondrial genome for LOW88 SK695. Deamination profile: Red: C→T transitions, Blue: G→A transitions, Purple: Others

6.2 Geographic Distribution of Haplogroups

Mitochondrial haplotypes are defined by a series of diagnostic mutations (Phylotree.org) with the distribution of the major branches of the mitochondrial phylogeny closely recapitulating the expansion of modern humans Out of Africa (Figure 7). All four individuals had mitochondrial haplotypes belonging to macrohaplogroup N and within N to the R branch. Descendent lineages of macrohaplogroup N are found through the world, but are most common in Eurasia. The distribution of mitochondrial haplogroups is often difficult to interpret due to limited sampling of populations, incomplete mitochondrial genome sequencing, and the tendencies of many publications to preselect the genomes they sequenced fully based only on hyper variable region haplogroup assignments. Additionally, for the purposes of ancient DNA work, the distribution of mitochondrial haplogroups is known almost exclusively from modern populations and it is difficult to know what distributions may have looked like in the past. Furthermore, as many haplogroups have been defined on the presence of a single sequenced genome, it is often impossible to talk about the distribution of a very specific haplogroup but rather of the greater haplogroup that that lineage belongs to (e.g. haplogroup V rather than haplogroup V16). With those caveats in mind, we confidently say that the mitochondrial lineages of all four individuals are most frequent in modern day Western Eurasia (mitomap.org) but it is very difficult to assign a more specific geographic ancestry with any confidence or scientific rigor. That being said, it is also important to recall that the mitochondrion is passed exclusive from mother to child, and as such represents the origin of only a single female ancestor.

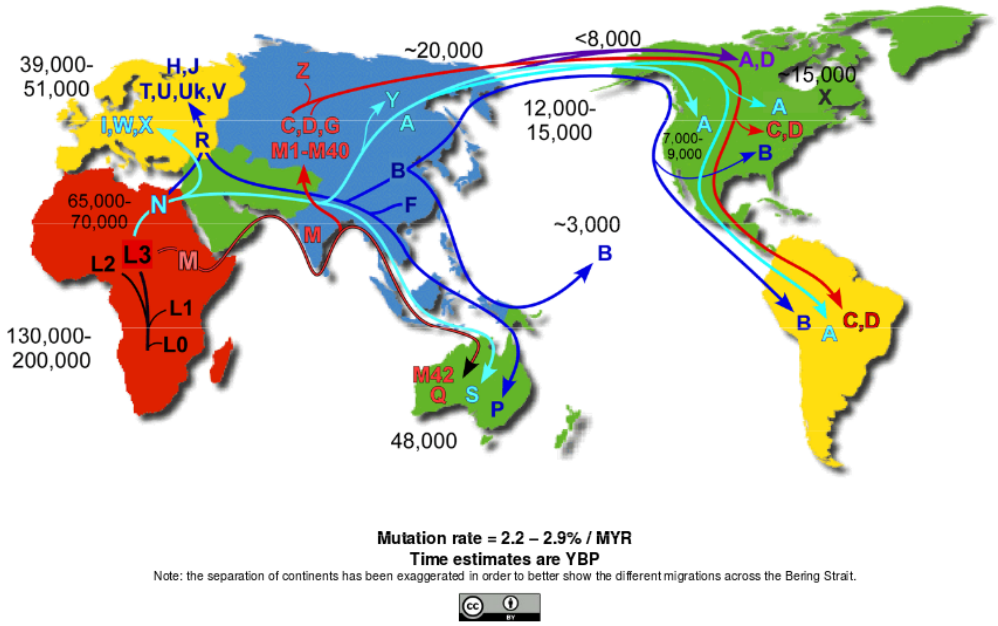

Figure 7: Human mtDNA Migrations (mitomap 2014)

6.3 HIrisPlex Genotype

HIrisplex genotyping was performed similarly to mitochondrial haplotyping (Section 6.1) with several alterations. Genotypes were estimated with the HaplotypeCaller using the "given alleles" genotyping mode rather than variant discovery with GATK v3.4-46 (McKenna et al. 2010). Genotypes were assigned based on the smallest phred-scaled genotype likelihood and were used to predict hair and eye color with the Erasmus MC HIrisPlex Database (Available Online: http://hirisplex.erasmusmc.nl/) (Walsh, Chaitanya, et al. 2014). The most likely phenotype was assigned using the step-wise model (IrisPlex & HirisPlex DNA Phenotyping Webtool User Manual Version 1.0). Mapping metrics for HIrisPlex loci are summarized in (Table 10) and initial predicted phenotypes are described in (Tables 11,12). Additional genotype filtering parameters were then included (FisherStrandBias, ReadPosRankSum, MQRankSum, and MappingQuality) to assess the effect of stringent filtering on all individuals (Tables 11,12). Fragment length distributions and damage profiles were not produced for HIrisPlex reads as these tools do not provide reliable estimates at this coverage depth.

MNL88 SK37 was predicted to most likely have had brown eyes with dark-brown or black hair. HR79 SK311 was initially classified as having brown eyes and black hair, however upon more stringent filtering, could not be firmly classified due to low coverage. LTU03 SK385's phenotype prediction faced similar limitations with an initial prediction of blonde with blue eyes but again was unable to be classified due to low coverage under the new stringent filtering measures we employed. LOW88 SK695.5 was predicted to have brown eyes and dark-brown or black hair.

Table 10: HIrisPlex Mapping Metrics

Sample ID	**HIrisPlex Mapped**	**Filtered Reads**	**Average Coverage Depth**
MNL88 SK37	3700	782	16.875
HR79 SK311	694	65	1.42
LTU03 SK385	712	47	0.61
LOW88 SK695.5	2839	332	8.22
Extraction Blank	0	0	0

Table 11: HIrisPlex Eye Color Predictions

Sample ID	**Eye Color (Initial)**			**Eye Color (Stringent)**			**Average Coverage Depth**
	Blue	Int.	Brown	Blue	Int.	Brown	
MNL88 SK37	0.094	0.148	**0.758**	0.187	0.151	**0.662**	16.875
HR79 SK311	0.066	0.081	**0.853**	NA	NA	NA	1.42
LTU03 SK385	**0.419**	0.371	0.211	NA	NA	NA	0.61
LOW88 SK695.5	0.02	0.05	**0.948**	0.027	0.232	**0.741**	8.22

Table 12: HIrisPlex Hair Color Predictions

Sample ID	**Hair Color (Initial)**				**Hair Color (Stringent)**				**Average Coverage Depth**
	Blonde	Brown	Red	Black	Blonde	Brown	Red	Black	
MNL88 SK37	0.264	0.348	0.001	**0.387**	0.226	**0.548**	0.001	0.225	16.875
HR79 SK311	0.134	0.269	0.000	**0.597**	NA	NA	NA	NA	1.42
LTU03 SK385	**0.532**	0.229	0.000	0.239	NA	NA	NA	NA	0.61
LOW88 SK695.5	0.02	0.252	0.000	**0.728**	0.133	**0.839**	0.029	0.000	8.22

6.4 Sex Estimation

Estimating genetic sex was performed by comparing the number and proportions of reads that aligned to the sex chromosome genes for each individual. Table 13 details the number of filtered reads which mapped to each reference and Table 14 describes the genetic sex estimation for each individual. Fragment length distributions and damage profiles were not produced for the sex gene reads as these tools do not provide reliable estimates at this coverage depth.

Table 13: Sex Chromosome Regions Metrics

Sample ID	Total Mapped	Filtered Reads					Average Coverage Depth
		AMELX	AMELY	SRY	SOX3	TSPY	
MNL88 SK37	3028	187	188	21	106	2546	36.33
HR79 SK311	263	9	12	0	10	232	2.57
LTU03 SK385	34	13	1	0	15	5	0.29
LOW88 SK695.5	1423	66	57	9	69	1222	17.19
Extraction Blank	0	0	0	0	0	0	0

MNL88 SK37 had a near equal number of reads mapping to AMELX and AMELY, 187 and 188 respectively, and a 10-fold increase in reads mapping to TSPY as would be expected due to the high-copy number of TSPY when the Y chromosome is present. The ratio of reads mapping to SRY and SOX3, 21 and 106 respectively, was skewed as SRY is under-represented in the bait set due to its short length (See Table 6). SRY is only represented by 266 nucleotides of probes, whereas its homologue is represented by 1479 nucleotides. Proportionally, we would therefore expect a ratio of approximately 5.56:1 SOX3 reads to SRY reads. This ratio is closely matched in MNL88 with 106 reads mapping to SOX3 and 21 reads mapping to SRY, with a ratio of 5.05:1. The presence of all three Y chromosome genes and a near 1:1 ratio of AMELX/AMELY reads provides strong evidence for the presence of both the X and Y chromosome.

HR79 SK311 had fewer reads mapping to the target sex genes as indicated by an average coverage depth of 2.57. The ratio of reads mapping to the AMELX and AMELY regions was 1:1.33. HR79 SK311 demonstrated a 20-fold increase in reads mapping to TSPY compared to the single-copy AMEL genes. Due to the relatively lower coverage of this individual, all AMELX and AMELY were BLAST-validated against the nt database (Altschul et al. 1990) to detect non-specific mapping. All reads aligned solely to their original mapped gene, suggesting the authentic presence of both the

X and the Y. SOX3 is present as indicated by 10 reads but no reads mapped to the SRY regions. As mentioned previously, the estimated proportion of SOX3 to SRY reads should follow a ratio of 5.56:1, resulting in an expected number of SRY reads to be 1.7 for this individual. The inability to detect SRY in HR79 SK311 is therefore more likely due to reduced coverage rather than the complete absence of the Y chromosome, given the presence of both AMELY and TSPY.

LTU03 SK385 was similarly limited by low coverage with only 34 reads mapping to all regions with an average coverage depth of 0.29. The ratio of AMELX and AMELY reads was strongly skewed to 13:1 and TSPY was only represented by 5 reads for this individual. Compared to the three other individuals, the lack of reads mapping to TSPY was drastically reduced, suggesting the potential absence of the Y chromosome. All reads for this individual were BLAST-validated and the results indicate that the 1 read mapping to AMELY shares sequence homology with AMELX, and only 1 exceptionally long read (131 bp) was specific to TSPY. Given the unlikelihood of observing such a long molecule in this low-coverage individual, only the X chromosome is rigorously represented in this individual.

LOW88 SK695.5's sex estimation benefits from higher average coverage (17.19) compared to HR79 SK311 and LTU03 SK385. The ratio of reads mapping to the AMELX and AMELY regions was 1.16 and the increase in TSPY reads compared to the single-copy AMEL genes was approximately 20-fold. SOX3 and SRY were both detected in this individual with 69 and 9 reads mapping to them respectively with a ratio of 7.67:1. The presence of AMELY and SRY and the expected fold-increase of the multi-copy TSPY suggests that both the X and Y chromosome are present.

Table 14: Genetic Sex Estimation

Sample ID	X	Y
MNL88 SK37	Present	Present
HR79 SK311	Present	Present
LTU03 SK385	Present	Absent
LOW88 SK695.5	Present	Present

7 Summary

The results of the genetic biography are summarized in Table 15. High coverage mitochondrial genomes were obtained from all individuals and corresponded to four different haplotypes predominantly found in Western Eurasia. Hair and eye color prediction could be robustly estimated from two individuals, MNL88 SK37 and LOW88 SK695.5, while HR79 SK311 and LTU03 were limited by low coverage under the new stringent filtering we used. Genetic sex was estimated for all four individuals based on the detection of X and Y-specific genes.

Table 15: Genetic Biography Summary

Sample ID	**Mitochondrial Haplotype**	**Eye Color**	**Hair Color**	**Genetic Sex**
MNL88 SK37	V16	Brown	Dark-Brown or Black	Male
HR79 SK311	H2a2a1	NA	NA	Male
LTU03 SK385	HV6	NA	NA	Female
LOW88 SK695.5	J1b1a1	Brown	Dark-Brown or Black	Male

References

Alikhan, N.-F., Petty, N. K., Zakour, N. L. B., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, *12*, 402.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Benoit, J.-N., Quatrehomme, G., Carle, G. F., & Pognonec, P. (2013). An alternative procedure for extraction of DNA from ancient and weathered bone fragments. *Medicine, Science and the Law*, *53*(2), 100–106. doi:10.1258/msl.2012.012026

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., ... Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, *104*(37), 14616–14621. doi:10.1073/pnas.0704665104. eprint: http://www.pnas.org/content/104/37/14616.full.pdf

Butler, E. & Li, R. (2014). Genetic markers for sex identification in forensic DNA analysis. *Journal of Forensic Investigation*, *2*(3), 10–19.

Campos, E. A., Pitta, D. R., Costa, F. A., Campos, V. M., Yela, D., & Fernandes, A. (2014). DNA extraction from filter-paper spots of vaginal samples collected after sexual violence. *International Journal of Gynecology & Obstetrics*, *126*(1), 23–27. doi:http://dx.doi.org/10.1016/j.ijgo.2014.02.010

dbSNP Build ID: 142. (2014). Database of single nucleotide polymorphisms (dbSNP). Available from: http://www.ncbi.nlm.nih.gov/SNP/. Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Münsterberg, A., ... Lovell-Badge, R. (1990). A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature*, *346*, 245–250.

Jacot, T. A., Zalenskaya, I., Mauck, C., Archer, D. F., & Doncel, G. F. (2013). TSPY4 is a novel sperm-specific biomarker of semen exposure in human cervicovaginal fluids; potential use in {HIV} prevention and contraception studies. *Contraception*, *88*(3), 387–395. doi:http://dx.doi.org/10.1016/j.contraception.2012.11.022

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., & Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, *29*(13), 1682–1684. doi:10.1093/bioinformatics/btt193. eprint: http://bioinformatics.oxfordjournals.org/content/29/13/1682.full.pdf+html

Kamodyová, N., Durdiaková, J., Celec, P., Sedláčková, T., Repiská, G., Sviežená, B., & Minárik, G. (2013). Prevalence and persistence of male DNA identified in mixed saliva samples after intense kissing. *Forensic Science International: Genetics*, *7*(1), 124–128. doi:http://dx.doi.org/10.1016/j.fsigen.2012.07.007

Katoh K, M. T. (1999). A heuristic approach of maximum likelihood method for inferring phylogenetic tree and an application to the mammalian SOX-3 origin of the testis-determining gene SRY. *FEBS Letters*, *463*, 129–132.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649.

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the illumina platform. *Nucleic Acids Research*, *40*(1), e3.

Kloss-Brandstätter, A., Pacher, D., Schönherr, S., Weissensteiner, H., Binna, R., Specht, G., & Kronenberg, F. (2011). HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Human Mutation*, *32*(1), 25–32. doi:10.1002/humu.21382

Krause, J., Good, J. M., Viola, B., Shunkov, M. V., Derevianko, A. P., & Paabo, S. (2010). The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, *464*, 894–897. doi:10.1038/nature08976

Krausz, C., Giachini, C., & Forti, G. (2010). TSPY and male fertility. *Genes*, *1*(2), 308–316.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi:10.1093/bioinformatics/btp324. eprint: http://bioinformatics.oxfordjournals.org/content/25/14/1754.full.pdf+html

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303.

Meyer, M. & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *6*. doi:10.1101/pdb.prot5448

mitomap. (2014). Retrieved February 24, 2014, from http://www.mitomap.org/pub/MITOMAP/MitomapFigures/WorldMigrations2013.pdf

Nakahori, Y., Hamano, K., Iwaya, M., & Nakagome, Y. (1991). Sex identification by polymerase chain reaction using X-Y homologous primer. *American Journal of Medical Genetics*, *39*(4), 472–473.

Nakahori, Y., Takenaka, O., & Nakagome, Y. (1991). A human X-Y homologous region encodes "amelogenin". *Genomics*, *9*, 264–269.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., ... Ostell, J. M. (2014). Refseq: an update on mammalian reference sequences. *Nucleic Acids Research*, *42*(D1), D756–D763. doi:10.1093/nar/gkt1114. eprint: http://nar.oxfordjournals.org/content/42/D1/D756.full.pdf+html

Renaud, G., Stenzel, U., & Kelso, J. (2014). leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Research*. doi:10.1093/nar/gku699. eprint: http://nar.oxfordjournals.org/content/early/2014/08/05/nar.gku699.full.pdf+html

Schwarz, C., Debruyne, R., Kuch, M., McNally, E., Schwarcz, H., Aubrey, A., ... Poinar, H. (2009). New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. *Nucleic Acids Research*, *37*(10), 3215–3229.

Seguin-Orlando, A., Schubert, M., Clary, J., Stagegaard, J., Alberdi, M. T., Prado, J. L., ... Orlando, L. (2013, October). Ligation Bias in Illumina Next-Generation DNA Libraries: Implications for Sequencing Ancient Genomes. *PLoS ONE*, *8*(10), e78575. doi:10.1371/journal.pone.0078575

Smit, A. F. A., Hubley, R., & Green, P. (2013-2015). Repeatmasker open-4.0. Available at: http://www.repeatmasker.org.

van Oven, M. & Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, *30*(2), E386–E394. doi:10.1002/humu.20921

Walsh, S., Chaitanya, L., Clarisse, L., Wirken, L., Draus-Barini, J., Kovatsi, L., ... Kayser, M. (2014). Developmental validation of the {HIrisPlex} system: DNA-based eye and hair colour prediction for forensic and anthropological usage. *Forensic Science International: Genetics*, *9*, 150–161. doi:http://dx.doi.org/10.1016/j.fsigen.2013.12.006

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., ... Kayser, M. (2013). The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Science International: Genetics*, *7*(1), 98–115. doi:http://dx.doi.org/10.1016/j.fsigen.2012.07.005

Yasmin, L., Takano, J.-i., & Sankai, T. (2015). Effective use of the TSPY gene-specific copy number in determining fetal DNA in the maternal blood of cynomolgus monkeys. *Animal Science Journal*, n/a–n/a. doi:10.1111/asj.12523