

MASSIVELY PARALLEL SEQUENCING OF FORENSIC MARKERS: SEQUENCE VARIATION AND FORENSIC APPLICATION

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

Tunde Ildiko Huszar

Department of Genetics and Genome Biology
College of Life Sciences, University of Leicester

September 2019



Abstract

MASSIVELY PARALLEL SEQUENCING OF FORENSIC MARKERS: SEQUENCE VARIATION AND FORENSIC APPLICATION

Tunde Ildiko Huszar

DNA analyses have been used to aid forensic investigations since 1985 and to date, human identification relies on the typing of polymorphic autosomal short tandem repeats (STRs), supplemented by the use of paternal (Y-chromosomal STRs) and maternal (mitochondrial DNA; mtDNA) lineage markers. To accurately measure the significance of matching genotypes/haplotypes their frequencies in relevant populations must be estimated.

Massively parallel sequencing (MPS) technologies became more cost-effective and therefore more accessible to forensic analysis in the last decade. Many markers and marker types can be simultaneously analysed to observe underlying sequence variants beyond simple length-variation. The distribution of sequence-level variants within and between populations also helps to illuminate population substructure.

In this study, STRs and mtDNA were analysed using MPS first in a diverse global set of samples to establish a framework of variants beyond any single population, followed by the population analysis of a set of 362 samples from the People of the British Isles (PoBI) collection, chosen to represent the core population with minimal admixture in the last two generations. Available data from PoBI also allowed Y-SNP-defined haplogroups to be studied.

Using MPS in these sample sets allowed the description of new STR variants, both within the sequence structure of repeat arrays and their flanking nucleotides. Analysis of the mtDNA control region following PCR with a multi-primer system revealed the importance of accounting for reference sequence bias in data analysis.

The PoBI sample provided an overview of sequence-level variants of forensic markers across The British Isles, and serves as a reference for future MPS-based forensic applications. While mtDNA and autosomal markers show no appreciable population substructure, Y-chromosomal variation revealed clear East-West differentiation, probably reflecting a male-mediated Anglo-Saxon cultural and demographic shift around 500 CE, and correlating with current Celtic-Germanic linguistic divisions.

Work described in Chapter 3 of this thesis has been published as:

Huszar, T.I., Jobling, M.A. and Wetton, J.H. 2018.

A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing.

Forensic Sci Int Genet, 35:97-106.

doi: [10.1016/j.fsigen.2018.03.012](https://doi.org/10.1016/j.fsigen.2018.03.012)

Work described in Chapter 5 of this thesis has been published as:

Huszar, T.I., Wetton, J.H. and Jobling, M.A. 2019.

Mitigating the effects of reference sequence bias in single-multiplex massively parallel sequencing of the mitochondrial DNA control region.

Forensic Sci Int Genet, 40:9-17.

doi: [10.1016/j.fsigen.2019.01.008](https://doi.org/10.1016/j.fsigen.2019.01.008)

Copies of both of these papers are included in the electronic appendix.

This research project was supervised by

Prof Mark A Jobling, Dr Jon H Wetton and Dr Celia A May.

Acknowledgements

First and foremost, I am sincerely grateful to my supervisors Prof Mark Jobling and Dr Jon Wetton for their continuous support, balanced with healthy challenges and constructive criticisms. They led me by example, both in science and life and made working on this project truly enjoyable. Also would like to express my appreciation to Dr Celia May and Dr Sandra Beleza for their humanity and support; they are truly inspiring women in science. I'm very grateful to Prof Denise Syndercombe Court and Dr Sandra Beleza for taking the time to thoroughly examine this thesis and providing helpful comments.

I would like to thank Dr Chiara Batini and Dr Pille Hallast for infecting me with the love of bioinformatics, sharing their experiences and encouraging me to find my own solutions. I thank Gurdeep Matharu Lall for her technical support in the lab and for her friendship through the years. I would also like to thank all the past and present lab members for their friendships, for all the interesting discussions on all subjects, not just science, and for creating a motivating environment: Jordan, Yahya, Nitikorn, Sam, Marwan, Margherita, Ettore, Orie, Jodie and others.

I thank Reshma Vaghela and Dr Nic Sylvius for sharing their knowledge on sequencing, and for trusting me fully with the MiSeq. I would like to thank Kasia Hutnik and Prof Sir Walter Bodmer for generously providing assistance and access to a subset of the PoBI samples. I am grateful for Promega's support of my project, especially to Dr Andy Hopwood and Dr Nikki Peake. My research was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) iCASE studentship BB/M016706/1 with Key Forensic Services as an industrial partner; I thank Paul Hackett of KFS for agreeing to the partnership and thus making the work described here possible.

Last, but not least, I am also grateful to those without whom I would not ended up here, where I am: Dr Klára Csete, Emese Lauly and Mária Déri; Dr Caro Head and Dr Mike Rogers; my family and friends.

Table of Contents

CHAPTER 1: Introduction	1
1.1 Forensic genetic markers.....	1
1.1.1 Short tandem repeats (STRs).....	3
1.1.1.1 The origin of STRs	3
1.1.1.2 Structure of STRs.....	4
1.1.1.3 Conventional amplification and detection of STRs	5
1.1.1.4 Typing STRs using MPS	10
1.1.2 Single nucleotide polymorphisms (SNPs) and short indels	12
1.1.2.1 Forensic use of SNPs	12
1.1.2.2 SNPs compared to STRs	13
1.1.2.3 Typing SNPs	14
1.1.2.4 Using SNP data and databasing	15
1.1.2.5 Mitochondrial DNA typing.....	16
1.1.2.6 Short insertions or deletions.....	19
1.1.2.7 Short indels around STRs	20
1.2 Massively parallel sequencing and its use in forensic genetics and genomics	21
1.2.1 Massively parallel sequencing	21
1.2.1.1 Massively parallel sequencing technologies	21
1.2.1.2 Massively parallel sequencing process	23
1.2.2 Application of MPS in forensic genetics and genomics	27
1.2.2.1 Advantages of MPS for forensic use	27
1.2.2.2 Improving STR typing.....	29
1.2.2.3 Forensic use of MPS beyond STRs	31
1.2.2.4 MPS options and challenges of implementation	31
1.2.2.5 Latest developments in MPS and its future use in the UK	32
1.3 Aims and objectives	33

CHAPTER 2: Materials and methods.....	35
2.1 Sample sets	35
2.1.1 Samples of global variation.....	35
2.1.2 Samples from the People of the British Isles.....	38
2.2 Laboratory experiments	45
2.2.1 Quantification of DNA	45
2.2.2 Sample preparation specific to Promega PowerSeq™ Auto/Mito/Y System prototype	45
2.2.2.1 Fragment generation by PCR.....	45
2.2.2.2 Purification of PCR products	48
2.2.2.3 Quantification of PCR products	48
2.2.2.4 Library preparation	49
2.2.3 Sample preparation specific to Promega PowerSeq™ CRM Nested System, Custom.....	53
2.2.3.1 Fragment generation and library preparation by PCR	53
2.2.3.2 Purification of amplified libraries.....	54
2.2.3.3 Quantification of amplified libraries.....	54
2.2.4 Sequencing pooled libraries on the MiSeq®	54
2.3 Data analysis	55
2.3.1 Output of sequencing runs.....	55
2.3.2 Quality check and control of raw FASTQ files.....	55
2.3.2.1 Trimming of reads	56
2.3.2.2 Optional error correction.....	56
2.3.2.3 Quality controlled FASTQ files	57
2.3.3 Analysis of mtDNA sequence data	57
2.3.3.1 Variant calling pipeline	57
2.3.3.2 Detection of nuclear mtDNA insertions (numts).....	57
2.3.3.3 Detection and quantification of indels in mtDNA.....	58
2.3.3.4 Overarching Read Enrichment Option (OREO)	58
2.3.3.5 Additional data analysis steps relating to the PowerSeq™ CRM Nested System.....	60
2.3.3.6 Other options for mtDNA analysis or primer removal.....	61

2.3.3.7 Haplogroup assignments and phylogenetic relationships of samples	61
2.3.4 Analysis of STR sequence data.....	62
2.3.4.1 FSTools.....	62
2.3.4.2 STRait Razor	63
2.3.4.3 Command-line sequence data interrogation and standard variant calling	63
2.3.4.4 Relative read-depth ratio test for duplicated Y-STR alleles.....	64
2.3.4.5 Y-STR haplogroup assignments and phylogenetic relationships of the samples	66
2.3.5 Population genetic data analysis	66
2.3.6 Datasets used for concordance and validation of the results	67
2.3.6.1 Datasets for comparison with the global variation dataset.....	67
2.3.6.2 Datasets for comparison with the PoBI sample set.....	68
2.4 Data visualisation.....	68
 CHAPTER 3: Global Y-STR sequence variation in a phylogenetic context	70
 3.1 Introduction	70
3.2 Materials and methods.....	72
3.2.1 DNA samples	72
3.2.2 DNA quantitation and PCR amplification	72
3.2.3 Library preparation and sequencing	72
3.2.4 Data processing and analyses.....	73
3.3 Results.....	74
3.3.1 Phylogenetic diversity of the samples	74
3.3.2 Coverage ranges observed	77
3.3.3 Identified Y-STR alleles	78
3.3.4 Concordance of MPS data with CE-defined alleles.....	88
3.3.5 Diversity of observed alleles	90
3.3.6 Isometric alleles.....	98
3.3.7 Compiled variants of the dataset	99
3.3.8 Novel variants with implications for nomenclature.....	99
3.3.8.1 Nomenclature of DYS385a,b.....	100

3.3.8.2 Nomenclature of DYS481.....	102
3.3.8.3 Nomenclature of DYS390.....	102
3.3.8.4 Consideration for nomenclature detailing flanking region variants	102
3.3.9 Phylogenetic association of variants	103
3.3.9.1 Phylogenetic association of SNPs/indels.....	104
3.3.9.2 Phylogenetic association of RPs.....	104
3.4 Discussion	108
3.4.1 Highlights and Conclusions	111
 CHAPTER 4: Autosomal STR sequence variation	112
4.1 Introduction	112
4.2 Materials and methods.....	112
4.2.1 DNA samples and laboratory experiments.....	112
4.2.2 Targeted autosomal STRs.....	112
4.2.3. Data processing and analyses.....	113
4.3 Results.....	114
4.3.1 Coverage ranges observed	114
4.3.2 Identified Amelogenin and aSTR alleles	115
4.3.3 Diversity of observed alleles	125
4.3.4. Isometric heterozygote alleles and observed heterozygosity	127
4.3.5 Heterozygote balance.....	130
4.3.6 Variants that are novel to the STRSeq database	133
4.4 Discussion	135
4.4.1 Highlights and Conclusions	137
 CHAPTER 5: mtDNA analysis using an MPS workflow.....	138
5.1 Introduction	138
5.2 Materials and Methods.....	142
5.2.1 DNA samples	142
5.2.2 DNA quantitation	142

5.2.3 PCR amplification, library preparation and sequencing.....	142
5.2.4 Data processing and analyses.....	143
5.3 Results.....	145
5.3.1 Coverage ranges observed	147
5.3.2 Calling variants in the control region.....	154
5.3.2.1 Considering primer sequences.....	154
5.3.2.2 An alternative data processing tool to primer removal	155
5.3.3 Validation of PowerSeq™ MPS mtDNA data	156
5.3.4 Performance of mtDNA amplification across the phylogeny.....	157
5.3.5 Detection and quantitation of heteroplasmy	158
5.3.6 Improved heteroplasmy detection with the CRM Nested System.....	163
5.3.7 Detection of numt sequences	177
5.4 Discussion	180
5.4.1 Highlights and Conclusions	184
 CHAPTER 6: Massively parallel sequencing of forensic markers in the People of the British Isles	 185
6.1 Introduction	185
6.1.1 History of the British Isles	185
6.1.2 Summary of earlier studies of genetic diversity in the British Isles	189
6.1.3 The PoBI project.....	189
6.1.4 Other studies from the British Isles	193
6.1.5 Current ethnic composition of the British Isles	194
6.1.6 Terminology of the British Isles.....	197
6.2 Materials and Methods.....	198
6.2.1 DNA samples	198
6.2.2 Options for dividing samples.....	198
6.2.2.1 Thirty-seven regions.....	199
6.2.2.2 Ten large geographic, administrative regions	200
6.2.2.3 Divisions following the five main PoBI clusters	202
6.2.2.4 Division following regions with Celtic languages.....	203
6.2.2.5 Division following the Danelaw region	204

6.2.2.6 Division congruent with the PoBI fineSTRUCTURE-based Central/South England cluster	204
6.2.2.7 Division representing the East.....	205
6.2.3 DNA quantitation and PCR amplification	206
6.2.4 Library preparation and sequencing	206
6.2.5 Data processing and analyses.....	206
6.3 Results.....	207
6.3.1 Run statistics and coverage values	207
6.3.1.1 Run statistics.....	207
6.3.1.2 Coverage values	208
6.3.2 MtDNA control region population data in the UK.....	212
6.3.2.1 Observed variants and haplotypes	212
6.3.2.2 Median-joining (MJ) networks of mtDNA CR variants.....	216
6.3.2.3 Testing population differentiation using mtDNA CR variants	220
6.3.3 Y-STR sequence variation population data in the UK	222
6.3.3.1 Observed variants and haplotypes	222
6.3.3.2 Median-joining (MJ) networks of Y-STR variants.....	229
6.3.3.3 Testing population differentiation using Y-STR variants	245
6.3.3.4 Summary of Y-STR forensic statistics	258
6.3.4 Population data of aSTR sequence variation in the UK	259
6.3.4.1 Observed variants	259
6.3.4.2 Testing population differentiation using aSTR variants.....	274
6.3.4.3 Summary of aSTR forensic statistics.....	276
6.4 Discussion	277
6.4.1 Highlights and conclusions	279
CHAPTER 7: Discussion and future directions	280
7.1 MPS technology in forensic DNA analysis	280
7.2 Summary of the results	280
7.3 Considerations relating to MPS and future directions.....	282
7.3.1 Concordance between CE and MPS	282
7.3.2 Nomenclature adjustments to MPS	283

7.3.3 Advantages and disadvantages of MPS over CE.....	284
7.3.4 Future directions for MPS and beyond.....	285

The list of tables

CHAPTER 1

Table 1.1 Summary of standard autosomal forensic markers used globally.

CHAPTER 2

Table 2.1 Selection of the 100 global samples.

Table 2.2 Samples selected from the PoBI study.

Table 2.3 Details of the STR loci and the mtDNA regions amplified.

Table 2.4 Repeat motifs and genomic positions of the nuclear genomic loci amplified.

Table 2.5 Thermocycling parameters used with the Promega PowerSeq™ Auto/Mito/Y System prototype multiplex.

Table 2.6 Thermocycling parameters used with the Promega PowerSeq™ CRM Nested System, Custom multiplex.

CHAPTER 3

Table 3.1 Haplogroup information for the selected 100 samples.

Table 3.2 Sample coverage statistics.

Table 3.3 Example of visual summary of the structures of alleles.

Table 3.4 Allele ranges and count of variants observed for all 23 Y-STRs.

Table 3.5 Supernumerary alleles.

Table 3.6 Details of discordant alleles.

Table 3.7 Increase of allele diversity due to sequencing.

Table 3.8 Summary of isometric allele groups.

Table 3.9 List of novel Y-STR sequence variants defined by MPS.

Table 3.10 List of Y-STR markers with observed SNP or indel variants.

Table 3.11 Statistics relating to observed SNP or indel variants in Y-STR sequences.

Table 3.12 Summary of MPS sequence variants with variable length flanking region alleles.

CHAPTER 4

Table 4.1 Sample coverage statistics.

Table 4.2 Amelogenin and autosomal STR sequence alleles.

Table 4.3 Autosomal STRs with isometric heterozygote allele pairs.

Table 4.4 Autosomal STR loci in order of maximum increase in observed heterozygosity.

Table 4.5 50 sequence-based alleles identified as 'novel to STRSeq'.

CHAPTER 5

Table 5.1 Mean coverage statistics for each overlapping and non-overlapping segments of the amplicons.

Table 5.2 Substitution error rates calculated for each mtDNA kit type plus the Amelogenin marker for comparison.

Table 5.3 Testing thresholds for heteroplasmy calling.

Table 5.4 Discrepancies between the called homoplasmic substitutions and the reference datasets.

Table 5.5 Effects of data processing steps on 47 heteroplasmic variants.

Table 5.6 Effects of data processing steps on 102 heteroplasmic variants.

CHAPTER 6

Table 6.1 Self-reported ethnicity components of the British Isles.

Table 6.2 44 sampling regions grouped into 37 balanced composite regions.

Table 6.3 44 sampling regions assigned to ten large geographic and administrative regions.

Table 6.4 Run statistics for the sequencing runs.

Table 6.5 Coverage values for each marker type.

Table 6.6 Discordant haplogroup predictions.

Table 6.7	Frequencies of predicted haplogroups by MPS targeting the CR for the ten large geographical regions.
Table 6.8	MtDNA CR variants molecular statistics, Analysis of MOlecular VAriance (AMOVA) statistics and pairwise F_{ST} values.
Table 6.9	Length-based Y-STR allele frequencies from 362 samples.
Table 6.10	Y-STR allele duplications observed in 8 samples.
Table 6.11	Example of Y-STR sequence allele variants for the ten large geographical regions.
Table 6.12	Frequencies of predicted Y-haplogroups by MPS for the ten large geographical regions.
Table 6.13	Y-STR length and sequence variants compared using Analysis of MOlecular VAriance (AMOVA) statistics and pairwise R_{ST} values.
Table 6.14	Standard Y-STR loci forensic parameters.
Table 6.15	Length-based aSTR allele frequencies from 361 samples.
Table 6.16	Sequence-based aSTR allele frequencies from 361 samples.
Table 6.17	AMOVA statistics and pairwise R_{ST} values for aSTR variants.
Table 6.18	Standard aSTR forensic parameters.

The list of figures

CHAPTER 1

- Figure 1.1 An early DNA fingerprint.
- Figure 1.2 A schematic representation of an STR locus.
- Figure 1.3 Example of an electropherogram (EPG).
- Figure 1.4 Stutter peaks.
- Figure 1.5 The human mtDNA.
- Figure 1.6 Illumina®'s sequencing-by-synthesis method.
- Figure 1.7 Run analytics.
- Figure 1.8 Forensic markers.
- Figure 1.9 Effects of indels in an STR locus.

CHAPTER 2

- Figure 2.1 Geographical distribution of the 100 samples.
- Figure 2.2 Geographical distribution of the 362 PoBI samples.
- Figure 2.3 Dual or single index adapters.
- Figure 2.4 Differences between ligated and PCR-incorporated dual adapters and sequencing of the generated libraries.
- Figure 2.5 OREO selects reads containing variants intrinsic to the mtDNA.
- Figure 2.6 Relative read-depth ratio test.

CHAPTER 3

- Figure 3.1 Phylogenetic relationships of the analysed samples.
- Figure 3.2 Examples of allele duplications and somatic mutations.
- Figure 3.3 Length-based Y-STR allele phylogeny.
- Figure 3.4 Allele diversity increase by type of variants.
- Figure 3.5 Observed SNPs and indels in their phylogenetic context.
- Figure 3.6 Examples of observed RPVs in their phylogenetic contexts.

CHAPTER 4

- Figure 4.1 Increase of allele diversity by type of variants.
- Figure 4.2 Heterozygote balance in the 100 isometric heterozygote pairs across 18 loci.
- Figure 4.3 Heterozygote balance relative to allele spans across 22 aSTRs.
- Figure 4.4 Example of a decline of heterozygote balance in an STR locus over larger allele spans.

CHAPTER 5

- Figure 5.1 Schematic representation of a multiplex assay design over the mtDNA control region.
- Figure 5.2 Maximum Likelihood phylogenetic tree of the mtDNA control region sequences obtained from 101 samples.
- Figure 5.3 Non-uniform coverage of the multiplex assay design over the control region.
- Figure 5.4 Finding a compromise for heteroplasmy reporting thresholds.
- Figure 5.5 Distribution of apparent heteroplasmy across the control region.
- Figure 5.6 Overlap extension of PCR products.
- Figure 5.7 The Nested CRM design prevents extension of overlapping single-stranded PCR products and internalisation of primers.
- Figure 5.8 Read-length profiles compared for the prototype and the CRM kit.
- Figure 5.9 Contribution of short amplicons to non-uniform coverage over the control region.
- Figure 5.10 Distribution of confirmed heteroplasmy across the control region.
- Figure 5.11 Example of the effects of kit design and data processing in mitigating reference sequence bias.
- Figure 5.12 Low coverage of amplicon #3 and detection of the numt sequence.

CHAPTER 6

- Figure 6.1 Major events in the peopling of the British Isles (from Leslie et al. 2015).
- Figure 6.2 Autosomal-based fineSTRUCTURE clusters detected in the PoBI study (from Leslie et al. 2015).
- Figure 6.3 The main census-based ethnicity components of the British Isles.
- Figure 6.4 Terminology of the British Isles.
- Figure 6.5 The locations of the 44 mostly county-based sampling regions from The British Isles.
- Figure 6.6 The ten large geographic and administrative regions.
- Figure 6.7 Division by the five main autosomal fineSTRUCTURE-based clusters detected in the PoBI study.
- Figure 6.8 Division of the Celtic fringe regions from the rest of the British Isles.
- Figure 6.9 Division of the Danelaw region from the rest of the British Isles.
- Figure 6.10 The PoBI autosomal Central/South England cluster represented in the samples.
- Figure 6.11 Division of the East regions from the rest of the British Isles.
- Figure 6.12 Efficiency of the sequencing runs.
- Figure 6.13 Normalised coverage of mtDNA and STR markers sequenced.
- Figure 6.14 Geographical distribution of predicted mtDNA haplogroups across the British Isles.
- Figure 6.15 Median-joining network of samples with mtDNA CR variants.
- Figure 6.16 Increase in allele diversity of Y-STRs in 362 samples analysed by MPS.
- Figure 6.17 Geographical distribution of predicted Y-chromosomal haplogroups across the British Isles.
- Figure 6.18 Median-joining network of samples with Y-STR length variants.
- Figure 6.19 Median-joining network of samples with Y-STR sequence array variants.
- Figure 6.20 Median-joining network of samples with Y-STR sequence array, SNP and indel variants in the flanking region.

- Figure 6.21 Truncated version of the median-joining network focussing on R1a and R1b clusters.
- Figure 6.22 Median-joining network of sequence features and SNPs and indels of the sequenced Y-STRs.
- Figure 6.23 Multidimensional Scaling (MDS) plots of pairwise R_{ST} values using Y-STR length/sequence allele variants, by ten regions.
- Figure 6.24 Multidimensional Scaling (MDS) plots of pairwise R_{ST} values using Y-STR length/sequence allele variants, by 37 populations.
- Figure 6.25 Multidimensional Scaling (MDS) plots of pairwise R_{ST} values using Y-STR length/sequence allele variants, by language and historical regions.
- Figure 6.26 Principal Component Analysis (PCA) results plotted using Y-STR length/sequence allele variants, coloured by Y haplogroups.
- Figure 6.27 Quality of representation of PCA variables in length- and sequence-based analyses.

Abbreviations

A	Adenine
AIM	Ancestry Informative Marker
<i>Alu</i>	<i>Arthrobacter luteus</i>
AmelX	Amelogenin X form
AmelY	Amelogenin Y form
AMOVA	Analysis of Molecular Variance
aSTR	Autosomal Short Tandem Repeat
BAM	Binary Alignment Map
Bash	Bourne-Again Shell
bp	Base pairs
BR	Broad Range
BQ	Base quality score
BWA	Burrows-Wheeler Aligner
C	Cytosine
°C	Degrees Centigrade
CE	Capillary electrophoresis
CE	Common Era
CEPH	Centre d'Étude du Polymorphisme Humain
CEU	Utah Residents (CEPH) with Northern and Western European ancestry
CHB	Han Chinese in Beijing, China
CODIS	Combined DNA Index System
CR	Control region
dbSNP	Single Nucleotide Polymorphism Database
ddNTP	Dideoxy nucleotide triphosphate
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
dsDNA	Double-stranded DNA
DoC	Depth of Coverage

DTC	Direct-to-customer
EA	East
EB	Elution Buffer
EM	East Midlands
EMPOP	EDNAP mitochondrial DNA population database
EPG	Electropherogram
ESS	European Standard Set of loci
FBI	Federal Bureau of Investigation
FSP	Forensic Service Provider
F_{ST}	Fixation Index
FTDNA	FamilyTreeDNA
G	Guanine
GATK	The Genome Analysis Tool Kit
Gb	Gigabase
GD	Gene diversity
GRCh38	Genome Reference Consortium Human Build 38
GSK	Golden State Killer
H	Heterozygosity
HapMap	Haplotype Map Catalog
HD	Haplotype Diversity
He	Expected Heterozygosity
Hg	Haplogroup
HGDP	Human Genome Diversity Project
HGDP-CEPH	Human Genome Diversity Project - Centre d'Etude du Polymorphisme Humain
HMP	Haplotype Match Probability
H_{obs}	Observed Heterozygosity
HS	High sensitivity
HT	High throughput
HV1/2/3	Hypervariable Regions 1/2/3
HWE	Hardy Weinberg Equilibrium

IBD	Isolation by Distance, or Identity-by-Descent
IBS	Identity-by-State
IGV	Integrative Genomics Viewer
iiSNP	Identity informative single nucleotide polymorphism
Indel	Insertion/Deletion Polymorphism
IOM	Isle of Man
IRE	Ireland
IRE	Ireland and the Isle of Man
ISFG	International Society for Forensic Genetics
ISOGG	International Society of Genetic Genealogy
JPT	Japanese in Tokyo, Japan
kb	Kilobase
KYA	Thousand Years Ago
LCL	Lymphoblastoid Cell Line
LHP	Length Heteroplasmy
LINE	Long Interspersed Nuclear Element
LR	Likelihood Ratio
LT	Low Throughput
LTR	Long Terminal-Repeat
LUS	Longest Uninterrupted Stretch
LWK	Luhya from Webuye, Kenya
M	Million
MAF	Minor Allele Frequency
Mb	Megabase
MDS	Multi-Dimensional Scaling
MJ	Median-joining
MPS	Massively parallel sequencing
MQ	Mapping Quality
MSY	Male-specific region of the Y chromosome
MXL	Mexican Ancestry in Los Angeles, California

mtDNA	Mitochondrial DNA
NC	Not called
NCBI	National Center for Biotechnology Information
NDIS	National DNA Index System
NE	North East
ng	Nanogram
NGS	Next-Generation Sequencing
nM	Nanomolar
numts	Nuclear mitochondrial DNA sequence
nt	Nucleotide
NW	North West
OREO	Overarching Read Enrichment Option
PacBio	Pacific Biosciences
PAR	Pseudoautosomal Region
PC1	Principal Component 1
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PD	Power of Discrimination
PE	Paired-end
PE	Power of Exclusion
PF	Passing Filter
pg	Picogram
PIC	Polymorphism Information Content
PM	March Probability
PoBI	People of the British Isles
PRT	Paralogue ratio test
pType	Prototype
QC	Quality control
qPCR	Quantitative Polymerase Chain Reaction
QR	Quick Response

R1/2	Read1/2
rCRS	Revised Cambridge Reference Sequence
RFLP	Restriction Fragment Length Polymorphism
RFU	Relative Fluorescence Unit
RG	Read Group information
PHP	Point Heteroplasmy
RM	Rapidly Mutating
RMP	Random Match Probability
RPV	Repeat Pattern Variant
rs	Reference SNP Cluster ID
RUO	Research Use Only
Sal	Saliva
SAM2	String-based Search Algorithm 2
SAV	Sequencing Analysis Viewer
SBS	Sequencing-by-Synthesis
SCO	Scotland
SD	Standard Deviation
SE	South East
SE	Single-end
SID	Sequence Identifier
SINE	Short Interspersed Nuclear Element
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SPRI	Solid Phase Reversible Immobilisation
ssDNA	Single-stranded DNA
STR	Short Tandem Repeat
STRAF	STR Analysis for Forensics
STRAND	Short Tandem Repeat: Align, Name, Define
STRSeq	STR Sequence database
SV	Structural Variation
SW	South West

SWGDAM	Scientific Working Group on DNA Analysis Methods
T	Thymine
TPI	Typical Paternity Index
TSI	Toscani from Italy
TMRCA	Time to most recent common ancestor
UAS	Universal Analysis Software
UK	United Kingdom
UK NDNAD	United Kingdom National DNA Database
US	United States
USD	United States Dollar
WAL	Wales
WhB	Whole blood
WGS	Whole Genome Sequencing
WHG	Western European Hunter-Gatherer
WM	West Midlands
YHRD	Y Chromosome Haplotype Reference Database
YRI	Yoruba in Ibadan, Nigeria
Y-STR	Y Chromosome Short Tandem Repeat
μl	Microlitre
1KGP	1000 Genomes Project

CHAPTER 1: Introduction

Forensic genetic identification uses genetic markers to describe and compare individuals in assistance of legal procedures, such as criminal investigations, kinship and parentage testing, verification of identities of missing persons or victims of mass disasters (Jobling and Gill 2004). Analyses of the different types of genetic markers are interpreted in the context of the frequencies of the identified alleles in the relevant populations. This provides a statistical evaluation of the strength of the biological evidence found for the purpose of discriminating and identifying human or non-human subjects.

This thesis examines different forensic genetic markers at the level of DNA sequence using high-throughput sequencing technology, including a small but highly diverse global sample set, and a larger set of samples with good geographical representation of the British Isles. This introductory chapter sets the scene by considering the nature of forensic genetics and the DNA markers commonly used, and then introduces the recently developed methods for analysing such markers in massively-parallel fashion. The four later results chapters contain their own introductions that focus on background material for each specific topic explored.

1.1 Forensic genetic markers

Using DNA as an evidential or investigative tool is a relatively new branch of the forensic sciences. Prior to DNA, biological samples were investigated for the presence of polymorphic protein markers including blood groups and isozymes, but these gave relatively low discriminatory power, and some were suited only to specific sample types. DNA typing rapidly overtook the role of serology in forensic investigations because of its focus on polymorphic markers which when combined had the power to identify and discriminate between individuals and characterise their kinship with high probability (Jeffreys et al. 1991)

The first markers used in 1984 by Sir Alec Jeffreys, at the University of Leicester, were minisatellites, hypervariable markers with a high degree of length polymorphism (Jeffreys et al. 1985b). Using electrophoretic separation and multilocus probes (detecting many minisatellites simultaneously) generated autoradiographs with patterns as unique as fingerprints (Figure 1.1), that inspired the term 'DNA fingerprint' (Jeffreys et al. 1985a). Combination of two multilocus probes gave a random match probability (the chance of two random individuals sharing the same fingerprints) of 7×10^{-22} . The first legal use of the technology was in a kinship case, since the bands in a fingerprint segregate as Mendelian markers within pedigrees.

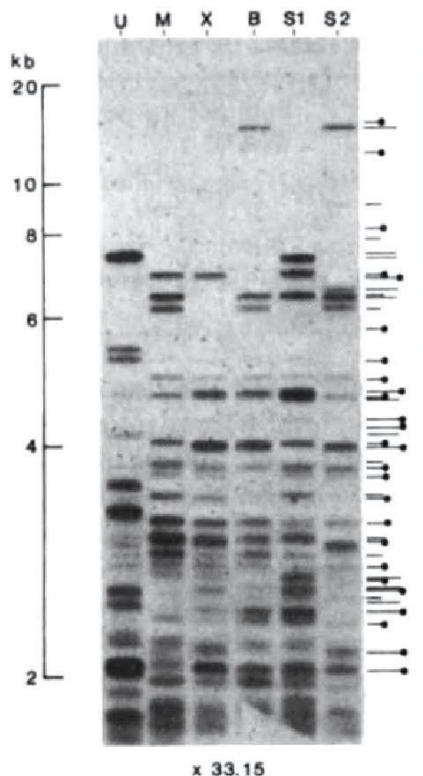


Figure 1.1
An early DNA fingerprint.

From Jeffreys et al. (1985a), where minisatellites were first used to prove kinship in an immigration test-case. The bands on the autoradiograph showing patterns of inheritance in the absence of a father, between the mother (M) her undisputed son (B) and daughters (S1,S2) and the boy in dispute (X).

However, despite its power, the general use of the technology was limited by the available quantity and quality of the DNA, which needed to be available in >100 ng amounts and at high molecular weight (Gill et al. 1985). The introduction of the polymerase chain reaction (PCR; Mullis 1990) allowed the amplification of specific target sequences from trace amounts of genetic material, and the discovery of

microsatellite markers (or short tandem repeats, STRs) provided a set of polymorphic markers suitable for PCR-based forensic DNA analysis (Reynolds et al. 1991). The short nature of these STRs allowed successful amplification from limited samples and those containing degraded DNA. Multiplexes amplifying many STRs simultaneously, and fluorescent detection after electrophoresis became available. The first national DNA database was launched in 1995 in the UK, storing genotype information for STR markers, which could be rapidly compared between individuals and which became the gold standard forensic markers (Wallace 2006).

1.1.1 Short tandem repeats (STRs)

1.1.1.1 The origin of STRs

The assembled reference sequence of the Human Genome (current build is GRCh38, released in Dec 2013, <https://www.ncbi.nlm.nih.gov/grc>) is a haploid composite reference of multiple genomes, not reflecting any one individual, however, the Y chromosome is mostly of a single source, which helped the assembly of its reference. The Y chromosome harbours significant variation in its non-recombining region, which is divergent from the X chromosome with high level of structural variations, including inversions, constitutive and polymorphic, segmental duplications, and deletions (Balaesque et al. 2008).

STRs are often associated with active retrotransposons, specifically non-LTR (long terminal-repeat) transposons, like the autonomous L1 or the non-autonomous Alu elements, which can give rise to new STRs by a seed repeat either located internally or in their 3' poly-A tail, where new STRs can "grow" by sequentially expanding, and incorporating nucleotide changes, hence creating new repeat motifs (Grandi and An 2013).

The stepwise mutation model describes a steady rate of evolution of microsatellites, where due to replication slippage the temporarily dissociated strand realigns at a neighbouring repeat unit, and thus, the net length of the repeat array expands or contracts by a complete repeat. However, it was shown (Brinkmann et al. 1998;

Sun et al. 2012) that the actual mutation rates are more complicated and the direction of change depends on allele size, where shorter arrays grow stepwise, while longer alleles rather shorten via deletions, generating a balanced allele range. The mutation rate increases with the array length and this can be further affected by the sequence structure of the array: e.g. a SNP interrupting the homogeneity of a long array slows its mutation rate down.

1.1.1.2 Structure of STRs

Commonly analysed STRs consist of arrays of 2-7 base pair (bp) long repeat units, and show high degrees of polymorphism, with a sufficiently narrow allelic range (typically of 8-20 repeat units) to avoid allelic imbalance due to preferential amplification of heterozygotes during PCR. STR markers are classified based on their repeat patterns (Urquhart et al. 1994): simple STRs, which contain uniform arrays of identical repeats (e.g. TPOX or DYS456); compound STRs which contain more than one adjacent simple array of repeats of the same unit length (e.g. D12S391 or DYS389I/II); and complex STRs which contain several arrays of interspersed and variable repeats of different unit lengths (e.g. D21S11). STRs have been widely used as markers not just in forensic genetics for human identification, but in linkage (Hearne et al. 1992) and population-genetic studies (Kim and Sappington 2013) as well. Their short lengths compared to minisatellites led to their rapid gain in popularity as marker types ideal for amplification from degraded forensic samples - most currently used STRs can be amplified on fragments of less than ~400 bp, while miniSTRs, designed to produce especially short amplicons, even reduce this to 200-300 bp (Butler et al. 2003).

STRs are found across all human chromosomes. Like minisatellites in a DNA fingerprint, the autosomal STRs are inherited independently in a Mendelian fashion and therefore in combination lead to very low random match probabilities, making them powerful in individual identification (Jobling and Gill 2004). Sex-chromosomal STRs have the same kinds of structures and variability as autosomal STRs, but have different inheritance patterns. Y chromosomes are passed down exclusively

from father to son without recombination (except from the short pseudoautosomal regions that are homologous to the X chromosome) and therefore the STRs typed along the male-specific region of the Y chromosome (MSY) cannot be considered independent markers, but rather linked as a Y-chromosomal haplotype and inherited together along the male lineage. Therefore haplotypes of uniparentally-inherited markers, like the Y chromosome or mitochondrial DNA (mtDNA), are not individually identifying, as these are shared along the respective parental lineages (Jobling et al. 1997). This also means that haplotypes are much less variable than genotypes based on independently inherited autosomal STRs. Each sex has at least one X chromosome, and while in males there is no possibility of recombination with another X chromosome it is passed to his daughter as a haplotype, without change apart from mutations. In females, by contrast, the two X chromosomes similarly to the autosomes bear their STRs as independently recombining markers provided these are sufficiently separated on the chromosome. Y-STR profiling helps the examination and comparison of male lineages, can characterise male contributors in male-female mixtures, can contribute to the exclusion of direct paternal relationship or can connect distant male lineage relatives. X-chromosomal STRs are not widely used, but can inform about sex-related bias in populations when compared to Y-chromosomes and can provide additional clarification in extended kinship testing or identification of remains (Bennett 2000; Butler 2005; Tamaki and Jeffreys 2005).

1.1.1.3 Conventional amplification and detection of STRs

PCR amplification of STRs targets the non-variable flanking regions of a locus with primers, and thus all length variants originating from the array can be captured. However, occasionally deletions and insertions outside the array can affect the length of the amplicon, or can decrease amplification success or even prevent successful amplification. To overcome the effect of known sequence variation at the primer binding sites (SNPs and short indels), which can affect the annealing of primers to the template, most commercial kits likely use degenerate primer mixtures to facilitate amplification (Leibelt et al. 2003). Even in the presence of these

variants, however, unexpected rare nucleotide changes can still interfere with the amplification and could reduce the efficiency of PCR, or even prevent amplification, thus creating “null alleles”.

Standard length-based separation of fragments by capillary electrophoresis (CE) uses the detection of a fluorescent signal (Figure 1.2). Each amplified fragment is labelled by incorporating the primers, one of which is fluorescently labelled. The emission from the fluorophore-tagged strand then gets registered after laser excitation as it passes the detection window.

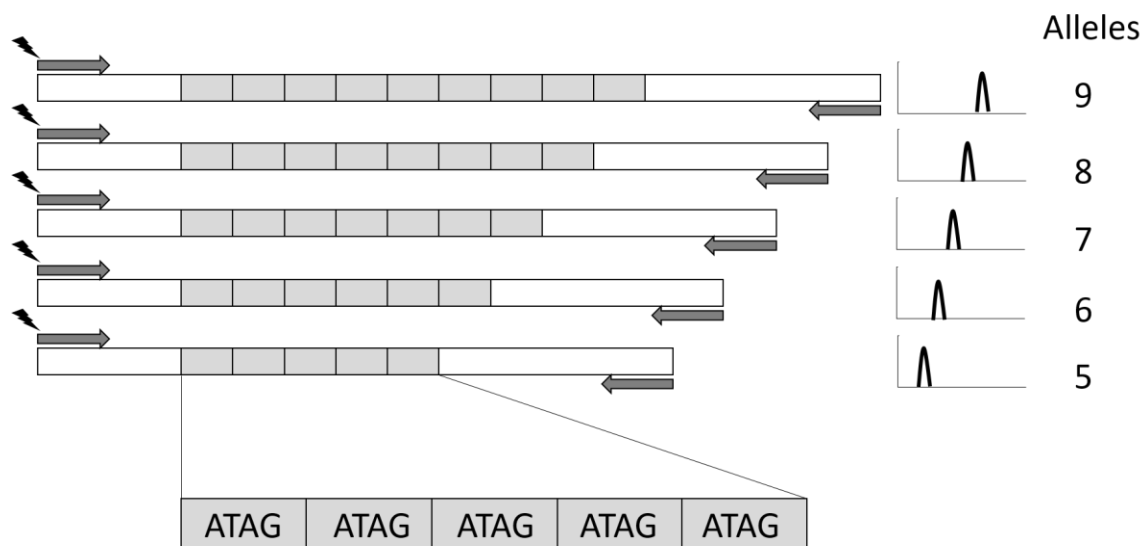


Figure 1.2

A schematic representation of an STR locus

STRs are amplified during PCR using primers (arrows) targeting unique sequences in their flanking regions. One primer of the pair is fluorescently labelled (bolts) to allow detection by CE. The alleles are detected as peaks on the electropherograms and named according to their size, determined by the assumed number of repeat units based on fragment length, here ATAG repeats.

To allow several STRs to be efficiently co-amplified in one reaction, multiplex PCR reactions have been designed commercially. To maximise output these use several different fluorophores to allow distinction of overlapping size range fragments by colour. The number of available electropherogram colour channels is a limitation on

the number of loci that can be included, as is the need to keep fragment sizes sufficiently small (<~450 bp) to allow genotyping from degraded material. In practice, the largest CE multiplexes contain 27 markers detected across six dye channels (PowerPlex® Fusion 6C and VersaPlex™ 27PY Systems, Promega; Yfiler™ Plus PCR Amplification Kit, Applied Biosystems Thermo Fisher).

STR alleles were selected as forensic markers due to their length polymorphism, and allelic variants of STRs are identified and classified by their overall length. The separation of amplified allelic variants by length is achieved by electrophoresis, where the negatively charged DNA molecules subjected to an electric field travel towards the positive electrode (anode) at a rate according to their molecular size through a gel state matrix. Slab-gel separation was originally used as a separation method, but in forensic DNA typing it was replaced by the automated detection of fluorescently labelled variable length fragments in the polymer-filled capillaries of a CE machine. In CE the migration of the labelled DNA fragments is facilitated by the current provided by electrokinetic injection to separate the fragments, and the time to migrate through the fixed length capillary is used to measure alleles, as opposed to their migrated distance during a fixed time as in classic gel electrophoresis. CE is optimised for small quantities and highly sensitive detection by performing separation within a small volume of matrix inside the capillary. Labelled fragments passing through the detection window generate a trace of fluorescence intensity against time, rendered as base-pairs, known as an electropherogram (EPG; Figure 1.3).

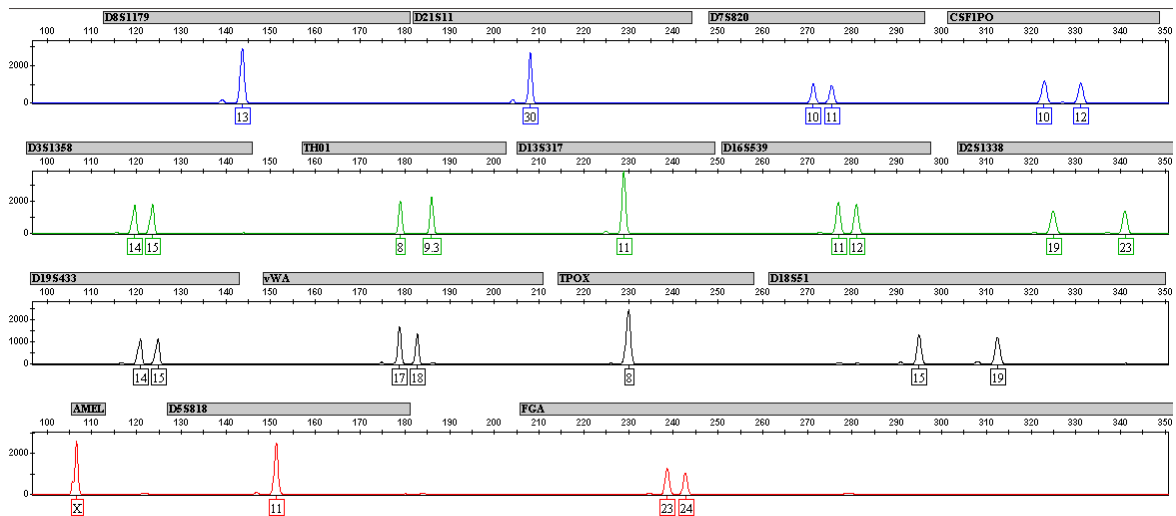


Figure 1.3
Example of an electropherogram (EPG).

EPG of a complete DNA profile from a single source. Markers are grouped into colours of their respective fluorophores (except the yellow one is represented by black colour for better visibility), placed along the x-axes representing length of fragments; y-axes show the relative fluorescence values. One or two peaks represent homo- and heterozygous alleles for each of the loci marked in the grey boxes representing their range by length. In this system, a fifth dye channel (not shown) contains an in-capillary size standard.

Somatic mutations that occur during development can create multiallelic profiles for the affected loci, if tissues with different alleles are sampled and typed together. This can cause unsuspected discordance between different sample sources from the body affecting forensic casework and kinship testing results (Rolf et al. 2002).

Similar replication slippage on the template and the lack of proofreading mechanisms *in vitro* during PCR amplification generates stutter products. Stutter products are smaller artefactual peaks adjacent to the main alleles, usually detected most clearly in the n-1 position of allele n, but sometimes seen in n-2 and n+1 positions in CE (Figure 1.4). Mean stutter proportions can be identified and adjusted for, either locus by locus, or by setting a broad stutter filter for all loci at around 10 to 20% (Leclair et al. 2004). Stutter peaks are usually easily identifiable in reference samples, but can interfere with genotyping in complex mixtures.

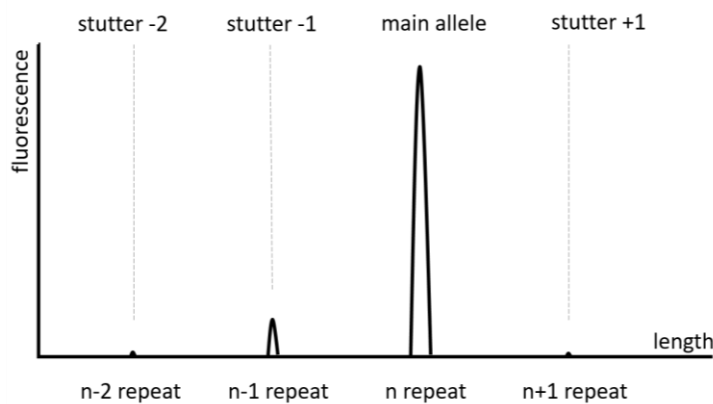


Figure 1.4
Stutter peaks.

Stutter peaks are characterised by their length difference relative to the main allele as seen on this schematic diagram.

Current standard practice uses a multiplex-PCR approach (Table 1.1) together with CE to detect length variants of STR alleles, where the alleles at each locus are referred to by, and correspond to the variable lengths of these STRs. While this approach includes several well-described polymorphic loci scattered along the genome, it does not explore the underlying sequence variation present in these alleles, and thus ignores internal variants that could increase discriminative power (Butler 2007; Yang et al. 2014). This thesis explores such sequence variants, which are discussed below.

Table 1.1**Summary of standard autosomal forensic markers used globally.**

	European Standard Set - ESS (incl. the UK since July 2014)	US core loci	US expanded core loci from January 2017	Autosomal STRs used in this study
D1S1656	✓	-	✓	✓
D2S1338	✓	-	✓	✓
D2S441	✓	-	✓	✓
D3S1358	✓	✓	-	✓
D8S1179	✓	✓	-	✓
D10S1248	✓	-	✓	✓
D12S391	✓	-	✓	✓
D16S539	✓	✓	-	✓
D18S51	✓	✓	-	✓
D19S433	✓	-	✓	✓
D21S11	✓	✓	-	✓
D22S1045	✓	-	✓	✓
FGA	✓	✓	-	✓
TH01	✓	✓	-	✓
VWA	✓	✓	-	✓
SE33	✓	-	-	-
Amelogenin	✓	✓	-	✓
CSF1PO	-	✓	-	✓
TPOX	-	✓	-	✓
D5S818	-	✓	-	✓
D7S820	-	✓	-	✓
D13S317	-	✓	-	✓
PentaD	-	-	-	✓
PentaE	-	-	-	✓

1.1.1.4 Typing STRs using MPS

Typing STRs by their sequence has gained traction due to affordable custom-made forensic solutions designed for different massively parallel sequencing (MPS) platforms that became available in the past decade. When STRs are sequenced, the size of amplicons can generally be reduced, as the restriction of the colour-channels and length-ranges of the CE assay does not apply. Furthermore, the number of amplified loci in one multiplex reaction can be increased, and combined with different types of markers as well.

Sequenced PCR amplicons are analysed as reads, stacked into histograms (equivalent of peaks in an EPG) generating a representation of genotype calls

similar to customary CE. PCR-slippage artefacts are also observed in the stutter positions, with even greater sensitivity, down to the single-read level. The sequenced stutter alleles, however, have a better potential to be differentiated from minor components, as sequence variants may be observed between these, even if they may be indistinguishable by length. Setting a custom sequence-specific stutter filter has now become possible through knowledge of internal STR allele structures, parts of which may contribute differently to overall stutter level.

One of the main advantages of sequencing for STR detection is the ability to separate isoalleles - STR alleles with the same length, but different sequences. STR detection with MPS therefore can improve forensic parameters for the markers by increasing allele diversity, increasing heterozygosity, decreasing relative allele frequencies, lowering random match probabilities, and overall increasing the discrimination power of the loci typed (de Knijff 2019b).

1.1.2 Single nucleotide polymorphisms (SNPs) and short indels

1.1.2.1 Forensic use of SNPs

Another class of markers that can be used for forensic purposes are single nucleotide variants (SNV), which are traditionally referred to as single nucleotide polymorphisms (SNP) when their frequency reaches 1% in the population, although it is generally accepted to refer to any SNV as a SNP (Budowle and van Daal 2008; International HapMap Consortium 2003).

SNPs are highly abundant - on average, a human genome contains 3-4 million SNP differences when compared to any other (1000 Genomes Project Consortium 2015) - and are widely used for linkage analysis (Ott et al. 2015) and medical genetic studies (Rabbani et al. 2014), pharmacogenetics and toxicology (Yucesan and Ozten 2019) and genetic genealogy and citizen science (Jobling and Tyler-Smith 2017). In forensic use, SNPs on the mitochondrial DNA (mtDNA) are used to define haplotypes that can provide matches or exclusions between mitotypes (Parson et al. 2014); they also define lineages that form a phylogeny and allow classification into SNP-based haplogroups (van Oven and Kayser 2009). SNPs on the male-specific region of the Y-chromosome (MSY) likewise define haplogroups that form a phylogeny, and while not much used as a primary tool in forensic genetics, provide an evolutionary framework that allow STR haplotypes to be better understood (Jobling 2001).

Autosomal SNPs are used in several different contexts in forensic DNA analysis (Budowle and van Daal 2008). Ancestry informative SNPs provide a probabilistic approach to biogeographical ancestry for investigative leads (Phillips 2015), with low heterozygosity and high F_{ST} (fixation index) values, meaning they are less variable and can be nearly fixed for the considered population (Frudakis et al. 2003). Phenotypically informative SNPs can be used with a probabilistic approach

to estimate phenotypic traits, such as hair, eye and skin colours for investigative leads (Chaitanya et al. 2018).

Sets of identity informative SNPs (iiSNPs) were selected for high global heterozygosity and low F_{ST} values (being very variable and not population-specific), and can be used as an alternative to STRs to differentiate and identify individuals (Sanchez et al. 2006). These are discussed further below.

1.1.2.2 SNPs compared to STRs

SNPs are stable markers with relatively low mutation rates compared to STRs (Sobrino and Carracedo 2005). The iiSNPs are ideal when STR markers fail due to non-ideal sample quality (Budowle and van Daal 2008), as amplicons targeting these single nucleotide sequence variants can be shorter (<100 bp) than STR amplicons (up to ~400 bp), and therefore amplifying SNPs is more likely to be successful from degraded samples, for example identification of unknown remains in missing person cases (Sobrino et al. 2005). Due to their mostly biallelic nature, SNPs are less individually variable than STRs, and therefore the number of SNP markers to be typed to determine identity is higher than the corresponding number of STRs (Gill 2001). In order to reach comparable discrimination power an estimated ~100 highly heterozygous SNPs can be the equivalent of ~16 STRs .

SNPs can be multiplexed to higher levels for CE analysis, up to ~50-plex, but there are also other detection methods which can increase the number of analysed SNPs, such as massively parallel sequencing targeting virtually unlimited number of SNPs in one experiment or hybridisation to microarray chips which allow detection of over a million SNPs at once. Detection of SNPs at this scale is automated and interpretation of genotypes is not affected by artefacts like stutter in the case of STR typing. The limited allele diversity compared to STRs makes SNPs less informative for mixture deconvolution, and SNPs in forensic genetics are excellent tools to extend, but not to replace the capability of currently routinely typed STR markers (Butler and Hill 2012; Pontes et al. 2015).

When considering STRs and SNPs within the phylogeny of the Y chromosome, the mutation rate of SNPs ($\sim 10^{-8}$) is lower compared to STRs (in general $\sim 10^{-3}$ - 10^{-4} or as high as $\sim 10^{-2}$ for rapidly mutating Y-STRs). Since the phylogeny reflects sequential changes to the Y-chromosomes, the observed SNPs reflective of certain branches of the Y-chromosome tree could be associated with broader biogeographic origin (Jobling 2001).

Since STRs are mutating at a higher rate on a slower changing SNP-based framework of the Y chromosome, the observed allelic ranges of Y-STRs within a clade with shared ancestry are not random, and therefore Y-STR haplotypes can be used to infer membership of a clade, which is the basis of haplogroup prediction.

1.1.2.3 Typing SNPs

Several detection technologies were established in the last decade using different approaches: allele-specific hybridisation, primer extension, ligation or cleavage (Sobrino et al. 2005; Sobrino and Carracedo 2005); some of these require large amounts of input DNA, and some are limited in their ability to multiplex.

Multiplexed typing of SNPs for general forensic purposes can practically be achieved via capillary electrophoresis using SNaPshot assays (Tully et al. 1996), where markers are targeted with locus-specific primers adjacent to the SNP which are then extended with fluorescently labelled nucleotides, specific either to the ancestral or derived alleles. For clear spatial separation long-tailed primers are used and the genotypes of the samples are interpreted from the generated electropherograms.

SNPs can also be determined by direct sequencing of the region of interest; however, using classic Sanger sequencing is impractical when large numbers of targeted SNPs are scattered along the genome. Instead, this method is used when SNP detection is confined to a shorter region, for example sequencing mtDNA to define a SNP haplotype.

A genomewide SNP-typing alternative is a relatively affordable and accurate way to type even millions of SNPs at once using microarrays also known as `SNP chips` (Ragoussis 2009). The technique uses an array of short (~50-nt) oligonucleotide probes, to which whole-genome amplified DNA fragments anneal and serve as templates for the extension of these short probes as primers, incorporating labelled terminating nucleotides reflecting the intrinsic variants. The variants are detected by measuring fluorescence intensities, which are then converted to genotype data.

Though this method is very reproducible and effective, unaccounted variants can interfere with the annealing effectiveness and can miss typing certain variants. Therefore, with the increasing affordability of whole genome sequencing (WGS), the most reliable way to type SNP variants unbiased is to sequence the genome and extract the SNP variants from it.

With the spread of massively parallel sequencing the classical Sanger sequencing method can be replaced by faster and affordable high-throughput options, with nearly unlimited multiplexing of amplicons targeting hundreds of diverse set of SNPs into one MPS reaction. SNPs can also be combined with other markers in these approaches to increase informativeness. The interpretation of the sequence reads from MPS is often aided by automated software to determine the genotypes.

1.1.2.4 Using SNP data and databasing

SNP-based identity genotypes derived from degraded DNA cannot be searched against STR-based national databases and so requires a separate SNP-based database of reference samples to be created for practical forensic use (e.g. missing persons, unidentified remains, disaster victim identification).

Some information of forensic relevance can be gleaned from the Single Nucleotide Polymorphism Database (dbSNP) at NCBI which catalogs known SNP variation including short insertion/deletions (indels) and provides each with a unique identifier in the form of a Reference SNP (rs) number. Frequency information in a number of

large-scale population studies and information about the variants known to affect phenotypes are also available.

Large publicly available datasets, such as the 1000 Genomes Project (<https://www.internationalgenome.org/data>), the Human Genome Diversity Panel (HGDP; <http://www.cephb.fr/hgdp/>), the Simons Genome Diversity Project (<https://reichdata.hms.harvard.edu/pub/datasets/sgdp/>), and the Personal Genome Project (<https://www.personalgenomes.org/>) also allow access to a large collection of human SNP genotypes typed in populations chosen to survey diverse biogeographic variation, usually the available SNP data is array-typed or sequenced using whole genome sequencing (WGS).

1.1.2.5 Mitochondrial DNA typing

Analysis of mitochondrial variants is a long-standing branch of forensic SNP analysis; due to its relatively small-sized circular genome of ~16.6 kb and its abundance in the cell, it was sequenced early on (Anderson et al. 1981) and has been extensively used in forensic typing of heavily degraded (Torroni et al. 2006) or ancient remains (Krings et al. 1997).

MtDNA is a lineage marker, shared by maternal relatives and passed down maternally without change, except mutations. The high level of oxidative stress in this organelle together with the unique characteristics of nucleic acid organisation and repair systems accounts for its elevated mutation rate, which is at least one order of magnitude higher than that of nuclear DNA (Alexeyev et al. 2013; Just et al. 2015).

MtDNA sequence variants are the result of mutations which accumulate sequentially creating diverging lineages (haplotypes) which are clustered within phylogenetically related groups, haplogroups, derived by descent from a common ancestral mtDNA bearing distinctive SNPs (Torroni et al. 2006). In practice, this picture is complicated by the particularly high mutation rates of some sites, where

back and recurrent mutation is relatively common; these can lead to reticulate structures within a network of mtDNA haplotypes.

The mtDNA control region is 1122 bp long and contains three hypervariable regions (HV1, HV2 and HV3) that show particularly high degrees of variation between individuals (Figure 1.5). Originally only these specific regions were targeted in forensic analysis, since they contain sufficient sequence variants to be informative for most forensic applications, but currently sequencing of the whole control region or the whole mtDNA is regarded as a better approach to avoid the introduction of errors from artificial recombination (Parson and Bandelt 2007; Parson et al. 2014).

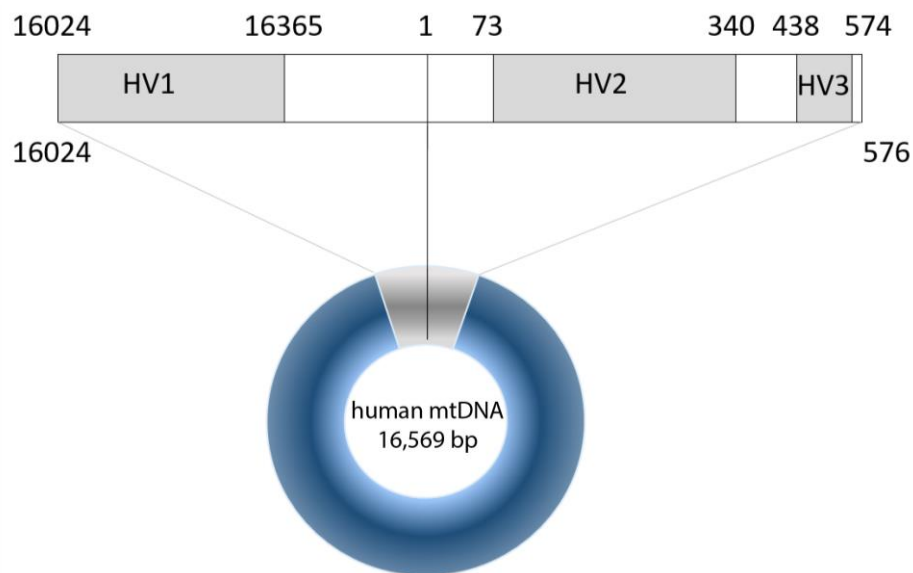


Figure 1.5
The human mtDNA.

Schematic representation of the molecule, highlighting the control region and the hypervariable segments (HV1, HV2, HV3) with their bounds given by nucleotide coordinates according to the rCRS.

Heteroplasmy refers to more than one type of mtDNA being detected within a single-source sample. Heteroplasmic sites can be defined as point heteroplasmy (PHP) affecting a single nucleotide or length heteroplasmy (LHP), where the length of an array of nucleotides is variable due to insertions and deletions. PHP is

recognised as a mix of two bases at a single position, usually at the hypermutable sites that form reticulations within mtDNA phylogenies, while LHP is observed in relation to long homopolymeric tracts, such as polyC tracts in HV1 and HV2. Reporting of LHP is often omitted as detection is dependent on technical conditions, and therefore it is also ignored in database searches. The extent of heteroplasmy between and within tissues can vary (Parson and Bandelt 2007; Parson et al. 2014), as can the limit of detection, but generally Sanger sequencing is able to detect heteroplasmic minor components at about 10-20% (Just et al. 2015).

MtDNA is a single locus, and its SNPs are linked within a haplotype. Sequencing full mitochondrial genomes with classical Sanger-sequencing to explore these variations is costly and time consuming, therefore forensic laboratories which have implemented techniques to utilise the information from the extranuclear genome have generally compromised by sequencing only the most variable regions of the mtDNA, in the non-coding control region. This region is under less constraint due to not encoding proteins, and therefore can accumulate more mutations than the coding region, where changes can affect the function of the organelle and the cells containing it.

Cases that can benefit from mtDNA analysis are usually those with limited access to nuclear DNA, such as hair shafts, bones, teeth and highly degraded samples, including charred or decomposed remains in unidentified and missing persons cases, and any other kinship testing that may require comparison of distant relatives within the maternal lineage (Just et al. 2015; King et al. 2014; Parson et al. 2014).

Apart from Sanger-sequencing, other avenues to analyse variation in the mtDNA include SNaPshot-based screening to genotype selected SNPs by primer extension (Weiler et al. 2016) or the use of MPS for mtDNA typing in a high throughput form making it more affordable, while also simplifying the analysis (Irwin et al. 2011; Parson and Dur 2007; Parson et al. 2014). This approach capitalises on the massively parallel nature of this sequencing technology, where high depth of coverage can be achieved at relatively low cost to raise the sensitivity of detection

of minor components and authentic heteroplasmic sites even below the previously often imperceptible 5% limit and help them to be distinguished from 'noise' (Just et al. 2015).

Interpretation of mtDNA variants is based on alignment to the reference sequence (revised Cambridge Reference Sequence, rCRS; Andrews et al. 1999) and can be controversial at times. To minimise ambiguous typing of variants nomenclature unification guidelines for forensic casework and population studies were provided by the DNA commission of the International Society of Forensic Genetics (ISFG) (Carracedo et al. 2000; Parson and Dur 2007; Parson et al. 2014). Online resources, such as EMPOP (www.empop.org), were created to provide a centrally curated, quality controlled forensic mtDNA nomenclature and database, a global collection of mtDNA haplotypes and associated tools (Parson and Dur 2007). A tool of EMPOP, SAM2, specifically provides a phylogenetically corrected realignment of the observed variants to avoid ambiguous alignments resulting in false exclusions, unidentified matches and biased frequency estimates for forensic mtDNA typing (Huber et al. 2018). This approach shows the variants which may not be the most parsimonious to the reference sequence, but the most closely related to their neighbours on the mtDNA phylogeny, therefore the described variants are more reflective of the actual sequential changes on the mtDNA, than transformed onto a derived form of reference sequence. The latter approach is more prone to generate false variant calls; for example the variant 16193T near the C-stretch region of HV1 is instead expressed as a longer string, but phylogenetically more accurate variants of 16191.1C 16192T 16193del. This is particularly important for haplogroups where the signature mutation is 16192T and the additional variants have occurred on this background (Huber et al. 2018).

1.1.2.6 Short insertions or deletions

While indels can affect large (up to multi-megabase) segments of the DNA, short indels from a single nucleotide up to ~25 nucleotides are often analysed together with SNPs. Though indels are heterogenous in origin and characteristics, their

mutation rates are more comparable to those of SNPs (Montgomery et al. 2013). The presence of the same indel in a non-structured sequence is likely the result of shared ancestry (identity by descent). Indels are also biallelic markers and similarly to SNPs a large number of them need to be analysed to be reasonably informative (Zidkova et al. 2013). Similarly to STRs, the population frequencies of indels need to be surveyed for them to be used for identification purposes (LaRue et al. 2012). Due to the length difference between alleles these are also ideal to be analysed by electrophoresis.

Though it cannot be considered strictly as an indel, a particularly relevant polymorphism is the 6-bp difference between the gametologous X and Y chromosome forms of the Amelogenin marker, AmelX and AmelY, the former bearing the deletion and therefore being shorter, a difference that is easily distinguishable by CE (Sullivan et al. 1993). Amelogenin therefore became the standard sex-typing marker present in most commercial multiplexes, although additional markers are often included considering that Y chromosomes with different deletions affecting the AmelY form of the marker exist (Jobling et al. 2007) and therefore their sex could be typed incorrectly as female (Santos et al. 1998; Thangaraj et al. 2002). Many of these deletions share common ancestry, for example the AmelY drop-outs related to the Indian subcontinent are known to belong to a J2b2-M241 Y-chromosomal lineage (Cadenas et al. 2007). Rarely, primer binding site mutations can cause either of the AmelX or AmelY forms of the marker to be typed as absent (Roffey et al. 2000; Shadrach et al. 2004). When included, additional Y-STR typing, can confirm the sex status together with Amelogenin (Butler 2005).

1.1.2.7 Short indels around STRs

Allelic STR length-variants mostly have different numbers of repeats compared to the genomic reference sequence; although these are often described as a special class of indels (Montgomery et al. 2013), here they are not regarded as such. An alignment-based view of STR amplicons is justified in the flanking regions, but when

variants are observed within the array, such reference-based comparison is not biologically sensible or practical, and in any case nucleotide changes or indels can rarely be pinpointed to exact genomic coordinates.

In a forensic context the relevance of indels is not only their potential to interrupt primer binding sites, thus affecting genotyping via missing (null) alleles, but also the potential for them to occur within an amplified region, thus affecting the amplified fragment length.

Short indels can also occur within the repeat arrays of STRs, and are represented by unusual allele lengths out of the expected register for uninterrupted n-mer repeats ('bins'). These intermediate alleles have been investigated by sequencing to explain the reason for such occurrences; some originate from the insertion or deletion of a part of a repeat unit, thus creating a partial remainder, which is then expressed in the allele name by adding the number of orphan nucleotides after a dot (e.g. 9.3, 12.1, 31.2). The other class of indels occurs outside the arrays, in the flanking regions, and depending on the different primer binding sites used by different multiplex kits may cause discordant results between the CE-typed alleles at the affected locus.

1.2 Massively parallel sequencing and its use in forensic genetics and genomics

1.2.1 Massively parallel sequencing

1.2.1.1 Massively parallel sequencing technologies

Classical Sanger-sequencing is a trusted standard method, in which the DNA sequence is determined during the synthesis of a DNA chain using the incorporation of nucleotide-specific fluorophores attached to chain-terminating modified nucleotides. The electrophoretic separation and detection of the generated labelled fragments reveals the DNA sequence of the sample (Heather and Chain 2016). The

method is relatively slow and expensive for sequencing large segments of the genome due to its serial nature, even though it has low error rates. An affordable alternative when sequencing many targets (different markers and individuals), is next-generation sequencing (NGS), which has evolved in the last two decades.

Various platforms offer different advantageous features, and together these technologies are referred to as next-generation sequencing or more accurately as massively parallel sequencing, MPS (and sometimes as second-generation sequencing). Another term, 'third-generation sequencing', is used for the subset of these methods which uses long-read sequencing on single DNA molecules, such as Oxford Nanopore Technology and Pacific Biosciences sequencing.

Though the underlying technologies vary, these all use massively parallel approaches to generate large quantities of sequence data (in the range of millions of reads) from multiple samples during the course of a single experiment. Such robust capacity and high throughput decrease the cost per nucleotide sequenced: by early 2017 the cost of sequencing a human-sized genome came close to 1000 USD (<https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>).

While the length of generated reads for some platforms is still on the short end of the scale, for example ~300 bp for Illumina®, or ~400 bp for Ion Torrent, some platforms allow highly increased read lengths, over 2 Mb with Oxford Nanopore ultralong reads and an average of 10-15 kb long with PacBio sequencing (Rhoads and Au 2015). MPS achieves a high sensitivity by generating large number of multiple reads concurrently during a run, resulting in a high coverage that mitigates the relatively high sequencing error rates, upon which these platforms are constantly improving, to provide more and more reliable alternatives to Sanger sequencing.

MPS made sequencing more accessible and effective not just for forensics, but many research fields from microbial metagenomics to medical diagnostics (Fuller et al. 2009; Goodwin et al. 2016; Quail et al. 2012; Reuter et al. 2015; van Dijk et al. 2014).

1.2.1.2 Massively parallel sequencing process

In MPS analysis the process of preparing samples prior to sequencing is referred to as 'library preparation'. For short-read approaches this requires initial tailoring of the insert DNA to the desired fragment length, which can be achieved either by enzymatic fragmentation, physical shearing or using PCR to generate amplicons; for long-read approaches, the DNA can be left in its native unfragmented state. The generated DNA fragments are prepared by adding terminal adapters, which contain specific sequences for sequencing initiation and indices (barcode sequences) giving the ability to multiplex several samples in one experiment.

Sequencing-by-synthesis (SBS) methods detect a signal at the incorporation of nucleotides into a growing strand. Depending on the platform, this can be by visual detection of different fluorescent signals (Illumina® platforms), or by measuring quasi-proportional electrochemical signals (Ion Torrent platforms) that occur during incorporation events. In both approaches the sequences of nucleotides can be recorded in parallel for millions of clusters each consisting of thousands of clonal copies of single molecules (Goodwin et al. 2016; Heather and Chain 2016; Reuter et al. 2015).

The focus here will be on the technology offered by Illumina®, since this dominates global sequencing production today, and was the technology used in this thesis. Illumina® offers a range of platforms from benchtop to production-scale, with different sequencing capacities, and these all use the SBS technology with cyclic reversible termination and optical detection. This, in principle, is similar to the Sanger-method, but different in that the termination of synthesised DNA chains is reversed in each cycle.

The sequencing starts by clonal amplification to increase signal strength (Figure 1.6). This is performed by 'bridge amplification' on the surface of a glass sequencing chip, which is preloaded with forward and reverse primers that are complementary to parts of the adapter sequences; the templates bind to these sequences forming bridge-like structures, along which amplification takes place. The newly generated

products act as templates themselves, binding to nearby fixed primers and the exponential amplification generates a positionally defined cluster of thousands of clonal copies of the original template. Sequencing of these amplified molecules synchronously in a cluster give detectable uniform visual signal during the subsequent SBS steps.

In the case of the Illumina® MiSeq® device (used in this thesis), during the sequencing phase the machine uses four different fluorophores to label and also temporarily block bases in each cycle. These nucleotides hybridise to their complementary sequences in the clusters of templates, and when excited emit a colour specific to the base incorporated. In each sequencing cycle these signals are imaged, then the fluorophores are cleaved off, permitting the next cycle of nucleotide addition, until sequencing is completed (Goodwin et al. 2016; Reuter et al. 2015).

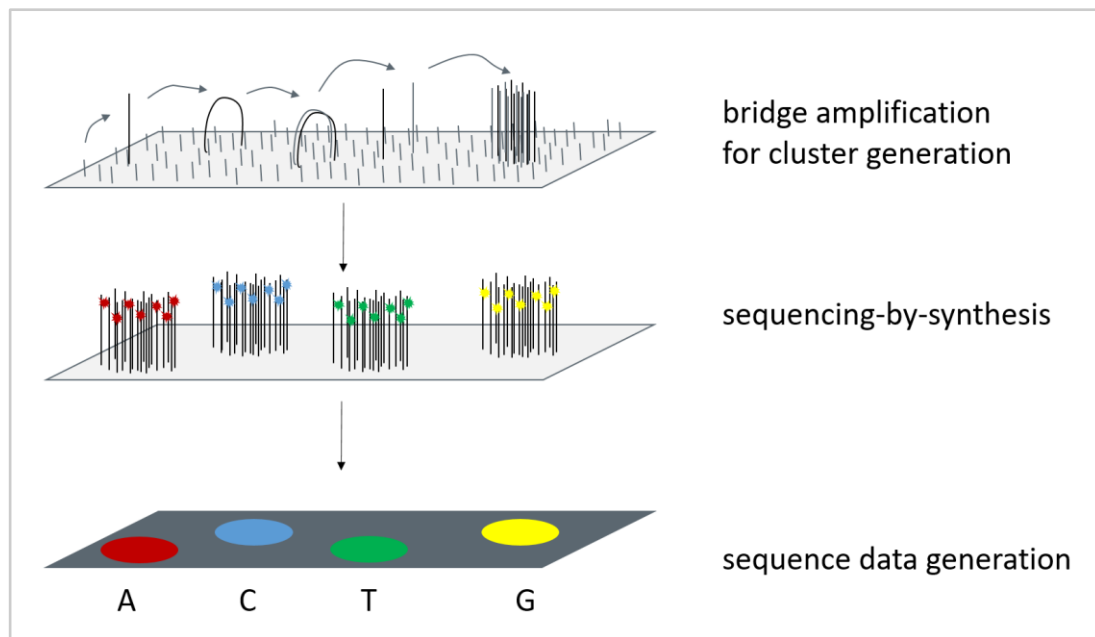


Figure 1.6
Illumina®'s sequencing-by-synthesis method.

(adapted from: www.illumina.com)

Once sequencing is complete, the device's software demultiplexes the reads according to the index sequences and assigns the detected reads to each analysed sample. Reads are compiled per sample and per read direction: read1 (R1) in single-end (SE) and read1 and read2 (R1, R2 respectively) in paired-end (PE) sequencing modes, which sequence respectively one or both strands of the bound library molecules; the latter improves quality of basecalls. The MiSeq® itself performs a level of adapter sequence removal, but this process is often incomplete and can be improved by standalone software as detailed in Chapter 2, Section 2.3.2, on quality control of raw FASTQ files. The output files of each run provide details of the sequencing and samples analysed and the index combination used to identify them. The run information also contains metrics and statistics that can inform the user about distribution of reads along the sequencing chip, quality of reads in the form of Q scores (a quality score scale representing the probability of erroneous base calls, for example Q40 means a probability of 1 error for every 10,000 nucleotides), detailed run-specific metrics for sequencing read and index cycles and the balance of multiplexed samples according to their index sequences. These can be reviewed independently from the sequencer using the raw output files with the free software Illumina® Sequencing Analysis Viewer (SAV v2.4.7), which generates a graphical output as shown on Figure 1.7.

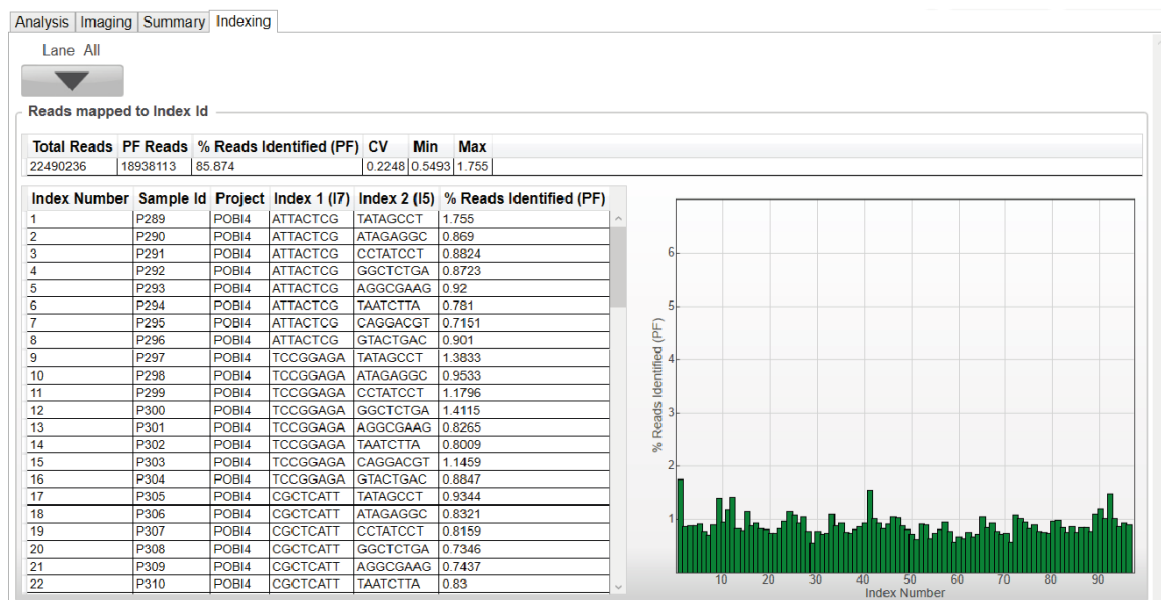
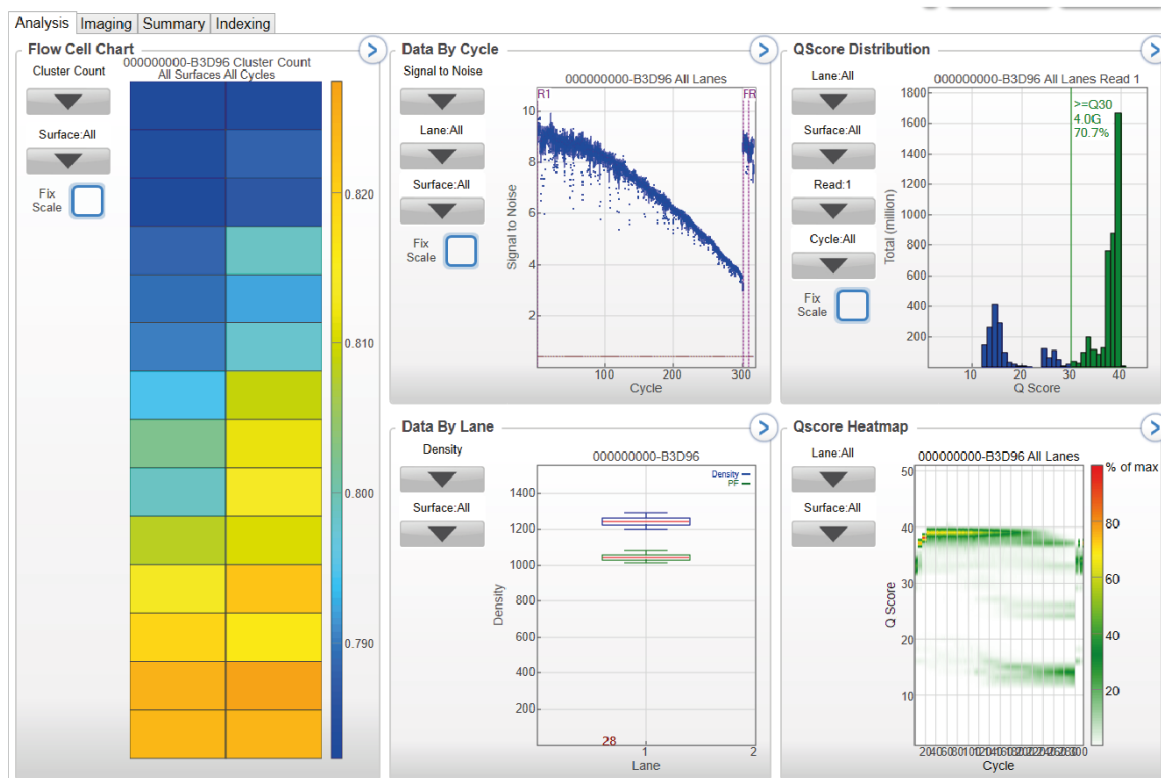


Figure 1.7
Run analytics.

Examples of run analytics (cluster count, density, Q score and adapter counts) provided by Illumina Sequencing Analysis Viewer software (SAV v2.4.7).

The generated raw files are provided in a compressed FASTQ format, which besides the read and sequence information (which are also found in FASTA files) contains the respective quality information as well for each of the called bases in the reads. The generated FASTQ files can be analysed by the MiSeq FGx[®] machine's integrated ForenSeq[™] Universal Analysis Software (UAS), or can also be used as an input for other third-party software to analyse the sequencing results.

1.2.2 Application of MPS in forensic genetics and genomics

While MPS approaches have been around for more than a decade, previously they could not meet the requirements and the limited budgets of forensic DNA analyses and therefore were no competition to the established CE-based method of STR typing, or the Sanger-type sequencing standard (Borsting and Morling 2015). With the technological advances of the last decade and developments in chemistries, some MPS platforms now offer truly affordable alternatives to CE in forensic DNA analysis (Yang et al. 2014).

1.2.2.1 Advantages of MPS for forensic use

MPS has many appealing qualities to the forensic community: it is a high-throughput technique using a massively parallel approach to generate large quantities of sequence data in a single experiment, where the unique tagging and pooling of individual samples allows the reduction of cost per analysed marker (Borsting and Morling 2015). Compared to CE, MPS has the potential to provide results for a wider range of markers, in a relatively shorter time and with more resolution to the analysis; can process up to 96 samples at once and could offer to combine different types of markers in a single experiment for maximum information gain with minimal sample usage (Figure 1.8). MPS techniques are able to sequence millions of molecules at a time and the generated data can provide from high to extremely high coverage of the regions examined, depending on the set-up and marker types; they are therefore able to generate very accurate consensus sequence data and resolution of variants beyond what is provided by CE analysis (Yang et al. 2014).

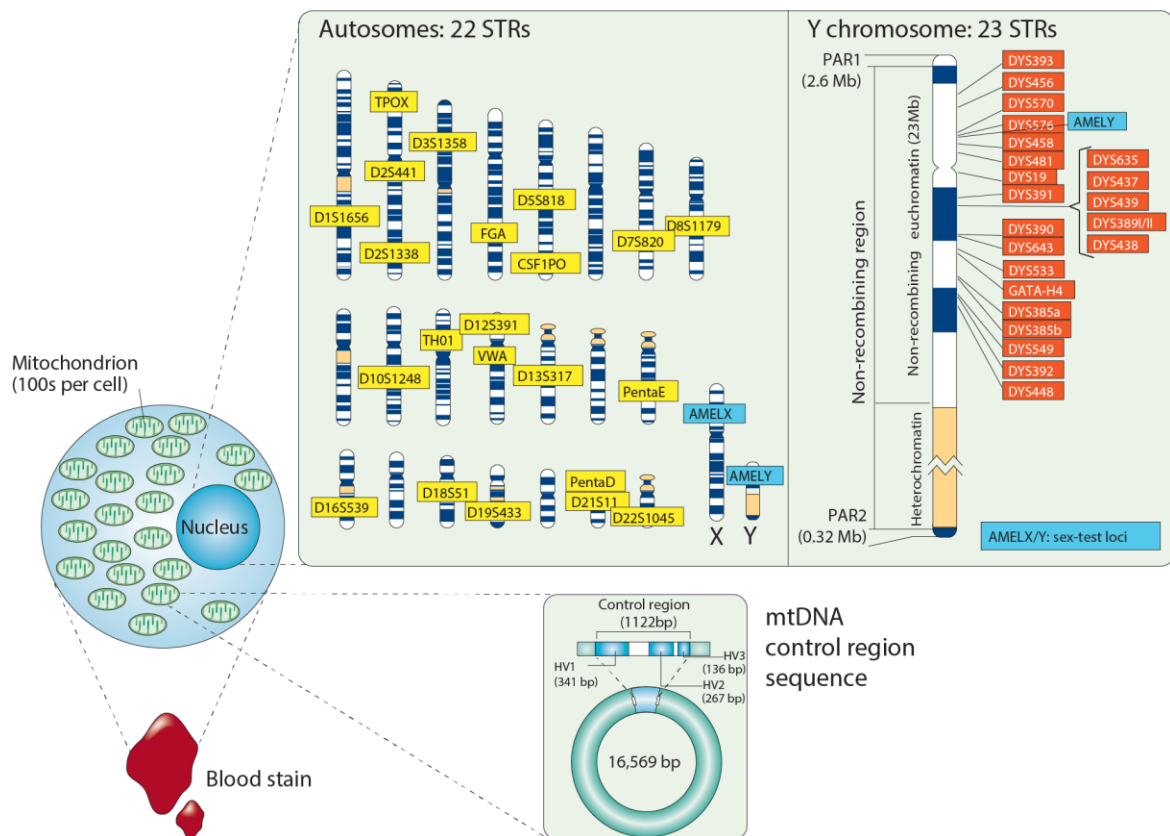


Figure 1.8
Forensic markers.

Forensic markers targeted in a single reaction in this study using the Promega PowerSeq™ Auto/Mito/Y System prototype MPS kit. Adapted from Jobling and Gill (2004).

MPS can distinguish between the alleles of isometric heterozygotes or isoalleles, alleles with the same length, but different sequence variants (Gettings et al. 2015). These allele types are shown as homozygotes by CE-typing, and therefore resolving them as separate alleles by their sequences can be particularly useful in mixture deconvolution, typing low-template DNA and degraded samples, or could help differentiate stutter peaks from minor contributors, when these are different sequence variants (Yang et al. 2014). In MPS, STRs of the autosomes and the Y

and X chromosomes can be sequenced in the same single reaction or combined with parts of the mtDNA (Warshauer et al. 2015b). While not every case requires such extensive analysis, it generally gives better value to have these generated at the same time, in the same workflow, from the same amount of sample; when a case requires lineage markers to be investigated by CE, these take up extra time and labour, incur further cost, and use up more of the sample (Willems et al. 2014; Zhao et al. 2015).

1.2.2.2 Improving STR typing

CE typing utilises only the size difference between STR alleles, while MPS analysis offers an insight into sequence level variation, which has the potential to increase the number of alleles observed at a locus, thus increasing power of discrimination for forensic markers. There has been a tendency worldwide to increase the number of STR loci, for example the introduction of the DNA-17 set of markers into the UK NDNAD (UK National DNA Database) in 2014 and the extension of CODIS (Combined DNA Index System) to store and compare 20 STR markers implemented in the US in 2017. These extensions were introduced to further decrease the probability of finding a random match to a profile, to compensate for some markers often left untyped in compromised-quality samples, and to aid international compatibility of profiles where different national databases were based on different marker sets. Using MPS can also potentially offer a similar increase in power of discrimination of already used markers without necessarily introducing new loci. MPS can offer better resolution of variants when comparing genotypes generated with this same technique; however, when comparing to previously CE-typed length-based profiles this advantage is lost. The sequenced alleles convey length information and thus MPS alleles are backward compatible and comparable to CE-derived alleles, but previously typed samples (if available), need to be retyped with MPS to make comparisons at the sequence level.

One of the common sources of length-based discordance between different kits (CE or MPS) used for typing STRs, is the presence of flanking region indels, due to

different primer positions, therefore whether the primers include flanking indels or not can have an effect on the length-based allele designations.

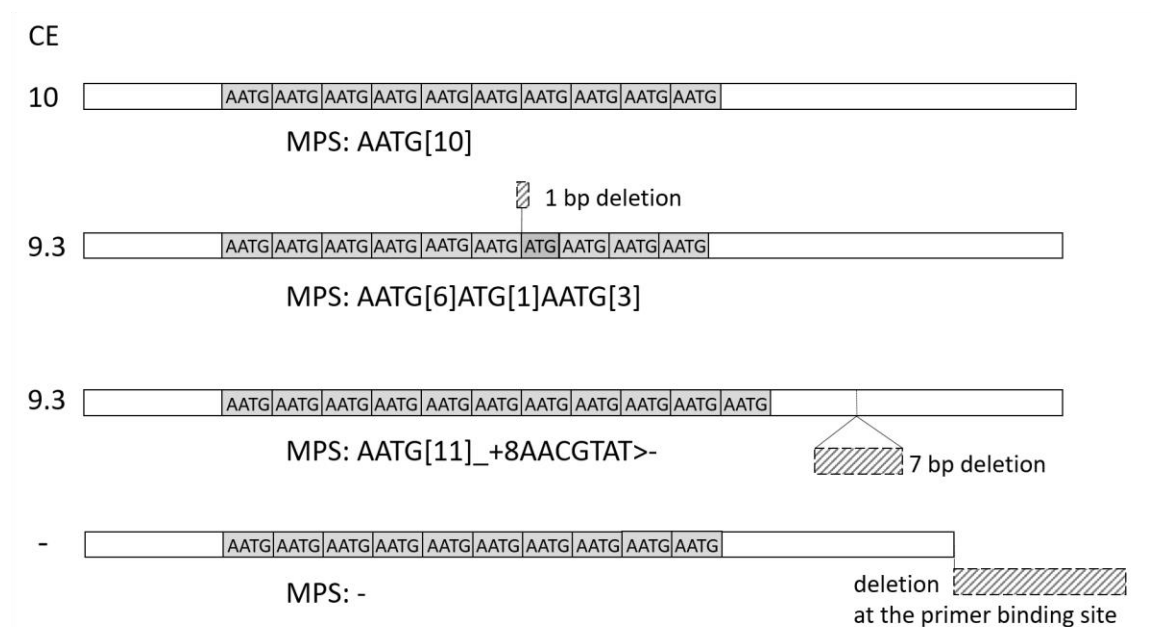


Figure 1.9
Effect of indels in an STR locus.

Indels internal to the array or in the flanking region may result in the same allele by length (here 9.3), but different by sequence. Deletions affecting the primer binding sites may result in failed amplification.

As an example, in Figure 1.9 an allele 9.3 for an AATG repeat locus could mean nine whole repeat units plus three remaining ATG nucleotides; however, it could also be an allele 9.3 but with 11 whole repeats and a 7-bp deletion in the flanking region enclosed by the primers; their lengths are identical, but the difference can only be observed by sequencing.

The positions of MPS primers are important, similarly to differences between different CE typing kits explained in Section 1.1.2.7. Variation in the positioning of primers within and between MPS- and CE-based detection kits from different suppliers can lead to them targeting different parts of the flanking region, potentially affecting the overall length-based allele designation if indels are present in or

around the repeat arrays. If in Figure 1.9 a primer were placed between the 7-bp flanking indel and the array, the length of the allele would be typed as 11, rather than 9.3, which could create a discordance with primers placed further 3' including the indel, resulting in a 9.3 length-based allele as shown in the Figure. Unfortunately, when indels affect the primer binding sites this can result in unsuccessful amplification in both CE and MPS kits.

In conclusion, sequence-level differences and the exact source of intermediate variant alleles remain hidden with length-based detection; however, sequencing of STRs resolves these differences and help to identify where indels occurred within an amplified region. Sequencing can also resolve occasional discordant results and their origins between different kits.

1.2.2.3 Forensic use of MPS beyond STRs

MPS can also reveal information on identity informative SNPs in the same single reaction, as in the ForenSeq™ DNA Signature Prep Kit of Verogen. Although these are not usually asked for in a case, typing these SNPs together with the STRs can be useful for providing a way to identify heavily decomposed remains, when STRs fail due to the level of degradation affecting the DNA. Specific SNPs can also offer investigative leads by the prediction of possible phenotypic traits or biogeographical origin of a person of interest.

1.2.2.4 MPS options and challenges of implementation

Commercial companies are offering different platforms that are adapted to forensic use: Verogen offers a fully validated ready-to-use system, the MiSeq FGx® Forensic Genomics System, and Thermo Fisher offers its Ion Chef™ and Ion S5™ platform.

Some countries are in the process of implementing MPS to supplement CE-based analysis. Currently a limitation to the use of MPS platforms in most countries is legislative, as this type of data needs to be considered and approved ethically, and its use be validated approved and regulated. The change in technology is also

inhibited by the logistical adjustment and cost that is required from laboratories using CE to implement MPS and its specialist equipment into their workflows, and also by the lack of large databases that can accept, store and compare sequence-based designations (de Knijff 2019b).

To facilitate the transition from length-based to sequenced STR alleles a unified nomenclature would be an ideal solution. However, this is not available to date, in spite of suggestions and considerations on the matter (Bodner et al. 2016; Gettings et al. 2015; Parson et al. 2016; Rajeevan et al. 2012; Warshauer et al. 2015b; Willems et al. 2014; Zhao et al. 2015). Currently there is no unified consensus or agreement as to how to extend the current system so it can capture, analyse and report the sequence variants in an unambiguous manner. In 2019 a meeting of stakeholders facilitated by the STRAND ('Short Tandem Repeat: Align, Name, Define') working group set the suggested formats for reporting sequenced alleles and highlighted the potential pitfalls based on user experiences (Gettings et al. 2019). A harmonised way of reporting is essential for comparison of results between different platforms and different laboratories.

1.2.2.5 Latest developments in MPS and its future use in the UK

In 2019 the first criminal conviction using MPS-derived data was secured in the Netherlands in a sexual assault case (<https://verogen.com/news/first-criminal-conviction-with-next-gen-forensic-dna/>; de Knijff 2019a), and in the US the FBI approved the first products that can be used to generate MPS-derived data that can be uploaded to the National DNA Index System (NDIS).

In the UK, MPS is currently considered only for providing investigative leads; one of the largest Forensic Service Providers (FSPs), Cellmark Forensic Services, is set to offer MPS-based analysis for the UK police forces (<https://www.cellmarkforensics.co.uk/news/next-generation-forensic-dna-analysis>). To introduce this technique to routine casework, population databases are needed which are tailored to the UK and its main regions and ethnic groups, which would be able to highlight any substructure that may affect the prevalence of

certain variants according to their ancestry. These would be able to give statistical weighting to the generated profiles, as do currently available CE-based STR allele frequency and haplotype frequency databases. For the UK, Devesse et al. (2018) generated sequence-based autosomal STR population data representing White British and British Chinese populations within the UK. In this work this is supplemented with autosomal and Y-STR data, as well as mtDNA control region information for 362 samples across the British Isles, a sample set selected for minimal admixture, represented by donors whose ancestors were living in rural areas within 50 miles from one another in the last two generations.

1.3 Aims and objectives

- To use MPS to analyse Y-STRs in a diverse set of samples based on a known phylogeny of Y chromosomes in order to capture a wide variety of Y-STR sequences and place these in an evolutionary context
 - To use the PowerSeq™ Auto/Mito/Y System kit to generate data for 23 Y-STRs in a panel of Y chromosomes whose phylogenetic relationships are known from high-resolution sequencing
 - To catalogue alleles and compare these with known Y-STR variants
 - To consider repeat-motif and flanking sequence variation within the context of phylogenetic relationships and mutation rates of different variant types
 - To ask if rare repeat-region variants affect the nomenclature of allele types
- To use MPS to analyse mtDNA control region diversity in the same samples and ask how mtDNA variant typing behaves within the same workflow.
 - To use the PowerSeq™ Auto/Mito/Y System kit to generate sequence data for the mtDNA control region in a panel of highly diverse global samples
 - To assess the phylogenetic diversity in the sample set

- To consider how single-multiplex typing affects the reliability of variant calling and the assessment of heteroplasmy
- To use MPS to analyse autosomal STRs in the same samples
 - To use the PowerSeq™ Auto/Mito/Y System kit to generate data for 22 autosomal STRs in a panel of highly diverse global samples
 - To catalogue alleles and compare these with known autosomal STR variation
- To use the MPS methods described above to analyse sequences for forensic markers in a set of samples from the British Isles
 - To use the PowerSeq™ Auto/Mito/Y System kit to generate sequence data for Y-STRs, autosomal STRs and the mtDNA control region in a panel of indigenous samples from the People of the British Isles cohort
 - To catalogue Y-STR and autosomal STR alleles
 - To use population genetic methods to compare the geographical distributions of Y, autosomal and mitochondrial diversity within the British Isles, and ask if any effects of sex-biased historical processes can be observed
 - To provide a useful dataset of well-characterised indigenous samples for reference use.

CHAPTER 2: Materials and methods

2.1 Sample sets

This study uses two sample sets, each selected from previously published larger collections of samples. The first set represents global variation to maximise the captured sequence variability; and the second set represents the indigenous United Kingdom (UK) population.

2.1.1 Samples of global variation

One hundred male DNAs were selected to represent global variation from the 448 samples from worldwide populations analysed by Hallast et al. (2015), who used sequence capture of 3.7 Mb of DNA from the male-specific region of the Y-chromosome to identify 13,261 SNPs and form a maximum parsimony tree rooted using great ape sequences. They also typed the same samples by CE for the 23 Y-STRs with the Promega PowerPlex® Y23 kit.

The 100 samples for the current study were selected by ensuring that every major clade and deep-rooting node of the Y-chromosomal phylogeny was represented to maximise the captured Y-STR sequence variation. Figure 2.1 shows the geographic distribution of these samples and Table 2.1 lists these samples, detailing their use for each marker analysis. Sample information can be viewed interactively at this Microreact project (Argimon et al. 2016; <https://microreact.org/project/U91JiX69f>).



Figure 2.1
Geographical distribution of the 100 samples.

The sample proportional map is generated using Microreact (Argimon et al. 2016).

Table 2.1
Selection of the 100 global samples.

List of the selected 100 samples with their geographical origins and previously published (Hallast et al. 2015) names.

Sample name	Population	Metapopulation	Sample name	Population	Metapopulation
aus-m119	Australian	Australian	JPT-NA18940	Japanese (JPT)	Asian
bak-25	Baka	African	JPT-NA18974	Japanese (JPT)	Asian
bak-33	Baka	African	kun-m82	!Kung	African
bak-35	Baka	African	LWK-NA19334	Luhya (LWK)	African
bak-41	Baka	African	man-231	Mandara	African
bak-55	Baka	African	mbe-237P	Mbenzele	African
bas-19	Basque	European	mbu-13	Mbuti	African
bav-13	Bavarian	European	mbu-29	Mbuti	African
bav-55	Bavarian	European	mbu-33	Mbuti	African
bhu-0957	Bhutanese	Asian	mbu-5	Mbuti	African
bhu-0984	Bhutanese	Asian	MXL-NA19664	Mexican (MXL)	American
bhu-1000	Bhutanese	Asian	nep-0171	Nepalese	Asian
bhu-1142	Bhutanese	Asian	nep-0172	Nepalese	Asian
bhu-1150	Bhutanese	Asian	nep-0186	Nepalese	Asian
bhu-1151	Bhutanese	Asian	nep-0273	Nepalese	Asian
bhu-1157	Bhutanese	Asian	nep-0387	Nepalese	Asian
bhu-1564	Bhutanese	Asian	nep-0809	Nepalese	Asian
bhu-1611	Bhutanese	Asian	nep-0902	Nepalese	Asian
bhu-1892	Bhutanese	Asian	ngo-98	Ngoumba	African
bkl-46	Bakola	African	nor-16	Norwegian	European
CEU-NA06994	French (CEU)	European	ork-007	Orcadian	European
CEU-NA11829	French (CEU)	European	ork-026m	Orcadian	European
CEU-NA11992	French (CEU)	European	ork-036m	Orcadian	European
CEU-NA12003	French (CEU)	European	pal-4919	Palestinian	Near and Middle Eastern
CEU-NA12716	French (CEU)	European	pal-4922	Palestinian	Near and Middle Eastern
CHB-NA18558	Chinese (CHB)	Asian	pal-4929	Palestinian	Near and Middle Eastern
CHB-NA18561	Chinese (CHB)	Asian	pal-5225	Palestinian	Near and Middle Eastern
CHB-NA18608	Chinese (CHB)	Asian	pal-5232	Palestinian	Near and Middle Eastern
CHB-NA18612	Chinese (CHB)	Asian	pal-5366	Palestinian	Near and Middle Eastern
CHB-NA18620	Chinese (CHB)	Asian	saa-15	Saami	European
CHB-NA18621	Chinese (CHB)	Asian	ser-19	Serbian	European
CHB-NA18622	Chinese (CHB)	Asian	ser-21	Serbian	European
CHB-NA18632	Chinese (CHB)	Asian	spa-20	Spanish	European
CHB-NA18637	Chinese (CHB)	Asian	TSI-NA20510	Italian (TSI)	European
den-190	Danish	European	TSI-NA20527	Italian (TSI)	European
eng-GB1778	English	European	TSI-NA20543	Italian (TSI)	European
eng-hgQ-1	English	European	TSI-NA20588	Italian (TSI)	European
eng-hgQ-2	English	European	TSI-NA20805	Italian (TSI)	European
gre-10	Greek	European	tur-1	Turkish	Near and Middle Eastern
gre-12	Greek	European	tur-13	Turkish	Near and Middle Eastern
gre-192	Greek	European	tur-16	Turkish	Near and Middle Eastern
gre-2	Greek	European	tur-4	Turkish	Near and Middle Eastern
gre-4	Greek	European	tur-7	Turkish	Near and Middle Eastern
gre-77	Greek	European	tur-9	Turkish	Near and Middle Eastern
hun-29	Hungarian	European	YRI-NA18504	Yoruba (YRI)	African
hun-47	Hungarian	European	YRI-NA18507	Yoruba (YRI)	African
hun-5	Hungarian	European	YRI-NA18853	Yoruba (YRI)	African
ire-0068	Irish	European	YRI-NA18856	Yoruba (YRI)	African
ire-0114	Irish	European	YRI-NA19098	Yoruba (YRI)	African
ire-94	Irish	European	YRI-NA19175	Yoruba (YRI)	African

2.1.2 Samples from the People of the British Isles

The samples representing the UK were selected from the males of the People of the British Isles (PoBI) sample set (Leslie et al. 2015; Winney et al. 2012) provided by Prof Sir Walter Bodmer from the Weatherall Institute of Molecular Medicine, at the John Radcliffe Hospital, Oxford, UK. The People of the British Isles project (Leslie et al. 2015; Winney et al. 2012) captured the autosomal genetic variation of rural areas of the UK, based on SNP chip analysis of 2039 individuals. The PoBI set aimed to represent local ancestry, relatively unaffected by recent migrations, by recruiting individuals with ancestors known to have inhabited the same local areas for at least two generations.

Limited by available sequencing capacity for the project, it was not possible to analyse all PoBI samples by MPS, therefore selection of a subset was required. The aim was to sample across all 45 PoBI regions available, planning to use a maximum of ten samples per region. All PoBI samples were from donors for whom all four grandparents were known to be local, defined by the mean of their birthplaces falling within a 50-mile radius. To emphasise the importance of Y chromosome locality, the selection was aimed especially at samples with known local paternal grandfathers. When the number of available samples meeting this criteria exceeded ten, priority was given to those for whom typed SNP array information was available. When all these criteria were found to be equal, samples were selected randomly.

Three hundred and eighty-three males were chosen by this process originally; however, five individuals were shown to be female based on genotypes at Amelogenin and the Y-STRs, and a further sixteen were excluded from analysis due to overall low quality and degradation affecting the recovered data. This led to 362 individuals being analysed for Y-STR markers and mtDNA control region variants. One further individual was excluded from autosomal STR analysis due to lower read counts potentially affecting accurate genotyping of these diploid markers, and therefore 361 samples were eventually analysed for autosomal STR

markers. Figure 2.2 shows the sample-sizes in proportion and their geographic distribution across the British Isles. Table 2.2 lists the samples and their populations of origin with the main geographic regions noted as SCO: Scotland; IRE; Ireland; IOM: Isle of Man; NW: North West; NE: North East; WAL: Wales, WM: West Midlands; EM: East Midlands; EA: East; SW: South West; SE: South East. Sample information can be viewed interactively at this Microreact project (Argimon et al. 2016; <https://microreact.org/project/jlcJm9Kgn>).



Figure 2.2
Geographical distribution of the 362 PoBI samples.

The sample map is generated using Microreact (Argimon et al. 2016).

Table 2.2
Samples selected from the PoBI study.

*The 362 samples analysed for Y- and autosomal STRs and the mtDNA control region.
The sample marked '*' was not analysed for autosomal STRs.*

PoBI ID	Population	Region	PoBI ID	Population	Region
ARG002	Argyll and Bute	SCO	CHE036	Cheshire	NW
ARG025	Argyll and Bute	SCO	CHE503	Cheshire	NW
ARG026	Argyll and Bute	SCO	CHE507	Cheshire	NW
ARG028	Argyll and Bute	SCO	CHE511	Cheshire	NW
ARG032	Argyll and Bute	SCO	CHE512	Cheshire	NW
ARG051	Argyll and Bute	SCO	COR012	Cornwall	SW
ARG053	Argyll and Bute	SCO	COR040	Cornwall	SW
ARG308	Argyll and Bute	SCO	COR043	Cornwall	SW
ARG309	Argyll and Bute	SCO	COR048	Cornwall	SW
BAN008	Banff and Buchan	SCO	COR055	Cornwall	SW
BAN012	Banff and Buchan	SCO	COR057	Cornwall	SW
BAN501	Banff and Buchan	SCO	COR093	Cornwall	SW
BAN517	Banff and Buchan	SCO	COR096	Cornwall	SW
BAN526	Banff and Buchan	SCO	COR243	Cornwall	SW
BAN529	Banff and Buchan	SCO	COR260	Cornwall	SW
BAN539	Banff and Buchan	SCO	CUM014	Cumbria	NW
BAN546	Banff and Buchan	SCO	CUM056	Cumbria	NW
BAN549	Banff and Buchan	SCO	CUM065	Cumbria	NW
BAN551	Banff and Buchan	SCO	CUM128	Cumbria	NW
BED001	Bedfordshire	EA	CUM151	Cumbria	NW
BER001	Berkshire	SE	CUM210	Cumbria	NW
BER002	Berkshire	SE	CUM242	Cumbria	NW
BER004	Berkshire	SE	CUM386	Cumbria	NW
BER014	Berkshire	SE	CUM525	Cumbria	NW
BER015	Berkshire	SE	CUM582	Cumbria	NW
BER022	Berkshire	SE	DER502	Derbyshire	EM
BUC001	Buckinghamshire	SE	DER508	Derbyshire	EM
CAM502	Cambridgeshire	EA	DER513	Derbyshire	EM
CAM507	Cambridgeshire	EA	DER515	Derbyshire	EM
CAM509	Cambridgeshire	EA	DER516	Derbyshire	EM
CAM511	Cambridgeshire	EA	DEV011	Devon	SW
CHE004	Cheshire	NW	DEV014	Devon	SW
CHE010	Cheshire	NW	DEV056	Devon	SW
CHE018	Cheshire	NW	DEV066	Devon	SW
CHE022	Cheshire	NW	DEV084	Devon	SW
CHE034	Cheshire	NW	DEV095	Devon	SW

cont.

PoBI ID	Population	Region
DEV501	Devon	SW
DEV504	Devon	SW
DEV508	Devon	SW
DEV513	Devon	SW
DOR002	Dorset	SW
DOR008	Dorset	SW
DOR012	Dorset	SW
DOR015	Dorset	SW
DOR021	Dorset	SW
DOR025	Dorset	SW
DOR028	Dorset	SW
DOR029	Dorset	SW
DOR033	Dorset	SW
ESS001	Essex	EA
ESS002	Essex	EA
ESS003	Essex	EA
ESS012	Essex	EA
FOD014	Forest of Dean	SW
FOD022	Forest of Dean	SW
FOD035	Forest of Dean	SW
FOD047	Forest of Dean	SW
FOD049	Forest of Dean	SW
FOD054	Forest of Dean	SW
FOD057	Forest of Dean	SW
FOD061	Forest of Dean	SW
FOD107	Forest of Dean	SW
GLO004	Gloucestershire	SW
GLO005	Gloucestershire	SW
GLO015	Gloucestershire	SW
GLO027	Gloucestershire	SW
GLO035	Gloucestershire	SW
GLO038	Gloucestershire	SW
GLO044	Gloucestershire	SW
GLO047	Gloucestershire	SW
GLO049	Gloucestershire	SW
GLO051	Gloucestershire	SW
HAM004	Hampshire	SE
HAM005	Hampshire	SE
HAM006	Hampshire	SE
HAM010	Hampshire	SE
HAM088	Hampshire	SE
HAM100	Hampshire	SE

PoBI ID	Population	Region
HAM112	Hampshire	SE
HAM126	Hampshire	SE
HAM130	Hampshire	SE
HAM136	Hampshire	SE
HER003	Herefordshire	WM
HER017	Herefordshire	WM
HER027	Herefordshire	WM
HER038	Herefordshire	WM
HER051	Herefordshire	WM
HER060	Herefordshire	WM
IOM023	Isle of Man	IOM
IOM049	Isle of Man	IOM
IOM075	Isle of Man	IOM
IOM079	Isle of Man	IOM
IOM084	Isle of Man	IOM
IOM091	Isle of Man	IOM
IOM094	Isle of Man	IOM
IOM111	Isle of Man	IOM
KEN004	Kent	SE
KEN024	Kent	SE
KEN039	Kent	SE
KEN051	Kent	SE
KEN055	Kent	SE
KEN062	Kent	SE
KEN077	Kent	SE
KEN082	Kent	SE
KEN084	Kent	SE
KEN099	Kent	SE
LAN009	Lancashire	NW
LAN010	Lancashire	NW
LAN502	Lancashire	NW
LAN516	Lancashire	NW
LAN518	Lancashire	NW
LAN521	Lancashire	NW
LAN527	Lancashire	NW
LAN531	Lancashire	NW
LAN540	Lancashire	NW
LAN545	Lancashire	NW
LEI002	Leicestershire	EM
LEI007	Leicestershire	EM
LEI025	Leicestershire	EM
LEI028	Leicestershire	EM

cont.

PoBI ID	Population	Region	PoBI ID	Population	Region
LEI041	Leicestershire	EM	NIR092	Northern Ireland	IRE
LEI042	Leicestershire	EM	NOR002	Norfolk	EA
LEI046	Leicestershire	EM	NOR094	Norfolk	EA
LEI050	Leicestershire	EM	NOR149	Norfolk	EA
LEI066	Leicestershire	EM	NOR163	Norfolk	EA
LEI067	Leicestershire	EM	NOR164	Norfolk	EA
LIN009	Lincolnshire	EM	NOR181	Norfolk	EA
LIN508	Lincolnshire	EM	NOR193	Norfolk	EA
LIN650	Lincolnshire	EM	NOR196	Norfolk	EA
LIN653	Lincolnshire	EM	NOR197	Norfolk	EA
LIN679	Lincolnshire	EM	NOR524	Norfolk	EA
LIN700	Lincolnshire	EM	NOT020	Nottinghamshire	EM
LIN703	Lincolnshire	EM	NOT021	Nottinghamshire	EM
LIN719	Lincolnshire	EM	NOT028	Nottinghamshire	EM
LIN730	Lincolnshire	EM	NOT039*	Nottinghamshire	EM
LIN733	Lincolnshire	EM	NOT061	Nottinghamshire	EM
MWA009	Mid Wales	WAL	NOT071	Nottinghamshire	EM
MWA053	Mid Wales	WAL	NOT078	Nottinghamshire	EM
MWA068	Mid Wales	WAL	NOT086	Nottinghamshire	EM
MWA072	Mid Wales	WAL	NOT096	Nottinghamshire	EM
MWA093	Mid Wales	WAL	NPE005	North Pembrokeshire	WAL
MWA127	Mid Wales	WAL	NPE012	North Pembrokeshire	WAL
MWA129	Mid Wales	WAL	NPE023	North Pembrokeshire	WAL
MWA133	Mid Wales	WAL	NPE029	North Pembrokeshire	WAL
MWA139	Mid Wales	WAL	NPE049	North Pembrokeshire	WAL
NEA059	North East	NE	NPE051	North Pembrokeshire	WAL
NEA090	North East	NE	NPE055	North Pembrokeshire	WAL
NEA144	North East	NE	NPE056	North Pembrokeshire	WAL
NEA165	North East	NE	NPE062	North Pembrokeshire	WAL
NEA197	North East	NE	NPE064	North Pembrokeshire	WAL
NEA295	North East	NE	NTH007	Northamptonshire	EM
NEA506	North East	NE	NTH019	Northamptonshire	EM
NEA520	North East	NE	NTH032	Northamptonshire	EM
NIR005	Northern Ireland	IRE	NTH042	Northamptonshire	EM
NIR010	Northern Ireland	IRE	NTH044	Northamptonshire	EM
NIR022	Northern Ireland	IRE	NTH045	Northamptonshire	EM
NIR067	Northern Ireland	IRE	NTH067	Northamptonshire	EM
NIR073	Northern Ireland	IRE	NTH076	Northamptonshire	EM
NIR074	Northern Ireland	IRE	NTH080	Northamptonshire	EM
NIR075	Northern Ireland	IRE	NTH081	Northamptonshire	EM
NIR088	Northern Ireland	IRE	NWA010	North Wales	WAL
NIR091	Northern Ireland	IRE	NWA015	North Wales	WAL

cont.

PoBI ID	Population	Region	PoBI ID	Population	Region
NWA016	North Wales	WAL	SHR503	Shropshire	WM
NWA061	North Wales	WAL	SPE007	South Pembrokeshire	WAL
NWA070	North Wales	WAL	SPE010	South Pembrokeshire	WAL
NWA083	North Wales	WAL	SPE023	South Pembrokeshire	WAL
NWA090	North Wales	WAL	SPE025	South Pembrokeshire	WAL
NWA111	North Wales	WAL	SPE045	South Pembrokeshire	WAL
NWA121	North Wales	WAL	SPE053	South Pembrokeshire	WAL
NWA126	North Wales	WAL	SPE062	South Pembrokeshire	WAL
ORK503	Orkney	SCO	STA504	Staffordshire	WM
ORK508	Orkney	SCO	STA505	Staffordshire	WM
ORK515	Orkney	SCO	STA510	Staffordshire	WM
ORK529	Orkney	SCO	STA512	Staffordshire	WM
ORK540	Orkney	SCO	STA513	Staffordshire	WM
ORK544	Orkney	SCO	STA517	Staffordshire	WM
ORK549	Orkney	SCO	STA522	Staffordshire	WM
ORK550	Orkney	SCO	STA524	Staffordshire	WM
ORK555	Orkney	SCO	STA526	Staffordshire	WM
ORK584	Orkney	SCO	SUF049	Suffolk	EA
OXF021	Oxfordshire	SE	SUF073	Suffolk	EA
OXF044	Oxfordshire	SE	SUF079	Suffolk	EA
OXF050	Oxfordshire	SE	SUF163	Suffolk	EA
OXF084	Oxfordshire	SE	SUF174	Suffolk	EA
OXF130	Oxfordshire	SE	SUF181	Suffolk	EA
OXF160	Oxfordshire	SE	SUF217	Suffolk	EA
OXF168	Oxfordshire	SE	SUF249	Suffolk	EA
OXF170	Oxfordshire	SE	SUF251	Suffolk	EA
OXF174	Oxfordshire	SE	SUF512	Suffolk	EA
OXF175	Oxfordshire	SE	SUS014	Sussex	SE
RIR003	Republic of Ireland	IRE	SUS019	Sussex	SE
RIR006	Republic of Ireland	IRE	SUS026	Sussex	SE
RIR015	Republic of Ireland	IRE	SUS027	Sussex	SE
RIR016	Republic of Ireland	IRE	SUS034	Sussex	SE
SCO008	Scotland	SCO	SUS035	Sussex	SE
SCO010	Scotland	SCO	SUS421	Sussex	SE
SCO017	Scotland	SCO	SUS431	Sussex	SE
SCO043	Scotland	SCO	SUS443	Sussex	SE
SCO045	Scotland	SCO	SUS469	Sussex	SE
SCO047	Scotland	SCO	WAL001	Wales	WAL
SCO050	Scotland	SCO	WAL008	Wales	WAL
SCO062	Scotland	SCO	WAL011	Wales	WAL
SCO065	Scotland	SCO	WAL012	Wales	WAL
SHR501	Shropshire	WM	WAL023	Wales	WAL

cont.

PoBI ID	Population	Region
WAL042	Wales	WAL
WAL043	Wales	WAL
WAL046	Wales	WAL
WAL048	Wales	WAL
WAL049	Wales	WAL
WAR002	Warwickshire	WM
WAR004	Warwickshire	WM
WAR005	Warwickshire	WM
WAR015	Warwickshire	WM
WIL002	Wiltshire	SW
WIL007	Wiltshire	SW
WIL009	Wiltshire	SW
WIL011	Wiltshire	SW
WIL014	Wiltshire	SW
WIL017	Wiltshire	SW
WIL023	Wiltshire	SW
WIL025	Wiltshire	SW
WIL027	Wiltshire	SW
WIL030	Wiltshire	SW

PoBI ID	Population	Region
WOR011	Worcestershire	WM
WOR013	Worcestershire	WM
WOR018	Worcestershire	WM
WOR023	Worcestershire	WM
WOR034	Worcestershire	WM
WOR042	Worcestershire	WM
WOR046	Worcestershire	WM
YOR008	Yorkshire	NE
YOR069	Yorkshire	NE
YOR097	Yorkshire	NE
YOR129	Yorkshire	NE
YOR139	Yorkshire	NE
YOR202	Yorkshire	NE
YOR221	Yorkshire	NE
YOR227	Yorkshire	NE
YOR564	Yorkshire	NE
YOR599	Yorkshire	NE
RAN138	Buckinghamshire	SE
RAN147	Shropshire	WM

2.2 Laboratory experiments

2.2.1 Quantification of DNA

Quantities of double-stranded DNA were measured using a Qubit® 2.0 fluorometer (Thermo Fisher Scientific) with the Qubit® dsDNA HS (high sensitivity) assay kit and the dsDNA BR (broad range) assay kit following the manufacturer's recommended protocol. These assays are highly selective for double-stranded DNA over the ranges of 10 pg/µl – 100 ng/µl (HS) and 100 pg/µl – 1000 ng/µl (BR). The HS kit was used primarily to quantify DNA samples, while the BR kit was used mainly to quantify amplification products post-PCR during library preparation.

2.2.2 Sample preparation specific to Promega PowerSeq™ Auto/Mito/Y System prototype

2.2.2.1 Fragment generation by PCR

Prior to any library preparation, fragments of the desired size need to be generated; in this study, the fragments were generated by PCR. Targeted STRs and the control region of the mtDNA were amplified in a single multiplex using the Promega PowerSeq™ Auto/Mito/Y System prototype reagents following the manufacturer's recommended protocol (Promega 2015) adding 0.5 ng genomic DNA as template to the multiplex PCR reaction that contained a buffered master mix and primer pair mix for the amplification of the loci listed in Table 2.3 below.

Table 2.3**Details of the STR loci and the mtDNA regions amplified.**

Fragment size ranges of the STR loci and the mitochondrial regions amplified in the Promega PowerSeq™ Auto/Mito/Y System prototype multiplex.

Adapted from Promega (2015)

aSTRs		Y-STRs		mtDNA control region		
Locus	Size(bp)	Locus	Size(bp)	Region	Size(bp)	Amplified region
CSF1PO	185–229	DYS19	168–204	F15989M	164	15989–16152
D10S1248	135–179	DYS385a,b	202–303	F16094	155	16094–16248
D12S391	202–254	DYS389I/II	258–294	F16197	237	16197–16433
D13S317	209–257	DYS390	204–248	F16363	147	16363–16509
D16S539	198–246	DYS391	147–178	F16450	172	16450–52
D18S51	190–277	DYS392	143–164	F16533	217	16533–180
D19S433	193–253	DYS393	224–256	F109	185	109–293
D1S1656	161–208	DYS437	181–197	F220M	170	220–389
D21S11	203–273	DYS438	202–242	F317	144	317–460
D22S1045	129–176	DYS439	204–224	F402	218	402–619
D2S1338	197–269	DYS448	213–255			
D2S441	168–204	DYS456	141–165			
D3S1358	192–240	DYS458	171–199			
D5S818	191–239	DYS481	139–184			
D7S820	211–255	DYS533	242–284			
D8S1179	203–255	DYS549	189–230			
FGA	176–268	DYS570	157–217			
PentaD	192–265	DYS576	155–203			
PentaE	179–284	DYS635	155–179			
TH01	220–264	DYS643	150–210			
TPOX	196–244	Y-GATA-H4	231–251			
vWA	202–262					

sex typing	
Locus	Size(bp)
Amelogenin	192,198

Further information about the STR loci tested, their structure and genomic positions, is given in Table 2.4.

Table 2.4
Repeat motifs and genomic positions of the nuclear genomic loci amplified.

aSTRs		GRCh38 reference primary assembly			
Locus	Main repeats	GenBank ID	Chromosome	Amplified region	Allele in GRCh38
CSF1PO	(ATCT) _n	CM000667.2	chr5	150076289-150076505	13
D10S1248	(GGAA) _n	CM000672.2	chr10	129294165-129294319	13
D12S391	(AGAT) _n (AGAC) _o (AGAT) _p	CM000674.2	chr12	12296979-12297129	19
D13S317	(TATC) _n	CM000675.2	chr13	82147933-82148165	12
D16S539	(GATA) _n	CM000678.2	chr16	86352547-86352772	11
D18S51	(AGAA) _n	CM000680.2	chr18	63281581-63281814	18
D19S433	(CCTT) _n ccta CCTT cttt CCTT	CM000681.2	chr19	29926137-29926360	16
D151656	(TCTA) _n	CM000663.2	chr1	230769538-230769727	16
D21S11	(TCTA) _m (TCTG) _n (TCTA) _o ta (TCTA) _p tca (TCTA) _q tccata (TCTA) _r	CM000683.2	chr21	19181941-19182163	29
D22S1045	(ATT) _n ACT (ATT) _o	CM000684.2	chr22	37140213-37140371	17
D2S1338	(GGAA) _n (GGCA) _o	CM000664.2	chr2	218014771-218015019	23
D2S441	(TCTA) _n	CM000664.2	chr2	68011842-68012025	12
D3S1358	TCTA (TCTG) _m (TCTA) _n	CM000665.2	chr3	45540620-45540843	16
D5S818	(ATCT) _n	CM000667.2	chr5	123775494-123775704	11
D7S820	(TATC) _n	CM000669.2	chr7	84160125-84160367	13
D8S1179	(TCTA) _n	CM000670.2	chr8	124894824-124895047	13
FGA	(GGAA) _m GGAG (AAAG) _n AGAA AAAA (GAAA) _o	CM000666.2	chr4	154587633-154587846	22
PentaD	(AAAGA) _n	CM000683.2	chr21	43636113-43636357	13
PentaE	(TCTTT) _n	CM000677.2	chr15	96830885-96831063	5
TH01	(AATG) _n	CM000673.2	chr11	2171059-2171294	7
TPOX	(AATG) _n	CM000664.2	chr2	1489510-1489721	8
vWA	(TAGA) _n (CAGA) _o (TAGA) _p	CM000674.2	chr12	5983938-5984167	17

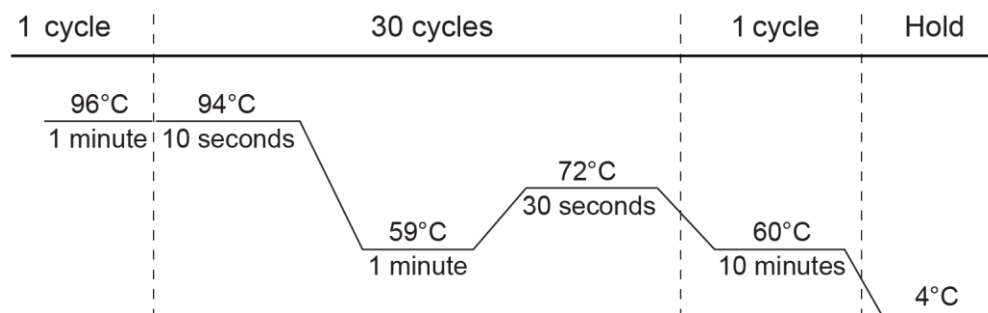
		GRCh38 reference primary assembly			
Locus	Motif	GenBank ID	Chromosome	Amplified region	Allele in GRCh38
Amelogenin X	6 bp deletion	CM000685.2	chrX	11296795-11296986	X
Amelogenin Y	-	CM000686.2	chrY	6869840-6870037	Y

Y-STRs		GRCh38 reference primary assembly			
Locus	Main repeats	GenBank ID	Chromosome	Amplified region	Allele in GRCh38
DYS19	(TCTA) _n ccta (TCTA) _o	CM000686.2	chrY	9684333-9684526	15
DYS385a	(AAGG) ₆ (GAAA) _n	CM000686.2	chrY	18639681-18639915	11
DYS385b	(AAGG) ₆ (GAAA) _n	CM000686.2	chrY	18680473-18680719	14
DYS389I	(TCTG) _o (TCTA) _p	CM000686.2	chrY	12500370-12500523	12
DYS389II	(TCTG) _m (TCTA) _n n ₄₈ (TCTG) _o (TCTA) _p	CM000686.2	chrY	12500370-12500643	29
DYS390	(TAGA) _n (CAGA) _o (TAGA) _p (CAGA) _q (TAGA) _r	CM000686.2	chrY	15163020-15163251	24
DYS391	(TCTA) _n	CM000686.2	chrY	11982052-11982218	11
DYS392	(ATA) _n	CM000686.2	chrY	20471962-20472125	13
DYS393	(AGAT) _n	CM000686.2	chrY	3263088-3263205	12
DYS437	(TCTA) _n (TCTG) _o (TCTA) _p	CM000686.2	chrY	12346237-12346424	15
DYS438	(TTTTC) _n	CM000686.2	chrY	12825857-12826006	12
DYS439	(GATA) _n	CM000686.2	chrY	12403484-12403609	12
DYS448	(AGAGAT) _n n ₄₂ (AGAGAT) _o	CM000686.2	chrY	22218890-22219114	19
DYS456	(AGAT) _n	CM000686.2	chrY	4402899-4403052	15
DYS458	(GAAA) _n	CM000686.2	chrY	7999799-7999981	16
DYS481	(CTT) _n	CM000686.2	chrY	8558282-8558432	22
DYS533	(TATC) _n	CM000686.2	chrY	16281222-16281486	12
DYS549	(GATA) _n	CM000686.2	chrY	19358203-19358424	13
DYS570	(TTTTC) _n	CM000686.2	chrY	6993116-6993287	17
DYS576	(AAAG) _n	CM000686.2	chrY	7185294-7185472	17
DYS635	(TAGA) _n (TACA) _o (TAGA) _p (TACA) _q (TCTA) _r	CM000686.2	chrY	12258797-12258975	23
DYS643	(CTTTT) _n	CM000686.2	chrY	15314051-15314249	11
Y-GATA-H4	(TCTA) _n	CM000686.2	chrY	16631648-16631894	12

Amplification was performed using a DNA Engine Tetrad2 (MJ Research), and the parameters used are shown in Table 2.5.

Table 2.5
Thermocycling parameters used with the Promega PowerSeq™ Auto/Mito/Y System prototype multiplex.

The steps of amplification follow manufacturer's recommendations. Adapted from (Promega 2015)



2.2.2.2 Purification of PCR products

Amplified products were purified using the MinElute® PCR purification kit (Qiagen) following manufacturer's recommendations, with the addition of an extra washing step, resulting in a total of two washes with MinElute® PE buffer. To maximise product recovery the elution step was performed twice in 15 µl - a total of 30 µl pre-warmed (~60°C) EB elution buffer.

2.2.2.3 Quantification of PCR products

Purified amplicons were quantified using the Qubit® dsDNA BR assay kit on a Qubit® 2.0 fluorometer (Thermo Fisher Scientific) following the manufacturer's recommended protocol, and ~500 ng amplicons for each sample were taken into the library preparation steps.

2.2.2.4 Library preparation

Library preparation was performed using Illumina® TruSeq™ DNA PCR-free library preparation LT and HT kits according to the manufacturer's protocols, with modifications suggested by the Promega (2015) protocol, which adjusts the general process to their specific product and workflow respectively.

2.2.2.4.1 The characteristics of PCR-free library preparation

This form of library preparation workflow has some unique features and advantages: this process is PCR-free, meaning the already amplified PCR products do not go through another amplification, which is generally a standard step to incorporate the adapters to the end of fragments. Instead, adapters are ligated to the fragments after the steps of fragment end-repair and A-tailing that prepares the fragments to accept the adapters in the subsequent ligation step (Figure 2.3). For this, the library preparation requires a large input, which is suitable for PCR-generated amplicon sequencing. Ligated adapters are specific to this type of library preparation, and designed with a special Y-shaped structure, shown in Figure 2.3.

The addition of adapters to the prepared fragment library is necessary to allow the molecules to hybridise to the preloaded, surface-bound complementary sequences on the sequencing chip, but also to add the option to multiplex more than one sample, by using adapters that include index sequences to allow sample identification after demultiplexing the results.

The TruSeq™ DNA PCR-free LT or HT library preparation kits use two approaches to indexing the samples. The low throughput (LT) option offers two times twelve unique 8-nucleotide single index tags within the adapters to be added to one end of the fragment, allowing the multiplexing of up to 24 samples. The high throughput (HT) option offers multiplexing up to 96 samples using dual indexing - the combination of 8 and 12 different 6-nucleotide index tags, one at each end of the fragments, as shown in Figure 2.3. The difference between the two methods results

in different numbers of cycles to be used for index sequencing - 8 cycles for LT, and 12 cycles for HT kits.

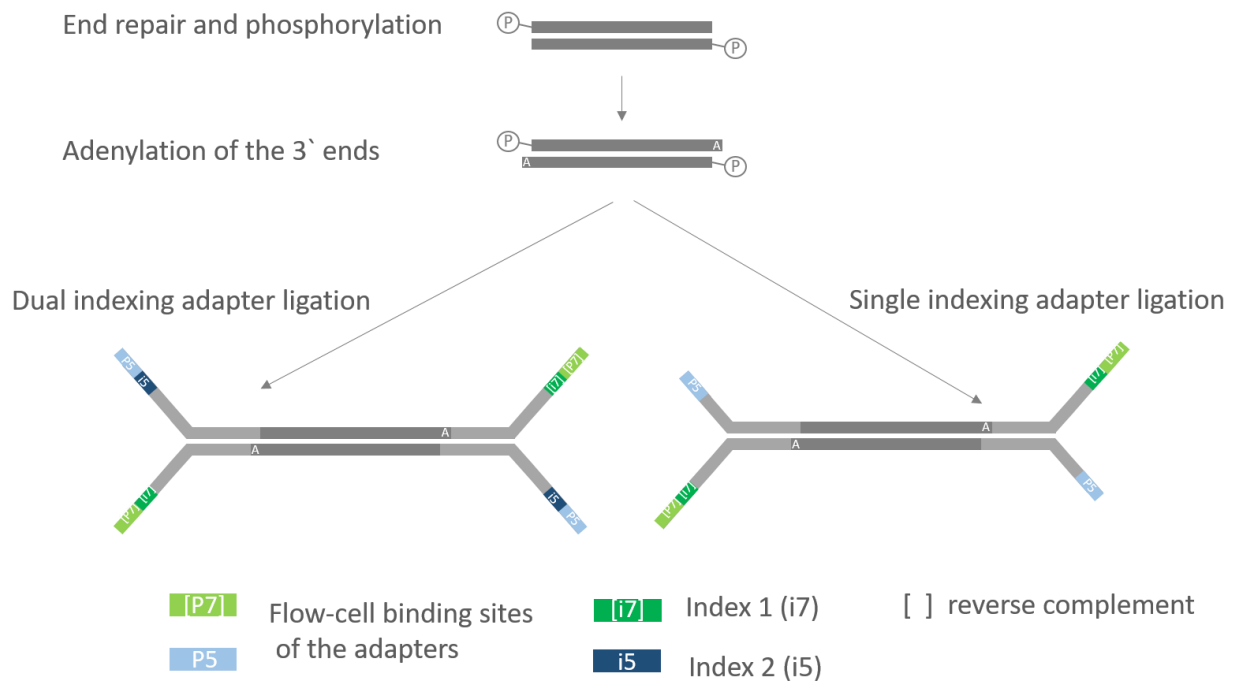


Figure 2.3.
Dual or single index adapters.

Adapters are ligated to the end-repaired, A-tailed amplicons creating sequence-ready products (adapted from: www.illumina.com).

A disadvantage of these adapters is that once added to the end of the fragments they change their mobility in separation, for example when using a Bioanalyzer; for this reason, libraries containing these type of adapters cannot be reliably quality checked by such methods, for example for size range or quantity. The advantage of adapter ligation, however, is that the method allows the fragments to be randomly inserted in either orientation with respect to the adapters, as opposed to PCR-based adapter incorporation, where adapters are tied to primers targeting the specific ends of the fragment (Figure 2.4). The latter method results in a fixed directionality with

regard to the P5 and P7 adapters, which mark specific ends of the molecules and are directing sequencing orientation. The latter method, combined with the often accompanying bead-normalisation which allows only one strand (the strand not physically bound to the SPRI normalisation beads, for example in ForenSeq™ DNA Signature Prep Kit, Verogen) to be sequenced, causes the same directionality to be carried over to the sequence reads. Thus, R1 always starts from the same end of the fragment, which means that the quality of base-calls decreases as sequencing progresses along the same direction of the fragment (Kim 2016). Therefore, to achieve good base-calling quality throughout the sequenced region it is necessary to apply paired-end sequencing, where R2 in the reverse direction will produce the same quality gradient in the opposite direction and therefore can compensate for the quality drop, thus generating an acceptable overall quality.

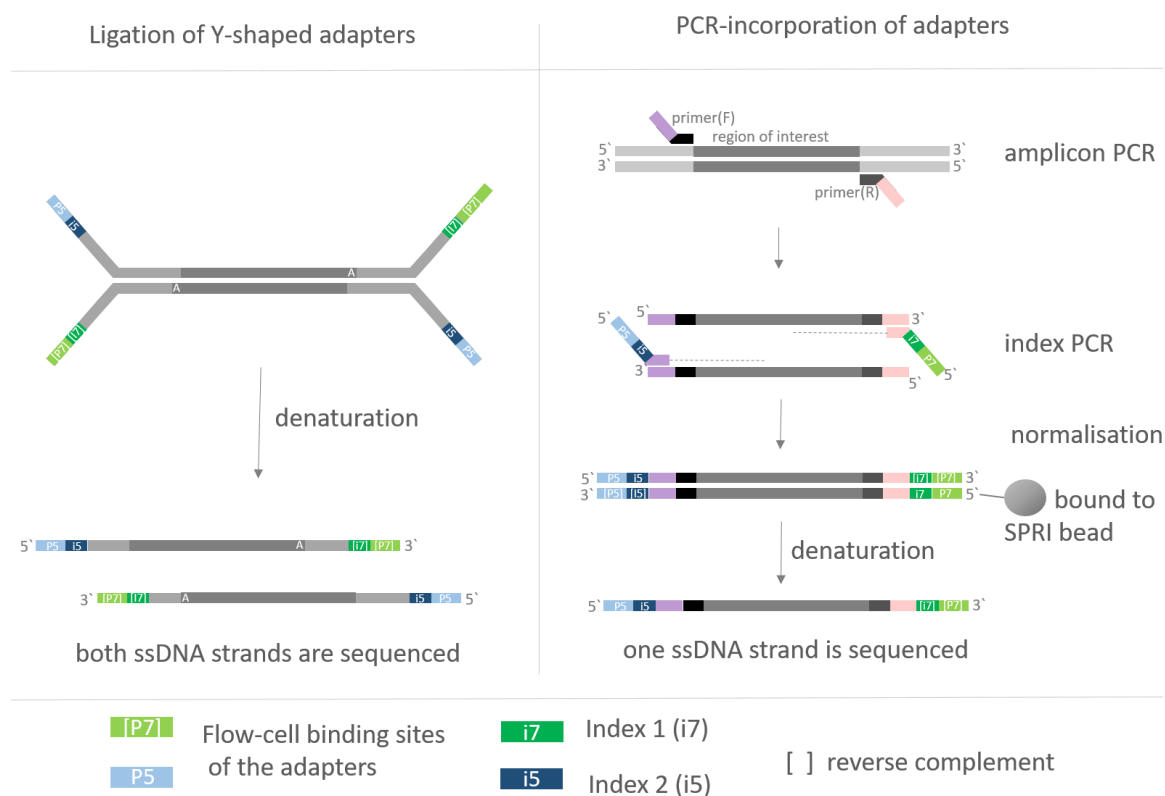


Figure 2.4
Differences between ligated and PCR-incorporated dual adapters and sequencing of the generated libraries.

If adapter-ligated libraries were subjected to bead normalisation and therefore stripped from their second strand prior to sequencing, these would still preserve both the directions as their second strand would just as likely be the forward strand as the reverse.

In the adapter-ligation protocol used here, this issue does not arise due to the random addition of the adapters, and sequencing of both strands, which even in R1 alone generates the same overall base-calling quality across the sequenced region.

2.2.2.4.2 Size selection of libraries

Selection of the ideal size range libraries followed the recommended protocol using magnetic bead-based selection; this step aims to remove any empty adapter-dimers and similarly unproductive elements on the basis of their smaller size.

2.2.2.4.3 Quantification of libraries

Prepared libraries were quantified using the KAPA Library Quantification Kit for Illumina® platforms (Kapa Biosystems, later acquired by the Roche Group) with the LightCycler®480 (Roche) real-time PCR system following the manufacturer's recommended protocol. Triplicate measures were used to define the mean concentration values, or identify and exclude outliers. Libraries were size-corrected to reflect actual concentration following the manufacturer's and Promega PowerSeq™ Auto/Mito/Y System's prototype protocol (Promega 2015).

The KAPA Library Quantification Kit is designed for measuring the concentration of those molecules in libraries that are able to bind to the surface of Illumina® sequencing chips, to form clusters and thus to generate sequence information, and therefore is considered the most accurate way to measure and to balance samples prior to sequencing. The libraries were normalised to 4 nM concentrations and were then pooled using equal 5 µl volumes. The pooled libraries were re-quantified using the same kit and method to confirm dilution and overall concentration.

2.2.3 Sample preparation specific to Promega PowerSeq™ CRM Nested System, Custom

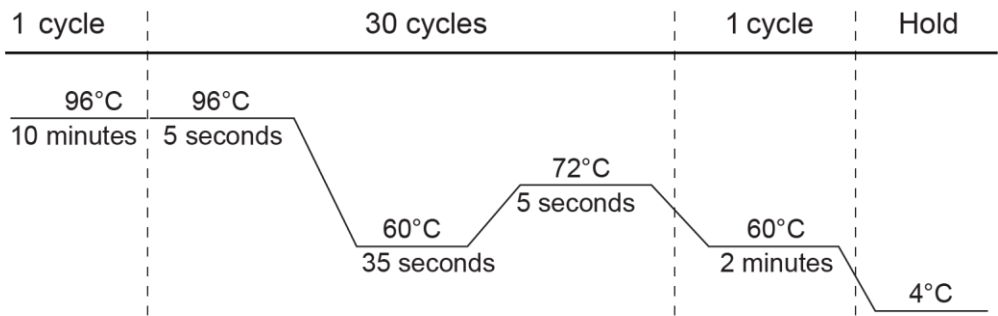
2.2.3.1 Fragment generation and library preparation by PCR

This library preparation is an application-specific method, in the form of a simplified single-step PCR reaction that amplifies the target regions of the mtDNA, while also adding the required adapter sequences to the amplified fragments in a single reaction. This step was performed following the manufacturer’s recommended protocol (Promega 2018) using 0.5 ng genomic DNA as template in a multiplex PCR comprising a buffered master mix, primer pair mix and index primer pairs for the amplification of the mtDNA control region as ten overlapping amplicons at the same time as dual indexing these samples.

Amplification was performed using a DNA Engine Tetrad2 (MJ Research), with the parameters shown in Table 2.6.

Table 2.6
Thermocycling parameters used with the Promega PowerSeq™ CRM Nested System, Custom multiplex.

*The steps of amplification follow the manufacturer’s recommendations.
Adapted from (Promega 2018)*



2.2.3.2 Purification of amplified libraries

The amplified products were purified and size-selected using an AMPure® XP magnetic bead-based purification method following the recommendations of the manufacturer.

2.2.3.3 Quantification of amplified libraries

Prepared libraries were quantified following the manufacturer's recommended protocol using the Promega PowerSeq™ Quant MS System qPCR kit with a LightCycler®480 (Roche) real-time PCR system. Triplicate measures were taken to define the mean concentration values or identify and exclude outliers.

The libraries were normalised to 4 nM concentrations and pooled using equal 5 µl volumes. The pooled libraries were re-quantified using the same kit and method to confirm dilution and overall concentration.

2.2.4 Sequencing pooled libraries on the MiSeq®

Pooled libraries were prepared following manufacturer's recommendations; this includes the steps of denaturation, the addition of an internal sequencing PhiX Control DNA, and the dilution of libraries to 10-12.5 pM with a 10-15% PhiX 'spike' to compensate for low complexity libraries, following the protocols (Promega 2015, 2018), and the manufacturer's protocols.

Sequencing of the prepared libraries from Promega PowerSeq™ Auto/Mito/Y Systems prototype experiments was performed on an Illumina® MiSeq®/MiSeq FGx® sequencer in 'research use only' (RUO) mode, using v2 (300 cycles) sequencing reagent kits, with the machine's 'Generate FASTQ' workflow and with the SE sequencing method.

Sequencing of the prepared libraries from Promega PowerSeq™ CRM Nested System, Custom experiments was performed on an Illumina® MiSeq FGx® sequencer in 'research use only' (RUO) mode, using v3 (600 cycles) sequencing

reagent kits, with the machine's 'Generate FASTQ' workflow and with the PE sequencing method.

2.3 Data analysis

2.3.1 Output of sequencing runs

At the end of the sequencing runs, the MiSeq® de-multiplexes the reads according to the index sequences and allocates each read to one of the samples, or otherwise marks them as unallocated. The reads are collated into raw compressed FASTQ files for each sample and for each read direction: read1 (R1) in single-end (SE) and read1 and read2 (R1, R2 respectively) in paired-end (PE) sequencing modes.

Additional output files of each sequencing run, containing run information, metrics and statistics were reviewed independently from the sequencer using the Illumina® Sequencing Analysis Viewer (SAV v2.4.7) software.

The raw compressed FASTQ files were downloaded and used as an input for other software.

2.3.2 Quality check and control of raw FASTQ files

The raw FASTQ files were checked using the FastQC v0.11.2 programme (Andrews 2010), which informs about quality of base-calling in the reads, sequence length distribution, overrepresented sequences or motifs, nucleotide content, and general quality metrics. This tool was used to monitor the files in raw and improved (QC-ed) state. Some specific features of a targeted MPS study can automatically raise quality flags; for example, sequencing PCR amplicons naturally causes high levels of duplicate reads, mtDNA sequences are always flagged as overrepresented due to their higher copy number, and sequenced STRs often show skewed nucleotide balance in reads due to their non-random, structured sequences. In respect of this type of study such flags are expected, and therefore can be ignored.

2.3.2.1 Trimming of reads

To improve the quality metrics, adapter removal and quality-dependent end trimming was performed by using Trimmomatic v 0.32 (Bolger et al. 2014). This software removes low-quality base-calls from the reads, and with user-set parameters that include a defined list of adapters in FASTA format, removes any remaining adapters that escaped this process within the MiSeq®.

For example, the arguments: PE ILLUMINACLIP:adapters.fa:2:40:15:1:true LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36, specify the software to: use the provided list of adapters in FASTA format to look in the PE sequence data for seed matches to these sequences, while allowing two mismatches; extend and clip these if a score of 40 is reached; remove the shortest fragments of adapters while keeping both reads; remove leading and trailing low quality bases below quality 3; and scan the reads with a 4-base window, cutting when average quality drops below 15. After performing these steps the program is instructed to remove any remaining reads that are 36 bases long.

2.3.2.2 Optional error correction

The trimmed reads can also be subject to error correction; however, this is only applied to reference sample typing for STRs. As very low component reads in mixtures, or in mtDNA heteroplasmy studies, could potentially be affected by this quality improvement method, it was not applied here prior to mtDNA analysis.

Reads were screened for their error profiles and identified random errors were corrected using SOAPec v2.01 KmerFreq_AR and Corrector_AR tools (Li et al. 2008). These tools screen the sample to identify random error in the reads compared to other sequence reads at that position and subsequently correct these errors using the rest of the samples as templates.

2.3.2.3 Quality controlled FASTQ files

The FastQC programme was again used to confirm the improvement of the FASTQ files. The output files of these programmes were saved as quality controlled FASTQ files with a designated extension (.qc.fq) and were used as input for other data analysis software and command line pipelines.

2.3.3 Analysis of mtDNA sequence data

2.3.3.1 Variant calling pipeline

A standard variant calling pipeline adapted for amplicon sequencing (omitting duplicate removal steps) was used to detect variants in the mtDNA control region from quality-controlled FASTQ files. This pipeline used the tools BWA v0.7.12 (Li and Durbin 2009) for alignment, SAMtools v1.3.1 (Li et al. 2009) for generating, sorting and indexing BAM files and calling variants, Picard v2.1.0 (<http://broadinstitute.github.io/picard>) for adding read groups, GATK v3.4-0 (Van der Auwera et al. 2013) for local realignment, base calibration and coverage calculations, and VCFtools v0.1.15 (Danecek et al. 2011) for filtering variants. The reads were aligned to the revised Cambridge Reference Sequence (rCRS, GenBank accession: NC_012920.1; Andrews et al. 1999). Variants were also visualised and confirmed using the Integrative Genomics Viewer (IGV) tool (Robinson et al. 2011).

2.3.3.2 Detection of nuclear mtDNA insertions (numts)

In-house Bash scripts were used to filter reads amplified from numt sequences by focusing on numt-specific variants identified by clusters of differences to known current human mtDNA variation. For example the numt sequence described in section 5.3.7 was identified using an approximate regular expression search for GaccCaTccCcccttaaT/AttaagggGggAtGggtC sequences which covers the numts version of the rCRS from positions 16284-16301; the 5 mismatches to rCRS are

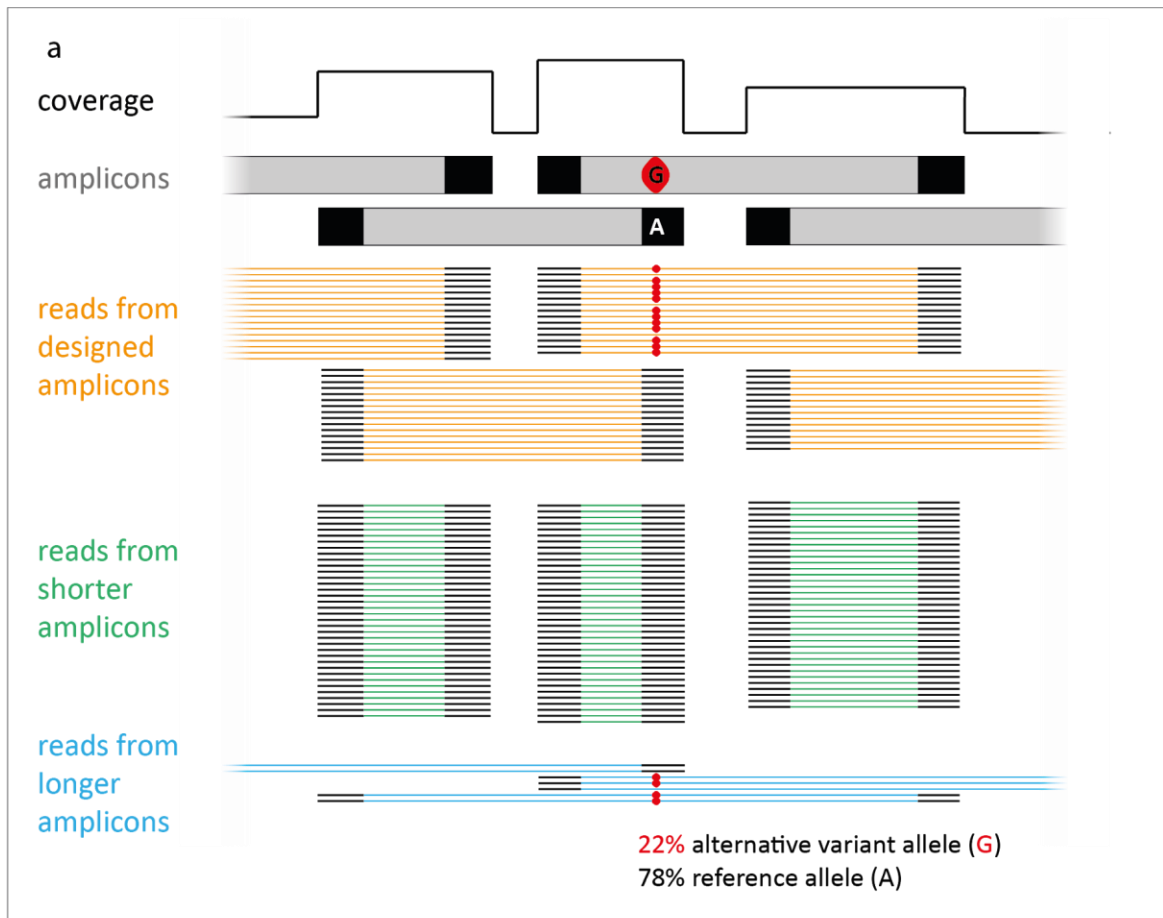
indicated by the capitalised letters. The reads identified as representing variants from numts were excluded from variant calling.

2.3.3.3 Detection and quantification of indels in mtDNA

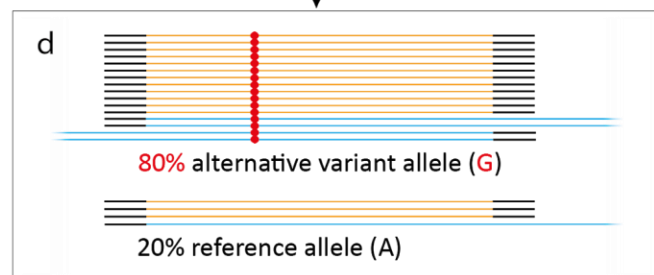
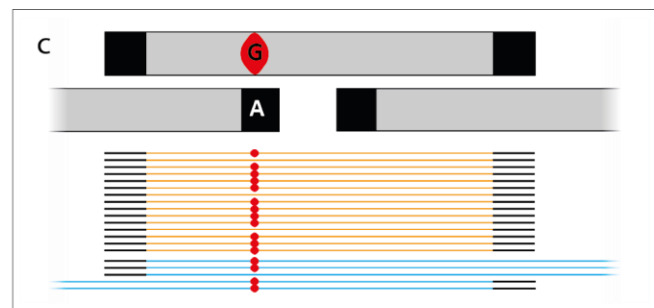
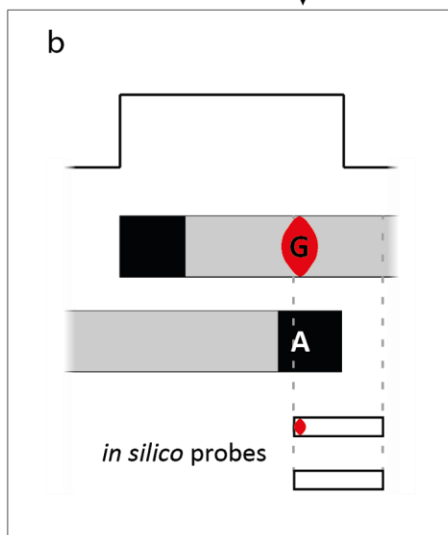
In-house Bash scripts were also used to attempt assessments of the level of length heteroplasmy and to interpret insertions and deletions and other variants found in ambiguous regions, such as homopolymeric tracts and the AC-repeat region; detailed examples are given in Appendix E.

2.3.3.4 Overarching Read Enrichment Option (OREO)

Due to the proprietary nature of primer sequences in commercial kits, perfect primer removal from the reads is not possible without losing information on the cognate mtDNA data as well, therefore to find an alternative way to remove the primer interference from the results a further variant refinement option was introduced. Overarching Read Enrichment Option (OREO) is a Bash script tool designed for use with single multiplexes of overlapping amplicons, which identifies sequences that perfectly match a probe that extends beyond the 3' end of putative primer binding regions with the aim of retaining only those reads which do not end in primer sequences, but overarch the queried region, as shown in Figure 2.5.



overarching read enrichment option
(OREO) ↓



Cont.

Figure 2.5

OREO selects reads containing variants intrinsic to the mtDNA.

Improving estimation of heteroplasmy using Overarching Read Enrichment Option (OREO). a) When overlapping amplicons (grey bars, top) are used to generate sequence data from the control region an alternative data processing approach is required because exact primer-derived sequences (black boxes) are not known. The sequence reads derive from the designed amplicons (orange), short overlapping amplicons (green) and longer reads (blue). b) The overarching read enrichment option (OREO) filters reads for the presence of specific in silico 'probe' sequences (white boxes). The probes are designed to enrich for reads spanning the primer site, thus excluding reads that carry variants copied from the primer sequences rather than from the mtDNA itself. c, d) Among the retained reads the proportion of reference and alternative alleles is measured and provides a less biased estimate of the level of heteroplasmy.

The detected level of heteroplasmy is affected by the primer-derived sequences (Figure 2.5a). These overarching reads are selected by the script (Figure 2.5b and c) and are parsed into ancestral and derived categories (Figure 2.5d) according to their match to a tested pairs of probes with ancestral and derived alleles of the tested variants. From the proportions of these read categories, an accurate quantitation of heteroplasmy levels become possible, which is not affected by the bias from primer-derived sequences present at the end of reads.

To increase the confidence of lower-level heteroplasmy detection and quantitation within the putative primer-binding regions, OREO was applied to the FASTQ files to filter the variants falling within the putative primer binding regions, and verify if the called variants were indeed of cognate mtDNA origin.

2.3.3.5 Additional data analysis steps relating to the PowerSeq™

CRM Nested System

As a consequence of the use of an overlapping amplicon design within a single reaction, shorter amplicons appear together with the intended size amplicons. These small fragments, having the size of the overlap only, mainly consist of primer sequences and still occupy many reads; however, they do not contribute new

information, and even increase any potential primer bias. The option of removing these from data analysis can be beneficial in reducing processing time, and in lessening any primer-derived bias. The primer-derived reference-bias phenomenon and the effect of shorter amplicons and primer internalisation are detailed in Chapter 5.

To remove the short amplicons having only the size of the overlap, Trimmomatic v0.32 software was used this time as an *in silico* size selection step. FastQC software v0.11.2 generated a read-length profile plot for each sample, and after examining these a 95-bp cut-off limit was determined to safely remove reads below this length by Trimmomatic v0.32.

2.3.3.6 Other options for mtDNA analysis or primer removal

Two tools designed for the analysis of mtDNA MPS data were also tested: mtDNAServer (Weissensteiner et al. 2016a); and the trial version of the commercial software GeneMarkerHTS v1.2.2.1338 (SoftGenetics, LLC, 2018).

To account for unknown primer sequences three methods were tested: Cutadapt (Martin 2011) and seqtk (available from: <https://github.com/lh3/seqtk>) to trim reads, and BAMClipper (Au et al. 2017) to trim BAM files. All three methods require the user to know, or alternatively to estimate, the primer lengths.

2.3.3.7 Haplogroup assignments and phylogenetic relationships of samples

Called variants were checked for correct nomenclature in a phylogenetic alignment context using the SAM2 tool of the EMPOP mtDNA database (Huber et al. 2018). Haplogroups were predicted and their relationships visualised using HaploGrep2 HaploGrep2 (Weissensteiner et al. 2016b) and confirmed by the EMPOP EMMA tool (Rock et al. 2013).

A maximum-likelihood phylogenetic tree of the control region sequences (from position 15,989 to 619) was constructed using MEGA v6.06 software (Tamura et

al. 2013) and was visualised using the software FigTree v1.4.3 (available at: <http://tree.bio.ed.ac.uk/software/figtree/>).

A median-joining network (Bandelt et al. 1999) was built using control region sequence variants using the Network 5.0.1.1 freely available software and annotated using Network Publisher 2.1.2.5 commercial software (both from Fluxus Technology Ltd., available at: <http://www.fluxus-engineering.com/>).

2.3.4 Analysis of STR sequence data

2.3.4.1 FDS Tools

Quality-controlled FASTQ files were analysed primarily with the software FDS Tools v1.1.1 (Hoogenboom et al. 2017) for autosomal and Y-STR sequence data. FDS Tools is a circular self-teaching tool, allowing the identification of completely new variants as they are described. This flexibility provided by FDS Tools was the most suitable for the data set of global samples which displays a wide range of variants. FDS Tools is a bundle of Python tools with versatile packages that can be used modularly, or by a predefined pipeline.

This analysis used a custom script utilising the following tool packages: 'tssv' for identifying sequence variants, 'stuttermark' to mark variant classes, 'seqconvert' to provide string and different bracketed formats of the results, and 'vis' to generate an html output visual summary of the results. During the analysis the 'library files' (files to modularly describe possible variation) developed organically with every observed new variant.

The advantage of FDS Tools over other approaches is its flexibility, not just regarding what is known or as yet undescribed variation, but also regarding the flexibility of boundaries between flanking region and arrays. Flanking regions are often considered to be invariant, showing only fixed repeats; however, when looking at global variation, motifs may show variability which is not observed in population-level sample sets. FDS Tools can identify elements with variable numbers of repeats

in the flanking region and describe these as such, rather than missing them, or marking them as indels.

2.3.4.2 STRait Razor

For the global variation dataset, quality-controlled FASTQ files were analysed for autosomal and Y-STRs using STRait Razor 2.0 (Warshauer et al. 2015b; Warshauer et al. 2013) using a modified configuration file with extended allele ranges to include extreme alleles present in the sample set. For the PoBI dataset, STRait Razor 3.0 (Woerner et al. 2017) was used to analyse autosomal and Y-STR sequence data. STRait Razor output files were used as input for the supplementary Microsoft Excel workbooks provided (<https://www.unthsc.edu/graduate-school-of-biomedical-sciences/molecular-and-medical-genetics/laboratory-faculty-and-staff/strait-razor/>) and the generated files were further used in comparison and representation of the results in Microsoft Excel.

2.3.4.3 Command-line sequence data interrogation and standard variant calling

Standard command-line tools and scripts (for example to search, parse, filter and summarise data in the quality-controlled FASTQ files) were also used to verify sequence variants of STR arrays and flanking regions. This was required particularly when any two software/analysis options gave discordant results, and therefore the raw data were investigated and used to verify the source of discrepancy in the software.

To verify variants in the flanking regions, sequences related to STR markers were also subject to the standard variant calling pipeline as described above in 2.3.3.1. The alignment in this case was to a custom reference file containing the genomic reference sequences (version GRCh38) of all targeted loci with 500-bp padding on each side of the amplified segment. The variants identified were used to confirm other software-based designations of flanking region variants and the Integrative Genomics Viewer (IGV) tool (Robinson et al. 2011) was used for visualisation.

Linux-based data analysis and application command-line software were used for this research on the SPECTRE High Performance Computing Facility of the University of Leicester.

2.3.4.4 Relative read-depth ratio test for duplicated Y-STR alleles

To distinguish between Y-STR alleles resulting from somatic mutation and constitutive allele duplications, stutter-adjusted sequence read-depths for different PCR products were considered. This test (Figure 2.6) is analogous to the semi-quantitative analysis of peak heights in CE and similar in concept to the paralogue ratio test (PRT) used for copy number evaluation (Armour et al. 2007). The test assumes that similar size-range STRs in a multiplex reaction amplify and are detected proportionally to their dosage. When finding an additional allele (putative duplication) at a given STR in a sample, coverage values of the test locus and a selected reference locus (another similar size-range marker amplified in the same reaction) were compared in all the other analysed samples; this gave a range of expected relative read-depth ratios for that pair of loci. The same comparison was then applied to each of the alleles of the test locus against the same reference locus in the test sample. The test indicated whether the two alleles were indeed duplicated (together displaying approximately double the expected read-depth ratio), or if the second allele is a likely result of somatic mutation (the summed ratios of both alleles lying in the expected range of a single-dose allele).

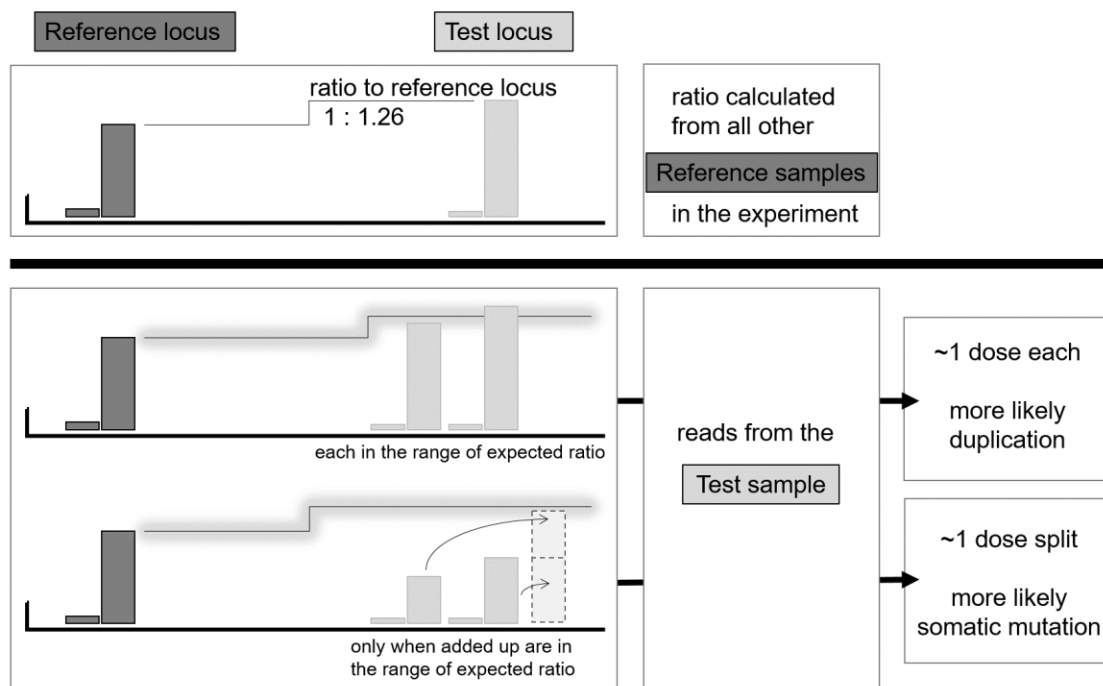


Figure 2.6
Relative read-depth ratio test.

A schematic explanation of the relative read-depth ratio test to assess whether the supernumerary alleles are more likely to represent true duplications or somatic mutations. The test uses a reference locus comparable in size and co-amplified in the same reaction, and uses all non-affected samples to calculate the ratio expected. In test samples the ratio is defined for both tested alleles and the dosage determines the class of these alleles.

To avoid confusion with stutter products, somatic mutants were only called when they did not lie in the 'n-1' stutter position compared to the 'n-repeat' main allele.

In theory, this principle is applicable to autosomal STRs as well; however, it is difficult to gauge relative read-depth either for a normal range or for a tested sample when more than one main allele is present with its accompanying stutter derivatives within a close range of alleles.

2.3.4.5 Y-STR haplogroup assignments and phylogenetic relationships of the samples

Based on Y-STR haplotypes, haplogroups were predicted using NevGen Y-DNA Haplogroup Predictor (Gentula and Nevski, Serbian DNA Project, 2015, available at: <https://www.nevgen.org/>).

Median-joining networks were built from both Y-STR length and sequence variants using the Network 5.0.1.1 software and annotated using Network Publisher 2.1.2.5 software (both from Fluxus Technology Ltd., available at: <http://www.fluxus-engineering.com/>).

2.3.5 Population genetic data analysis

Population summary statistics were generated using the Arlequin v 3.5.2.2 software (Excoffier and Lischer 2010), including molecular diversity indices, F-statistics for population comparisons and differentiation.

To visualise genetic distances of the samples, pairwise F_{ST} or R_{ST} values were used to construct multidimensional scaling (MDS) plots using the R software 'MASS' package and principal component analysis (PCA) plots using the R 'stats' package and 'factoextra' package to visualise the outputs (R Core Team 2014).

The browser-based application STRAF 1.0.5 was used to analyse STR data for standard forensic and population genetic parameters (Gouy and Zieger 2017).

2.3.6 Datasets used for concordance and validation of the results

2.3.6.1 Datasets for comparison with the global variation dataset

2.3.6.1.1 Datasets for Y-STR comparison

Sequence-derived repeat array lengths were compared to previously determined CE-based PowerPlex® Y23 data (Hallast et al. 2015). In addition there are 29 samples in the global dataset that are also part of the 1000 Genomes Project (1KGP).

2.3.6.1.2 Datasets for autosomal-STR comparison

The 29 samples from the 1000 Genomes Project were also used to check concordance with the global dataset.

2.3.6.1.3 Comparison of mtDNA variants

Validation of the results was performed by comparing SNP calls against a total of 65/101 independently sequenced samples: 58 from previously published data (Batini et al. 2017), of which 23 were also sequenced by the 1000 Genomes Project (1000 Genomes Project Consortium 2015), thereby providing a three-way comparison; and six additional 1000 Genomes Project samples;

For the operator control sample the whole mtDNA was sequenced commercially (GenBank accession: MG551929) and was used for direct comparison.

.

2.3.6.2 Datasets for comparison with the PoBI sample set

2.3.6.2.1 Comparison for mtDNA

Defined mtDNA CR variants were compared to the same samples analysed by array-based SNP-typing of the whole mtDNA, carried out as part of the PoBI project. Unpublished data were supplied by Sir Walter Bodmer. Predicted haplogroups based on the two different approaches were compared.

2.3.6.2.2 Datasets for STR comparisons

There are eight samples in the PoBI dataset analysed here that are also part of the 1000 Genomes Project, in the GBR population.

2.4 Data visualisation

Sample sets and their distinct features were encoded in Microreact (Argimon et al. 2016) to allow an interactive online surface for interrogation of sample details and features in a geographically coded context.

The Integrative Genomics Viewer (IGV) tool (Robinson et al. 2011) was used to visualise called variants, as part of the verification of variants for both mtDNA amplicons and STR flanking regions.

A maximum-likelihood phylogenetic tree of the mtDNA control region constructed using MEGA v6.06 (Tamura et al. 2013) was visualised using FigTree v1.4.3 (available at: <http://tree.bio.ed.ac.uk/software/figtree/>).

Median-joining networks were used to visualise phylogenetic relationships of samples using the Network 5.0.1.1 software and furthered organised, coloured, and annotated using the Network Publisher 2.1.2.5 software (both from Fluxus Technology Ltd., available at: <http://www.fluxus-engineering.com/>).

To visualise the samples based on their genetic distances pairwise F_{ST} or R_{ST} (for microsatellites) values were used to construct multidimensional scaling (MDS) plots

using the R software 'MASS' package. Principal component analysis (PCA) plots were used to visually capture the main components of the variation between populations or regions based on the genotypes of their loci, using the R 'stats' package and the outputs were visualised by the 'factoextra' package.

Graphs and diagrams were created or outputs of other software described here were manipulated using Microsoft Office and the R software (R Core Team 2014). Vector graphics were edited in Adobe Illustrator (v21.1.0, 2017).

CHAPTER 3: Global Y-STR sequence variation in a phylogenetic context

Work described in this chapter has been published as:

Huszar, T.I., Jobling, M.A. and Wetton, J.H. 2018. **A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing.** *Forensic Sci Int Genet*, 35:97-106.

doi: [10.1016/j.fsigen.2018.03.012](https://doi.org/10.1016/j.fsigen.2018.03.012)

The paper itself is included in the electronic appendices (Appendix A).

3.1 Introduction

Classically, STRs are divided into simple, compound, complex or even complex hypervariable types, reflecting the increasing complexity of the length, sequence and intermittent elements of building blocks (Butler 2010). However, conventional analysis of STR variation via CE considers only overall length variation at such markers. Now that MPS is being implemented in forensic typing, STRs are also becoming characterised by the richer range of variation displayed at the DNA sequence level, and this allows a more nuanced understanding of their diversity and the underlying mutation processes that generate this diversity.

One indication that increased allelic diversity is likely to be observed via MPS-based analysis of an STR is the complexity of the array (Gettings et al. 2016), since repeat pattern variation (RPV) can arise from different numbers of repeat blocks with the same allele length (isometric alleles). SNPs and indels within repeat arrays can also contribute to diversity. While single nucleotide changes typically have very low mutation rates ($\sim 10^{-8}$ per base per generation; Nachman and Crowell 2000) and therefore are unlikely to be observed as independent recurrences, the RPV in STRs mainly results from a more rapid ($\sim 10^{-3}$ per repeat array per generation; Weber and Wong 1993) mutation process driven by replication slippage, so that the same

variants can arise multiple times independently. SNPs and indels are not restricted to the repeat array, but are also found in the flanking regions, providing further basis for discrimination.

While autosomal STRs assort independently and are therefore uncorrelated, STRs on the male-specific region of the Y chromosome (MSY) are permanently linked together into a haplotype. This reduces the overall diversity that a Y-STR profile provides (Jobling et al. 1997), but also means that Y-STR sequence diversity can be considered within the framework of a robust phylogeny of haplogroups defined by single-nucleotide polymorphisms. Indeed, this relationship forms the basis of various methods that have been developed to predict MSY haplogroups from Y-STR haplotypes (Athey 2005, 2006; Schlecht et al. 2008; Seman et al. 2012). Because of the high degree of population structure among Y chromosomes (Jobling and Tyler-Smith 2003), studies of individual populations tend to capture a limited range of haplogroup diversity. Choosing samples for MPS-based Y-STR analysis to maximise haplogroup diversity, rather than on a population basis, should permit a broad survey of Y-STR sequence diversity to be undertaken efficiently. In addition, the phylogenetic framework should allow the degree of mutational recurrence of observed variants to be understood, with slow-mutating SNPs and indels tending to occur only once in the tree (monophyletic), and more rapidly-mutating RPs showing recurrence (polyphyletic).

Here a set of 100 diverse samples was selected in which MSY resequencing previously defined a highly resolved SNP-based phylogeny (Hallast et al. 2015), and MPS was used to sequence 23 Y-STRs in each. The observed variants are described, some improvements to MPS allele designations are suggested, and the different classes of variants are placed in their phylogenetic contexts.

3.2 Materials and methods

3.2.1 DNA samples

One hundred male DNA samples were selected from a previously described set of 448 (Hallast et al. 2015). Sample details were given in Table 2.1 in Chapter 2, Section 2.1.1.

3.2.2 DNA quantitation and PCR amplification

Quantities of double-stranded DNA were verified prior to PCR using the Qubit® 2.0 fluorometer (Thermo Fisher Scientific) with the Qubit® dsDNA HS kit. Twenty-three Y-STRs (DYS19, DYS385a,b, DYS389I/II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS481, DYS533, DYS549, DYS570, DYS576, DYS635, DYS643 and Y-GATA-H4) were amplified from 0.5 ng template DNA using the prototype PowerSeq™ Auto/Mito/Y System (Promega) following the manufacturer's recommended protocol (Promega 2015).

3.2.3 Library preparation and sequencing

Amplified products were purified using the MinElute® PCR purification kit (Qiagen), then quantified using the Qubit® dsDNA BR kit on the Qubit® 2.0 fluorometer.

Library preparation was performed on ~500 ng product per sample using the TruSeq™ DNA PCR-free LT (24-plex) and HT (96-plex) sample preparation reagents (Illumina®). The manufacturer's protocol was used, with an adjustment for the PowerSeq™ System (Promega), namely the use of the MinElute® PCR purification kit for size selection of amplicons.

Prepared libraries were quantified using the KAPA Library Quantification Kit for Illumina® platforms (KAPA Biosystems, now acquired by Roche) with the LightCycler®480 (Roche) real-time PCR system following the manufacturer's recommendations. All indexed libraries were normalised to 4 nM, pooled at equal

volumes and re-quantified using the same method to confirm pooled library concentration.

Pooled libraries were prepared for sequencing following the manufacturer's protocol, diluting to 12 pM for loading and using a higher (15%) PhiX internal control library 'spike', as recommended for sequencing low-complexity libraries. Sequencing was performed on a MiSeq FGx[®] (Illumina[®]) sequencer in 'research use only' (RUO) mode, via the "Generate FASTQ" workflow with "FASTQ Only" application and single-end (SE) method using MiSeq[®] v2 (300 cycles) reagent kits.

3.2.4 Data processing and analyses

Raw compressed FASTQ files were transferred from the MiSeq[®] for external analysis. Quality checking was done by trimming any remaining known adapter sequences and low-quality read ends with Trimmomatic v0.32 (Bolger et al. 2014) and SOAPec v2.01 (Li et al. 2008) software. The resulting improvement in quality was confirmed using the FastQC v0.11.5 (Andrews 2010) programme.

The open-source software FDS Tools v1.1.1 (Hoogenboom et al. 2017) was used to analyse reads spanning the STR repeat regions and their flanking regions. Allele calls were confirmed using STRait Razor 2.0 (Warshauer et al. 2015b) using a modified configuration file to include all alleles present in the sample set and command line tools and standard variant calls to resolve occasionally different designations.

Discovered variants were compared to the human genome reference sequence (GRCh38) and queried in dbSNP (Database of Single Nucleotide Polymorphisms, build 151, Available from: <http://www.ncbi.nlm.nih.gov/SNP/>). Repeat pattern variants were compared to the existing literature and the database STRBase (Ruitberg et al. 2001); strbase.nist.gov, accessed 02-Nov-2017).

For duplicated alleles a relative read-depth ratio test was used to distinguish between alleles resulting from somatic mutation and constitutive allele duplications.

3.3 Results

3.3.1 Phylogenetic diversity of the samples

In order to capture a wide range of Y-STR variants a phylogenetic approach was taken, choosing a subset of one hundred DNA samples from a previously analysed set (Hallast et al. 2015). The published analysis had used massively-parallel sequencing of ~3.7 Mb of DNA in each of 448 diverse Y chromosomes, and constructed a maximum-parsimony tree based on a total of 13,261 SNPs. The subset here was selected to ensure that major clades and deep-rooting nodes of the tree were represented. The phylogenetic relationships of the analysed samples with true branch lengths is shown in Figure 3.1.

In Table 3.1 the short forms of the MSY haplogroups of the samples are given to supplement the earlier Table 2.1. The haplogroup designations were defined by publications (Hallast et al. 2015; Poznik et al. 2016; van Oven et al. 2014) as described in detail together with the complete table in Appendix C.

Samples were selected to establish a framework for maximum diversity, rather than to represent populations, and therefore classical population statistics are not applicable to these results.

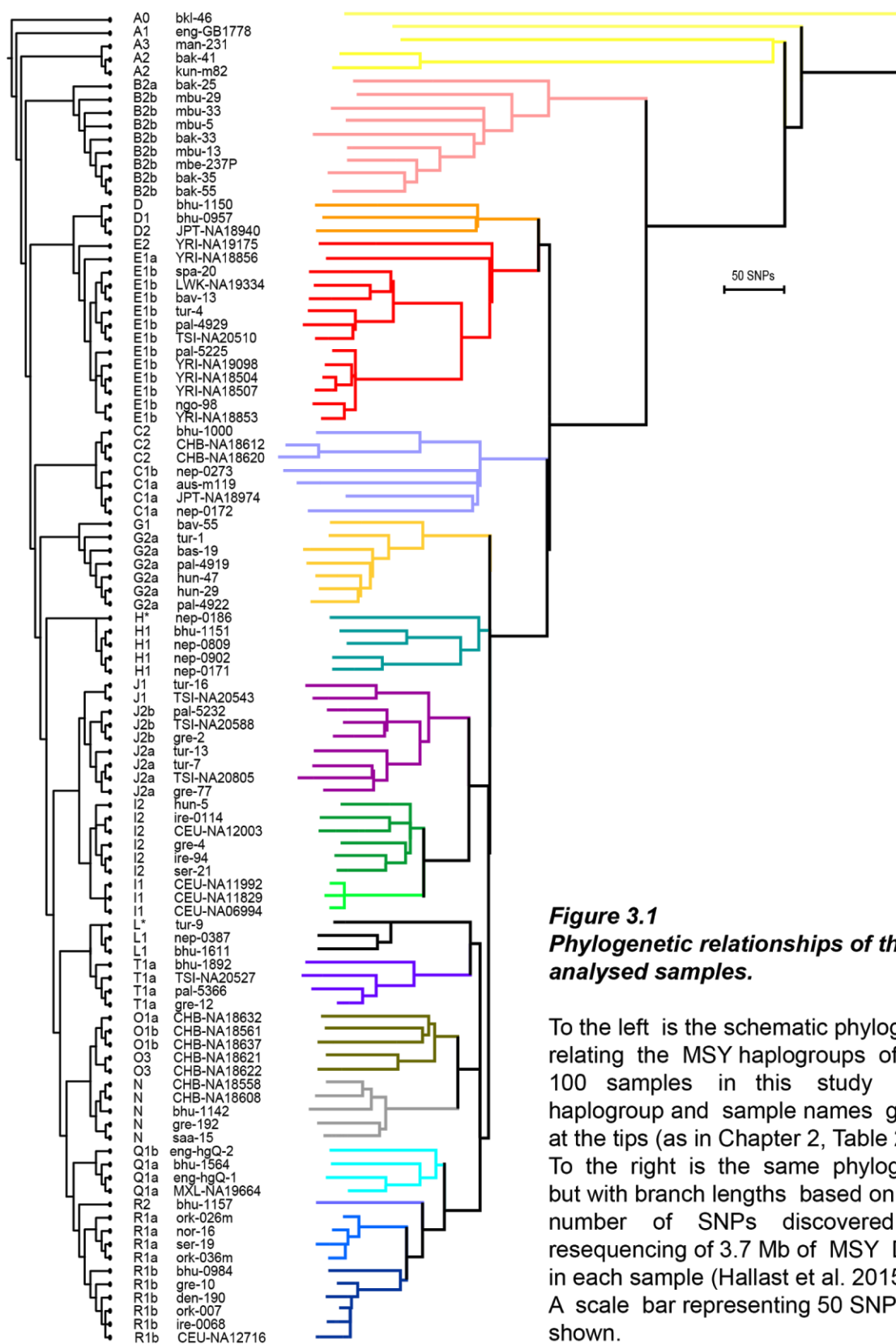


Figure 3.1
Phylogenetic relationships of the analysed samples.

To the left is the schematic phylogeny relating the MSY haplogroups of the 100 samples in this study with haplogroup and sample names given at the tips (as in Chapter 2, Table 2.1). To the right is the same phylogeny, but with branch lengths based on the number of SNPs discovered in resequencing of 3.7 Mb of MSY DNA in each sample (Hallast et al. 2015). A scale bar representing 50 SNPs is shown.

Table 3.1
Haplogroup information for the selected 100 samples.

Sample name	Y haplogroup short form	Sample name	Y haplogroup short form	Sample name	Y haplogroup short form
bkl-46	A0	ngo-98	E1b	tur-9	L*
eng-GB1778	A1	YRI-NA18853	E1b	nep-0387	L1
bak-41	A2	YRI-NA19098	E1b	bhu-1611	L1
kun-m82	A2	YRI-NA19175	E2	bhu-1142	N
man-231	A3	bav-55	G1	CHB-NA18558	N
bak-25	B2a	bas-19	G2a	CHB-NA18608	N
bak-33	B2b	hun-29	G2a	gre-192	N
bak-35	B2b	hun-47	G2a	saa-15	N
bak-55	B2b	pal-4919	G2a	CHB-NA18632	O1a
mbu-29	B2b	pal-4922	G2a	CHB-NA18561	O1b
mbu-33	B2b	tur-1	G2a	CHB-NA18637	O1b
mbu-5	B2b	nep-0186	H*	CHB-NA18621	O3
mbe-237P	B2b	bhu-1151	H1	CHB-NA18622	O3
mbu-13	B2b	nep-0171	H1	bhu-1564	Q1a
aus-m119	C1a	nep-0902	H1	eng-hgQ-1	Q1a
JPT-NA18974	C1a	nep-0809	H1	MXL-NA19664	Q1a
nep-0172	C1a	CEU-NA06994	I1	eng-hgQ-2	Q1b
nep-0273	C1b	CEU-NA11829	I1	nor-16	R1a
bhu-1000	C2	CEU-NA11992	I1	ork-026m	R1a
CHB-NA18612	C2	hun-5	I2	ork-036m	R1a
CHB-NA18620	C2	ire-0114	I2	ser-19	R1a
bhu-1150	D	CEU-NA12003	I2	bhu-0984	R1b
bhu-0957	D1	gre-4	I2	CEU-NA12716	R1b
JPT-NA18940	D2	ire-94	I2	den-190	R1b
YRI-NA18856	E1a	ser-21	I2	gre-10	R1b
bav-13	E1b	TSI-NA20543	J1	ire-0068	R1b
LWK-NA19334	E1b	tur-16	J1	ork-007	R1b
spa-20	E1b	gre-77	J2a	bhu-1157	R2
tur-4	E1b	TSI-NA20805	J2a	bhu-1892	T1a
pal-4929	E1b	tur-13	J2a	gre-12	T1a
TSI-NA20510	E1b	tur-7	J2a	pal-5366	T1a
pal-5225	E1b	gre-2	J2b	TSI-NA20527	T1a
YRI-NA18504	E1b	pal-5232	J2b		
YRI-NA18507	E1b	TSI-NA20588	J2b		

3.3.2 Coverage ranges observed

Promega's prototype PowerSeq™ Auto/Mito/Y System kit was used to generate MPS data for 23 Y-STRs in the 100 samples. With the analytical threshold set to 20 × coverage, a minimum-to-maximum per-allele sequence coverage of 251 - 11,600 × was observed for 24-plex library preparation, and 72 - 11,906 × for 96-plex library preparation. Per-sample, per-STR and per-run statistics are described in Table 3.2.

Table 3.2

Sample coverage statistics.

Values are provided for both library pooling complexity approaches, low throughput (LT) and high throughput (HT) given as overall values (a) and per locus per multiplexity level (b and c).

a.

	LT 24-plex	HT 96-plex
Mean coverage	69,717	22,068
Standard deviation	10,671	6,218
Median coverage	73,308	20,830
Minimum coverage	48,122	12,061
Maximum coverage	84,111	43,410

b. LT 24-plex						c. HT 96-plex					
Y-STR	Mean coverage	Standard deviation	Median coverage	Min coverage	Max coverage	Y-STR	Mean coverage	Standard deviation	Median coverage	Min coverage	Max coverage
DYS19	2041.0	544.7	2203.5	1059	2805	DYS19	963.6	285.7	901.5	517	2490
DYS385a,b	699.3	337.9	620.0	251	1645	DYS385a,b	286.8	172.8	241.0	72	1158
DYS389I	2022.5	515.2	1907.0	1325	2860	DYS389I	539.2	161.8	521.0	256	1010
DYS389II	1276.8	286.0	1284.5	785	1657	DYS389II	454.4	141.5	428.0	169	843
DYS390	2019.0	474.5	1978.5	1290	2694	DYS390	581.3	175.1	551.0	282	1160
DYS391	5882.5	1071.7	6216.0	3278	7066	DYS391	1312.2	551.7	1193.0	552	3002
DYS392	1448.9	745.6	1210.0	581	3218	DYS392	667.0	268.0	624.0	128	1502
DYS393	3263.5	951.0	3329.0	1530	4842	DYS393	1390.6	476.8	1268.5	462	3378
DYS437	6701.6	2013.5	6277.0	4279	11600	DYS437	2873.1	1923.3	2398.0	595	11906
DYS438	2889.3	604.2	2891.0	1970	3956	DYS438	1117.4	396.5	1058.0	410	2443
DYS439	4543.1	923.9	4652.5	2615	5774	DYS439	1456.2	507.8	1377.5	581	4123
DYS448	1765.5	651.0	1515.5	1125	3212	DYS448	430.6	162.8	406.5	116	927
DYS456	7560.5	1328.3	7663.0	4841	9246	DYS456	1598.8	633.6	1530.5	652	4064
DYS458	2070.9	363.9	2038.5	1438	2679	DYS458	645.3	215.0	605.0	302	1347
DYS481	4147.7	558.9	4201.0	3275	5191	DYS481	958.9	450.3	794.5	390	2950
DYS533	2042.7	431.8	2155.5	1207	2741	DYS533	597.9	203.7	572.0	121	1135
DYS549	1921.4	555.4	1888.0	1049	3063	DYS549	848.4	330.2	795.0	205	1751
DYS570	2265.5	601.4	2270.5	982	3032	DYS570	815.7	307.5	720.0	306	1857
DYS576	2896.7	576.3	3066.0	1620	3520	DYS576	1010.0	337.7	941.0	547	2033
DYS635	5546.1	1186.2	5548.5	3371	7206	DYS635	1561.9	496.7	1465.0	562	3780
DYS643	2894.2	738.9	2905.5	1958	4369	DYS643	845.5	318.7	775.5	277	2042
Y-GATA-H4	3189.2	590.6	3215.5	1900	4097	Y-GATA-H4	876.4	305.5	845.5	461	2276

3.3.3 Identified Y-STR alleles

A total of 2311 Y-STR alleles were analysed in the 100 samples; as well as the expected 23 alleles per sample, this included eleven additional alleles, which were interpreted as five allele duplications and six somatic mutant alleles. For calculations at the DYS385a,b loci for each homoallelic combination the presence of two alleles was assumed.

To represent the observed sequence-level variation in a visually comprehensible way, Microsoft Excel was used to build a compressed and uniform summary of the allele range and internal structure of each of the Y-STRs. An example of this is shown in Table 3.3 and the complete table listing all loci and their ranges and sequence variations are provided in Appendix C.

Table 3.3

Example of visual summary of the structures of alleles.

For an example locus, *DYS458*, alleles are listed according to increasing overall length (as in CE) and sequence features are detailed by a bracketed format of allele names, including internal and flanking region SNPs and indels. Allele counts are given for each allele as observed in the set of 100 samples. MPS alleles are heat-scaled for variable number of repeat units and SNPs are highlighted in green and indels in orange.

CE13-20_GAAA[13-19]GGAA[0-1]		Count	CE	MPS	
DYS458	CE13_GAAA[13]	1	13	13	
	CE14_GAAA[13]GGAA[1]	1	14	13	1
	CE14_GAAA[14]	3	14	14	
	CE15_GAAA[14]GGAA[1]	1	15	14	1
	CE15_GAAA[15]	20	15	15	
	CE16_GAAA[15]GGAA[1]	1	16	15	1
	CE16_GAAA[16]	25	16	16	
	CE17_GAAA[17]	23	17	17	
	CE17_GAAA[17]_+32T>C rs549572931 @7,999,934 M11097	1	17	17	
	CE17.2_GAAA[15]AA[1]GAAA[2]	1	17.2	15	1 2
	CE18_GAAA[18]	19	18	18	
	CE19_GAAA[19]	2	19	19	
	CE19_GAAA[19]_+32T>C rs549572931 @7,999,934 M11097	1	19	19	
	CE19_GAAG[1]GAAA[18]	1	19	18	1
	CE19.2_GAAA[17]AA[1]GAAA[2]	1	19.2	17	1 2
	CE20_GAAA[19]GGAA[1]	1	20	19	1


















The allele ranges with detailed sequence variants for all 23 loci were collated according to their CE-equivalent alleles and are shown in Table 3.4.

Table 3.4**Allele ranges and counts of variants observed for all 23 Y-STRs.**

Allele structures and their counts are listed for each locus according to increasing length alleles.

	CE12-17_TCTA[9-14]CCTA[0-1]TCTA[3]	Count	CE
DYS19	CE12_TCTA[9]CCTA[1]TCTA[3]	2	12
	CE12_TCTA[13]	1	12
	CE13_TCTA[10]CCTA[1]TCTA[3]	11 ■	13
	CE14_TCTA[11]CCTA[1]TCTA[3]	32 ■■	14
	CE15_TCTA[12]CCTA[1]TCTA[3]	36 ■■■	15
	CE16_TCTA[13]CCTA[1]TCTA[3]	15 ■■	16
	CE17_TCTA[14]CCTA[1]TCTA[3]	4 ■	17
	CE9-22_AAGG[5-8]GAAA[9-22]	Count	CE
DYS385a,b	CE9_AAGG[5]GAAA[10]	1	9
	CE9_AAGG[6]GAAA[9]	3	9
	CE10_AAGG[6]GAAA[10]	7 ■	10
	CE11_AAGG[6]GAAA[11]	20 ■■	11
	CE12_AAGG[6]GAAA[12]	17 ■■	12
	CE13_AAGG[5]GAAA[14]	1	13
	CE13_AAGG[6]GAAA[13]	28 ■■	13
	CE14_AAGG[6]GAAA[14]	38 ■■■	14
	CE15_AAGG[5]GAAA[16]	1	15
	CE15_AAGG[6]GAAA[15]	18 ■■	15
	CE15_AAGG[8]GAAA[13]	1	15
	CE16_AAGG[6]GAAA[16]	17 ■■	16
	CE16_AAGG[6]GAAA[2]TAAA[1]GAAA[13]	2	16
	CE16_AAGG[8]GAAA[14]	1	16
	CE17_AAGG[5]GAAA[18]	1	17
	CE17_AAGG[6]GAAA[17]	24 ■■	17
	CE18_AAGG[6]GAAA[18]	11 ■■	18
	CE18_AAGG[7]GAAA[17]	2	18
	CE19_AAGG[6]GAAA[19]	5 ■	19
	CE20_AAGG[6]GAAA[20]	1	20
	CE21_AAGG[6]GAAA[21]	1	21
	CE22_AAGG[6]GAAA[22]	1	22

cont.

DYS389I	CE11-15_TAGA[8-12]CAGA[2-3]	Count	CE
	CE11_TAGA[8]CAGA[3]	2	11
	CE12_TAGA[9]CAGA[3]	26 	12
	CE13_TAGA[10]CAGA[3]	48 	13
	CE13_TAGA[11]CAGA[2]	1	13
	CE14_TAGA[11]CAGA[3]	22 	14
	CE15_TAGA[12]CAGA[3]	2	15
DYS389II	CE27-34_TAGA[8-12]CAGA[2-3]N[48]TAGA[10-15]CAGA[4-6]	Count	CE
	CE27_TAGA[8]CAGA[3]N[48]TAGA[11]CAGA[5]	1	27
	CE27_TAGA[9]CAGA[3]N[48]TAGA[11]CAGA[4]	1	27
	CE28_TAGA[10]CAGA[3]N[48]TAGA[10]CAGA[5]	1	28
	CE28_TAGA[8]CAGA[3]N[48]TAGA[12]CAGA[5]	1	28
	CE28_TAGA[9]CAGA[3]N[48]TAGA[11]CAGA[5]	8 	28
	CE28_TAGA[9]CAGA[3]N[48]TAGA[12]CAGA[4]	2	28
	CE29_TAGA[10]CAGA[3]N[48]TAGA[10]CAGA[6]	1	29
	CE29_TAGA[10]CAGA[3]N[48]TAGA[11]CAGA[5]	11 	29
	CE29_TAGA[10]CAGA[3]N[48]TAGA[12]CAGA[4]	4 	29
	CE29_TAGA[11]CAGA[3]N[48]TAGA[10]CAGA[5]	2	29
	CE29_TAGA[11]CAGA[3]N[48]TAGA[11]CAGA[4]	1	29
	CE29_TAGA[9]CAGA[3]N[48]TAGA[11]CAGA[6]	2	29
	CE29_TAGA[9]CAGA[3]N[48]TAGA[12]CAGA[5]	7 	29
	CE30_TAGA[10]CAGA[3]N[48]TAGA[11]CAGA[6]	5 	30
	CE30_TAGA[10]CAGA[3]N[48]TAGA[12]CAGA[5]	17 	30
	CE30_TAGA[10]CAGA[3]N[48]TAGA[13]CAGA[4]	3	30
	CE30_TAGA[11]CAGA[2]N[48]TAGA[13]CAGA[4]	1	30
	CE30_TAGA[11]CAGA[3]N[48]TAGA[11]CAGA[5]	8 	30
	CE30_TAGA[11]CAGA[3]N[48]TAGA[12]CAGA[4]	3	30
	CE30_TAGA[12]CAGA[3]N[48]TAGA[10]CAGA[5]	1	30
	CE30_TAGA[9]CAGA[3]N[48]TAGA[12]CAGA[6]	2	30
	CE30_TAGA[9]CAGA[3]N[48]TAGA[13]CAGA[5]	4 	30
	CE31_TAGA[10]CAGA[3]N[48]TAGA[11]CAGA[1]TAGA[1]CAGA[5]	1	31
	CE31_TAGA[10]CAGA[3]N[48]TAGA[12]CAGA[6]	2	31
	CE31_TAGA[10]CAGA[3]N[48]TAGA[13]CAGA[5]	1	31
	CE31_TAGA[10]CAGA[3]N[48]TAGA[14]CAGA[4]	1	31
	CE31_TAGA[11]CAGA[3]N[48]TAGA[11]CAGA[6]	1	31
	CE31_TAGA[11]CAGA[3]N[48]TAGA[12]CAGA[5]	2	31
	CE31_TAGA[11]CAGA[3]N[48]TAGA[13]CAGA[4]	1	31
	CE31_TAGA[12]CAGA[3]N[48]TAGA[11]CAGA[5]	1	31
	CE32_TAGA[11]CAGA[3]N[48]TAGA[13]CAGA[5]	4 	32
	CE34_TAGA[10]CAGA[3]N[48]TAGA[15]CAGA[6]	1	34
DYS390	CE19-26_TAGA[4-5]CAGA[1]TAGA[8-13]CAGA[4-10]TAGA[1-3]	Count	CE
	CE19_TAGA[4]CAGA[1]TAGA[10]CAGA[4]TAGA[2]	1	19
	CE20_TAGA[4]CAGA[1]TAGA[8]CAGA[7]TAGA[2]	1	20
	CE21_TAGA[4]CAGA[1]TAGA[8]CAGA[8]TAGA[2]	8 	21
	CE21_TAGA[4]CAGA[1]TAGA[9]CAGA[7]TAGA[2]	3	21
	CE22_TAGA[14]CAGA[8]TAGA[2]	1	22
	CE22_TAGA[4]CAGA[1]TAGA[10]CAGA[7]TAGA[2]	2	22
	CE22_TAGA[4]CAGA[1]TAGA[9]CAGA[8]TAGA[2]	20 	22
	CE23_TAGA[4]CAGA[1]TAGA[10]CAGA[8]TAGA[2]	21 	23
	CE23_TAGA[4]CAGA[1]TAGA[11]CAGA[7]TAGA[2]	2	23
	CE23_TAGA[5]CAGA[1]TAGA[9]CAGA[8]TAGA[2]	1	23
	CE24_TAGA[4]CAGA[1]TAGA[10]CAGA[10]TAGA[1]	1	24
	CE24_TAGA[4]CAGA[1]TAGA[10]CAGA[9]TAGA[2]	2	24
	CE24_TAGA[4]CAGA[1]TAGA[11]CAGA[7]TAGA[3]	1	24
	CE24_TAGA[4]CAGA[1]TAGA[11]CAGA[8]TAGA[1]GAGA[1]	1	24
	CE24_TAGA[4]CAGA[1]TAGA[11]CAGA[8]TAGA[2]	26 	24
	CE25_TAGA[4]CAGA[1]TAGA[11]CAGA[9]TAGA[2]	1	25
	CE25_TAGA[4]CAGA[1]TAGA[12]CAGA[8]TAGA[2]	4 	25
	CE26_TAGA[4]CAGA[1]TAGA[12]CAGA[9]TAGA[2]	2	26
	CE26_TAGA[4]CAGA[1]TAGA[13]CAGA[8]TAGA[2]	2	26

cont.

DYS391	CE8-12_TCTA[8-12]	Count	CE
	CE8_TCTA[8]_+50C>A rs112815242 @11,982,182 M8738/CTS1866	2	8
	CE9_TCTA[9]	6	9
	CE9_TCTA[9]_+50C>A rs112815242 @11,982,182 M8738/CTS1866	2	9
	CE10_TCTA[10]	58	10
	CE10_TCTA[10]_+50C>A rs112815242 @11,982,182 M8738/CTS1866	2	10
	CE11_TCTA[11]	23	11
	CE11_TCTA[11]_+50C>A rs112815242 @11,982,182 M8738/CTS1866	2	11
	CE11_TCTG[1]TCTA[10]	1	11
	CE12_TCTA[12]	3	12
	CE12_TCTA[12]_+50C>A rs112815242 @11,982,182 M8738/CTS1866	1	12
DYS392	CE7-16_ATA[7-16]	Count	CE
	CE7_ATA[7]	1	7
	CE10_ATA[10]	3	10
	CE11_ATA[11]	61	11
	CE12_ATA[12]	3	12
	CE13_ATA[13]	15	13
	CE14_ATA[14]	13	14
	CE15_ATA[15]	2	15
	CE16_ATA[16]	1	16
DYS393	CE10-15_AGAT[10-15]	Count	CE
	CE10_AGAT[10]	1	10
	CE11_AGAT[11]	4	11
	CE12_AGAT[12]	16	12
	CE13_AGAT[13]	48	13
	CE13_CGAT[1]AGAT[12]	4	13
	CE14_AGAT[14]	24	14
	CE15_AGAT[15]	3	15
DYS437	CE13-17_TCTA[7-10]TCTG[1-3]TCTA[4]	Count	CE
	CE13_TCTA[7]TCTG[2]TCTA[4]_3C>T rs9786886 @12,346,264 M4790	1	13
	CE14_TCTA[8]TCTG[2]TCTA[4]	54	14
	CE14_TCTA[8]TCTG[2]TCTA[4]_3C>T rs9786886 @12,346,264 M4790	5	14
	CE15_TCTA[9]TCTG[2]TCTA[4]	17	15
	CE15_TCTG[1]TCTA[8]TCTG[2]TCTA[4]	1	15
	CE15_TCTA[10]TCTG[1]TCTA[4]	2	15
	CE16_TCTA[10]TCTG[2]TCTA[4]	18	16
	CE16_TCTA[6]TCTG[1]TCTA[3]TCTG[2]TCTA[4]	1	16
	CE17_TCTA[10]TCTG[3]TCTA[4]	1	17
DYS438	CE8-14_TTTTC[8-14]	Count	CE
	CE8_TTTTC[8]_+21T>C rs761843885 @12,825,969 Z10613	1	8
	CE9_TTTTC[9]	10	9
	CE10_TTTTC[1]TTTTC[1]TTTTC[8]	2	10
	CE10_TTTTC[10]	48	10
	CE10_TTTTC[10]_+7A>C rs760613324 @12,825,955 L255/PF4706	1	10
	CE11_TTTTC[11]	27	11
	CE11_TTTTC[11]_+7A>C rs760613324 @12,825,955 L255/PF4706	1	11
	CE12_TTTTC[12]	9	12
	CE13_TTTTC[13]	1	13
	CE14_TTTTC[14]	1	14

cont.

DYS439	CE10-14_GATA[10-14]	Count	CE
	CE10_GATA[10]	12	10
	CE11_GATA[11]	34	11
	CE11_GATA[11]_+3A>T SNP @12,403,567	1	11
	CE12_GATA[12]	35	12
	CE13_GATA[13]	14	13
DYS448	CE14_GATA[14]	4	14
	CE13-23_AGAGAT[5-15]N[42]AGAGAT[6-10]	Count	CE
	CE13_AGAGAT[5]N[42]AGAGAT[8]	1	13
	CE17_AGAGAT[10]N[42]AGAGAT[7]	2	17
	CE18_AGAGAT[10]N[42]AGAGAT[8]	7	18
	CE18_AGAGAT[11]N[42]AGAGAT[7]	6	18
	CE19_AGAGAT[10]N[42]AGAGAT[9]	1	19
	CE19_AGAGAT[11]N[42]AGAGAT[8]	26	19
	CE19_AGAGAT[12]N[42]AGAGAT[7]	6	19
	CE19_AGAGAT[13]N[42]AGAGAT[6]	1	19
	CE20_AGAGAT[11]N[42]AGAGAT[9]	4	20
	CE20_AGAGAT[12]N[42]AGAGAT[8]	20	20
	CE20_AGAGAT[13]N[42]AGAGAT[7]	1	20
	CE20.4_AGAGAT[3]AGAT[1]AGAGAT[9]N[42]AGAGAT[8]	1	20.4
	CE21_AGAGAT[12]N[42]AGAGAT[9]	9	21
	CE21_AGAGAT[13]N[42]AGAGAT[8]	10	21
	CE22_AGAGAT[13]N[42]AGAGAT[9]	2	22
	CE22_AGAGAT[14]N[42]AGAGAT[8]	1	22
	CE23_AGAGAT[13]N[42]AGAGAT[10]	1	23
	CE23_AGAGAT[14]N[42]AGAGAT[9]	1	23
	CE23_AGAGAT[15]N[42]AGAGAT[8]	1	23
DYS456	CE13-17_AGAT[13-17]	Count	CE
	CE13_AGAT[13]	4	13
	CE14_AGAT[14]	17	14
	CE15_AGAT[15]	46	15
	CE16_AGAT[16]	24	16
	CE17_AGAT[17]	9	17
DYS458	CE13-20_GAAA[13-19]GGAA[0-1]	Count	CE
	CE13_GAAA[13]	1	13
	CE14_GAAA[13]GGAA[1]	1	14
	CE14_GAAA[14]	3	14
	CE15_GAAA[14]GGAA[1]	1	15
	CE15_GAAA[15]	20	15
	CE16_GAAA[15]GGAA[1]	1	16
	CE16_GAAA[16]	25	16
	CE17_GAAA[17]	23	17
	CE17_GAAA[17]_+32T>C rs549572931 @7,999,934 M11097	1	17
	CE17.2_GAAA[15]AA[1]GAAA[2]	1	17.2
	CE18_GAAA[18]	19	18
	CE19_GAAA[19]	2	19
	CE19_GAAA[19]_+32T>C rs549572931 @7,999,934 M11097	1	19
	CE19_GAAG[1]GAAA[18]	1	19
	CE19.2_GAAA[17]AA[1]GAAA[2]	1	19.2
	CE20_GAAA[19]GGAA[1]	1	20

cont.

	CE19-30_CTG[0-2]CTT[19-30]	Count	CE
DYS481	CE19_CTG[1]CTT[19]	1	19
	CE20_CTG[1]CTT[20]	2	20
	CE21_CTG[1]CTT[21]	10	21
	CE21_CTG[2]CTT[20]	6	21
	CE22_CTG[1]CTT[22]	11	22
	CE22_CTG[1]CTT[22]_-13G>A rs368663163 @8,558,321 L266/PF6108	1	22
	CE22_CTG[2]CTT[21]	1	22
	CE23_CTG[1]CTT[23]	13	23
	CE24_CTG[1]CTT[24]	13	24
	CE24_CTT[25]	1	24
	CE25_CTG[1]CTT[25]	14	25
	CE25_CTG[2]CTT[24]	1	25
	CE26_CTG[1]CTT[26]	10	26
	CE26_CTG[2]CTT[25]	1	26
	CE26_CTT[27]	1	26
	CE27_CTG[1]CTT[27]	8	27
	CE27_CTT[28]	2	27
	CE28_CTG[1]CTT[28]	1	28
	CE28_CTG[1]CTT[3]CCT[1]CTT[24]	1	28
	CE29_CTG[1]CTT[29]	1	29
	CE30_CTG[1]CTT[30]	1	30
	CE9-15_TATC[9-15]	Count	CE
DYS533	CE9_TATC[9]	4	9
	CE10_TATC[10]	10	10
	CE11_TATC[11]	37	11
	CE11_TATC[7]TGTC[1]TATC[3]	1	11
	CE12_TATC[12]	42	12
	CE13_TATC[13]	3	13
	CE14_TATC[14]	1	14
	CE14.1_TATC[11]_-48.1->CTCTTCTAACTAT indel @16,281,301	1	14.1
	CE15_TATC[15]	1	15
	CE10-15_GATA[10-15]	Count	CE
DYS549	CE10_GATA[10]	5	10
	CE11_GATA[11]	23	11
	CE12_GATA[12]	42	12
	CE13_GATA[13]	26	13
	CE14_GATA[14]	3	14
	CE15_GATA[15]	1	15
	CE14-21_TTTC[14-21]	Count	CE
DYS570	CE14_TTTC[14]	1	14
	CE15_TTTC[15]	2	15
	CE16_TTTC[16]	17	16
	CE16_TTTC[16]_+4T>G rs763920632 @6,993,261 PH250	1	16
	CE17_TTCC[1]TTTC[16]	1	17
	CE17_TTTC[15]CTTC[1]TTTC[1]	1	17
	CE17_TTTC[17]	22	17
	CE18_TTTC[18]	23	18
	CE19_TTTC[19]	13	19
	CE19_TTTC[5]TCTC[1]TTTC[13]	1	19
	CE20_TTTC[20]	11	20
	CE21_TTTC[21]	7	21

cont.

DYS576	CE14-21_AAAG[14-21]	Count	CE
	CE14_AAAG[14]	3	14
	CE15_AAAG[15]	10	15
	CE16_AAAG[16]	25	16
	CE17_AAAG[17]	21	17
	CE17.1_AAAG[18]_+3AAA>- indel @7,185,388	1	17.1
	CE18_AAAG[18]	21	18
	CE19_AAAG[19]	15	19
	CE20_AAAG[20]	3	20
	CE21_AAAG[21]	1	21
DYS635	CE17-26_TAGA[7-15]TACA[2]{TAGA[2]TACA[2]}{1-2}TAGA[4]	Count	CE
	CE17_TAGA[7]TACA[2]TAGA[2]TACA[2]TAGA[4]	2	17
	CE18_TAGA[8]TACA[2]TAGA[2]TACA[2]TAGA[4]	3	18
	CE19_TAGA[9]TACA[2]TAGA[2]TACA[2]TAGA[4]	4	19
	CE20_TAGA[10]TACA[2]TAGA[2]TACA[2]TAGA[4]	11	20
	CE20_TAGA[8]CAGA[1]TAGA[1]TACA[2]TAGA[2]TACA[2]TAGA[4]	1	20
	CE21_TAGA[11]TACA[2]TAGA[2]TACA[2]TAGA[4]	38	21
	CE21_TAGA[9]CAGA[1]TAGA[1]TACA[2]TAGA[2]TACA[2]TAGA[4]	1	21
	CE21.3_TAGA[2]TGA[1]TAGA[5]TACA[2]TAGA[2]TACA[2]TAGA[2]TACA[2]TAGA[4]	1	21.3
	CE22_TAGA[12]TACA[2]TAGA[2]TACA[2]TAGA[4]	13	22
	CE22_TAGA[8]TACA[2]TAGA[2]TACA[2]TAGA[2]TACA[2]TAGA[4]	3	22
	CE23_TAGA[13]TACA[2]TAGA[2]TAGA[2]TAGA[4]	7	23
	CE23_TAGA[9]TACA[2]TAGA[2]TAGA[2]TAGA[2]TACA[2]TAGA[4]	9	23
	CE24_TAGA[10]TACA[2]TAGA[2]TACA[2]TAGA[2]TACA[2]TAGA[4]	1	24
	CE24_TAGA[14]TACA[2]TAGA[2]TACA[2]TAGA[4]	4	24
	CE25_TAGA[14]TACA[3]TAGA[2]TACA[2]TAGA[4]	1	25
	CE25_TAGA[15]TACA[2]TAGA[2]TACA[2]TAGA[4]	1	25
	CE26_TAGA[12]TACA[2]TAGA[2]TACA[2]TAGA[2]TACA[2]TAGA[4]	1	26
DYS643	CE7-15_CTTTT[7-15]	Count	CE
	CE7_CTTTT[7]	1	7
	CE8_CTTTT[8]	2	8
	CE9_CTTTT[9]	14	9
	CE10_CTTTT[10]	26	10
	CE11_CTTTT[11]	23	11
	CE11_CTTTT[11]_-7A>G SNP @15,314,125	1	11
	CE12_CTTTT[12]	21	12
	CE13_CTTTT[13]	9	13
	CE14_CTTTT[14]	3	14
	CE15_CTTTT[15]	1	15
Y-GATA-H4	CE8-13_TCTA[8-13]	Count	CE
	CE8_TCTA[8]	2	8
	CE9_TCTA[9]	1	9
	CE10_TCTA[10]	12	10
	CE11_TCTA[11]	40	11
	CE12_TCTA[12]	42	12
	CE13_TCTA[13]_+36A>G SNP @16,631,756 Y15322/Z34275	1	13
	CE13_TCTA[13]	3	13

Eleven supernumerary alleles in ten samples were deduced from read-depth values, as shown for two examples in Figure 3.2, and all are detailed in Table 3.5 using relative read-depth ratios. Five of these showed approximately double-dose coverage, suggesting duplication, whereas six had about a single-dose coverage split between the two alleles, and are therefore more likely to represent somatic mutations.

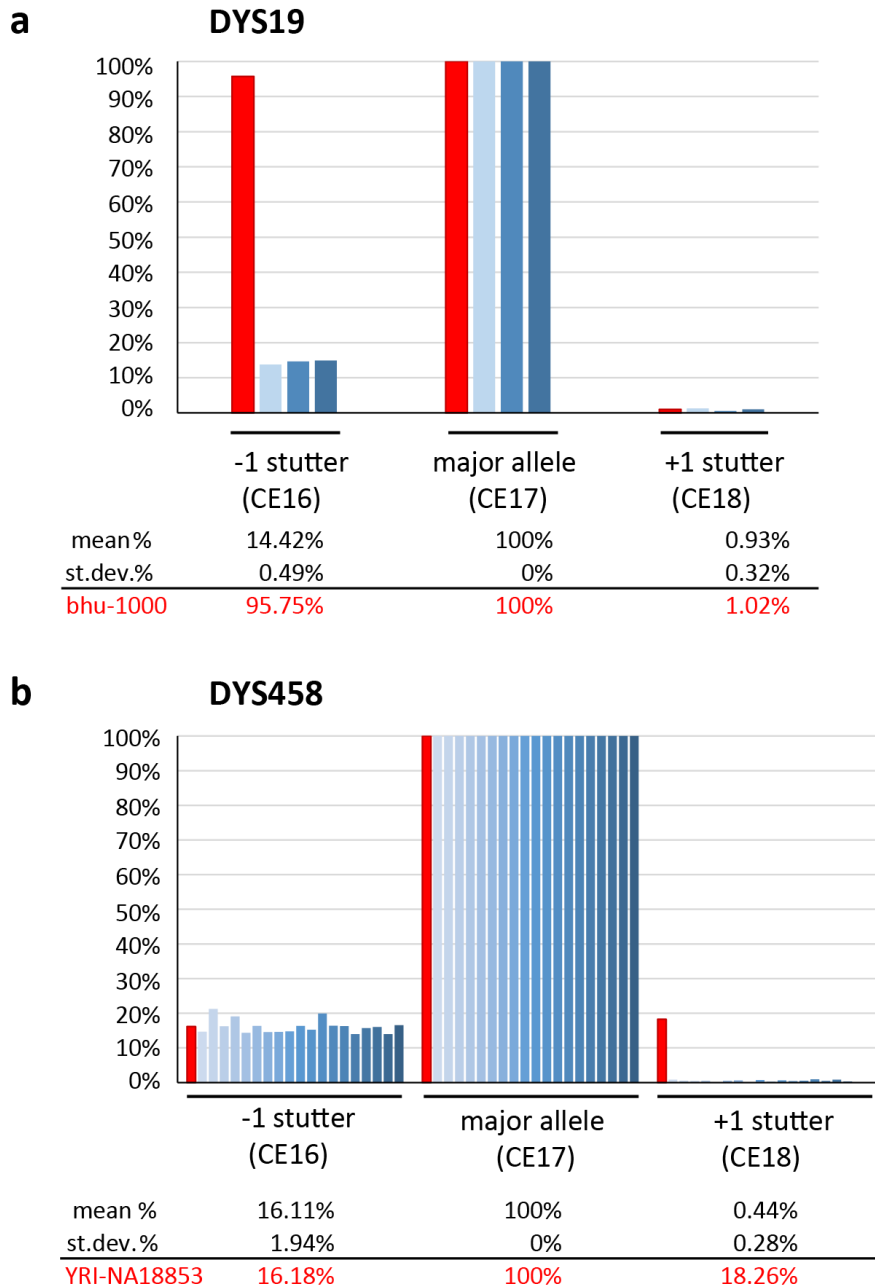


Figure 3.2

Examples of allele duplications and somatic mutations.

Read-depth of the two DYS19 alleles (a) in sample bhu-1000 (in red), compared to same-size alleles in other samples (blue) and their corresponding stutters. Read-depth of the two DYS458 alleles (b) in sample YRI-NA18853 (red), compared to same-size alleles in other samples (blue) and their corresponding stutters. The sample has an allele in +1 position with read-depth much greater than the +1 stutter in other samples (blue) and their corresponding stutters.

Table 3.5
Supernumerary alleles.

Eleven cases of Y-STRs showing supernumerary alleles (a) listed with their sequences, relative read depths (%) and position compared to the major alleles. Mean stutter percentages were calculated from the rest of the samples with the same MPS alleles and used for comparison (as shown in Figure 3.2) and read-depth values were adjusted for relevant stutter. Comparison between the tested loci and their comparable-size reference pair (b) shows whether the tested supernumerary alleles are more likely to be a result of duplications or somatic mutations by gauging an expected ratio range between the two markers.

a

Sample	Type	Hg	Y-STR affected	CE	MPS alleles	to main allele		samples (=MPS)	mean stutter %		stutter adjusted %
						%	relative position		(-1)	(+1)	
gre-12	whB	T1a	DYS385	14	CE14_AAGG[6]GAAA[14]	100%		5	17.56%	0.33%	100%
				13	CE13_AAGG[6]GAAA[13]	98%	-1				91%
				15	CE15_AAGG[6]GAAA[15]	71%	+1				80%
bhu-1000	whB	C2	DYS19	17	CE17_TCTA[14]CCTA[1]TCTA[3]	100%		3	14.42%	0.93%	100%
				16	CE16_TCTA[13]CCTA[1]TCTA[3]	96%	-1				82%
nor-16	sal	R1a	GATA-H4	11	CE11_TCTA[11]	100%		39	7.92%	1.16%	100%
				13	CE13_TCTA[13]	77%	+2				77%
YRI-NA18856	LCL	E1a	DYS448	19	CE19_AGAGAT[11]N[42]AGAGAT[8]	100%		25	2.49%	0.00%	100%
				20	CE20_AGAGAT[11]N[42]AGAGAT[9]	96%	+1				98%
CHB-NA18608	LCL	N	DYS458	16	CE16_GAAA[16]	100%		24	15.21%	0.27%	100%
				17	CE17_GAAA[17]	75%	+1				85%
bhu-1564	whB	Q1a	DYS389II	30	CE30_TAGA[11]CAGA[3]N[48]TAGA[11]CAGA[5]	100%		7	19.96%	0.19%	100%
				31	CE31_TAGA[12]CAGA[3]N[48]TAGA[11]CAGA[5]	25%	+1				26%
bhu-1564	whB	Q1a	DYS389I	14	CE14_TAGA[11]CAGA[3]	100%		21	9.78%	0.18%	100%
				15	CE15_TAGA[12]CAGA[3]	26%	+1				27%
JPT-NA18974	LCL	C1a	DYS643	11	CE11_CTTTT[11]	100%		22	3.04%	0.14%	100%
				11	CE11_CTTTT[11]_-7A>G SNP @15,314,125	63%	=				63%
YRI-NA18853	LCL	E1b	DYS458	17	CE17_GAAA[17]	100%		21	15.86%	0.44%	100%
				18	CE18_GAAA[18]	18%	+1				18%
ork-007	LCL	R1b	DYS438	11	CE11_TTTTC[11]	100%		26	3.89%	0.08%	100%
				12	CE12_TTTTC[12]	61%	+1				63%
gre-77	whB	J2a	DYS635	22	CE22_TAGA[12]TACA[2]TAGA[2]TACA[2]TAGA[4]	100%		12	10.81%	0.63%	100%
				23	CE23_TAGA[13]TACA[2]TAGA[2]TACA[2]TAGA[4]	70%	+1				76%

Sample types: whB - whole blood; sal - saliva; LCL - lymphoblastoid cell line
The corresponding stutter % and extra alleles % are underlined.

b

Sample	Y-STR		relative read-depth ratio test statistics				test ratios		summed test ratios/ mean	result more likely
	test	reference	MEAN	MEDIAN	MIN	MAX	summed	each		
gre-12	DYS385	GATA-H4	0.62	0.63	0.13	1.33	1.42	0.53 0.48 0.42	2.31	duplication
bhu-1000	DYS19	DYS643	1.13	1.11	0.52	2.26	2.32	1.27 1.04	2.04	duplication
nor-16	GATA-H4	DYS448	2.09	2.02	0.94	3.64	3.99	2.25 1.73	1.90	duplication
YRI-NA18856	DYS448	GATA-H4	0.53	0.49	0.27	1.07	0.91	0.46 0.45	1.72	duplication
CHB-NA18608	DYS458	DYS576	0.66	0.64	0.40	1.09	0.93	0.50 0.43	1.41	duplication
bhu-1564	DYS389II	DYS533	0.78	0.76	0.33	1.49	0.93	0.74 0.19	1.18	somatic mutation
bhu-1564	DYS389I	DYS456	0.36	0.35	0.14	0.73	0.41	0.33 0.09	1.15	somatic mutation
JPT-NA18974	DYS643	DYS19	0.96	0.90	0.44	1.92	1.06	0.65 0.41	1.10	somatic mutation
YRI-NA18853	DYS458	DYS576	0.66	0.64	0.40	1.09	0.59	0.49 0.09	0.89	somatic mutation
ork-007	DYS438	DYS481	1.20	1.16	0.43	2.19	0.90	0.55 0.35	0.75	somatic mutation
gre-77	DYS635	DYS570	2.08	2.03	0.92	3.43	1.56	0.89 0.67	0.75	somatic mutation

In these cases, with the exception of sample bhu-1564, only one Y-STR was affected, and even in this sample the loci are physically one (DYS389 I and II). This, and the fact that a non Y-chromosomal comparison (between the Amelogenin X and Y sequences) also showed about equal dosage between the sex chromosomes in these males, suggests that these are probably localised duplications.

To consider sample type in relation to the duplications and somatic mutations, the biological origins of samples were compared in the ten cases presenting these eleven cases of supernumerary Y-STR alleles. Five lymphoblastoid cell lines (LCLs), four whole-blood and one saliva sample were among these, but these types do not correlate in an obvious way with the classes. The saliva was deemed to be a duplication, as were two of the five LCL samples, and two of the four whole-blood samples. LCLs (particularly long-established ones) are known to present genomic instability in STRs (Lee et al. 2013), resulting in somatic mutations and showing multi-allelic patterns, but LCLs were not found to be overrepresented in the samples with supernumerary alleles. Contrary to expectations, no direct correlation was

found between LCLs and observed somatic mutations, possibly due to the relatively low sample size.

3.3.4 Concordance of MPS data with CE-defined alleles

Sequence-derived repeat array lengths were compared to previously-determined CE-based PowerPlex® Y23 data (Hallast et al. 2015). Four of 2311 alleles (0.17%) were found to be discordant between the two methods as shown in Table 3.6, and as highlighted in Figure 3.3. Of these, one could be resolved by examining full-length sequence (an insertion of 13 bp in the flanking DNA), one by a SNP-based mobility shift that has been previously noted elsewhere (Lee et al. 2016), and the remaining two by possible differences in the positions of proprietary PCR primers for MPS and CE kits.

Table 3.6
Details of discordant alleles.

Discordances observed between length-based (CE) and sequence-derived (MPS) calls and the details of the nature of discordance.

Sample	Locus	Meta-pop.	CE	MPS call		Discordance details	Bracketed sequence string
				array only	full length		
nep-0172	DYS533	Asi.	14.1	11	14.1	11 repeats +13 bp insertion in the 5' flank	ATCTACCTAATATTTATCTATATCATTCTAATTATGT CTCTTCTAACTATATAACTATGTCTCTCTAACTATA TTATCTATCAATCTTCTACCTATCATCTTTCTAGCTA GCTATCATC(1)TATC(11)ATCTATCATCTTCTATTGT TTGGTTGAGTTAAGAACTGATCATGAATAAATACAT TTCATTGGTGATCTC(1)
bhu-1157	DYS481	Asi.	22.1	22	22	rs368663163 mobility shift in CE Lee et al. (2016)	TAAAAGGAATGTGGCTAACACTGTTTCAGCATG(1) CTG(1)CTT(22)TTTTGAGTCT(1)
CHB-NA18637	DYS635	Asi.	21.3	22	22	deletion ? outside of MPS primers	GCCCAAATATCCATCAATCAATGAATGGATAAAGA AAATGTGA(1)TAGA(12)TACA(2)TAGA(2)TACA(2) TAGA(4)GATT(1)
mbe-237P	DYS392	Afr.	11	-	-	null allele likely PBSM	-

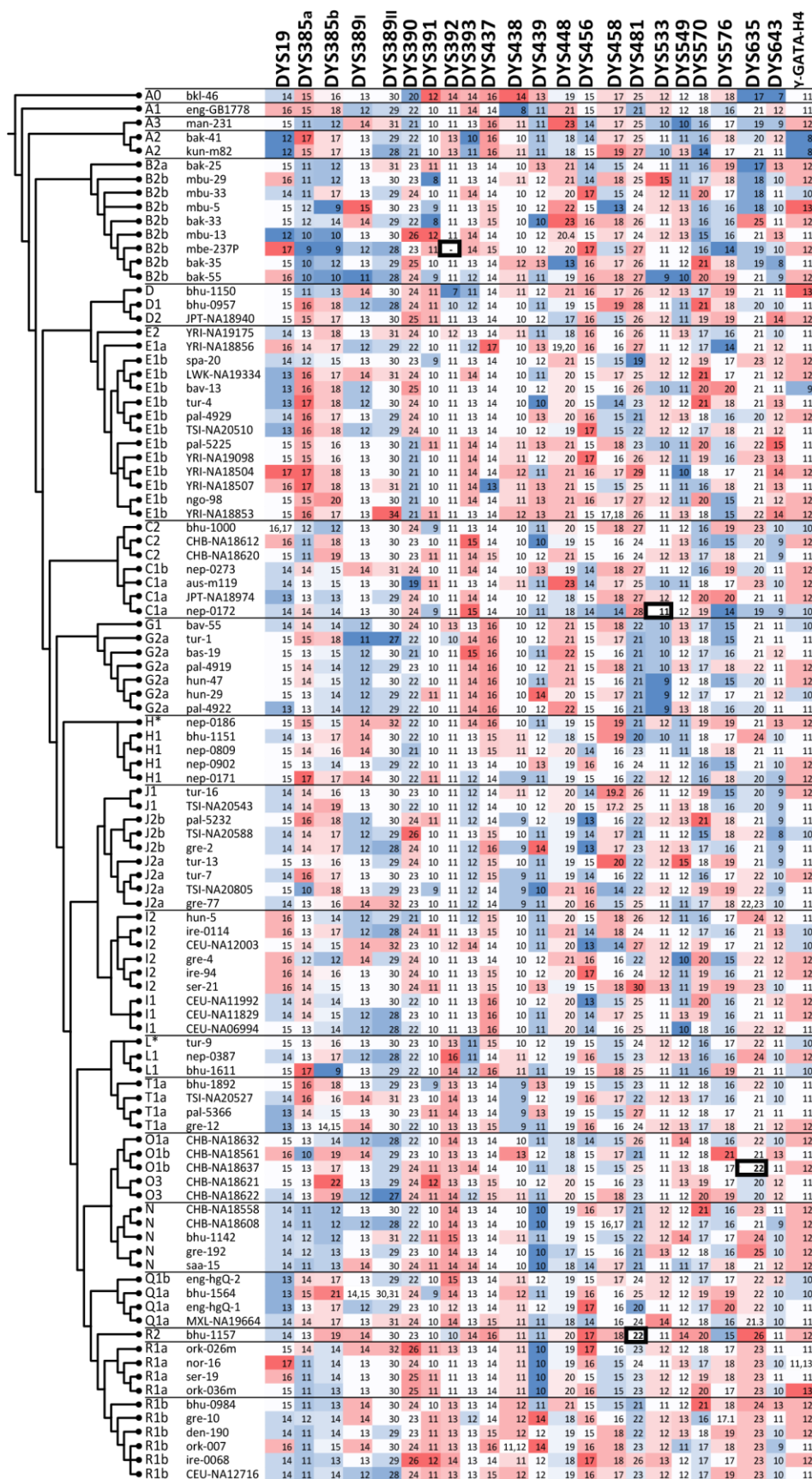


Figure 3.3
Length-based Y-STR allele phylogeny.

Length based CE-equivalent alleles derived from the MPS data in the 100 samples displayed in their phylogenetic context. The alleles are heat-scaled by their length. The four discordant alleles are bolded and boxed with thick black lines.

3.3.5 Diversity of observed alleles

The samples contain a total of 267 distinct sequence-based Y-STR alleles, an overall 58% increase from the 169 length-based alleles distinguishable by CE as detailed in Table 3.7 and in Figure 3.4.

Table 3.7
Increase of allele diversity due to sequencing.

Comparison of number of alleles for each Y-STR based on length only (as using CE) and by full sequence information (by MPS).

Y-STR	Count of length-based alleles	Count of sequence-based alleles	Increase in number of alleles (%)	Novel sequence variants in this study
DYS389II	7	32	357.1	4
DYS390	8	19	137.5	6
DYS448	9	19	111.1	5
DYS391	5	10	100.0	6
DYS437	5	9	80.0	2
DYS481	12	21	75.0	3
DYS458	10	16	60.0	9
DYS385a,b	14	22	57.1	7
DYS635	11	17	54.5	4
DYS570	8	12	50.0	4
DYS438	7	10	42.9	2
DYS389I	5	6	20.0	0
DYS439	5	6	20.0	1
DYS19	6	7	16.7	1
DYS393	6	7	16.7	0
GATAH4	6	7	16.7	1
DYS533	8	9	12.5	2
DYS643	9	10	11.1	2
DYS392	8	8	0.0	0
DYS456	5	5	0.0	0
DYS549	6	6	0.0	0
DYS576	9	9	0.0	1
Total	169	267	58.0	60

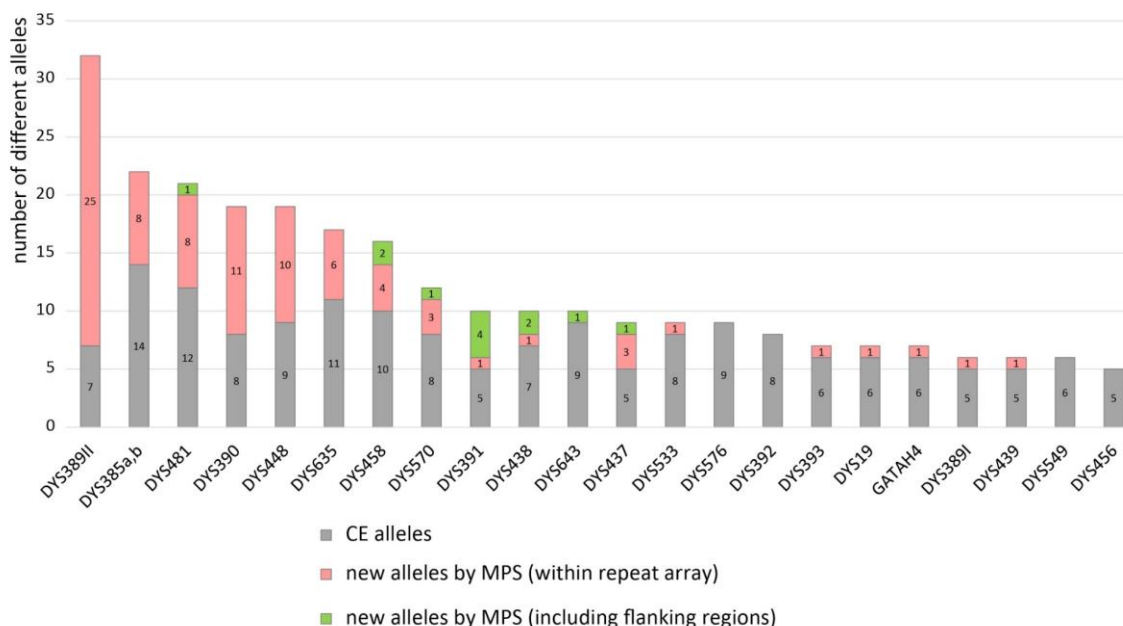


Figure 3.4
Allele diversity increase by type of variants.

Increase in number of different alleles for each Y-STR using sequence information from MPS analysis.

All but four Y-STRs showed increased allelic diversity when analysed by MPS. Isometric allele groups are the alleles observed with the same fragment length, but showing different sequences; the observed isometric allele groups are summarised in Table 3.8. Most of these (68.3%) are represented by just two sequence-based alleles having the same length; however, only a few groups consisted of from 5 to 9 sequence-based alleles per single length class - the latter was observed at DYS389II for length allele 30. The number of loci showing such high numbers of sequence variants per length was limited, and the majority of them had isometric allele groups with three or fewer sequence-based alleles per CE allele.

Table 3.8**Summary of isometric allele groups.***Summary of isometric allele groups among 23 Y-STRs analysed by MPS.*

	# of MPS alleles found per single CE allele							
	2	3	4	5	6	7	8	9
total # of isometric allele groups	41	13	2	1	–	1	1	1
# of Y-STRs with isometric allele groups	19	10	2	1	–	1	1	1

Although an online resource to collect STR sequence variation under an international collaboration is hosted by NCBI as STRSeq BioProject (Gettings et al. 2017), data for the Y-Chromosomal STR loci sub-project (Accession: PRJNA380347) are limited for query (<https://www.ncbi.nlm.nih.gov/bioproject/380347>, at the time of writing). Therefore the sequence variants were also compared to a comprehensive list of previously published literature (Churchill et al. 2016; D'Amato et al. 2010; Forster et al. 1998; Just et al. 2017; Kwon et al. 2016; Lee et al. 2016; Novroski et al. 2016; Redd et al. 2002; Ruitberg et al. 2001; Warshauer et al. 2015b; Wendt et al. 2016; Wendt et al. 2017; Zhao et al. 2015), (detailed in Appendix C). Fifty-eight of the 60 novel Y-STR variants in phase with their flanking sequences detailed in Table 3.9 were not reported elsewhere, and about a year after publication, only two of the 60 novel alleles were given accession numbers (as noted in Table 3.9) in the STRSeq Y-STR-specific BioProject.

Table 3.9

List of novel Y-STR sequence variants defined by MPS.

STRSeq bioproject accession numbers are noted in green font for two alleles, one for DYS389II and another for DYS643, as they recently became catalogued by STRSeq.

Y-STR	Y-STR definition; Novel sequence variants	Observed #	Aspects of novelty
DYS19	<i>[TCTA]_a ccta [TCTA]_b</i> (Parson et al. 2016)		
	CE12_TCTA[13] _{a+b} ccta[0]	1	SNP internal to repeat array, allele name is a+b-1 for compatibility to CE
DYS385a,b	<i>DYS385a [TTTC]_a / DYS385b [GAAA]_a</i> (Parson et al. 2016)		
	<i>DYS385a,b [aagg]₅₋₉ [GAAA]_a</i> (this study)		
	CE9_AAGG[5]GAAA[10]	1	new combination of repeat units; upstream flanking region previously considered non-variable, but shows high level of variation in number of repeats; therefore here considered part of the repeat array as AAGG[5-9] (Novroski et al. 2016 also found AAGG[9])
	CE13_AAGG[5]GAAA[14]	1	
	CE15_AAGG[5]GAAA[16]	1	
	CE15_AAGG[8]GAAA[13]	1	
	CE16_AAGG[8]GAAA[14]	1	
	CE17_AAGG[5]GAAA[18]	1	
	CE18_AAGG[7]GAAA[17]	2	
DYS389II	<i>[TAGA]_a [CAGA]_b N48 [TAGA]_c [CAGA]_d</i> (Parson et al. 2016)		
	CE30_TAGA[11]CAGA[2]N[48]TAGA[13]CAGA[4]	1	shorter first CAGA array
	CE30_TAGA[9]CAGA[3]N[48]TAGA[12]CAGA[6]	2	new combination of repeat units; MK990451
	CE31_TAGA[10]CAGA[3]N[48]TAGA[11]CAGA[1]TAGA[1]CAGA[5]	1	SNP internal to repeat array
	CE34_TAGA[10]CAGA[3]N[48]TAGA[15]CAGA[6]	1	longer second TAGA array
DYS390	<i>[TAGA]_a [CAGA]_b [TAGA]_c [CAGA]_d</i> (Parson et al. 2016)		
	<i>[TAGA]_a [CAGA]_b [TAGA]_c [CAGA]_d [taga]₁₋₃</i> (this study)		
	CE22_TAGA[14] _{a+c} CAGA[0]CAGA[8]TAGA[2]	1	SNP internal to repeat array
	CE23_TAGA[5]CAGA[1]TAGA[9]CAGA[8]TAGA[2]	1	longer first TAGA array
	CE24_TAGA[4]CAGA[1]TAGA[10]CAGA[10]TAGA[1]	1	longer second CAGA / shorter third TAGA array
	CE24_TAGA[4]CAGA[1]TAGA[11]CAGA[7]TAGA[3]	1	longer third TAGA array
	CE24_TAGA[4]CAGA[1]TAGA[11]CAGA[8]TAGA[1]GAGA[1]	1	SNP internal to repeat array
	CE26_TAGA[4]CAGA[1]TAGA[12]CAGA[9]TAGA[2]	2	new combination of repeat units
DYS391	<i>[TCTA]_a</i> (Parson et al. 2016)		
	CE8_TCTA[8] ₊ +50C>A rs112815242 @11,982,182 M8738/CTS1866	2	SNP in the flanking region
	CE9_TCTA[9] ₊ +50C>A rs112815242 @11,982,182 M8738/CTS1866	2	SNP in the flanking region
	CE10_TCTA[10] ₊ +50C>A rs112815242 @11,982,182 M8738/CTS1866	2	SNP in the flanking region
	CE11_TCTA[11] ₊ +50C>A rs112815242 @11,982,182 M8738/CTS1866	2	SNP in the flanking region
	CE11_TCTG[1]TCTA[10]	1	SNP internal to repeat array
	CE12_TCTA[12] ₊ +50C>A rs112815242 @11,982,182 M8738/CTS1866	1	SNP in the flanking region
DYS437	<i>[TCTA]_a [TCTG]_b [TCTA]₄</i> (STRBase, accessed on 03 Nov 2017)		
	CE15_TCTG[1]TCTA[8]TCTG[2]TCTA[4]	1	SNP internal to repeat array
	CE16_TCTA[6]TCTG[1]TCTA[3]TCTG[2]TCTA[4]	1	SNP internal to repeat array
DYS438	<i>[TTTTTC]_a</i> (Parson et al. 2016)		
	CE8_TTTTC[8] ₊ +21T>C rs761843885 @12,825,969 Z10613	1	shorter array; SNP in the flanking region
	CE11_TTTTC[11] ₊ +7A>C rs760613324 @12,825,955 L255/PF4706	1	SNP in the flanking region
DYS439	<i>[GATA]_a</i> (Parson et al. 2016)		
	CE11_GATA[11] ₊ +3A>T SNP @12,403,567	1	SNP in the flanking region
DYS448	<i>[AGAGAT]_a N42 [AGAGAT]_b</i> (Parson et al. 2016)		
	CE13_AGAGAT[5]N[42]AGAGAT[8]	1	shorter first AGAGAT array
	CE19_AGAGAT[13]N[42]AGAGAT[6]	1	shorter second AGAGAT array
	CE20.4_AGAGAT[3]AGAT[1]AGAGAT[9]N[42]AGAGAT[8]	1	indel in the repeat array
	CE23_AGAGAT[14]N[42]AGAGAT[9]	1	new combination of repeat units
	CE23_AGAGAT[15]N[42]AGAGAT[8]	1	longer first AGAGAT array

Cont.

Y-STR	Y-STR definition; Novel sequence variants	Observed #	Aspects of novelty
DYS458	<u>/GAAA/a</u> (Redd et al. 2002)		
	CE14_GAAA[13]GGAA[1]	1	SNP internal to repeat array
	CE15_GAAA[14]GGAA[1]	1	SNP internal to repeat array
	CE16_GAAA[15]GGAA[1]	1	SNP internal to repeat array
	CE17_GAAA[17]_+32T>C rs549572931 @7,999,934 M11097	1	SNP in the flanking region
	CE17.2_GAAA[15]AA[1]GAAA[2]	1	indel in the repeat array
	CE19_GAAA[19]_+32T>C rs549572931 @7,999,934 M11097	1	SNP in the flanking region
	CE19_GAAG[1]GAAA[18]	1	SNP internal to repeat array
	CE19.2_GAAA[17]AA[1]GAAA[2]	1	indel in the repeat array
	CE20_GAAA[19]GGAA[1]	1	SNP internal to repeat array
DYS481	<u>/CTT/a</u> (Parson et al. 2016)		
	<u>/ctg/0-2</u> <u>/CTT/a</u> (this study)		
	CE26_CTG[0]CTT[27]	1	new combination of repeat units
	CE27_CTG[0]CTT[28]	2	new combination of repeat units
DYS533	<u>/TATC/a</u> (Parson et al. 2016)		
	CE14.1_TATC[11]_-48.1->CTCTCTAACTAT indel @16,281,301	1	indel in the flanking region
	CE15_TATC[15]	1	longer repeat unit in array
DYS570	<u>/TTTC/a</u> (Parson et al. 2016)		
	CE16_TTTC[16]_+4T>G rs763920632 @6,993,261 PH250	1	SNP in the flanking region
	CE17_TTCC[1]TTTC[16]	1	SNP internal to repeat array
	CE17_TTTC[15]CTTC[1]TTTC[1]	1	SNP internal to repeat array
	CE19_TTTC[5]TCTC[1]TTTC[13]	1	SNP internal to repeat array
DYS576	<u>/AAAG/a</u> (Parson et al. 2016)		
	CE17.1_AAAG[18]_+3AAA> indel @7,185,388	1	indel in the flanking region
DYS635	<u>/TAGA/a</u> <u>/TACA/b</u> <u>/TAGA/c</u> <u>/TACA/d</u> <u>/TAGA/e</u> <u>/TACA/f</u> <u>/TAGA/g</u> (Parson et al. 2016)		
	CE18_TAGA[8]TACA[2]TAGA[2]TACA[2]TAGA[4]	3	new combination of repeat units
	CE20_TAGA[8]CAGA[1]TAGA[1]TACA[2]TAGA[2]TAGA[2]TAGA[4]	1	SNP internal to repeat array
	CE21_TAGA[9]CAGA[1]TAGA[1]TACA[2]TAGA[2]TAGA[2]TAGA[4]	1	SNP internal to repeat array
	CE25_TAGA[14]TACA[3]TAGA[2]TACA[2]TAGA[4]	1	SNP internal to repeat array
DYS643	<u>/CTTTT/a</u> (Parson et al. 2016)		
	CE11_CTTT[11]_-7A>G SNP @15,314,125	1	SNP in the flanking region
	CE15_CTTT[15]	1	longer repeat unit in array; MK990441
Y-GATA-H4	<u>/TCTA/a</u> (Parson et al. 2016)		
	CE13_TCTA[13]_+36A>G SNP @16,631,756 Y15322/Z34275	1	SNP in the flanking region

Newly arising Y-STR variants may result from single nucleotide changes, insertions or deletions affecting the repeats themselves, or the flanking regions. This study found 22 different SNPs or indels in the repeat regions in 27 distinct alleles of 15 Y-STRs, in 27 of the 100 samples. It is of paramount importance to analyse full-length sequences, rather than solely the repeat region, because the flanking regions contribute to the analysed length, and their omission can therefore lead to discordance with CE-based allele calls (as seen, for example, for *DYS533* in Table 3.6). Therefore twelve different flanking region SNPs or indels are also described here in 19 distinct alleles of 11 Y-STRs; such flanking region variants are observed in 26 of the 100 analysed samples. Altogether, 34 different SNPs or indels were

detected in 46 distinct alleles of 19 Y-STRs, observed in 43 of the 100 samples. These variations are shown in Figure 3.5 together with their schematic phylogenetic relationships, and their details are provided in Table 3.10. The ‘multiple SNPs’ internal to DYS635 (alternatively regarded as an RPV) are found in 85/100 samples because the GRCh38 reference assembly (Skaletsky et al. 2003) carries the same derived state as superhaplogroup P, and hence all deeper-rooting clades bearing the ancestral state are considered as ‘alternative’ rather than ‘reference’ variants. The variants rs370750300 and rs375658920 are DYS481-associated SNPs and thus included in the Figure; however, as suggested in Section 3.3.8.2, these can be alternatively regarded as RPVs.

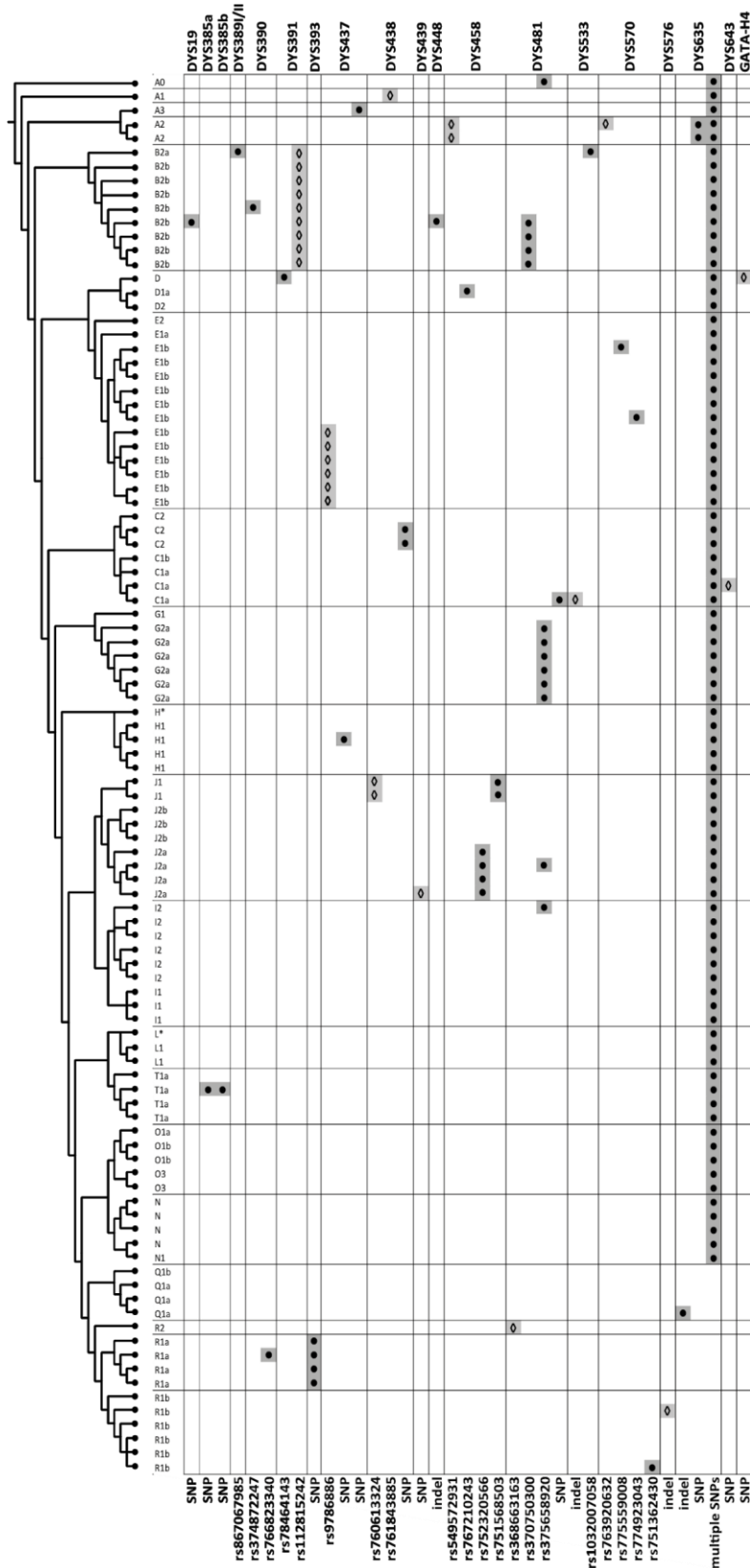


Figure 3.5
Observed SNPs and
indels in their
phylogenetic context.

The phylogenetic tree to the left represents the relationships among 100 diverse Y chromosomes, based on 13,261 high-confidence Y-SNPs previously described (Hallast et al. 2015). Y-chromosome haplogroups are given in their shorthand formats (Table 3.1) to the right of the tree. Y-STR names are listed above. Variants are shaded in grey and represented by filled circles if internal to the repeat array, or unfilled diamonds if in the flanking region. Variants are described by rs# where available, or otherwise as 'SNP' or 'indel' (Table 3.10).

Table 3.10**List of Y-STR markers with observed SNP or indel variants.**

Multiple variants alternatively considered RPs are highlighted in grey. Grey font in the haplogroup association column marks singletons.

Locus	GRCh38 (hg38)	Type	Reference	Alternative	Ancestral	rs#	Count (in 100)	Hg association	Variant previously reported by
DYS19	9,684,428	internal	C	T	C	SNP	1	1/8 B2b	
DYS385a	18,639,748	internal	C	A	C	SNP	2	1/4 T1a	Churchill et al. 2016,
DYS385b	18,680,640	internal	G	T	G	SNP			Just et al. 2017,
DYS389II	12,500,584	internal	T	C	T	rs867067985	1	1/1 B2a	
DYS390	15,163,083	internal	C	T	C	rs374872247	1	1/8 B2b	
	15,163,167	internal	T	G	T	rs766823340	1	1/4 R1a	
DYS391	11,982,092	internal	A	G	A	rs78464143		1/1 D	
	11,982,182	flanking	C	A	C	rs112815242	9	9/9 B2	
DYS393	3,263,111	internal	A	C	A	SNP	4	4/4 R1a	Warshauer et al. 2015
	12,346,264	flanking	C	T	C	rs9786886	6	6/6 E1b1a	Novroski et al. 2016
DYS437	12,346,270	internal	A	G	A	SNP	1	1/4 H1	
	12,346,294	internal	A	G	A	SNP	1	1/1 A3	
	12,825,955	flanking	A	C	A	rs760613324	2	2/2 J1	Novroski et al. 2016
DYS438	12,825,969	flanking	T	C	T	rs761843885	1	1/1 A1	Zhao et al. 2015, Kwon et al. 2016, Novroski et al. 2016
	12,825,898	internal	C	A	C	SNP	2	2/2 C2c	
DYS439	12,403,567	flanking	A	T	A	SNP	1	1/4 J2a	
DYS448	22,218,944	internal	GA	-	GA	SNP	1	1/8 B2b	
	7,999,934	flanking	T	C	T	rs549572931	2	2/2 A2	
DYS458	7,999,842	internal	G	G	A	rs767210243	1	1/1 D1	
	7,999,900	internal	G	G	A	rs752320566	4	4/4 J2a	
	7,999,891	internal	GAAA	GAAAAA	GAAA	rs751568503	2	2/2 J1	
								1/1 R2	
	8,558,321	flanking	G	A	G	rs368663163	1	(.1 alleles)	Lee et al. 2016
DYS481	8,558,336	internal	G	T	G	rs370750300	4	4/8 B2b	Novroski et al. 2016 Warshauer et al. 2015, Kwon et al. 2016, Just et al. 2017, Novroski et al. 2016
	8,558,339	internal	T	G	T	rs375658920	9	6/6 G2a, + sporadic A0, J2a, I2	
	8,558,347	internal	T	C	T	SNP	1	1/3 C1a	
DYS533	16,281,301	flanking	-	CTCTTCTAACTAT	-	SNP	1	1/3 C1a	
	16,281,382	internal	A	G	A	rs1032007058	1	1/1 B2a	Novroski et al. 2016
	6,993,261	flanking	T	G	T	rs763920632	1	1/2 A2	
DYS570	6,993,192	internal	T	C	T	rs775559008	1	1/6 E1b1b	
	6,993,211	internal	T	C	T	rs774923043	1	1/6 E1b1b	Warshauer et al. 2015
	6,993,250	internal	T	C	T	rs751362430	1	1/6 R1b	Novroski et al. 2016
DYS576	7,185,388	flanking	AAA	-	AAA	SNP	1	1/6 R1b	
	12,258,869	internal	TAGA	TGA	TAGA	SNP	1	1/3 Q1a	Butler et al. 2006
	12,258,888	internal	T	C	T	SNP	2	2/2 A2	
DYS635	12,258,880		C	G	G	rs200443815			
	12,258,884	internal	C	G	G	rs201083483	85	85/85 non P (Q, R)	Zhao et al. 2015, Kwon et al. 2016
DYS643	15,314,125	flanking	A	G	A	SNP	1	1/3 C1a	
Y-GATA-H4	16,631,756	flanking	A	G	A	SNP	1	1/1 D	

The other class of variants is defined by repeat pattern variation (RPV), in which arrays with more than one block of repeats present different combinations of units adding up to the same overall length, and therefore indistinguishable by CE (isometric alleles). This study finds 145 distinct alleles showing RPV affecting nine Y-STRs; such alleles are observed in all analysed samples.

All observed variants with SNPs, indels and RPV variations in the sequences of Y-STRs either internal to the arrays or in the flanking regions are summarised in Table 3.11.

Table 3.11
Statistics relating to observed SNP or indel variants in Y-STR sequences.

	# of SNPs or indels	# of alleles with variant	# of STRs	# of samples
SNPs & indels	34	46	19	43
<i>Flanking SNPs & indels</i>	12	19	11	26
SNPs	10	17	9	24
indels	2	2	2	2
<i>Internal SNPs & indels</i>	22	27	15	27
SNPs	19	23	14	24
indels	3	4	3	4
RPVs	-	145	9	100
RPV	-	132	9	100
alternatively indel	1	1	1	1
alternatively SNP	4	12	2	29

3.3.6 Isometric alleles

While Y-STRs studied here, with the exception of DYS385a,b, are expected to present only one allele, in the sample set analysed several examples were observed showing more than one allele (which could be either duplications or somatic mutations; Tables 3.5 a and b), one of which was only detected by MPS. In a haplogroup C1a sample, two isometric alleles of DYS643 were detected (Tables 3.5 a and b), and distinguished by a flanking A to G SNP seven bases upstream of the 11 CTTTT repeats in one allele, but not in the other. The allele with the flanking SNP showed 63% read depth of the allele without the SNP: this was

clearly higher than either of the neighbouring mean stutters $n-1$ 3% or $n+1$ 0.1%, based on 22 other samples bearing allele 11 at DYS643. The size-comparable reference locus was DYS19, and the mean ratio of this pair of loci was 0.96, with minimum and maximum values of 0.44 and 1.92. According to the relative read-depth ratio test, the test ratios for the two alleles were 0.65 and 0.41 each (if these were likely duplications, each would have a value comparable to the mean), and when considered together their test ratio was 1.06 (which is closer to the mean values in this case). Therefore, inferring from the dosage this pair of alleles is more likely to result from somatic mutation than from true duplication. In this case it is even possible to tell that the likely newly formed allele is the one with the flanking SNP, while the other major allele is the ancestral form.

3.3.7 Compiled variants of the dataset

To make the described variants easily searchable for comparisons, Appendix C lists all the variants for each allele and sample, together with sample and haplogroup information in a bracketed format and using additional heat-scale colouring for relative visualisation of the repeat structures.

To allow comparison between different nomenclature systems it is best practice to provide the sequence string, as well as the applied allele designations (the 'bracketed format'). All variants for each allele are listed with the complete reported sequence strings in Appendix C.

3.3.8 Novel variants with implications for nomenclature

This study focused on capturing a wide range of sequence variants through MPS analysis of Y-STRs, rather than taking a population-based approach (Kwon et al. 2016; Novroski et al. 2016; Zhao et al. 2015). The consequent observation of rare variants suggests a broader framework of sequence-level variation that is not always obvious in population studies. Considering rare variants within this framework allows improvements in the MPS-based reporting of alleles to be

suggested for three Y-STRs - DYS385a,b and DYS481 (both previously considered simple repeats), and DYS390.

3.3.8.1 Nomenclature of DYS385a,b

For DYS385a,b, nomenclature is complicated by the fact that the two copies of the STR lie on opposite strands, and the ISFG's last published consideration (Parson et al. 2016) is to report sequences based only on the forward strand direction, leading to different repeat designations for the 'a' and 'b' copies. However, current commercial kits do not distinguish between the two forms, so in order to minimise confusion, here it was decided to follow a description based on the 'b' copy (forward strand), because the GRCh38 human genome reference sequence for DYS385b is GAAA[14] (and to include the flanking repeats: AAGG[6]GAAA[14]), which is consistent with the classical, pre-MPS era repeat designation of GAAA[n] for DYS385. However, while the majority of samples analysed here indeed carry alleles containing six AAGG flanking repeats, examples are also observed showing variation in this block (Table 3.12). This, together with variants observed by others (Novroski et al. 2016), suggests a structure better described as AAGG[5-9]GAAA[n], where the flanking variants are not counted (for continuity) but are nonetheless specified to avoid ambiguity regarding this part of the flanking region.

Table 3.12
Summary of MPS sequence variants with variable length flanking region alleles.

Y-STR	Allele	# of observations in this study	General structure of alleles including variable flanking sequences	CE allele name designation derived from repeat units	Examples in study
DYS385a,b	canonical	193	AAGG[6]GAAA[n]	n	CEU-NA12716
	variant	4	AAGG[5]GAAA[n]	n-1	kun-m82
	variant	2	AAGG[7]GAAA[n]	n+1	CE15_AAGG[5]GAAA[16], CE17_AAGG[5]GAAA[18] TSI-NA20805
	variant	2	AAGG[8]GAAA[n]	n+2	bkl-46
	variant		AAGG[9]GAAA[n]	n+3	CE15_AAGG[8]GAAA[13], CE16_AAGG[8]GAAA[14] in Novroski et al. (2016)
DYS481	canonical	87	CTG[1]CTT[n]	n	CEU-NA12716
	variant	9	CTG[2]CTT[n]	n+1	tur-1
	variant	4	CTG[0]CTT[n]	n-1	CE21_CTG[2]CTT[20] bak55
DYS390	canonical	97	TAGA[n]CAGA[o]TAGA[p]CAGA[q]TAGA[2]	(n+o+p+q)	CEU-NA12716
	variant	2	TAGA[n]CAGA[o]TAGA[p]CAGA[q]TAGA[1]	(n+o+p+q)-1	bhu-1150
	variant	1	TAGA[n]CAGA[o]TAGA[p]CAGA[q]TAGA[3]	(n+o+p+q)+1	bav-55

The most frequent allele variants are denoted 'canonical'; repeat units that show additional polymorphism are shown in bold.

3.3.8.2 Nomenclature of DYS481

For DYS481, the GRCh38 reference assembly contains an array of 22 CTT repeats, preceded by the trinucleotide CTG as part of the flanking region. However, in the analysed sample set, sequence-based alleles were observed lacking this CTG, and also alleles containing two CTG copies (Table 3.12). Similar variants have been reported before (Just et al. 2017; Novroski et al. 2016; Warshauer et al. 2015a), but were described in terms of SNP variants. It seems logical (and biologically meaningful) to suggest applying the same principle as above, and to report sequence variants at DYS481 as CTG[0-2]CTT[n], where the flanking variants are not counted for continuity but specified to avoid ambiguity about this part of the flanking region specifically.

3.3.8.3 Nomenclature of DYS390

DYS390 is already considered to be a compound Y-STR (Forster et al. 1998) and in the GRCh38 reference assembly is represented as TAGA[4]CAGA[1]TAGA[11]CAGA[8] followed by a TAGATAGA flanking sequence that is considered non-variable. In the sequenced alleles of this locus, most of the samples carry alleles similar to the reference in the latter respect; however, the flanking sequence was also observed to present as a variable number of TAGA repeats, TAGA[1-3], and applying this principle the reference would be designated as having TAGA[2] (Table 3.12). DYS390 sequence variants would thus be described as TAGA[n]CAGA[o]TAGA[p]CAGA[q]TAGA[1-3], where the flanking variants are not counted for continuity, but are specified to avoid ambiguity about this part of the flanking region specifically.

3.3.8.4 Consideration for nomenclature detailing flanking region variants

When such flanking region variants are omitted from reporting, this assumes a structure matching that of the reference sequence; however, this assumption might not always be correct, depending on the reagent kit used to type these loci.

In summary, therefore, it is suggested that these flanking region sequences that are presented as variable number repeat units are specified by adding them to the bracketed alleles in MPS-based reporting to avoid ambiguity for loci DYS385a,b, DYS481 and DYS390, even when the flanking region matches the reference. These additional repeat units, however, should remain uncounted in length-based definition for continuity with the CE allele names and existing nomenclature.

3.3.9 Phylogenetic association of variants

In this sequenced dataset, Y-STRs can be classified into two groups. Certain simple (DYS391, DYS392, DYS393, DYS438, DYS439, DYS456, DYS458, DYS533, DYS549, DYS570, DYS576 and DYS643) and compound (DYS19 and Y-GATA-H4) STRs contain only one variable-length array of repeats, which is the source of the overall length variation. In these STRs, sequence variants result from SNPs and indels either within the array or in the flanking regions (Table 3.10). By contrast, DYS385a,b, DYS389I, DYS389II, DYS390, DYS437, DYS448, DYS481 and DYS635 all contain combinations of more than one variable-length array of repeats, which combine to generate the overall length variation (Appendix C). Sequence variants can therefore result not only from SNPs and indels, but also from RPVs in which isometric alleles differ in the numbers of each repeat component.

Different variant types have different underlying mutation processes and rates. While SNPs and short indels have low mutation rates (for SNPs, $\sim 10^{-8}$ per base per generation (Helgason et al. 2015), and slower for indels (Besenbacher et al. 2016), the replication-slippage-based mechanisms that affect STR repeat arrays have much higher rates: these are length-dependent, but are typically five orders of magnitude greater than those of SNPs (Ballantyne et al. 2010; Gusmao et al. 2005). Therefore, variant alleles involving SNPs and indels are expected to show clearer phylogenetic coherence than those involving RPVs.

3.3.9.1 Phylogenetic association of SNPs/indels

Previous studies have described a number of Y-STR sequence variants that are associated with particular haplogroups, and some of these associations are also confirmed in this study (Figure 3.5, Table 3.10). One example is the shortening of a CAGA repeat block within DYS390 (Forster et al. 1998) (corresponding to block 'q' in the notation given in 3.3.8.3 above, and also known as the 'DYS390.1 deletion'), previously reported to be associated with a sub-haplogroup of C (Kayser et al. 2001). A second example is an indel within the DYS458 repeat array, generating intermediate (.2) alleles, and associated with haplogroup J1 (Myres et al. 2007).

The additional SNPs and indels observed here also include several novel haplogroup associations, and a low degree of recurrent mutation, as expected (Figure 3.5, Table 3.10). Examples include a DYS391 flanking SNP (rs112815242) seen in all nine haplogroup B2 samples in this study, and the presence of a DYS393 internal SNP (A to C at the first base of the AGAT[n] repeat array) in all four haplogroup R1a samples (Figure 3.5, Table 3.10): this was also seen in a R1a individual analysed in a previous study (Warshauer et al. 2015a).

3.3.9.2 Phylogenetic association of RPVs

Despite the relatively high mutation rates of Y-STRs, allele lengths are well-known to be non-randomly associated with the phylogeny, and this was observed in this study (Figure 3.3). Similarly, some associations between RPVs and particular haplogroups are detectable here.

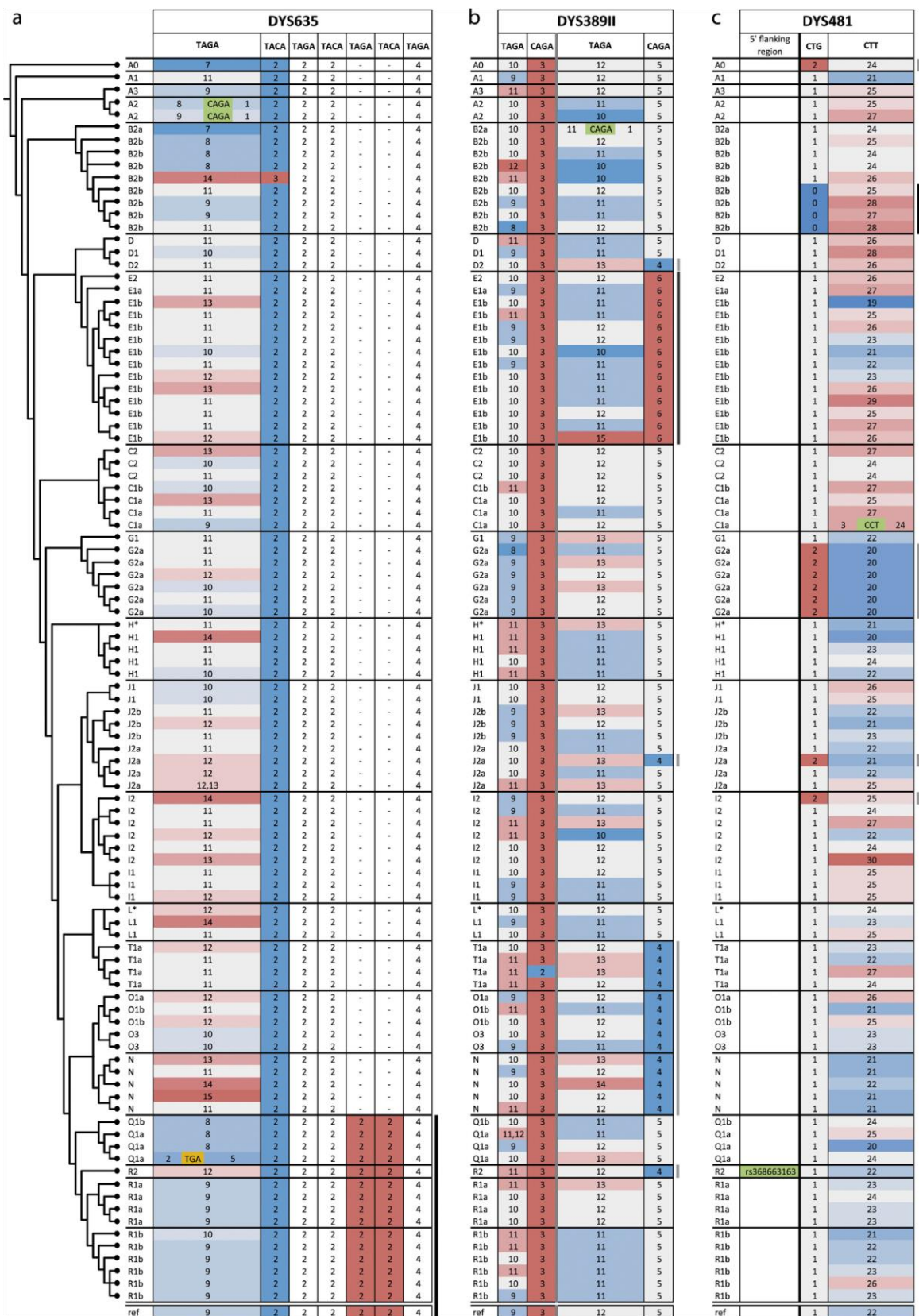


Figure 3.6
Examples of observed RPVs in their phylogenetic contexts.

A phylogenetic tree is shown to the left, as in Figure 3.5. Allele structures for DYS635 in all 100 samples are shown in (a). Repeat unit sequences are shown above, and boxes below contain the number of repeat units in each block, coloured by heat-map from blue (shortest) to red (longest). Invariant blocks are not coloured. SNPs and indels are highlighted by green and orange boxes, respectively. Bars on the right mark features coloured black for monophyletic, or grey for polyphyletic examples. The reference sequence allele structure ('ref.') in GRCh38 chrY is represented below the analysed alleles. Allele structures for DYS389II shown in (b) and for DYS481 in (c).

One clear example is seen in the exclusive association of an RPV in the compound STR DYS635 with the fifteen superhaplogroup P (encompassing Hgs Q, R) samples (Figure 3.6a): this variant, which features two additional repeat blocks compared to more ancestral haplogroups, is unlikely to arise independently multiple times. A haplogroup Q1a sample with a DYS635 21.3 allele carrying an internal indel on the background of this RPV (Figure 3.6a) allows the observation of these two types of variants relative to each other, and indicates that the RPV occurred prior to this indel. Previous sequencing of intermediate '.3' alleles (Butler et al. 2006) has not revealed any other underlying structure for these variants apart from that described here, therefore Y chromosomes with such CE alleles are most likely to belong to the same phylogenetic lineage as the Q1a case.

DYS389II provides a second example, where one short repeat block has a narrow range of variation (CAGA[4-6]), and hence a probable low mutation rate; in this sample set of 100 chromosomes, the presence of 6-repeat blocks appear monophyletic, being seen only in the fourteen haplogroup E samples, while 4-repeat blocks are polyphyletic, observed in all fourteen haplogroup T, O and N samples, but also appearing sporadically elsewhere in the phylogeny (Figure 3.6b).

A third example, DYS481, shows a monophyletic RPV (absence of the initial CTG repeat, which is considered part of the flanking region as detailed above in 3.3.8.2) in a sub-clade of haplogroup B2b (Figure 3.6c); by contrast, the presence of two

copies of this CTG repeat is polyphyletic, though its combination with CTT[20] is confined to haplogroup G2a in these samples.

3.4 Discussion

In this chapter, DNA sequence variation was described in the 23 Y-STRs of the prototype PowerSeq™ Auto/Mito/Y System within a set of 100 diverse Y chromosomes whose phylogenetic relationships have been previously determined via megabase-scale resequencing (Hallast et al. 2015). Of the 2311 STR alleles observed in this dataset, 267 are distinguishable by MPS analysis, compared to just 169 based on length-discrimination via CE (Table 3.7). Use of a phylogenetic framework enhances the observed STR sequence diversity compared to a typical population study (Appendix C), and indicates how variants arise via different mutation processes with different rates. It also provides a wider perspective to recognise additional variable sequences adjacent to classical arrays. The inclusion of these features in the reporting of sequence-based alleles should facilitate more harmonious nomenclature across different workflows and platforms.

One limitation of this study is its small overall sample size. This means that, while some haplogroups are represented multiple times and therefore provide evidence for coherent associations with particular Y-STR sequence variants, others are singletons, and therefore the status of observed variants is unclear (Figure 3.5, Table 3.10). In principle, these could also be true singletons, or they could be shared among a set of unobserved phylogenetically related Y chromosomes. Studies of larger sets of well-characterised Y chromosomes should address this. As an example, a recent study (Phillips et al. 2018) of the Human Genome Diversity Project (HGDP) panel samples with the Verogen ForenSeq™ DNA Signature Prep Kit, which analyses autosomal, X- and Y-STRs via MPS has so far only reported STR repeat region data, but the 929 highly diverse Y-chromosomes of the panel have been fully sequenced (Bergström et al. 2019), and should produce further insights into phylogenetic associations.

As in other recent MPS-based studies of forensically-relevant STRs (Gettings et al. 2016; Novroski et al. 2016), a positive relationship was observed between STR complexity and the number of sequence variants captured. Most of the newly-

described variants in this study originate from complex underlying structures (RPVs), while variants arising from SNPs and indels are independent of structure, and affect almost all the Y-STRs studied, regardless of complexity. These two main types of variants (RPVs, and SNPs/indels) were expected to present different patterns within the phylogeny due to their different likely mutation rates. This expectation was indeed realised (Figures 3.5 and Figure 3.6), with RPVs rarely corresponding to a single event, but several monophyletic occurrences being observed for SNPs or indels.

STR sequencing demonstrates the importance of flanking region variation: omitting the reporting of indels from these areas may result in CE/MPS discordance and could jeopardise the back-compatibility of allele calls. While differences in primer design may result in discordances due to inclusion/exclusion of indels (Table 3.6), another less obvious issue came to light in the dataset, namely a fragment mobility shift arising from flanking SNP variation (Table 3.6). This phenomenon has been described for other STRs (Fujii et al. 2016; Wang et al. 2012), but only recently for DYS481, (Aliferi et al. 2018; Lee et al. 2016). Here, the same flanking SNP was observed as described previously (Lee et al. 2016), resulting in the same discordance between sequence length and CE results (Table 3.6). This SNP (rs368663163, also known as L266 and PF6108) is phylogenetically associated with haplogroup R2 in the ISOGG tree (Y-DNA Haplogroup Tree 2017, Version: 12.320), and occurs in the single haplogroup R2 sample in this study. The mobility shift was noticed inconsistently in previous studies, due to different DYS481 primer designs: in some designs (and in the Yfiler® Plus and Yfiler® kits), a primer bridges the SNP, thus masking its CE mobility shift effect (Ballantyne et al. 2010; Cloete et al. 2010; D'Amato et al. 2010; Kayser et al. 2004; Leat et al. 2007; Shi et al. 2010; Vermeulen et al. 2009) while in others (and in the PowerPlex® Y23 and PowerSeq™ Y kits) the primers encompass the SNP, leading to a DYS481 '.1' allele 9 (Aliferi et al. 2018; Lee et al. 2016; Oh et al. 2015). One study (Lee et al. 2014) found 20 among 270 Pakistani males to carry DYS481 '.1' alleles, and used SNP typing to assign them all to haplogroup R2-M479. This haplogroup association can be further supported by surveying a large global PPY23 dataset (Purps et al. 2014), in which

all 26 samples carrying DYS481 '.1' alleles are predicted to belong to haplogroup R2 using the NevGen predictor, a tool whose accuracy has been recently assessed (Khubrani et al. 2018). These observations support the singleton finding, and suggest that rs368663163 is a strong indicator of haplogroup R2, and of the geographical regions of South and Central Asia (Balaesque et al. 2015; Sengupta et al. 2006) in which this lineage is prevalent.

Currently the most notable general effect of applying MPS to forensic STRs is the resulting increase in allele diversity, largely originating from RPs, and the resolution by sequence variants of a proportion of length-homozygous autosomal alleles as isometric heterozygotes. This study has shown that MPS-based analysis of STRs on the Y chromosome also increases observed allele diversity, and hence haplotype diversity, and that it has the potential to distinguish between isometric alleles of bilocal Y-STRs. Much effort has been devoted to elevating the discriminatory power of Y-STR typing by increasing the number of STRs analysed (Kayser et al. 2004), and by focusing on subsets that have particularly high mutation rates (rapidly mutating STRs; RM Y-STRs (Ballantyne et al. 2012; Ballantyne et al. 2014)). Applying MPS to additional STRs, including RM Y-STRs, is expected to improve discriminatory power as allele diversity increases. However, as these phylogenetically-based data show, within a patrilineage, additional variation from SNPs and indels is unlikely to be observed because of the associated low mutation rates of these events. Any additional variation at this scale will come from RPs which, while mutating more rapidly than SNPs and indels, appear to have mutation rates that are lower than the rate of overall STR length variation. If this is so, individual male identification via MSY analysis may not be greatly advanced by applying MPS approaches.

However, the association between SNPs and STRs is likely to be beneficial for the analysis of multi-male mixtures via MPS. If SNPs/indels prove to be phylogenetically restricted, as observed here, they will be associated with the characteristic Y-STR allele lengths, which have previously been exploited for haplogroup prediction (Athey 2005, 2006; Schlecht et al. 2008; Seman et al. 2012).

Knowledge of the apparent mixture ratio of the contributing haplogroups from SNP/indel variants may help with the deconvolution of mixtures when the two haplogroups have very distinct allele size ranges at particular loci. Further analysis of isometric alleles could provide insights into relative stutter ratios between pure and interrupted repeat array structures.

3.4.1 Highlights and Conclusions

Application of MPS resulted in an increase in allele diversity across 19 of the 23 typed Y-STRs, which mostly resulted from intra-array structural variation and less frequently from variants in the flanking region.

A phylogenetically-driven approach resulted in a wider range of allelic variants at the sequence level than is likely to be found in a population-based approach of comparable sample size, and identified 60 alleles (of which at the time of writing 58 are still not catalogued) not described elsewhere.

Variants resulting from SNPs, indels and RPVs showed different occurrence patterns within the phylogeny: SNPs and indels, with their lower mutation rates, were mostly monophyletic, while RPVs showed examples of polyphyletic distribution across the Y-phylogeny sampled here.

CHAPTER 4: Autosomal STR sequence variation

4.1 Introduction

While Y-STRs are used to define the male lineage, which is shared between direct male relatives, autosomal short-tandem repeats (aSTRs) are used to identify individuals through a profile containing a unique combination of alleles. Although the sample set of 100 males analysed in Chapter 3 was selected from the Y-phylogeny to maximise variability across Y-chromosomes, a compilation of such ethnically diverse samples is also a good source to investigate the sequence variability of autosomal STRs.

In this study the samples do not represent populations, and instead the focus is to describe the extent of variation in this global set and to identify any alleles not yet catalogued by the STRSeq database (Gettings et al. 2017).

4.2 Materials and methods

4.2.1 DNA samples and laboratory experiments

One hundred male DNA samples were selected as described in Chapter 2, Table 2.1. DNA quantitation, amplification library preparation and sequencing were described in Chapter 3, sections 3.2.2 and 3.2.3.

4.2.2 Targeted autosomal STRs

Twenty-two autosomal STRs (D1S1656, D2S1338, D2S441, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, CSF1P0, FGA, PentaD, PentaE, TH01, TPOX, vWA) and the Amelogenin XY-homologous sex-test loci were targeted using the prototype PowerSeq™ Auto/Mito/Y System (Promega) for MPS typing, and the results for these loci are described here.

4.2.3.Data processing and analyses

Autosomal and Y-STR data were analysed together as described in Chapter 3, Section 3.2.4.

The analysis primarily used the software FSTools v1.1.1 (Hoogenboom et al. 2017), and allele calls were confirmed using STRait Razor 2.0 (Warshauer et al. 2015b), with command-line tools and standard variant calling used to clarify allele designations.

Discovered variants were compared to the human genome reference sequence (GRCh38) and queried in dbSNP (build 151).

Allele sequence-variants were compared to the NCBI-hosted STRSeq BioProject (Gettings et al. 2017), an online resource to collect STR sequence variation under an international collaboration; data for the autosomal STR loci can be found in a sub-project with the accession: PRJNA380345 (<https://www.ncbi.nlm.nih.gov/bioproject/380345>). The sequence variants that were identified in the STRSeq database were marked by unique accession numbers, and alleles not yet described in this database were flagged as 'novel to STRSeq'.

4.3 Results

4.3.1 Coverage ranges observed

Promega's prototype PowerSeq™ Auto/Mito/Y kit was used to generate MPS data for 22 aSTRs and Amelogenin from 100 samples.

With the analytical threshold set to 20 × coverage, a minimum-to-maximum per-locus sequence coverage of 476 - 17,772 × was observed for 24-plex library preparation, and 384 - 6,358 × for 96-plex library preparation. Run-specific statistics were described in Chapter 3, Table 3.2. Summary of coverage statistics per library preparation complexity and per locus are provided in Table 4.1.

Table 4.1
Sample coverage statistics.

Values are provided for both library pooling complexity approaches, low throughput (LT) and high throughput (HT) given as overall values (a) and per locus per multiplexity level (b and c).

a.											
						LT		HT			
						24-plex		96-plex			
Mean coverage						100,003		35,714			
Standard deviation						50,198		8,482			
Median coverage						90,315		35,150			
Minimum coverage						32,518		20,382			
Maximum coverage						183,384		58,406			

b.						c.					
LT 24-plex						HT 96-plex					
Y-STR	Mean coverage	Standard deviation	Median coverage	Min coverage	Max coverage	Y-STR	Mean coverage	Standard deviation	Median coverage	Min coverage	Max coverage
Amel	9299.9	5636.2	8350.0	2561	17772	Amel	2256.5	690.8	2125.5	1339	3991
CSF1PO	3840.1	2533.4	3347.0	857	9897	CSF1PO	1204.8	333.2	1178.5	709	2051
D10S1248	4069.4	2219.1	3663.0	1124	7888	D10S1248	1108.8	341.2	1073.5	527	2011
D12S391	6193.0	3743.6	5177.0	1330	15725	D12S391	2045.3	708.8	1910.0	1377	3974
D13S317	3135.6	1628.2	2936.0	989	6987	D13S317	1527.9	417.7	1444.5	1044	2764
D16S539	4396.8	1909.7	3989.5	1855	8034	D16S539	1532.9	501.6	1513.0	831	2971
D18S51	2513.6	1697.0	2007.5	476	6169	D18S51	856.0	255.0	820.5	564	1480
D19S433	2529.4	1463.7	1813.5	811	4889	D19S433	949.1	272.1	903.5	437	1772
D151656	6135.8	2496.3	6526.0	3073	11173	D151656	2676.0	825.3	2551.5	1468	4697
D21S11	2661.6	1139.3	2839.0	1006	4639	D21S11	1613.4	415.1	1611.0	1084	2568
D22S1045	4377.2	2765.9	3799.0	935	10193	D22S1045	1196.4	422.6	1150.5	652	2314
D251338	3023.7	1673.9	2323.0	1174	6615	D251338	888.2	306.1	824.5	476	1813
D2S441	8231.8	3383.8	8152.5	3589	14963	D2S441	2401.4	1172.2	2030.5	1055	6358
D3S1358	2336.7	923.6	2235.5	1082	4130	D3S1358	1193.0	352.2	1154.5	649	2130
D5S818	5121.4	2856.9	4550.5	1300	9579	D5S818	2161.3	520.1	2081.0	1310	3343
D7S820	3403.2	1679.8	3188.5	1021	7295	D7S820	1673.4	469.2	1592.0	384	2970
D8S1179	5662.8	3201.6	4678.0	1627	11301	D8S1179	2047.3	513.4	2000.5	1489	3339
FGA	2714.9	1535.1	2376.5	851	6962	FGA	1194.7	306.9	1131.0	786	1898
PentaD	2941.2	1710.1	2510.5	832	7870	PentaD	1350.0	337.3	1340.5	823	2494
PentaE	3032.1	1728.7	2664.5	761	6388	PentaE	1551.3	515.7	1459.0	842	3152
TH01	3818.8	2346.9	3389.0	1027	7809	TH01	1043.4	332.9	1007.0	430	1962
TPOX	4069.3	1943.2	3112.0	1319	7382	TPOX	1545.6	512.1	1477.0	747	3344
vWA	6495.4	4001.6	5029.5	1882	12473	vWA	1697.2	640.7	1510.0	796	3651


















4.3.2 Identified Amelogenin and aSTR alleles

A total of 4600 alleles were analysed in 100 samples: this included 22 aSTRs with two alleles per sample, plus two distinguishable (X and Y-linked) sequences each for Amelogenin. By length, 241 distinct alleles were distinguished across all loci, including Amelogenin, and this was increased to 443 sequence-based alleles by using MPS.

The observed sequence-level variation is summarised in Table 4.2 by allele range for each locus, detailing the structure of each aSTR array and any flanking variations.

All aSTR sequence variants are provided in a bracketed format, but also as sequence strings (Appendix D) with exact genomic coordinates of the reported regions, to provide compatibility of the observed variants when typed with different platforms or methods (Parson et al. 2016).

Table 4.2. Amelogenin and autosomal STR sequence alleles.

Locus	CE	Sequence allele	STRSeq acc#	Allele count
Amelo	X	X	-	100 
	Y	Y	-	100 
				200
Locus	CE	Sequence allele	STRSeq acc#	Allele count
CSF1PO	8	CE8_ATCT[8]	MH085178.1	1 
	9	CE9_ATCT[9]	MH085179.1	5 
	10	CE10_ATCT[10]	MH085181.1	57 
	11	CE11_ATCT[11]	MH085186.1	54 
	12	CE12_ATCT[12]	MH085189.1	73 
		CE12_ATCT[12]_+86C>A_rs140751340	MH085188.1	1 
	13	CE13_ATCT[13]	MH085190.1	9 
range	8-13	ATCT[8-13]		200
Locus	CE	Sequence allele	STRSeq acc#	Allele count
D10S1248	10	CE10_GGAA[10]	MH167057.1	2 
	11	CE11_GGAA[11]	MH167058.1	2 
	12	CE12_GGAA[12]	MH167059.1	16 
	13	CE13_GGAA[13]	MH167061.1	49 
	14	CE14_GGAA[14]	MH167062.1	61 
	15	CE15_GGAA[15]	MH167063.1	36 
	16	CE16_GGAA[16]	MH167064.1	27 
	17	CE17_GGAA[17]	MH167065.1	7 
range	10-17	GGAA[10-17]		200

Cont.

Locus	CE	Sequence allele	STRSeq acc#	Allele count
D12S391	15	CE15_AGAT[8]AGAC[6]AGAT[1]	MH167108.1	7
	16	CE16_AGAT[9]AGAC[6]AGAT[1]	MH167111.1	4
	17	CE17_AGAT[9]AGAC[7]AGAT[1]	MH167113.1	2
		CE17_AGAT[10]AGAC[6]AGAT[1]	MH167114.1	13
	17.3	CE17_AGAT[11]AGAC[5]AGAT[1]	MH167115.1	3
		CE17.3_AGAT[1]GAT[1]AGAT[8]AGAC[7]AGAT[1]	MH167119.1	2
	18	CE18_AGAT[9]AGAC[8]AGAT[1]	MH167120.1	1
		CE18_AGAT[10]AGAC[7]AGAT[1]	MH167122.1	7
		CE18_AGAT[11]AGAC[6]AGAT[1]	MH167125.1	27
		CE18_AGAT[12]AGAC[5]AGAT[1]	MH167126.1	5
	18.3	CE18.3_AGAT[1]GAT[1]AGAT[9]AGAC[7]AGAT[1]	MH167128.1	2
	19	CE19_AGAT[11]AGAC[7]AGAT[1]	MH167131.1	5
		CE19_AGAT[12]AGAC[6]AGAT[1]	MH167134.1	29
		CE19_AGAT[12]AGAC[7]	MH167133.1	1
		CE19_AGAT[14]AGAC[4]AGAT[1]	novel to STRSeq	1
	20	CE20_AGAT[10]AGAC[9]AGAT[1]	MH167143.1	2
		CE20_AGAT[11]AGAC[8]AGAT[1]	MH167145.1	1
		CE20_AGAT[11]AGAC[9]	MH167144.1	2
		CE20_AGAT[12]AGAC[7]AGAT[1]	MH167147.1	9
		CE20_AGAT[12]AGAC[8]	MH167146.1	4
		CE20_AGAT[13]AGAC[6]AGAT[1]	MH167150.1	11
		CE20_AGAT[13]AGAC[7]	MH167149.1	1
		CE20_AGAT[14]AGAC[5]AGAT[1]	MH167152.1	1
	21	CE21_AGAT[11]AGAC[10]	MH167157.1	1
		CE21_AGAT[12]AGAC[8]AGAT[1]	MH167160.1	2
		CE21_AGAT[12]AGAC[9]	MH167159.1	5
		CE21_AGAT[13]AGAC[7]AGAT[1]	MH167162.1	3
		CE21_AGAT[13]AGAC[8]	MH167161.1	1
		CE21_AGAT[14]AGAC[6]AGAT[1]	MH167165.1	3
	22	CE22_AGAT[12]AGAC[10]	MH167169.1	2
		CE22_AGAT[12]AGAC[9]AGAT[1]	MH167170.1	1
		CE22_AGAT[13]AGAC[8]AGAT[1]	MH167172.1	3
		CE22_AGAT[13]AGAC[9]	MH167171.1	13
	23	CE22_AGAT[14]AGAC[8]	MH167173.1	1
		CE23_AGAT[12]AGAC[11]	MH167177.1	2
		CE23_AGAT[13]AGAC[9]AGAT[1]	MH167179.1	1
		CE23_AGAT[14]AGAC[8]AGAT[1]	MH167181.1	6
		CE23_AGAT[14]AGAC[9]	MH167180.1	5
	24	CE23_AGAT[15]AGAC[8]	MH167182.1	1
		CE24_AGAT[14]AGAC[10]	MH167184.1	2
		CE24_AGAT[14]AGAC[9]AGAT[1]	MH167185.1	2
		CE24_AGAT[15]AGAC[8]AGAT[1]	MH167187.1	1
	25	CE24_AGAT[15]AGAC[9]	MH167186.1	1
		CE25_AGAT[13]AGAC[11]AGAT[1]	novel to STRSeq	1
		CE25_AGAT[16]AGAC[8]AGAT[1]	MH167193.1	2
		CE25_AGAT[16]AGAC[9]	MH167192.1	1
range	15-25	AGAT[8-16]AGAC[4-11]AGAT[0-1]		200

Cont.

Locus	CE	Sequence allele	STRSeq acc#	Allele count
D13S317	8	CE8_TATC[8]AATC[2]ATCT[3]	MH167202.1	30
		CE8_TATC[9]AATC[1]ATCT[3]	novel to STRSeq	1
	9	CE9_TATC[9]AATC[2]ATCT[3]	MH167206.1	13
		CE9_TATC[10]AATC[1]ATCT[3]	MH167207.1	2
	10	CE10_TATC[9]AATC[3]ATCT[3]	novel to STRSeq	1
		CE10_TATC[10]AATC[2]ATCT[3]	MH167210.1	12
		CE10_TATC[11]AATC[1]ATCT[3]	MH167211.1	6
	11	CE11_TATC[6]TATT[1]TATC[5]AATC[1]ATCT[3]	novel to STRSeq	1
		CE11_TATC[11]AATC[2]ATCT[3]	MH167218.1	19
		CE11_TATC[12]AATC[1]ATCT[3]	MH167219.1	26
		CE11_TATC[12]AATC[1]ATCT[3]_-25C>T_rs73250432	MH167223.1	1
		CE11_TATC[12]AATC[1]ATCT[3]_-53G>A_rs73525369	MH167216.1	1
		CE11_TATC[13]ATCT[3]	MH167221.1	2
	12	CE12_TATC[12]AATC[2]ATCT[3]	MH167226.1	31
		CE12_TATC[13]AATC[1]ATCT[3]	MH167228.1	27
		CE12_TATC[13]AATC[1]ATCT[3]_-24G>A_rs146621667	MH167225.1	1
		CE12_TATC[13]AATC[1]ATCT[3]_-53G>A_rs73525369	MH167224.1	1
	13	CE13_TATC[13]AATC[2]ATCT[3]	MH167233.1	16
		CE13_TATC[14]AATC[1]ATCT[3]	MH167234.1	8
	14	CE14_TATC[15]AATC[1]ATCT[3]	MH167237.1	1
range	8-14	TATC[8-15]AATC[0-3]ATCT[3]		200
D16S539	5	CE5_GATA[5]	NEW/MH167240.1	1
	8	CE8_GATA[8]	MH167241.1	7
	9	CE9_GATA[9]	MH167243.1	32
		CE9_GATA[9]_-95A>C_rs1728369	novel to STRSeq	2
	10	CE10_GATA[10]	MH167249.1	16
	11	CE11_GATA[5]GACA[1]GATA[5]_-95A>C_rs1728369	novel to STRSeq	1
		CE11_GATA[11]	MH167251.1	42
		CE11_GATA[11]_-95A>C_rs1728369	MH167254.1	19
	12	CE12_GATA[12]	MH167259.1	27
		CE12_GATA[12]_-95A>C_rs1728369	MH167260.1	22
	13	CE13_GATA[13]	MH167261.1	13
		CE13_GATA[13]_-95A>C_rs1728369	MH167262.1	12
	14	CE14_GATA[14]	MH167264.1	3
		CE14_GATA[14]_-95A>C_rs1728369	MH167265.1	3
range	5-14	GATA[5-14]		200
D18S51	10	CE10_AGAA[10]	MH167284.1	2
	11	CE11_AGAA[11]	MH167285.1	3
	12	CE12_AGAA[12]	MH167287.1	23
	13	CE13_AGAA[13]	MH167288.1	32
	14	CE14_AGAA[14]	MH167290.1	25
		CE14_AGAA[1]AGCA[1]AGAA[12]	MH167291.1	1
	15	CE15_AGAA[15]	MH167293.1	27
	16	CE16_AGAA[16]	MH167296.1	28
	16.2	CE16.2_AGAA[16]AG[1]_+2A>G_rs535823682	MH167298.1	1
	17	CE17_AGAA[17]	MH167299.1	13
	18	CE18_AGAA[18]	MH167301.1	14
	19	CE19_AGAA[19]	MH167302.1	15
	20	CE20_AGAA[20]	MH167303.1	9
	21	CE21_AGAA[21]	MH167306.1	5
	22	CE22_AGAA[22]	MH167309.1	1
	23	CE23_AGAA[23]	MH167310.1	1
range	10-23	AGAA[10-23]		200

Cont.

Locus	CE	Sequence allele	STRSeq acc#	Allele count
D19S433	10	CE10_CCTT[8]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	MH174811.1	3
	11	CE11_CCTT[9]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	MH174812.1	2
	11.2	CE11.2_CCTT[10]CCTA[1]CCTT[1]TT[1]CCTT[1]	novel to STRSeq	1
	12	CE12_CCTT[10]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	MH174813.1	17
	12.2	CE12.2_CCTT[11]CCTA[1]CCTT[1]TT[1]CCTT[1]	MH174815.1	1
	13	CE13_CCTT[11]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	MH174816.1	53
	13.2	CE13.2_CCTT[12]CCTA[1]CCTT[1]TT[1]CCTT[1]	MH174819.1	6
	14	CE14_CCTT[12]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	MH174820.1	58
		CE14_CCTT[14]CTTT[1]CCTT[1]	MH174821.1	1
	14.2	CE14.2_CCTT[13]CCTA[1]CCTT[1]TT[1]CCTT[1]	MH174824.1	19
		CE14.2_CCTT[13]CCTA[1]CCTT[1]CTTT[1]CCTT[1]_-2CT>_rs74560776	novel to STRSeq	1
	15	CE15_CCTT[13]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	MH174825.1	14
	15.2	CE15.2_CCTT[14]CCTA[1]CCTT[1]TT[1]CCTT[1]	MH174827.1	15
	16	CE16_CCTT[14]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	MH174828.1	2
	16.2	CE16.2_CCTT[15]CCTA[1]CCTT[1]TT[1]CCTT[1]	MH174829.1	5
	17	CE17_CCTT[15]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	MH174830.1	1
	17.2	CE17.2_CCTT[16]CCTA[1]CCTT[1]TT[1]CCTT[1]	MH174831.1	1
range	10-17.2	CCTT[8-16]CCTA[0-1]CCTT[1]CTTT[0-1]TT[0-1]CCTT[1]		200
Locus	CE	Sequence allele	STRSeq acc#	Allele count
D1S1656	8	CE8_TCTA[8]	novel to STRSeq	2
	10	CE10_CCTA[1]TCTA[9]	MH174834.1	1
		CE10_TCTA[10]	MH174835.1	1
	11	CE11_CCTA[1]TCTA[10]	MH174836.1	1
		CE11_TCTA[11]	MH174837.1	13
	12	CE12_CCTA[1]TCTA[11]	MH174838.1	9
		CE12_TCTA[12]	MH174839.1	6
	13	CE13_CCTA[1]TCTA[12]	MH174840.1	7
		CE13_TCTA[13]	MH174842.1	14
	14	CE14_CCTA[1]TCTA[13]	MH174844.1	17
		CE14_TCTA[14]	MH174845.1	3
	15	CE15_CCTA[1]TCTA[14]	MH174848.1	36
		CE15_TCTA[15]	MH174850.1	1
	15.3	CE15.3_CCTA[1]TCTA[10]TCA[1]TCTA[4]_+6C>T_rs4847015	MH174851.1	4
		CE15.3_CCTA[1]TCTA[11]TCA[1]TCTA[3]_+6C>T_rs4847015	MH174852.1	2
	16	CE16_CCTA[1]TCTA[15]	MH174853.1	33
		CE16_TCTA[16]	MH174855.1	1
	16.3	CE16.3_CCTA[1]TCTA[11]TCA[1]TCTA[4]_+6C>T_rs4847015	MH174856.1	12
	17	CE17_CCTA[1]TCTA[16]	MH174858.1	14
	17.3	CE17.3_CCTA[1]TCTA[12]TCA[1]TCTA[4]_+6C>T_rs4847015	MH174861.1	15
	18	CE18_CCTA[1]TCTA[17]	MH174864.1	1
	18.3	CE18.3_CCTA[1]TCTA[13]TCA[1]TCTA[4]_+6C>T_rs4847015	MH174865.1	5
	19.3	CE19.3_CCTA[1]TCTA[14]TCA[1]TCTA[4]_+6C>T_rs4847015	MH174866.1	2
range	8-19.3	CCTA[0-1]TCTA[8-17]		200

Cont.

Locus	CE	Sequence allele	STRSeq acc#	Allele count
D21S11	25	CE25_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[7]	novel to STRSeq	1
	26	CE26_TCTA[6]TCTG[7]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[5]	novel to STRSeq	1
	27	CE27_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[9]	MH174720.1	3
		CE27_TCTA[5]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[9]	MH174717.1	1
	28	CE28_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	MH174728.1	21
		CE28_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[2]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	MH174725.1	1
		CE28_TCTA[6]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[9]	MH174722.1	3
	28.2	CE28.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[8]TA[1]TCTA[1]	MH174730.1	3
	29	CE29_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]	MH174737.1	26
		CE29_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	MH174736.1	7
		CE29_TCTA[6]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	MH174731.1	12
		CE29_TCTA[6]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[9]	MH174739.1	1
	29.2	CE29.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[2]TCA[1]TCTA[2]TCCATA[1]TCTA[10]TA[1]TCTA[1]	novel to STRSeq	1
	30	CE30_TCTA[4]TCTG[6]TCTA[1]TCTG[1]TCTA[1]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]	novel to STRSeq	1
		CE30_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]	MH174748.1	14
		CE30_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]	MH174747.1	10
		CE30_TCTA[6]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]	MH174745.1	20
		CE30_TCTA[7]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	MH174743.1	3
		CE30_TCTA[7]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[9]	MH174744.1	1
	30.2	CE30.2_TCTA[5]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]TA[1]TCTA[1]	MH174751.1	1
		CE30.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[2]TCA[1]TCTA[2]TCCATA[1]TCTA[11]TA[1]TCTA[1]	MH174752.1	2
		CE30.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]TA[1]TCTA[1]	MH174753.1	3
	31	CE31_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[13]	MH174761.1	3
		CE31_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]	MH174760.1	3
		CE31_TCTA[6]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]	MH174759.1	4
		CE31_TCTA[7]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	MH174757.1	1
	31.2	CE31.2_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]TA[1]TCTA[1]	MH174767.1	1
		CE31.2_TCTA[4]TCTG[7]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]TA[1]TCTA[1]	novel to STRSeq	1
		CE31.2_TCTA[5]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]TA[1]TCTA[1]	MH174764.1	2
		CE31.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]TA[1]TCTA[1]	MH174766.1	18
	32	CE32_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[13]	MH174773.1	1
		CE32_TCTA[6]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]	MH174771.1	1
	32.2	CE32.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]TA[1]TCTA[1]	MH174779.1	21
	33.2	CE33.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[13]TA[1]TCTA[1]	MH174789.1	4
	34	CE34_TCTA[10]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	novel to STRSeq	2
		CE34_TCTA[11]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	novel to STRSeq	1
	35.2	CE35.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[15]TA[1]TCTA[1]	novel to STRSeq	1
range	25-35.2	TCTA[4-11]TCTG[5-7]TCTA[1-3]TA[1]TCTA[2-3]TCA[1]TCTA[2]TCCATA[1]TCTA[5-15]		200

Cont.

Locus	CE	Sequence allele	STRSeq acc#	Allele count
D22S1045	10	CE10_ATT[7]ACT[1]ATT[2]	MH167268.1	6
	11	CE11_ATT[8]ACT[1]ATT[2]	MH167269.1	40
	12	CE12_ATT[9]ACT[1]ATT[2]	MH167270.1	2
	13	CE13_ATT[10]ACT[1]ATT[2]	MH167273.1	1
	14	CE14_ATT[11]ACT[1]ATT[2]	MH167274.1	10
	15	CE15_ATT[12]ACT[1]ATT[2]	MH167275.1	56
		CE15_ATT[12]ACT[1]ATT[2]_-15G>T_rs190864081	MH167276.1	1
	16	CE16_ATT[13]ACT[1]ATT[2]	MH167279.1	59
	17	CE17_ATT[14]ACT[1]ATT[2]	MH167280.1	25
range	10-17	ATT[7-14]ACT[1]ATT[2]		200

Locus	CE	Sequence allele	STRSeq acc#	Allele count
D2S1338	16	CE16_GGAA[10]GGCA[6]_-35C>A_rs6736691	MH105114.1	4
		CE16_GGAA[12]GGCA[4]_-35C>A_rs6736691	MH105112.1	1
	17	CE17_GGAA[10]GGCA[7]	MH105121.1	1
		CE17_GGAA[10]GGCA[7]_-35C>A_rs6736691	MH105119.1	1
		CE17_GGAA[11]GGCA[6]_-35C>A_rs6736691	MH105118.1	30
	18	CE18_GGAA[11]GGCA[7]	MH105129.1	14
		CE18_GGAA[12]GGCA[6]_-35C>A_rs6736691	MH105124.1	2
		CE18_GGAA[13]GGCA[5]_-35C>A_rs6736691	MH105123.1	1
		CE18_GGAA[15]GGCA[3]_-35C>A_rs6736691	MH105122.1	1
	19	CE19_GGAA[12]GGCA[7]	MH105135.1	14
		CE19_GGAA[13]GGCA[6]	MH105134.1	5
		CE19_GGAA[13]GGCA[6]_-35C>A_rs6736691	MH105132.1	1
		CE19_GGAA[14]GGCA[5]	MH105133.1	1
		CE19_GGAA[15]GGCA[4]_-35C>A_rs6736691	novel to STRSeq	1
	20	CE19_GGAA[16]GGCA[3]_-35C>A_rs6736691	MH105130.1	1
		CE20_GGAA[10]GAAA[1]GGAA[2]GGCA[7]	MH105142.1	1
		CE20_GGAA[12]GGGA[1]GGCA[7]	MH105148.1	2
		CE20_GGAA[13]GGCA[7]	MH105146.1	17
		CE20_GGAA[14]GGCA[6]	MH105145.1	2
		CE20_GGAA[16]GGCA[4]_-35C>A_rs6736691	MH105140.1	2
		CE20_GGAA[2]GGAC[1]GGAA[10]GGCA[7]	MH105151.1	2
		CE21_GGAA[13]GGCA[8]	MH105156.1	1
	21	CE21_GGAA[14]GGCA[7]	MH105155.1	2
		CE21_GGAA[15]GGCA[6]	MH105154.1	1
		CE21_GGAA[2]GGAC[1]GGAA[11]GGCA[7]	MH105159.1	3
		CE22_GGAA[14]GGCA[8]	MH105162.1	3
	22	CE22_GGAA[15]GGCA[7]	MH105161.1	2
		CE22_GGAA[2]GGAC[1]GGAA[12]GGCA[7]	MH105166.1	9
		CE22_GGAA[2]GGAC[1]GGAA[13]GGCA[6]	MH105165.1	4
		CE23_GGAA[16]GGCA[7]	MH105167.1	1
	23	CE23_GGAA[2]GGAC[1]GGAA[13]GGCA[7]	MH105171.1	22
		CE23_GGAA[2]GGAC[1]GGAA[14]GGCA[6]	MH105170.1	4
		CE24_GGAA[17]GGCA[7]	novel to STRSeq	1
	24	CE24_GGAA[2]GGAC[1]GGAA[14]GGCA[7]	MH105175.1	23
		CE24_GGAA[2]GGAC[1]GGAA[15]GGCA[6]	MH105174.1	3
		CE25_GGAA[2]GGAC[1]GGAA[14]GGCA[8]	MH105179.1	1
	25	CE25_GGAA[2]GGAC[1]GGAA[15]GGCA[7]	MH105178.1	11
		CE25_GGAA[2]GGAC[1]GGAA[16]GGCA[6]	MH105177.1	1
		CE26_GGAA[2]GGAC[1]GGAA[16]GGCA[7]	MH105182.1	4
range	16-26	GGAA[10-17]GGCA[3-8]		200

Cont.

Locus	CE	Sequence allele	STRSeq acc#	Allele count
D2S441	9	CE9_TCTA[9]	MH167314.1	1
	10	CE10_TCTA[10]	MH167317.1	12
		CE10_TCTA[10]_-25G>A_rs74640515	MH167316.1	3
		CE10_TCTA[8]TCTG[1]TCTA[1]	MH167318.1	20
	10.1	CE10.1_A[1]TCTA[10]_-25G>A_rs74640515	novel to STRSeq	1
	11	CE11_TCTA[11]	MH167320.1	57
		CE11_TCTA[11]_-25G>A_rs74640515	MH167319.1	1
		CE11_TCTA[8]TTTA[1]TCTA[2]	novel to STRSeq	1
		CE11_TCTA[9]TCTG[1]TCTA[1]	MH167321.1	4
	11.3	CE11.3_TCTA[4]TCA[1]TCTA[7]	MH167323.1	8
	12	CE12_TCTA[12]	MH167325.1	16
		CE12_TCTA[9]TTTA[1]TCTA[2]	MH167327.1	1
	12.3	CE12.3_TCTA[3]TCA[1]TCTA[9]	novel to STRSeq	3
	13	CE13_TCTA[10]TTTA[1]TCTA[2]	MH167331.1	6
		CE13_TCTA[13]	MH167329.1	1
	13.3	CE13.3_TCTA[3]TCA[1]TCTA[10]	MH167332.1	1
	14	CE14_TCTA[11]TTTA[1]TCTA[2]	MH167334.1	57
	15	CE15_TCTA[12]TTTA[1]TCTA[1]TCTG[1]	novel to STRSeq	1
		CE15_TCTA[12]TTTA[1]TCTA[2]	MH167337.1	3
	16	CE16_TCTA[13]TTTA[1]TCTA[2]	MH167338.1	3
range	9-16	TCTA[9-13]TTTA[0-1]TCTG[0-1]TCA[0-1]TCTA[0-10]		200

Locus	CE	Sequence allele	STRSeq acc#	Allele count
D3S1358	14	CE14_TCTA[1]TCTG[2]TCTA[11]	MH166965.2	16
	15	CE15_TCTA[1]TCTG[1]TCTA[13]	MH166968.2	8
		CE15_TCTA[1]TCTG[2]TCTA[12]	MH166969.2	50
		CE15_TCTA[1]TCTG[3]TCTA[11]	MH166971.1	5
	16	CE16_TCTA[1]TCTG[1]TCTA[14]	MH166974.1	6
		CE16_TCTA[1]TCTG[2]TCTA[13]	MH166975.1	33
		CE16_TCTA[1]TCTG[3]TCTA[12]	MH166976.1	18
		CE17_TCTA[1]TCTG[2]TCTC[1]TCTA[13]	MK990348.1	2
	17	CE17_TCTA[1]TCTG[2]TCTA[14]	MH166981.2	20
		CE17_TCTA[1]TCTG[3]TCTA[13]	MH166982.1	21
		CE18_TCTA[1]TCTG[1]TCTA[16]	MH166984.1	2
	18	CE18_TCTA[1]TCTG[2]TCTA[15]	MH166985.1	1
		CE18_TCTA[1]TCTG[3]TCTA[14]	MH166986.1	15
		CE19_TCTA[1]TCTG[1]TCTA[17]	novel to STRSeq	1
	19	CE19_TCTA[1]TCTG[2]TCTA[16]	MK990352.1	1
		CE19_TCTA[1]TCTG[3]TCTA[15]	MH166988.1	1
range	14-19	TCTA[1]TCTG[1-3]AATC[0-3]TCTA[11-17]		200

Locus	CE	Sequence allele	STRSeq acc#	Allele count
D5S818	7	CE7_CTCT[1]ATCT[7]_+13A>G_rs25768	MH166991.1	2
	8	CE8_ATCT[9]_+13A>G_rs25768	MH166992.1	2
	9	CE9_ATCT[10]_+13A>G_rs25768	MH166995.1	9
	10	CE10_ATCT[11]_+13A>G_rs25768	MH166997.1	9
		CE10_CTCT[1]ATCT[10]	MH166998.1	1
		CE10_CTCT[1]ATCT[10]_+13A>G_rs25768	MH166999.1	7
		CE10_CTCT[1]ATCT[10]_-32A>G_rs182073376_+13A>G_rs25768	MH167000.1	1
	11	CE11_ATCT[12]_+13A>G_rs25768	MH167001.1	8
		CE11_CTCT[1]ATCT[11]	MH167002.1	10
		CE11_CTCT[1]ATCT[11]_+13A>G_rs25768	MH167004.1	39
		CE12_ATCT[13]_+13A>G_rs25768	MH167005.1	20
	12	CE12_CTCT[1]ATCT[12]	MH167007.1	17
		CE12_CTCT[1]ATCT[12]_+13A>G_rs25768	MH167008.1	28
		CE12_CTCT[1]ATCT[12]_+13A>G_rs25768,+58T>G_rs146841551	novel to STRSeq	1
		CE13_ATCT[14]_+13A>G_rs25768	MH167009.1	6
	13	CE13_CTCT[1]ATCT[13]	MH167011.1	15
		CE13_CTCT[1]ATCT[13]_+13A>G_rs25768	MH167013.1	16
		CE13_CTCT[1]ATCT[13]_+13A>G_rs25768,+58T>G_rs146841551	MH167012.1	2
		CE13_CTCT[1]ATCT[3]ATGT[1]ATCT[9]_+13A>G_rs25768	MH167014.1	2
		CE13.3_CTCT[1]ATCT[12]ATC[1]ATCT[1]_+13A>G_rs25768	novel to STRSeq	1
	14	CE14_ATCT[15]_+13A>G_rs25768	MH167015.1	1
		CE14_CTCT[1]ATCT[14]_+13A>G_rs25768	MH167017.1	3
range	7-14	CTCT[0-1]ATCT[7-15]		200

Cont.

Locus	CE	Sequence allele	STRSeq acc#	Allele count
D7S820	7.3	CE7.3_TATC[8]_-22T>-_rs897512434	novel to STRSeq	1
	8	CE8_TATC[8]_-22T>A_rs7789995	MH167026.1	14
		CE8_TATC[8]_-22T>A_rs7789995_+9G>A_rs16887642	MH167025.1	12
		CE8_TATC[8]_-22T>A_rs7789995_-65A>C_rs7786079	MH167027.1	1
	9	CE9_TATC[9]	MH167031.1	1
		CE9_TATC[9]_-22T>A_rs7789995	MH167030.1	18
		CE9_TATC[9]_-22T>A_rs7789995_+9G>A_rs16887642	MH167029.1	6
		CE9_TATC[9]_-22T>A_rs7789995_-65A>C_rs7786079	MH167032.1	1
	9.1	CE9.1_TATC[9]_-22T>A_rs7789995_-65A>C_rs7786079_+0.1->T_rs1373477072	novel to STRSeq	1
	10	CE10_TATC[10]	MH167035.1	5
		CE10_TATC[10]_-22T>A_rs7789995_rs7789995	MH167034.1	37
		CE10_TATC[10]_-22T>A_rs7789995_+9G>A_rs16887642	MH167033.1	2
		CE10_TATC[10]_-22T>A_rs7789995_-65A>C_rs7786079	MH167037.1	8
	10.3	CE10.3_TATC[11]_-22T>-_rs897512434	MH167039.1	1
	11	CE11_TATC[11]	MH167043.1	8
		CE11_TATC[11]_-22T>A_rs7789995	MH167041.1	39
		CE11_TATC[11]_-22T>A_rs7789995_+9G>A_rs16887642	MH167040.1	3
		CE11_TATC[11]_-22T>A_rs7789995_-65A>C_rs7786079	MH167044.1	3
	12	CE12_TATC[12]	MH167048.1	7
		CE12_TATC[12]_-22T>A_rs7789995	MH167047.1	27
	13	CE13_TATC[13]	MH167051.1	1
		CE13_TATC[13]_-22T>A_rs7789995	MH167050.1	3
		CE13_TATC[13]_-22T>A_rs7789995_+9G>A_rs16887642	novel to STRSeq	1
range	7.3-13	TATC[8-13]		200
Locus	CE	Sequence allele	STRSeq acc#	Allele count
D8S1179	8	CE8_TCTA[8]	MH105186.1	1
	9	CE9_TCTA[9]	MH105187.1	1
	10	CE10_TCTA[10]	MH105188.1	18
	11	CE11_TCTA[11]	MH105190.1	15
		CE11_TCTA[2]TCTG[1]TCTA[8]	MH105192.1	1
	12	CE12_TCTA[12]	MH105194.1	18
		CE12_TCTA[1]TCTG[1]TCTA[10]	MH105196.1	9
		CE12_TCTA[2]TCTG[1]TCTA[9]	MH105195.1	4
	13	CE13_TCTA[13]	MH105197.1	16
		CE13_TCTA[1]TCTG[1]TCTA[11]	MH105201.1	38
		CE14_TCTA[14]	MH105204.1	1
	14	CE14_TCTA[1]TCTG[1]TCTA[12]	MH105206.1	27
		CE14_TCTA[2]TCTG[1]TCTA[11]	MH105205.1	11
		CE15_TCTA[1]TCTG[1]TCTA[13]	MH105211.1	5
	15	CE15_TCTA[2]TCTG[1]TCTA[12]	MH105209.1	21
		CE15_TCTA[2]TCTG[1]TCTA[12]_+56G>A_rs182593664	novel to STRSeq	1
		CE15_TCTA[2]TCTG[2]TCTA[11]	MH105210.1	1
	16	CE16_TCTA[2]TCTG[1]TCTA[13]	MH105214.1	8
	17	CE17_TCTA[2]TCTG[1]TCTA[14]	MH105217.1	2
		CE17_TCTA[2]TCTG[2]TCTA[13]	novel to STRSeq	1
	18	CE18_TCTA[2]TCTG[1]TCTA[15]	MH105219.1	1
range	8-18	TCTA[8-14]TCTG[0-2]TCTA[8-15]		200

Cont.

Locus	CE	Sequence allele	STRSeq acc#	Allele count
FGA	18	CE18_GGAA[2]GGAG[1]AAAG[10]AGAA[1]AAAA[1]GAAA[3]	MH232605.1	4
	18.2	CE18.2_GGAA[2]GGAG[1]AAAG[11]AA[1]AAAA[1]GAAA[3]	MH232606.1	1
	19	CE19_GGAA[2]GGAG[1]AAAG[11]AGAA[1]AAAA[1]GAAA[3]	MH232607.1	14
	19.2	CE19.2_GGAA[2]GGAG[1]AAAG[12]AA[1]AAAA[1]GAAA[3]	MH232608.1	1
	20	CE20_GGAA[2]GGAG[1]AAAG[12]AGAA[1]AAAA[1]GAAA[3]	MH232609.1	19
	21	CE21_GGAA[2]GGAG[1]AAAG[13]AGAA[1]AAAA[1]GAAA[3]	MH232611.1	26
	21.2	CE21.2_GGAA[2]GGAG[1]AAAG[14]AA[1]AAAA[1]GAAA[3]	MH232612.1	1
	22	CE22_GGAA[2]GGAG[1]AAAG[14]AGAA[1]AAAA[1]GAAA[3]	MH232613.1	29
	22.2	CE22.2_GGAA[2]GGAG[1]AAAG[15]AA[1]AAAA[1]GAAA[3]	MH232615.1	2
	23	CE23_GGAA[2]GGAG[1]AAAG[1]AAAC[1]AAAG[13]AGAA[1]AAAA[1]GAAA[3]	novel to STRSeq	1
		CE23_GGAA[2]GGAG[1]AAAG[15]AGAA[1]AAAA[1]GAAA[3]	MH232617.1	32
	24	CE24_GGAA[2]GGAG[1]AAAG[16]AGAA[1]AAAA[1]GAAA[3]	MH232622.1	35
	24.2	CE24.2_GGAA[2]GGAG[1]AAAG[17]AA[1]AAAA[1]GAAA[3]	MH232624.1	2
	25	CE25_GGAA[2]GGAG[1]AAAG[17]AGAA[1]AAAA[1]GAAA[3]	MH232626.1	19
	26	CE26_GGAA[2]GGAG[1]AAAG[18]AGAA[1]AAAA[1]GAAA[3]	MH232629.1	2
		CE26_GGAA[2]GGAG[1]AAAG[5]AAGG[1]AAAG[12]AGAA[1]AAAA[1]GAAA[3]	MH232631.1	2
	27	CE27_GGAA[2]GGAG[1]AAAG[17]AGAG[1]AAAG[1]AGAA[1]AAAA[1]GAAA[3]	novel to STRSeq	1
		CE27_GGAA[2]GGAG[1]AAAG[19]AGAA[1]AAAA[1]GAAA[3]	MH232632.1	3
		CE27_GGAA[2]GGAG[1]AAAG[5]AAGG[1]AAAG[13]AGAA[1]AAAA[1]GAAA[3]	MH232633.1	3
	28	CE28_GGAA[2]GGAG[1]AAAG[18]AGAG[1]AAAG[1]AGAA[1]AAAA[1]GAAA[3]	novel to STRSeq	1
		CE28_GGAA[2]GGAG[1]AAAG[5]AAGG[1]AAAG[5]GAAG[1]AAAG[8]AGAA[1]AAAA[1]GAAA[3]	novel to STRSeq	1
	45.2	CE45.2_GGAA[4]GGAG[1]AAAG[3]GAAG[3]AAAG[13]ACAG[3]AAAG[13]AA[1]AAAA[1]GAAA[4]	novel to STRSeq	1
range	18-45.2	GGAA[2-4]GGAG[1]AAAG[3-19]AGAA[0-1]AA[0-1]AAAA[1]GAAA[3-4]		200
Locus	CE	Sequence allele	STRSeq acc#	Allele count
PentaD	2.2	CE2.2_AAAGA[5]AAAAA[1]	MH232670.1	6
	3.2	CE3.2_AAAGA[6]AAAAA[1]	MH232671.1	1
	5	CE5_AAAGA[1]AAAA[1]AAAG[1]AAAGA[5]AAAAA[1]	MH232672.1	2
	6	CE6_AAAGA[1]AAAA[1]AAAG[1]AAAGA[6]AAAAA[1]	MH232673.1	1
	7	CE7_AAAGA[1]AAAA[1]AAAG[1]AAAGA[7]AAAAA[1]	MH232674.1	3
	8	CE8_AAAGA[1]AAAA[1]AAAG[1]AAAGA[8]AAAAA[1]	MH232675.1	12
		CE8_AAAGA[1]AAAA[1]AAAG[1]AAAGA[9]	novel to STRSeq	1
	9	CE9_AAAGA[1]AAAA[1]AAAG[1]AAAGA[9]AAAAA[1]	MH232676.1	34
		CE9_AAAGA[1]AAAA[1]AAAG[1]AAAGA[10]	MH232677.1	1
		CE9_AAAGA[1]AAAA[1]AAAG[1]AAAGA[9]AAAAA[1]_+57T>G_rs7279663	novel to STRSeq	3
	10	CE10_AAAGA[1]AAAA[1]AAAG[1]AAAGA[10]AAAAA[1]	MH232678.1	20
	11	CE11_AAAGA[1]AAAA[1]AAAG[1]AAAGA[11]AAAAA[1]	MH232681.1	33
		CE11_AAAGA[1]AAAA[1]AAAG[1]AAAGA[12]	novel to STRSeq	1
	12	CE12_AAAGA[1]AAAA[1]AAAG[1]AAAGA[12]AAAAA[1]	MH232685.1	34
		CE12_AAAGA[1]AAAA[1]AAAG[1]AAAGA[13]	MH232686.1	1
	13	CE13_AAAGA[1]AAAA[1]AAAG[1]AAAGA[13]AAAAA[1]	MH232687.1	31
	14	CE14_AAAGA[1]AAAA[1]AAAG[1]AAAGA[14]AAAAA[1]	MH232690.1	12
	15	CE15_AAAGA[1]AAAA[1]AAAG[1]AAAGA[15]AAAAA[1]	MH232692.1	3
	16	CE16_AAAGA[1]AAAA[1]AAAG[1]AAAGA[16]AAAAA[1]	MH232693.1	1
range	2.2-16	AAAGA[0-1]AAAA[0-1]AAAG[0-1]AAAGA[5-16]AAAAA[0-1]		200

Cont.

Locus	CE	Sequence allele	STRSeq acc#	Allele count
PentaE	5	CE5_TCTTT[5]	MH232642.1	16
	7	CE7_TCTTT[7]	MH232644.1	16
	8	CE8_TCTTT[8]	MH232645.1	8
	9	CE9_TCTTT[9]	MH232646.1	3
	10	CE10_TCTTT[10]	MH232647.1	8
	11	CE11_TCTTT[11]	MH232648.1	22
	12	CE12_TCTTT[12]	MH232649.1	25
	13	CE13_TCTTT[13]	MH232650.1	24
	14	CE14_TCTTT[14]	MH232651.1	8
	15	CE15_TATTT[1]TCTTT[14]	MH232652.1	1
		CE15_TCTTT[15]	MH232653.1	21
	16	CE16_TCTTT[16]	MH232657.1	11
	17	CE17_TCTTT[1]CCTTT[1]TCTTT[15]	novel to STRSeq	1
		CE17_TCTTT[17]	MH232660.1	13
	18	CE18_TCTTT[18]	MH232662.1	8
	19	CE19_TCTTT[19]	MH232663.1	8
	20	CE20_TCTTT[20]	MH232665.1	4
	21	CE21_TCTTT[21]	MH232666.1	2
	23	CE23_TCTTT[23]	MH232668.1	1
range	5-23	TCTTT[5-23]		200
Locus	CE	Sequence allele	STRSeq acc#	Allele count
TH01	6	CE6_AATG[6]	MH085115.1	31
	7	CE7_AATG[7]	MH085116.1	44
		CE7_AATG[7]_+85C>A_rsNA_@GRCh38,chr11:2171200	novel to STRSeq	2
	8	CE8_AATG[8]	MH085119.1	20
		CE8_AATG[8]_+129C>T_rs79373318	MH085120.1	3
	9	CE9_AATG[9]	MH085121.1	53
	9.3	CE9.3_AATG[6]ATG[1]AATG[3]	MH085124.1	39
	10	CE10_AATG[10]	MH085125.1	8
range	6-10	AATG[6-10]		200
Locus	CE	Sequence allele	STRSeq acc#	Allele count
TPOX	6	CE6_AATG[6]	MF044246.1	1
	7	CE7_AATG[7]	novel to STRSeq	1
		CE7_AATG[7]_-97C>T_rs115644759	MF044247.1	2
	8	CE8_AATG[8]	MF044248.1	82
		CE8_AATG[8]_-52G>T_rs149212737	MF044250.1	1
		CE8_AATG[8]_-96G>A_rs145426142	MG988075.1	1
		CE8_AATG[8]_-97C>T_rs115644759	MF044249.1	3
	9	CE9_AATG[9]	MF044251.1	26
		CE9_AATG[9]_-109C>A_rs13422969	MF044252.1	3
	10	CE10_AATG[10]	MF044253.1	15
	11	CE11_AATG[11]	MF044255.1	57
	12	CE12_AATG[12]	MF044256.1	7
	13	CE13_AATG[13]	MG988077.1	1
range	6-13	AATG[6-13]		200

Cont.

Locus	CE	Sequence allele	STRSeq acc#	Allele count
vWA	13	CE13_TGGA[2]TAGA[1]TGGA[1]TAGA[8]CAGA[4]TAGA[1]	MH167072.1	1
	14	CE14_TGGA[2]TAGA[1]TGGA[1]TAGA[9]CAGA[4]TAGA[1]	MH167076.1	7 ■
		CE14_TGGA[2]TAGA[1]TGGA[1]TAGA[10]CAGA[3]TAGA[1]	MH167077.1	1
		CE14_TGGA[4]TAGA[3]TGGA[1]TAGA[2]CAGA[5]TAGA[1]CAGA[1]	novel to STRSeq	1
		TAGA[1]_+72A>T_rs11063969,_+77C>T_rs11063970,_+90T>C_rs11063971		
		CE14_TGGA[4]TAGA[3]TGGA[1]TAGA[3]CAGA[4]TAGA[1]CAGA[1]	MH167078.1	32 ■
		TAGA[1]_+72A>T_rs11063969,_+77C>T_rs11063970,_+90T>C_rs11063971		
	15	CE15_TGGA[2]TAGA[1]TGGA[1]TAGA[10]CAGA[4]TAGA[1]	MH167081.1	11 ■
		CE15_TGGA[2]TAGA[1]TGGA[1]TAGA[11]CAGA[3]TAGA[1]	MH167082.1	8 ■
	16	CE16_TGGA[2]TAGA[1]TGGA[1]TAGA[9]CAGA[6]TAGA[1]	novel to STRSeq	1
		CE16_TGGA[2]TAGA[1]TGGA[1]TAGA[11]CAGA[4]TAGA[1]	MH167084.1	38 ■
		CE16_TGGA[2]TAGA[1]TGGA[1]TAGA[11]CAGA[5]	novel to STRSeq	1
		CE16_TGGA[2]TAGA[1]TGGA[1]TAGA[12]CAGA[3]TAGA[1]	MH167085.1	8 ■
		CE17_TGGA[2]TAGA[1]TGGA[1]TAGA[12]CAGA[4]TAGA[1]	MH167088.1	42 ■
		CE17_TGGA[2]TAGA[1]TGGA[1]TAGA[12]CAGA[5]	novel to STRSeq	1
	17	CE17_TGGA[2]TAGA[1]TGGA[1]TAGA[13]CAGA[3]TAGA[1]	MH167090.1	3 ■
		CE18_TGGA[2]TAGA[1]TGGA[1]TAGA[11]CAGA[6]TAGA[1]	MH167091.1	1
		CE18_TGGA[2]TAGA[1]TGGA[1]TAGA[12]CAGA[5]TAGA[1]	MH167092.1	1
		CE18_TGGA[2]TAGA[1]TGGA[1]TAGA[13]CAGA[4]TAGA[1]	MH167093.1	22 ■
	19	CE19_TGGA[2]TAGA[1]TGGA[1]TAGA[12]CAGA[6]TAGA[1]	MH167095.1	1
		CE19_TGGA[2]TAGA[1]TGGA[1]TAGA[13]CAGA[5]TAGA[1]	MH167096.1	1
		CE19_TGGA[2]TAGA[1]TGGA[1]TAGA[14]CAGA[4]TAGA[1]	MH167097.1	14 ■
		CE19_TGGA[2]TAGA[1]TGGA[1]TAGA[15]CAGA[3]TAGA[1]	MH167098.1	1
	20	CE20_TGGA[2]TAGA[1]TGGA[1]TAGA[15]CAGA[4]TAGA[1]	MH167101.1	4 ■
range	13-20	TGGA[2-4]TAGA[1-3]TGGA[1]TAGA[2-15]CAGA[3-6]TAGA[0-1]		200

4.3.3 Diversity of observed alleles

By sequencing STR alleles the number of allele variants grows, and this increase in allele diversity is presented in Figure 4.1, showing the numbers of length-based alleles and the increase delivered by sequence-based alleles.

To represent the proportion of repeat-array and flanking region variants, these are plotted separately in Figure 4.1. When both a variant internal repeat structure and a flanking variant(s) are present these are included in the array variants.

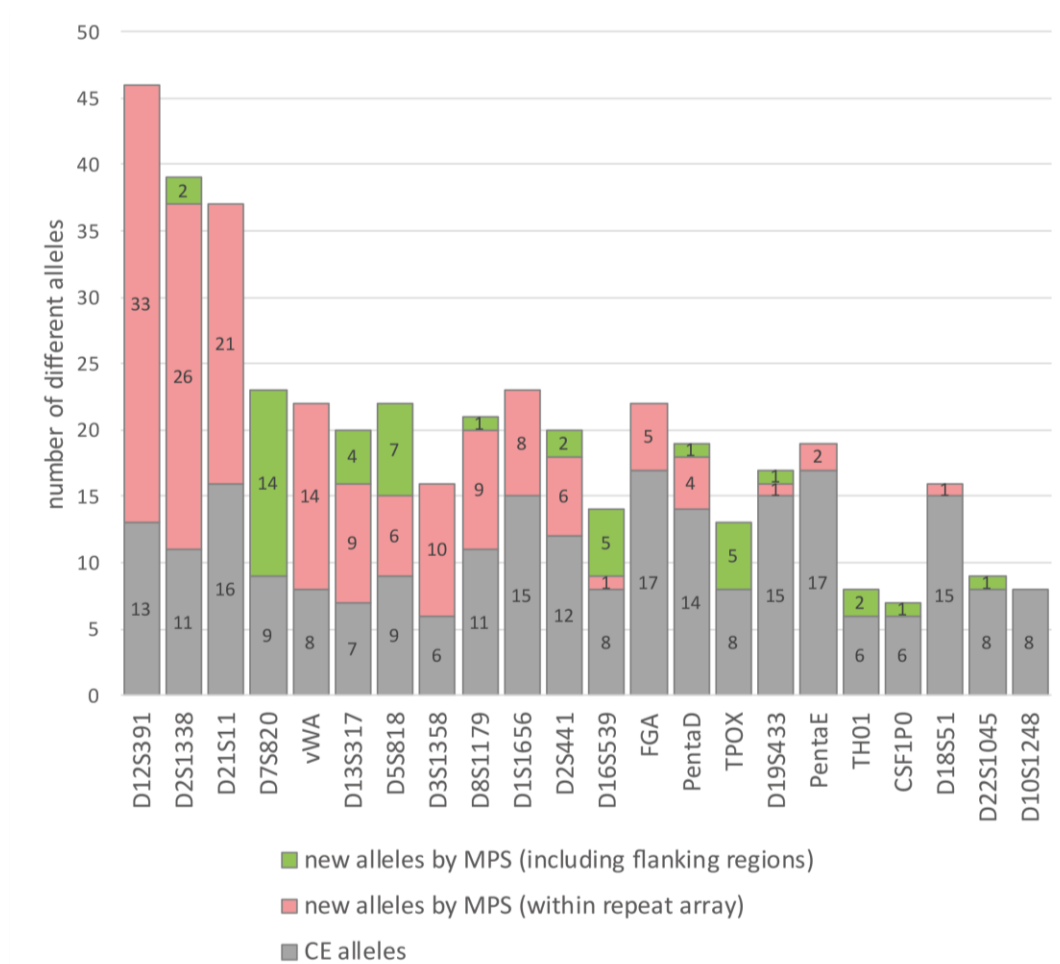


Figure 4.1
Increase of allele diversity by type of variants.

Increase in number of different alleles for each autosomal STR using sequence information from within the arrays and from the flanking regions.

All but one of the STRs (D10S1248) showed increased allelic diversity when analysed by MPS. The majority of new alleles are the result of internal SNPs, indels and repeat pattern variants (RPVs) within the STR arrays, and only 22.8% of MPS alleles are results of solely flanking region variation. However, loci with simple array structures, especially D7S820 and TPOX in this set, gained new alleles by MPS only due to flanking region variants.

4.3.4. Isometric heterozygote alleles and observed heterozygosity

An isometric heterozygote carries a pair of alleles that are indistinguishable by length, and which thus fall under the same CE allele category as isoalleles, but bear sequence variations, and therefore can be recognised as distinct alleles by sequence analysis.

In this set a total of 100 pairs of alleles were found to be homozygous by length, but heterozygous by sequence. The distribution of observed isometric heterozygote pairs across the autosomal STRs is summarised in Table 4.3.

Table 4.3
Autosomal STRs with isometric heterozygote allele pairs.

List of 18 autosomal STRs with isometric heterozygote allele pairs observed.

Locus	Isometric pairs	Length alleles (n)	Sequence alleles (p)	p/n ratio
D5S818	15	9	22	2.444
D13S317	14	7	20	2.857
D12S391	11	13	46	3.538
D7S820	9	9	23	2.555
D3S1358	9	6	16	2.667
D16S539	9	8	14	1.75
D8S1179	7	11	21	1.909
D21S11	7	16	37	2.313
vWA	5	8	22	2.75
D2S441	3	12	20	1.667
D2S1338	2	11	39	3.545
TH01	2	6	8	1.333
TPOX	2	8	13	1.625
PentaD	1	14	19	1.357
D1S1656	1	15	23	1.533
CSF1P0	1	6	7	1.167
D22S1045	1	8	9	1.125
FGA	1	17	22	1.294
D10S1248	0	8	8	1
D18S51	0	15	16	1.067
D19S433	0	15	17	1.133
PentaE	0	17	19	1.118
Total	100	239	441	

This means that 4.5% of the total of 2200 autosomal STR genotypes are isometric heterozygotes and 21.3% of length-based homozygous genotypes are resolved as heterozygous by sequence.

These allele pairs were observed in 18 of the STR loci, with the exception of D10S1248, D18S51 and PentaE, which are simple STRs with relatively narrow allele ranges, and D19S433 which has a complex structure, but with limited variation within alleles in the same length class.

Most of the isometric pairs were observed at D5S818, D13S317 and D12S391, all of which have a relatively narrow length-based allele range combined with several sequence-based variants, which also results in the increase in observed heterozygosity for these loci when typed with MPS (Table 4.4)

Table 4.4
Autosomal STR loci in order of maximum increase in observed heterozygosity.

List of the analysed autosomal STR loci in order of maximum increase in observed heterozygosity from length-based to sequence-based genotyping.

	observed heterozygosity		
	length-based	sequence-based	increase
D5S818	0.73	0.88	0.15
D13S317	0.77	0.91	0.14
D12S391	0.84	0.95	0.11
D16S539	0.75	0.84	0.09
D3S1358	0.77	0.86	0.09
D7S820	0.76	0.85	0.09
D21S11	0.82	0.89	0.07
D8S1179	0.82	0.89	0.07
vWA	0.80	0.85	0.05
D2S441	0.76	0.79	0.03
D2S1338	0.83	0.85	0.02
TH01	0.79	0.81	0.02
TPOX	0.62	0.64	0.02
D1S1656	0.89	0.90	0.01
D22S1045	0.73	0.74	0.01
FGA	0.83	0.84	0.01
CSF1P0	0.68	0.69	0.01
PentaD	0.81	0.82	0.01
D10S1248	0.73	0.73	0.00
D18S51	0.80	0.80	0.00
D19S433	0.85	0.85	0.00
PentaE	0.92	0.92	0.00
	0.786	0.832	0.045

4.3.5 Heterozygote balance

One of the reasons why STRs are preferred as markers in forensic DNA analysis is that although the two alleles differ in length in heterozygotes, the length difference has little effect on the amplification efficiency, and therefore heterozygous alleles amplify in a generally balanced way. In practice, heterozygotes show a slightly preferential amplification of the shorter allele due the size difference of the targeted templates; however, unless the allele span between the two alleles is especially large, the peak height (in CE) or reads (in MPS) of the alleles are usually roughly equal. Whether sequence structure itself affects heterozygote imbalance is unclear.

To compare the effect of allele span on heterozygote balance, Figure 4.2 shows the balance and its variance across loci for isometric heterozygotes (which are not affected by length difference), and for comparison Figure 4.3 shows the effect of allele span between heterozygotes on the observed heterozygote balance.

Considering the issue of sequence effects on balance, there appears to be no direct correlation between imbalance and the structure of the arrays or the largest number of longest uninterrupted repeats, as both complex and simple loci can show more imbalance as the span between the two alleles grows.

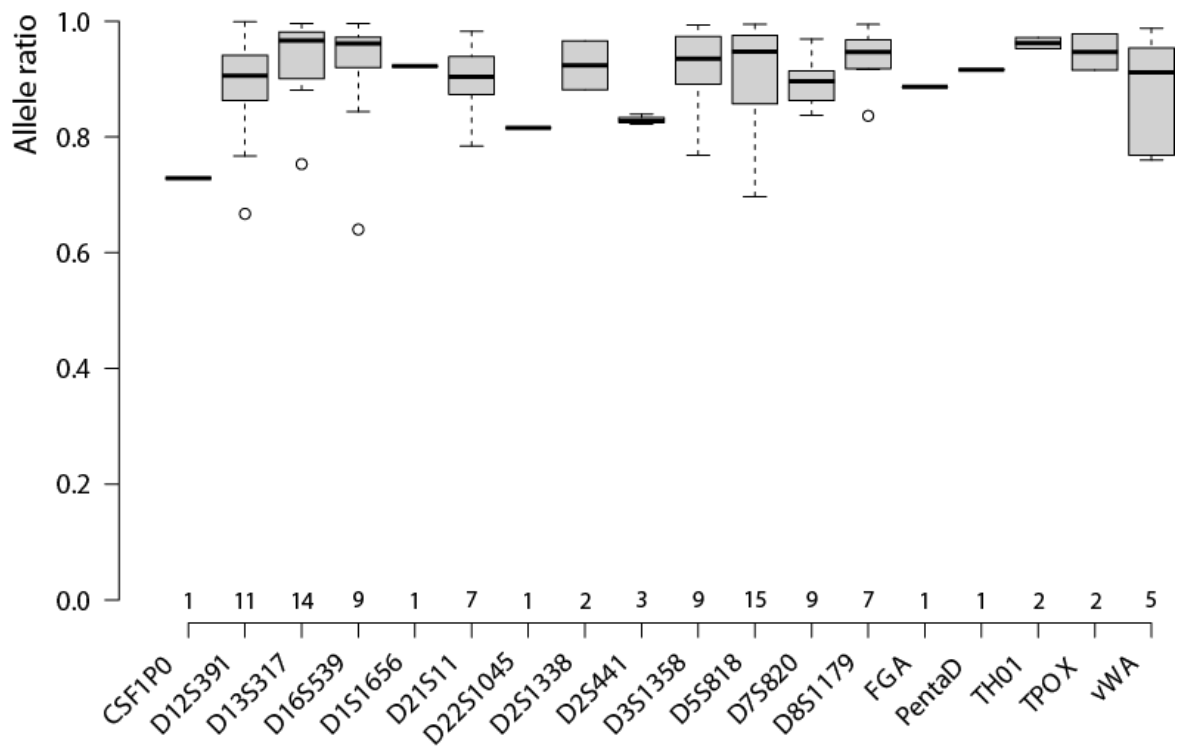


Figure 4.2
Heterozygote balance in the 100 isometric heterozygote pairs across 18 loci.

The boxplot shows mean heterozygote balance values for each locus. Numbers of observations are marked above the span axis for each plotbar. Centre lines indicate the medians; box limits are the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by circles.



The boxplot shows mean heterozygote balance values for each allele span across all loci from 0 to 23.2 repeat unit spans. Numbers of observations are marked above the span axis for each plotbar. Centre lines indicate the medians; box limits are the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by circles.

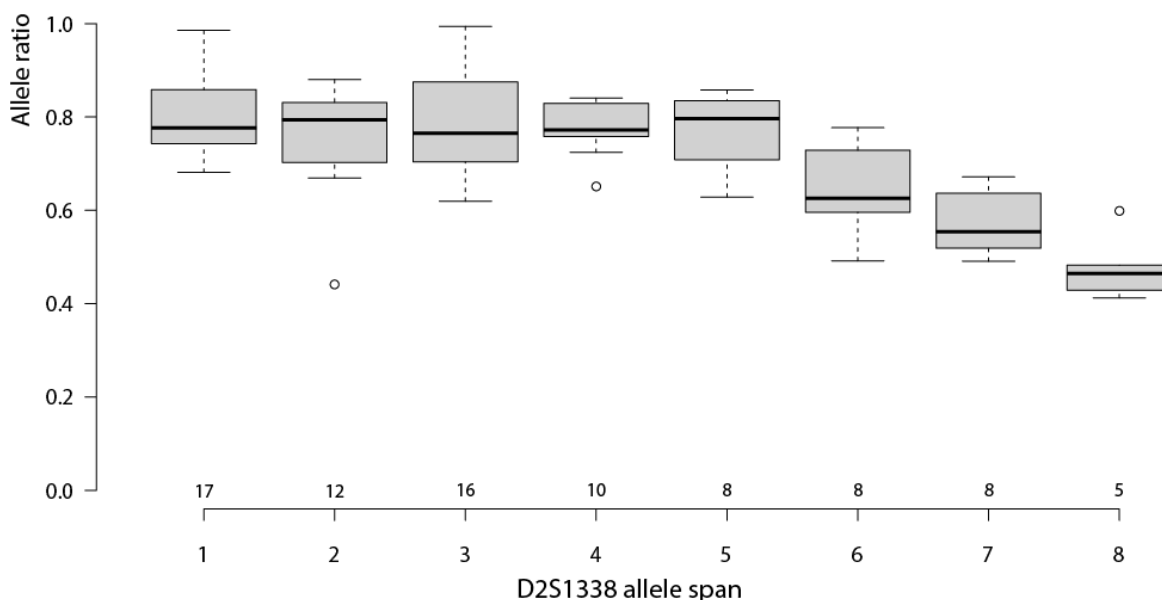


Figure 4.4

Example of a decline of heterozygote balance in an STR locus over larger allele spans.

The boxplot shows mean heterozygote balance values for locus D2S1338 for each allele span from 1 to 8 repeat units. Numbers of observations are marked above the span axis for each plotbar. Centre lines indicate the medians; box limits are the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by circles.

4.3.6 Variants that are novel to the STRSeq database

Sequence-based allele variants compiled in Table 4.2 were compared to the NCBI-hosted STRSeq BioProject (Gettings et al. 2017), and specifically its sub-project dedicated to cataloguing the most often used autosomal STR loci under the identifier: PRJNA380345 (<https://www.ncbi.nlm.nih.gov/bioproject/380345>). The sequence variants that were identified in the STRSeq database were marked by unique accession numbers in Table 4.2, and alleles not yet described in this database were flagged as ‘novel to STRSeq’. Fifty-five allele calls contained such alleles and the total of 48 unique alleles are listed separately in Table 4.5, which summarises these ‘novel to STRSeq’ variants, detailing their aspects of novelty.

Table 4.5**48 sequence-based alleles identified as 'novel to STRSeq'.**

Sequence structures are given with their aspects of novelty and origin. Aspects of novelty are marked with tick marks or crosses, representing the presence or absence of the feature, respectively.

Locus	Sequence adapted to the STRSeq format	RPV	iSNP	fSNP	length	indel	Metapop.
D12S391	19 [AGAT]14 [AGAC]4 AGAT	✓					African
	25 [AGAT]13 [AGAC]11 AGAT	✓					European
D13S317	8 [TATC]8 rs9546005	✓		✓			Asian
	10 [TATC]9 AATC	✓					African
	11 [TATC]6 TATT [TATC]4 rs9546005		✓				Australian
D16S539	5 [GATA]5			×	✓		African
	9 [GATA]9 rs1728369			✓	✓		African
	11 [GATA]5 GACA [GATA]5 rs1728369		✓				Asian
D19S433	11.2 [CCTT]10 ccta CCTT tt CCTT				✓	✓	African
	14.2 [CCTT]13 ccta CCTT cttt CCTT rs745607776					✓	African
D151656	8 [TCTA]8				✓		Asian
D21S11	25 [TCTA]4 [TCTG]6 [TCTA]3 ta [TCTA]3 tca [TCTA]2 tccata [TCTA]7				✓		NM Eastern
	26 [TCTA]6 [TCTG]7 [TCTA]3 ta [TCTA]3 tca [TCTA]2 tccata [TCTA]5	✓					African
	29.2 [TCTA]5 [TCTG]6 [TCTA]3 ta [TCTA]2 tca [TCTA]2 tccata [TCTA]10 TA TCTA	✓					NM Eastern
	30 [TCTA]4 [TCTG]6 TCTA TCTG TCTA ta [TCTA]3 tca [TCTA]2 tccata [TCTA]12		✓				NM Eastern
	31.2 [TCTA]4 [TCTG]7 [TCTA]3 ta [TCTA]3 tca [TCTA]2 tccata [TCTA]11 TA TCTA	✓					European
	34 [TCTA]10 [TCTG]6 [TCTA]3 ta [TCTA]3 tca [TCTA]2 tccata [TCTA]10	✓					African
	34 [TCTA]11 [TCTG]5 [TCTA]3 ta [TCTA]3 tca [TCTA]2 tccata [TCTA]10	✓					African
	35.2 [TCTA]5 [TCTG]6 [TCTA]3 ta [TCTA]3 tca [TCTA]2 tccata [TCTA]15 TA TCTA	✓					European
D251338	19 [GGAA]15 [GGCA]4 rs6736691	✓		✓			African
	24 [GGAA]17 [GGCA]7	✓					African
D2S441	10.1 A [TCTA]10 rs74640515					✓	Asian
	11 [TCTA]8 TTTA [TCTA]2		✓				Asian
	12.3 [TCTA]3 TCA [TCTA]9					✓	African
	15 [TCTA]12 TTTA TCTA TCTG		✓				Asian
D3S1358	19 TCTA TCTG [TCTA]17	✓					European
D5S818	12 [ATCT]12 rs25768 rs146841551	✓		✓			African
	13.3 [ATCT]12 ATC ATCT rs25768					✓	African
D7S820	7.3 [TATC]8 rs897512434	✓				✓	African
	9.1 [TATC]9 rs7789995 rs7786079 rs1373477072	✓		✓		✓	Asian
	13 [TATC]13 rs7789995 rs16887642	✓		✓			Asian
D8S1179	15 [TCTA]2 TCTG [TCTA]12 rs182593664	✓		✓			Asian
	17 [TCTA]2 [TCTG]2 [TCTA]13	✓					Asian
FGA	23 [GGAA]2 GGAG AAAG AAAC [AAAG]13 AGAA AAAA [GAAA]3		✓				European
	27 [GGAA]2 GGAG [AAAG]17 AGAG AAAG AGAA AAAA [GAAA]3		✓				Asian
	28 [GGAA]2 GGAG [AAAG]18 AGAG AAAG AGAA AAAA [GAAA]3		✓				Asian
	28 [GGAA]2 GGAG [AAAG]5 AAGG [AAAG]5 GAAG [AAAG]8 AGAA AAAA [GAAA]3		✓				African
	45.2 [GGAA]4 GGAG [AAAG]3 [GAAG]3 [AAAG]13 [ACAG]3 [AAAG]13 AA AAAA GAAA[4]	✓			✓		African
PentaD	8 [AAAGA]8 rs186259515	✓		✓			African
	9 [AAAGA]9 rs7279663	✓		✓			European, NM Eastern
	11 [AAAGA]11 rs186259515	✓		✓			European
PentaE	17 TCTTT CCTTT [TCTTT]15	✓	✓				African
TH01	7 [AATG]7 rsNA@GRCh38,chr11:2,171,200	✓		✓			Asian
TPOX	7 [AATG]7			×	✓		African
vWA	14 [TAGA]3 TGGG [TAGA]2 [CAGA]5 TAGA CAGA TAGA rs75219269 rs11063969 rs11063970 rs11063971		✓	✓			African
	16 [TAGA]9 [CAGA]6 TAGA	✓					African
	16 [TAGA]11 [CAGA]5		✓				Asian
	17 [TAGA]12 [CAGA]5		✓				European

4.4 Discussion

Here, DNA sequence variation was described in the 22 autosomal STRs and the XY-homologous Amelogenin loci amplified in the prototype PowerSeq™ Auto/Mito/Y System within a diverse set of 100 genomes, originally selected for Y-chromosomal diversity. As with the Y-STRs, a wide range of variable alleles were observed in the autosomal STRs. The amplified X and Y Amelogenin markers had no sequence variants different from their reference sequences in GRCh38.

General read statistics were comparable to the corresponding Y-STR values (as described in Chapter 3, Section 3.3.2) from the same experiments. Due to diploidy, mean coverage values were consistently higher for autosomal markers, as expected. Observed coverage exceeded the set analytical threshold of 20 × and allowed the calling of diploid alleles with certainty. Allele calls showed no discrepancies between calling methods.

As this dataset does not represent populations, the main focus was to set a wide framework of variants observed from global scale variation. Though the sample size of the study is limited, the geographical diversity is well represented (Figure 2.1, Table 2.1); however, a comprehensive geographical analysis is not the merit of this study, and have better been addressed recently by Phillips et al. (2018) sequencing 944 human genome diversity panel (HGDP-CEPH) samples from 51 globally distributed populations. A total of 4400 alleles of STRs were described, which accounts for 443 different sequence level alleles, an 83.8% increase from just 241 distinct alleles identified by length.

The additional sequence-based allele variants of autosomal STRs mostly originated from variation within the repeat arrays, a tendency similar to that observed among the Y-STR alleles. However, some loci, for example D7S820, presented new alleles only due to flanking region variation, a finding that correlates with the simple internal repeat array structures of such STRs. Simple repeats, like D7S820, TPOX or CSF1PO lack the compound structure which is the usual source of large numbers of repeat pattern variants, represented by combinations of different number of

building blocks of the compound and complex repeats. Due to their structures, most variants of the RPV type are seen in the loci D12S391, D2S1338 and D21S11. These gained the most from MPS typing, the first two with an increase in the number of distinct alleles of over 150% and 130% respectively, thus becoming the most diverse loci in this set.

One clear advantage of using MPS is the consequent increase in allele diversity and thus the added ability to resolve genotypes from length homozygotes as heterozygous by sequence. When isoalleles of the same length but different sequence present at a marker, these form an isometric heterozygote genotype. In this sample set 100 pairs of isometric heterozygote genotypes were observed at 18 loci. Those loci showing more (up to 15 pairs) tend to have more complex structures that facilitate a high number of RPV formations within the same overall allele length, while at the same time having a relatively narrow allele range in which most of the length-based alleles present more than one sequence-based alternatives. When looking at heterozygote balance based on sequence information, isometric heterozygotes represent a class with zero span between alleles, and compared to these the imbalance is increased by the size of the allele span between alleles in length-heterozygotes, while array structure types do not appear to have a direct affect on the imbalance.

When sequence-based allele variants were compiled, these were compared to the STRSeq BioProject (PRJNA380345) collating sequence variants for the most widely used autosomal STRs. Variants were identified by unique accession numbers; however, 48 alleles described here were not represented in STRSeq, and were therefore summarised as sequence-based alleles 'novel to STRSeq'. Their aspects of novelty were described, and in the majority of cases were due to new combinations of repeat units, which occasionally were caused by internally occurring SNPs or indels, but mostly by RPVs. Alleles were found to include some at the extremes of an allele range, shorter or longer than the catalogued alleles. Several variants contained a new combination of flanking region SNPs or indels

and the main array, while some represented rare flanking region variants not previously encountered.

Though comparison to length-based typing methods (CE) was not performed here, the presence of flanking region indels emphasises the importance of flanking region variation: reporting only the array structure while omitting the flanking areas could result in discordance between CE and MPS. Therefore here all aSTR sequence variants are not just provided in a bracketed format but also provided as sequence strings with exact genomic coordinates of the reported region, to provide compatibility of the observed variants when typed with different platforms or methods.

4.4.1 Highlights and Conclusions

Application of MPS resulted in an increase in allele diversity across 21 of the 22 typed aSTRs, which mostly resulted from structural variation within repeat arrays; however, some loci only presented flanking region variants.

Overall 4.5% of all genotypes were found to be isometric heterozygotes: homozygous by length, but heterozygous by sequence. Most of these pairs were observed at loci with relatively narrow length-based allele ranges combined with several sequence-based variants per length class.

A Y-chromosomal diversity-driven sample selection also provided a wide range of sequence variants at the autosomal STR loci, and identified 48 alleles not yet described in the STRSeq sequence allele database.

CHAPTER 5: mtDNA analysis using an MPS workflow

Work described in this chapter has been published as:

Huszar, T.I., Wetton, J.H. and Jobling, M.A. 2019. **Mitigating the effects of reference sequence bias in single-multiplex massively parallel sequencing of the mitochondrial DNA control region.** *Forensic Sci Int Genet*, 40:9-17.

doi: [10.1016/j.fsigen.2019.01.008](https://doi.org/10.1016/j.fsigen.2019.01.008)

The paper itself is included in the electronic appendices (Appendix B).

5.1 Introduction

When standard autosomal STR profiling fails because DNA in a biological sample is limited or highly degraded, forensically useful information can often be obtained via analysis of mitochondrial DNA (mtDNA) (Holland and Parsons 1999), thanks to its relatively high copy number in cells compared to the nuclear genome. Mitochondrial DNA is also the only viable source of genetic evidence from samples such as hair shafts (Higuchi et al. 1988), in which nuclear DNA is naturally depleted. Forensic practice has largely focused on analysing the 1122-bp control region, in which high levels of variation can be detected (Parson and Bandelt 2007; Parson et al. 2014), and which can be sequenced via Sanger technology from a single PCR amplicon when sample quality allows.

In particularly poor-quality samples, a set of overlapping shorter PCR amplicons spanning the control region can be used to generate sequence data; this was the approach taken in early ancient DNA analyses (Klings et al. 1999; Klings et al. 1997). However, the variability of the control region that makes it useful also makes PCR primer design challenging, because sequence variants within primer binding sites could inhibit the amplification due to mismatching (Eichmann and Parson 2008; Parson and Bandelt 2007). Surveying known variation in large reference mtDNA datasets has allowed rational primer design that avoids the most mutable

sites (Eichmann and Parson 2008). However, the multiple-amplicon approach to control region analysis remains vulnerable to variants within particular mtDNA haplogroups that may be unaccounted for in the primer design, and therefore could inhibit amplification in an amplicon- and sample-specific way in casework. If amplification of the primary source's mtDNA is reduced due to such mismatches, a non-mismatched minor source or nuclear mitochondrial DNA (numt) sequence might be amplified relatively efficiently, leading to complications of interpretation.

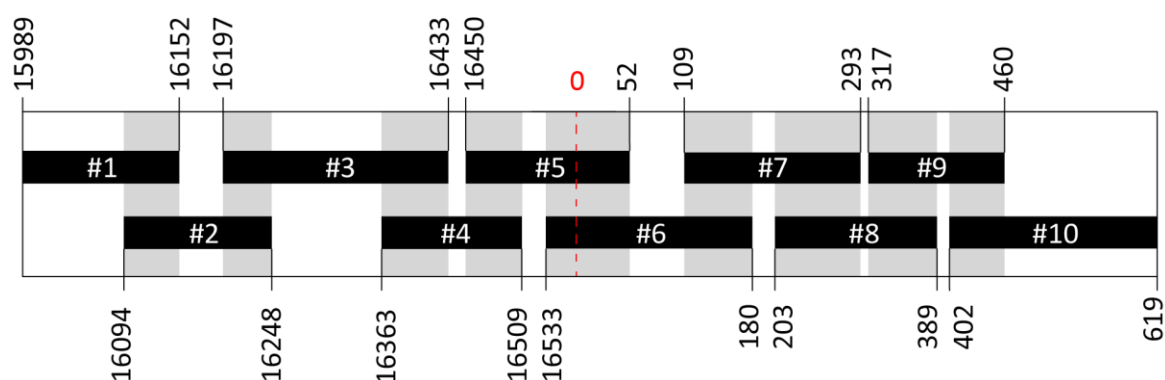


Figure 5.1
Schematic representation of a multiplex assay design over the mtDNA control region.

Amplicons are represented by the black bars, overlapping regions are highlighted in grey, vertical numbers are the start and end mtDNA positions.

Amplicons generated from within the control region (Figure 5.1; between 144 and 237 bp in length; Eichmann and Parson 2008) have traditionally been analysed individually by Sanger sequencing (Berger and Parson 2009; Eichmann and Parson 2008). However, the process can be made more efficient by multiplexing followed by massively parallel sequencing (Holland et al. 2011). Different commercial assay designs exist, either as two multiplexes, each containing alternate amplicons (e.g. Precision ID mtDNA Control Region Panel (Thermo Fisher); ForenSeq™ mtDNA Control Region (Verogen) - neither tested here), or as a single multiplex

PowerSeq™ CRM Nested System (Figure 5.1). The latter has the advantage of simplifying workflow and reducing sample usage. However, it also has the potential disadvantage that primer pairs will act in combinations other than those intended in the design and create many additional products that may be preferentially amplified because of their smaller size (also see Chapter 2, Figure 2.5a). In an MPS setting, such products consume sequencing capacity without contributing additional information about the mtDNA template and are expected to further increase the non-uniform coverage across the control region.

A further potential problem of targeted amplification of the control region is the introduction of bias, particularly reference sequence bias, into the reads. Primers are generally designed to primarily match the reference sequence, therefore in a sample carrying an alternative variant at the primer binding site in a position that is not prohibitive to amplification, the sequence reads could present a mix of an alternative allele originating from the cognate mtDNA, and the other allele matching the reference at a variant position originating from the primer sequences. Degenerate primer design can account for frequent and thus expected variants, resulting in more faithful amplification of cognate mtDNA, with reduced reference-bias from the primers. However, in the case of rare, unexpected variants unaccounted for by degenerate primers this reference bias may increase to the level where it interferes with accurate variant calling. In general, primer-derived sequences should be excluded from the reads, and if the primer sequences are known, these can be removed during data processing by trimming; however, in commercial kits this exact information is proprietary, and therefore unavailable. Alternatively, an attempt can be made to remove primer sequences by educated guesswork, but this may remove cognate mtDNA information as well, or retain some untrimmed primer sequences. Without any corrective process the primer sequences would remain untrimmed and thus likely to interfere with variant calling in the primer binding regions. The single-reaction design which gives rise to preferentially amplified short amplicons (here 52 bp to 91 bp) then further enhances the bias due to their large contribution to the coverage and to the fact that most of their content is primer-derived.

This bias is also expected to influence the accurate assessment of heteroplasmy - the presence of more than one control region sequence type due to mutation *in vivo* (Parson et al. 2014). If heteroplasmic variants lie within putative primer-binding sites, the level at which these are detected might be inaccurate (and probably mainly underestimated), because they are diluted by the presence of reads with primer-derived, and therefore reference-biased sequences. Under the reasonable assumption that commercial designs use degenerate primer sets to account for common variants that lie under the selected primer-binding sites, these alternative primers, even when incorporated with lower efficiency, could introduce false low-level variants not intrinsic to the samples at the affected positions. The presence of these low-level introduced variants will not usually cause any interference in typing reference-quality samples, but if the position is affected by genuine heteroplasmy or is a case-work sample with low-template DNA the stochastic effect in amplification can magnify these introduced variants to the level of detection. Considering this problem, a method is therefore required either during the library preparation (e.g. partial digestion of primers off the generated amplicons (Strobl et al. 2018) used in the Precision ID mtDNA Control Region Panel; Thermo Fisher) or during data processing, that accounts for the presence of the primer-derived sequences and minimises the effect of these on the data prior to reporting of variants.

In this chapter a highly diverse set of 101 DNA samples that cover most of the major clades of the mitochondrial DNA haplogroup tree was used to generate data using a prototype single-reaction mtDNA MPS multiplex. In addition, a subset of the same samples was retested with the commercially available PowerSeq™ CRM Nested System in order to assess its improved accuracy in detecting variants and heteroplasmy in the regions affected by primer-derived bias in the prototype version.

A bioinformatic method was developed to decrease the effects of non-uniform coverage and reference-sequence bias in both systems, thereby improving the detection of variants, and the quantification of heteroplasmy.

5.2 Materials and Methods

5.2.1 DNA samples

One hundred male DNA samples were selected as described in Chapter 2, Table 2.1. In addition, as a control for operator contamination, DNA was extracted from a buccal swab from the operator (TIH) using the DNA IQ™ System (Promega), giving a total of 101 analysed DNA samples.

5.2.2 DNA quantitation

Quantities of double-stranded DNA were verified prior to PCR using a Qubit® 2.0 fluorometer (Thermo Fisher Scientific) with the Qubit® dsDNA HS kit.

5.2.3 PCR amplification, library preparation and sequencing

A segment of mitochondrial DNA was amplified from position 15,989 to 619 (which includes the control region from 16,024 to 576) in a single reaction, generating ten overlapping amplicons from 0.5 ng template DNA.

All samples were amplified using the prototype multiplex PowerSeq™ Auto/Mito/Y System (Promega) and were prepared for sequencing in a separate library preparation step prior to sequencing on a MiSeq FGx® (Illumina®) sequencer using the parameters described in Chapter 2, Section 2.2.2. A subset of the samples was amplified using the PowerSeq™ CRM Nested System (Promega), and in the same single-step reaction generated libraries which were then sequenced using the parameters described in Chapter 2, Section 2.2.3. Both were performed following the manufacturer's recommended protocols. Apart from the library preparation, the two protocols differed in that the prepared libraries from the PowerSeq™ CRM Nested System were quantified with the PowerSeq™ Quant MS System (Promega), while the prototype system used the KAPA Library Quantification Kit for Illumina® platforms (Kapa Biosystems, later acquired by the Roche Group). The

sequencing parameters also differed for the two set-ups: the prototype used v2 (300 cycles) sequencing reagent kits, with the SE sequencing method, while the PowerSeq™ CRM Nested System used v3 (600 cycles) sequencing reagent kits with the PE sequencing method as detailed in Chapter 2, Section 2.2.4.

These differences are due to the different library preparation approaches used in the two kit types, as explained in Chapter 2, Section 2.2.2.4.

5.2.4 Data processing and analyses

Quality-checked FASTQ files were generated as described in Chapter 2 Section 2.3.2. Variants were called using a standard variant calling pipeline, adjusted to amplicon sequencing (omitting the duplicate removal step). The tools used in the pipeline are as described in Chapter 2 Section 2.3.3. The reads were aligned to the revised Cambridge Reference Sequence (Andrews et al. 1999; rCRS, GenBank accession: NC_012920.1). Variants were also visualised and confirmed using the Integrative Genomics Viewer (IGV) tool (Robinson et al. 2011).

In-house Bash scripts were used to seek the presence of sequence reads amplified from numt sequences, to attempt the assessment of the level of heteroplasmy, and to interpret insertions and deletions and other variants found in ambiguous regions, such as homopolymeric tracts and the AC-repeat region.

The ‘Overarching Read Enrichment Option’ (OREO), as explained in Chapter 2, Section 2.3.3.4 was applied to increase the confidence of lower-level heteroplasmy detection and quantitation within the putative primer-binding regions.

Two tools designed for the analysis of mtDNA MPS data were also tested: mtDNAServer (Weissensteiner et al. 2016a); and the trial version of the commercial software GeneMarker®HTS v1.2.2.1338 (SoftGenetics®, LLC).

Three methods were tested to account for unknown primer sequences: Cutadapt (Martin 2011) and seqtk (available from: <https://github.com/lh3/seqtk>) to trim reads, and BAMClipper (Au et al. 2017) to trim BAM files. For the analysis of the

commercially available PowerSeq™ CRM Nested System data, Trimmomatic v0.32 software (Bolger et al. 2014) was used for *in silico* size-selection as an additional step incorporated into data processing.

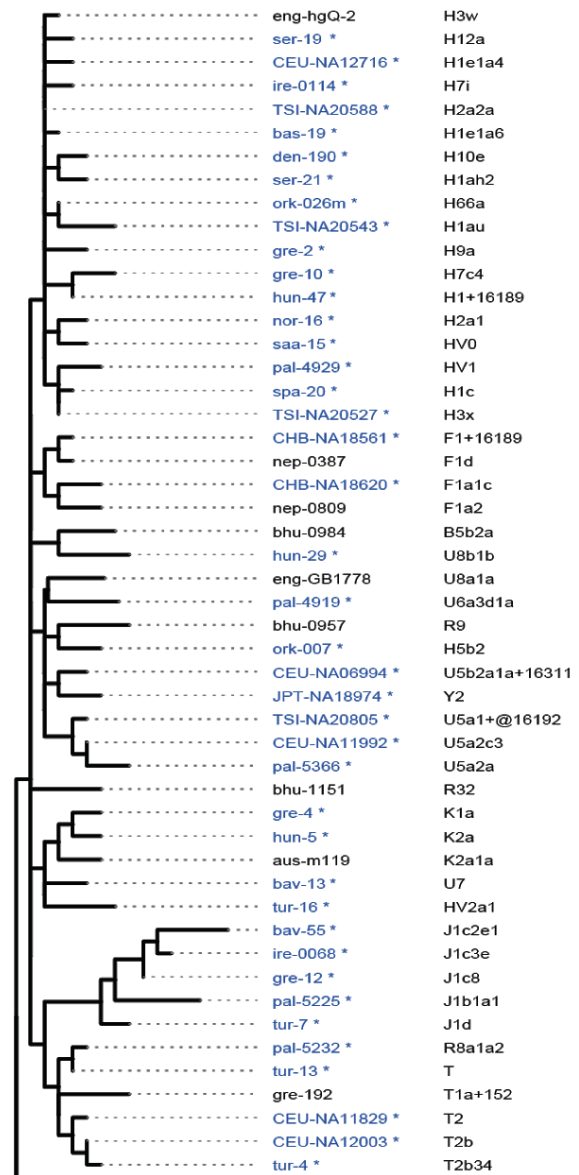
Validation of the results was performed by comparing SNP calls against a total of 65/101 independently sequenced samples, as follows: 58 from previously published data (Batini et al. 2017), of which 23 were also sequenced by the 1000 Genomes Project (1000 Genomes Project Consortium 2015), thereby providing a three-way comparison; six additional 1000 Genomes Project samples; and one operator (TIH) control sample sequenced commercially (GenBank accession: MG551929).

Called variants were checked for correct nomenclature in a phylogenetic alignment context using the SAM2 tool of the EMPOP mtDNA database, v4/R11 (Huber et al. 2018). Haplogroups were predicted and their relationships visualised using HaploGrep2 (Weissensteiner et al. 2016b). A maximum-likelihood phylogenetic tree of the control region (from position 15,989 to 619) was constructed using MEGA v6.06 (Tamura et al. 2013) then visualised using FigTree v1.4.3 (Rambaut 2006-2012).

Graphs and diagrams were created using Microsoft Office and R software (R Core Team 2014).

5.3 Results

A set of DNA samples was analysed that were selected previously to establish a phylogenetic framework for maximum diversity of the male-specific region of the Y chromosome (Huszar et al. 2018). These samples derive from ethnically diverse individuals including Europeans, Asians and Africans (Table 2.1) also representing many major mtDNA haplogroups as demonstrated by Figure 5.2 and Appendix E.



(Phylogenetic tree continues on the next page)

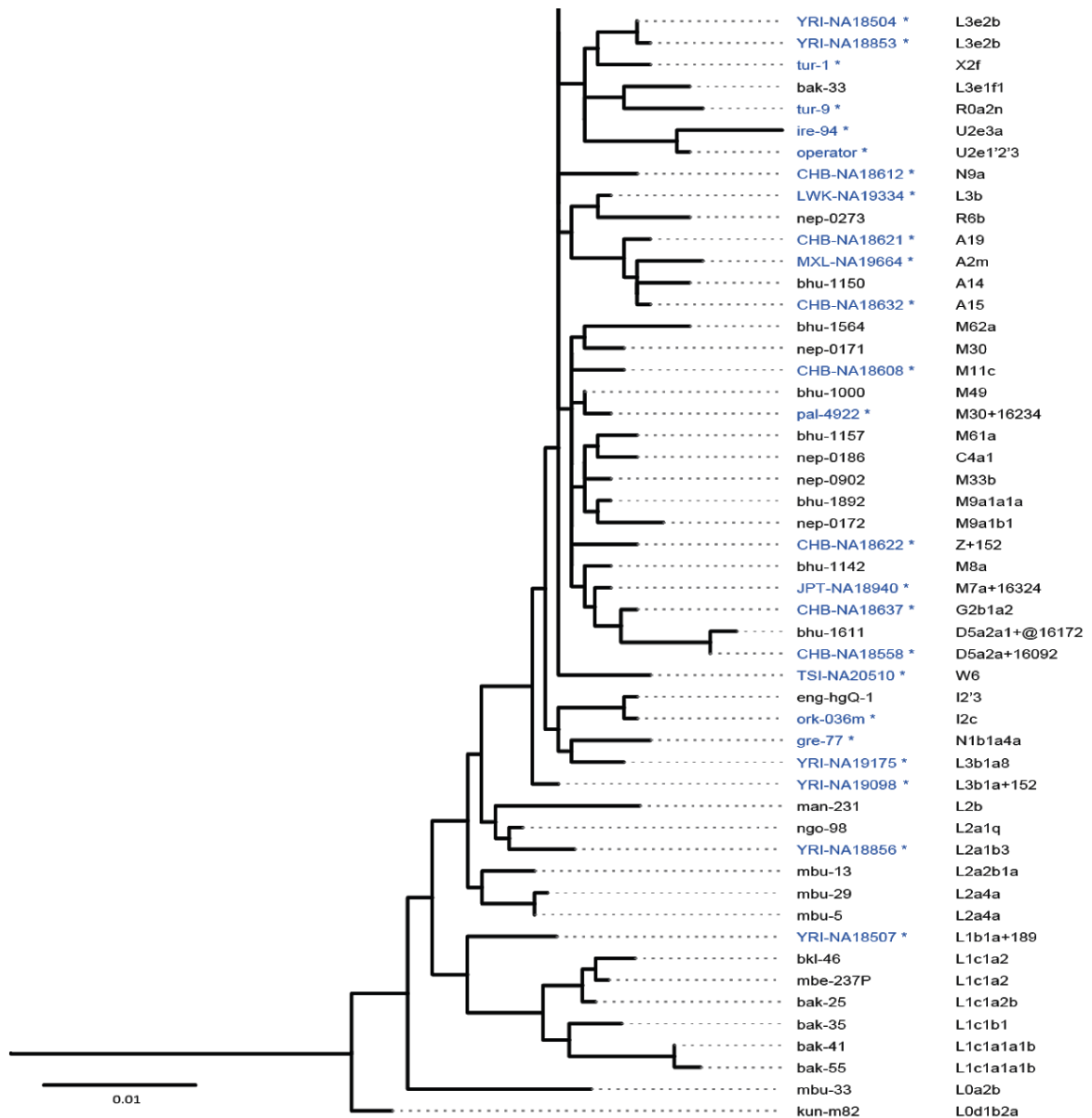


Figure 5.2. Maximum Likelihood phylogenetic tree of the mtDNA control region sequences obtained from 101 samples.

The tree was inferred by using the Maximum Likelihood method based on the Jukes-Cantor model using MEGA6 (Tamura et al. 2013) and edited using FigTree v.1.4.3 (Rambaut 2006-2012). Branch lengths measured in the number of substitutions per site. The tree was rooted to a Pan troglodytes mitochondrial sequence (NC_001643.1), equivalent to the region analysed (from position 15989 to 619), not shown here. Sample names are noted at the tips with haplogroups on the right as predicted by HaploGrep2 (Weissensteiner et al. 2016b). Samples validated against reference datasets are in blue font and marked with an asterisk.

5.3.1 Coverage ranges observed

Promega's prototype PowerSeq™ Auto/Mito/Y System was used to generate MPS data from the mitochondrial control region of each of the 101 samples. The mtDNA-specific components of this kit amplify ten overlapping PCR fragments in the size range of 144 - 237 bp that cover the control region (Figure 5.3).

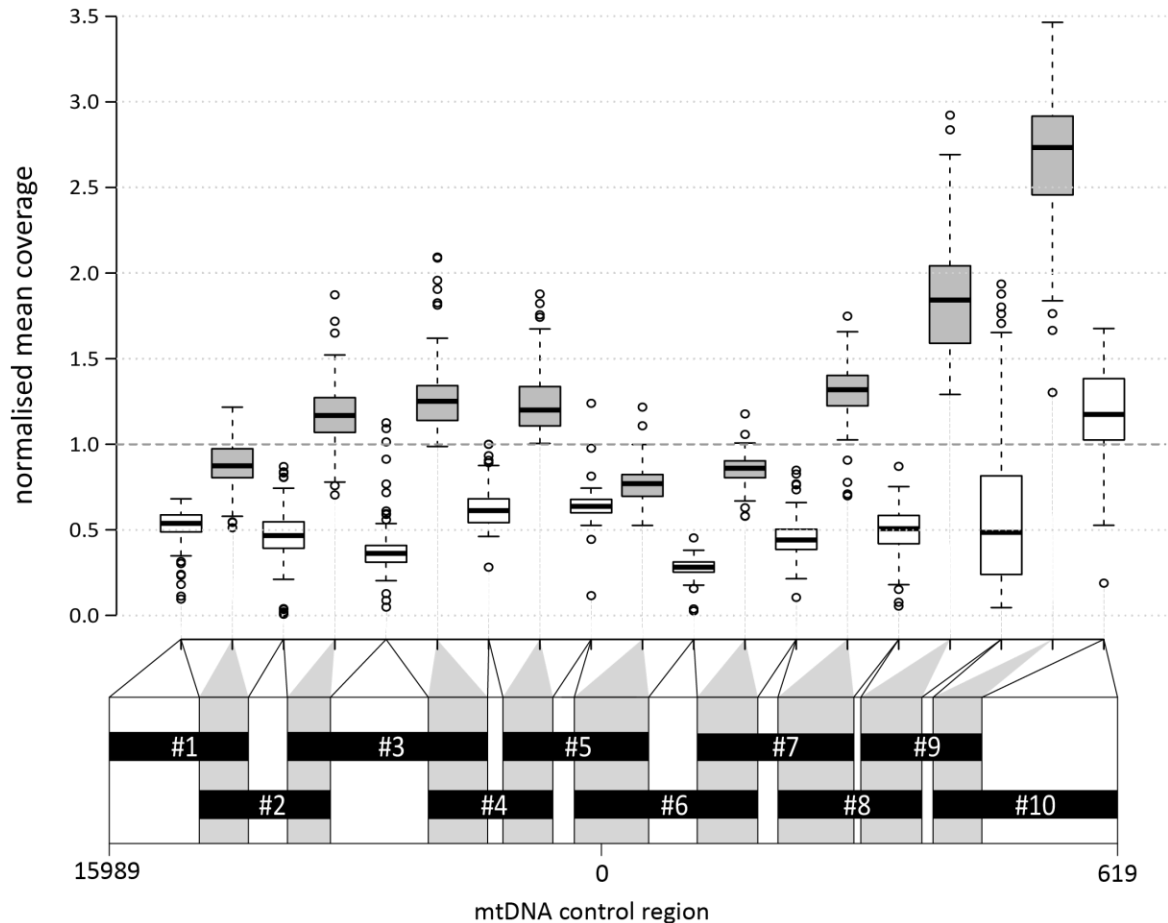


Figure 5.3. Non-uniform coverage of the multiplex assay design over the control region.

The schematic representation shows the ten designed amplicons at the bottom and their overlapping (grey) and non-overlapping (white) segments. Positions and sizes are approximately to scale. The boxplot above shows mean coverage for all 101 samples analysed, for each segment, normalised to sample means. Centre lines indicate the medians; box limits are the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by circles. Notably, the overlapping grey segments have relatively high coverage.

As expected given the design, coverage across the region is non-uniform: where the designed amplicons overlap ('overlapping segments') both amplicons contribute to elevated coverage; this imbalance may be increased further by short PCR products equivalent to the length of the overlap itself (as shown in Chapter 2, Figure 2.5). To reflect this, coverage and read statistics are calculated per segment as shown in Figure 5.3, and are described in Table 5.1.

Table 5.1
Mean coverage statistics for each overlapping and non-overlapping segments of the amplicons.

Values are shown for each multiplexing level (LT (a) and HT(b)) for the prototype kit and for both data processing levels (untrimmed (c) and trimmed (d)) of the CRM Nested kit. Minimum and maximum values are highlighted.

a.

PowerSeq™ Auto/Mito/Y System prototype				
24-plex	Amplicon(s)	Mean coverage	MIN coverage	MAX coverage
	#1	8364.0	3595	12258
	#1/#2	14129.8	8688	20125
	#2	8485.9	4202	14256
	#2/#3	18765.3	12326	26516
	#3	6170.7	3113	9481
	#3/#4	25304.5	20838	32397
	#4	13304.4	10899	15703
	#4/#5	22724.8	19333	28710
	#5	12545.5	9070	16760
	#5/#6	11070.7	8515	15895
	#6	4065.6	2341	6181
	#6/#7	14027.9	10191	22080
	#7	9790.0	6966	15092
	#7/#8	21660.8	12753	36000
	#8	7794.4	4317	10829
	#8/#9	29694.6	22023	42530
	#9	15110.2	1640	31915
	#9/#10	51972.9	35913	71124
	#10	24410.9	9472	32872

Cont.

b.

PowerSeq™ Auto/Mito/Y System prototype

96-plex	Amplicon(s)	Mean coverage	MIN coverage	MAX coverage
	#1	2886.1	472	5593
	#1/#2	4951.8	2208	9415
	#2	2577.3	57	6515
	#2/#3	6656.7	2718	14598
	#3	2184.3	159	6379
	#3/#4	7169.5	2321	16956
	#4	3445.6	943	9514
	#4/#5	6983.6	2301	18456
	#5	3509.1	688	11113
	#5/#6	4362.1	1538	9208
	#6	1555.0	227	3164
	#6/#7	4809.4	1616	9691
	#7	2466.2	756	5299
	#7/#8	7374.8	2715	14412
	#8	2786.9	294	6173
	#8/#9	10552.1	3212	23772
	#9	3259.7	272	14655
	#9/#10	14797.1	5615	27918
	#10	6401.4	815	15257

c.

PowerSeq™ CRM Nested System

untrimmed

59-plex	Amplicon(s)	Mean coverage	MIN coverage	MAX coverage
	#1	12522.3	621	28775
	#1/#2	20545.1	3218	54363
	#2	8763.9	837	26047
	#2/#3	31270.7	2598	73816
	#3	25086.3	2207	68694
	#3/#4	73413.7	20226	247010
	#4	16790.5	2647	46254
	#4/#5	37167.1	9526	121816
	#5	22132.5	3525	63259
	#5/#6	54816.1	15082	158081
	#6	24406.8	5054	38948
	#6/#7	60595.0	27984	98757
	#7	21083.6	1331	36973
	#7/#8	32729.4	12821	53635
	#8	12012.5	304	19912
	#8/#9	29190.8	6968	50367
	#9	22462.7	198	49428
	#9/#10	54892.7	23933	90423
	#10	26811.3	5091	50995

Cont.

d.

PowerSeq™ CRM Nested System
trimmed

59-plex	Amplicon(s)	Mean coverage	MIN coverage	MAX coverage
	#1	12468.9	610	28663
	#1/#2	20039.4	3174	54011
	#2	8267.3	831	25893
	#2/#3	30628.2	2576	70759
	#3	24894.1	2189	68179
	#3/#4	38883.7	6704	107482
	#4	16676.8	2630	45956
	#4/#5	30909.9	7677	90075
	#5	22047.7	3513	63020
	#5/#6	22457.8	5173	41307
	#6	24271.7	5024	38755
	#6/#7	40005.9	16617	63566
	#7	20997.7	1319	36827
	#7/#8	30654.1	11089	51559
	#8	10519.8	290	18467
	#8/#9	27879.6	5946	49763
	#9	22334.4	193	49126
	#9/#10	53847.1	23618	88636
	#10	26286.5	4990	50042

To set the thresholds for accurate heteroplasmy interpretation, first the level of background error was defined using conservative substitution error calculations based on mtDNA reads for all samples for the prototype and the CRM kits as per (Rathbun et al. 2017). The conservative approach meant that any non-major variants (even true heteroplasms) were included in the calculation for each position, thus overestimating the actual error rates, an approach that is suitable when aiming to set the analytical threshold sufficiently high. Furthermore, for the prototype, the same values were also calculated for Amelogenin reads as well, taking advantage of this co-amplified marker, which is not affected by stutter or heteroplasmy, and therefore ideal as an internal control to define the template-type-independent technical rate of substitution errors. Table 5.2 summarises the details of these substitution error rates.

Table 5.2

Substitution error rates calculated for each mtDNA kit type plus the Amelogenin marker for comparison.

Prototype, mtDNA-based substitution error rates		Prototype, Amelogenin-based substitution error rates	
A%	6.86E-02(±4.24E-01)	A%	3.75E-03 (±6.58E-03)
C%	1.99E-01(±1.42E+00)	C%	1.49E-02 (±2.40E-02)
G%	4.35E-02 (±6.34E-01)	G%	7.41E-03 (±1.89E-02)
T%	5.95E-02 (±5.55E-01)	T%	5.73E-03 (±8.25E-03)
Overall%	3.70E-01 (±8.33E-01)	Overall%	3.18E-02 (±1.71E-02)

CRM Nested, mtDNA-based substitution error rates	
A%	7.35E-02(±7.02E-01)
C%	2.43E-01(±1.35E+00)
G%	8.17E-02 (±3.79E-01)
T%	6.93E-02 (±5.80E-01)
Overall%	4.68E-01(±8.09E-01)

With the analytical threshold set to 20 × coverage, for 24-plex and 96-plex sample pooling during library preparations a mean of ~17,700 × and ~5500 × sequence coverage was observed over the control region, with lowest values for non-overlapping segments of 1640 ×, and 57 × respectively. The coverage range observed in the prototype kit for mtDNA-derived reads was suitable for calling SNPs or indels, and allowed the evaluation of heteroplasmic sites down to a conservative level of 10% defined as explained in Table 5.3. At heteroplasmic sites the minor component was also required to meet the 20 × minimum read depth criterion. Figure 5.4 demonstrated that the selected thresholds are set sufficiently above the background substitution error levels.

The results of testing different heteroplasmy calling thresholds are summarised in Table 5.3, for each kit type (the prototype PowerSeq™ Auto/Mito/Y System, and the PowerSeq™ CRM Nested System). This shows the tested scale of coverage requirements for each heteroplasmy detection level considered, together with the consequent loss of bases, and the proportion of samples affected by this loss of ability to call heteroplasmy to the defined levels.

Table 5.3
Testing thresholds for heteroplasmy calling.

For each experiment, applied thresholds for heteroplasmy calling were selected by finding a compromise (highlighted in green) of the lowest heteroplasmy detection levels with the minimum loss of reportable positions due to lower coverage, while considering the achieved coverage values in the experiments.

Prototype kit

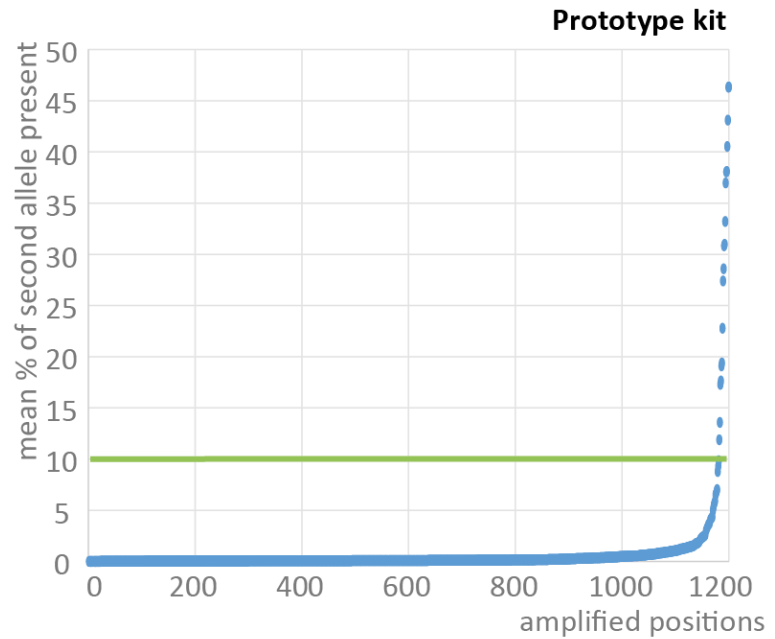
required min coverage values		threshold	bases below threshold		samples affected	
homoplasmic call	heteroplasmy major/minor		#	%	#	%
20	4000+/20					
20	4000/20	0.50%	50258	41.47%	99	98.02%
20	2000/20	1%	16306	13.45%	91	90.10%
20	1000/20	2%	3019	2.49%	39	38.61%
20	667/20	3%	1328	1.10%	27	26.73%
20	500/20	4%	783	0.65%	20	19.80%
20	400/20	5%	636	0.52%	16	15.84%
20	200/20	10%	380	0.31%	8	7.92%

CRM Nested kit

required min coverage values		threshold	bases below threshold		samples affected	
homoplasmic call	heteroplasmy major/minor		#	%	#	%
20	4000+/20					
20	4000/20	0.50%	2110	1.74%	39	68.42%
20	2000/20	1%	879	0.73%	25	43.86%
20	1000/20	2%	373	0.31%	12	21.05%
20	667/20	3%	288	0.24%	8	14.04%
20	500/20	4%	58	0.05%	6	10.53%
20	400/20	5%	37	0.03%	6	10.53%
20	200/20	10%	14	0.01%	4	7.02%

Prototype kit

bins	number of positions in bins
<0.5%	1000
0.5-1%	91
1-2%	52
2-3%	11
3-4%	8
4-5%	6
5-10%	11
>10%	21



CRM Nested kit

bins	number of positions in bins
<0.5%	986
0.5-1%	110
1-2%	40
2-3%	21
3-4%	9
4-5%	3
5-10%	12
>10%	19

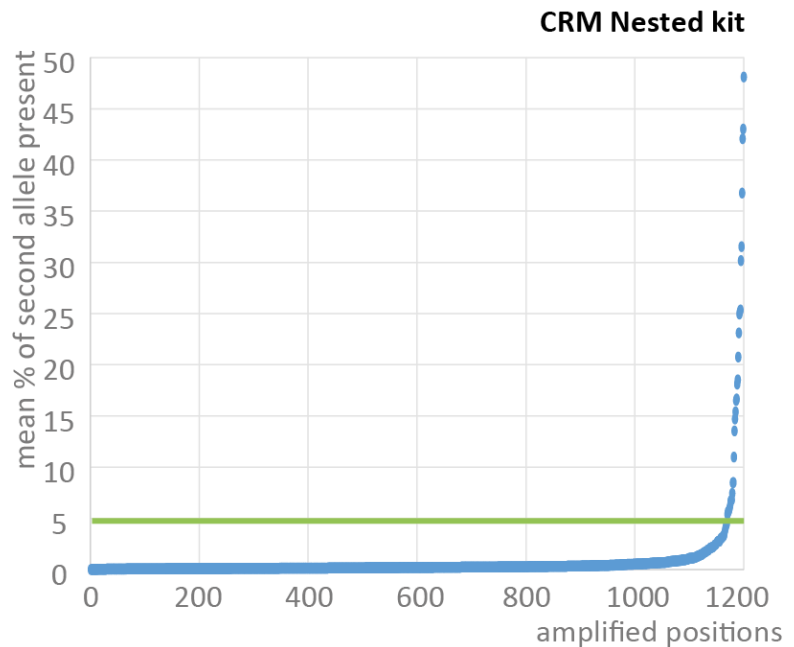


Figure 5.4 Finding a compromise for heteroplasmy reporting thresholds.

Summaries of the number of positions that fall in bins for % of second allele present ('error') for each kit type are shown on the left. The graphs on the right plot the distribution of mean error rates for each position amplified, in increasing % of second allele present. The green horizontal line representing the thresholds set using coverage values of the experiments.

5.3.2 Calling variants in the control region

An attempt was made to call variants using standard approaches (Chapter 2, Section 2.3.3), applying a 5% detection threshold and a set calling threshold of 10% (these limits are comparable to the approximate Sanger sequencing threshold at which homoplasmic and heteroplasmic calls can be distinguished). The achieved coverage values of an experiment limit the ability to lower the calling threshold where coverage values are too low to call heteroplasmies confidently (as shown in Table 5.3), while the threshold also have to be kept above the error rate (Table 5.2 and Figure 5.4). The quality differences between the two kits also required a higher threshold for the prototype; the number of positions above the limit are comparable.

5.3.2.1 Considering primer sequences

Given that a minimum of ten primer pairs are used to amplify a 1.2-kb segment, at least one third of the amplified region may harbour, and thus be affected by, primer-derived sequences. In order to reduce potential bias introduced by primers (Au et al. 2017; Kechin et al. 2017) at the ends of reads during variant detection, different conventional approaches to trimming were tried (tools listed in Chapter 2, Section 2.3.3.6). An attempt to remove 20-26-nucleotide sequences from the read ends, without creating coverage gaps, nonetheless still led to the detection of over 100 cases of apparent heteroplasmy in the samples. This is unexpectedly high compared with available data (Irwin et al. 2009), and therefore suggests that these 'blind' trimming efforts, without the knowledge of exact primer sequences, were still unsuccessful and that heteroplasmy levels were being inflated via the persistence of the unknown primer sequences in the data. Since primers are generally designed based on the reference sequence, this mainly constitutes 'reference sequence bias', but it is also possible that degenerate primers could contribute to the detection of low-level variants.

5.3.2.2 An alternative data processing tool to primer removal

As the removal of primer-derived sequences by trimming was not satisfactory, an alternative data-processing approach was developed to bypass primers without knowing their exact sequences, by seeking reads that spanned primer sites ('Overarching Read Enrichment Option'; OREO; Chapter 2, Figure 2.5). This position-specific approach was applied to variants by designing *in silico* 'probes', sequences 10–30 nucleotides in length (mean: 16 nt), covering the SNP of interest and extending outside the amplicon ends, which were identified by the sharp drops in the coverage track (Chapter 2, Figure 2.5). OREO requires reads to match probes exactly along their complete lengths, and therefore selects reads from the two overlapping amplicons only if those do not end in primer-derived sequences at a tested position. Variants lying in non-overlapping regions are unaffected by primer bias, and therefore do not require correction by OREO. Using one probe for each allele and stringent base-matching allows the specific enrichment and quantification of overarching reads containing reference or alternative SNP alleles. Applying a stringent filtering method is required to clearly distinguish the alleles at the queried position, but also means that reads with random sequencing errors at sites which the probes recognise are also excluded. To minimise this loss, probe length is kept to the minimum that permits selection of overarching reads, while remaining specific. Each probe in a pair has the same length, so any loss is proportional for both alleles. At any queried position the coverage of reads retained after OREO drops to the level of only one of the two amplicons, namely the one which lacks primer-derived sequences at the tested position. Therefore, the coverage loss is position- and sample dependent. Probe sequences, an example of use and the OREO Bash script is provided in text format in Appendix E.

To represent the significant reference bias removed by OREO, here is an example at position 16,234 in sample bak-41, where the alternative variant 'T' was detected in only 38% (2966/7824) of raw reads, but after OREO processing this increased to 97% (2909/2998) - and hence the variant 'T' was corrected from a minor heteroplasmy allele to technically a SNP under the applied thresholds. Thus,

applying OREO with the prototype kit identified SNPs (i.e. variants in >90% of the reads, using the 10% threshold for heteroplasmy calling) at 161 of the 1200 positions amplified, defining a total of 101 distinct and diverse mtDNA haplotypes. Prediction of haplogroups from sequences confirmed that the sample set includes most major clades of the mtDNA phylogeny (Figure 5.2). Since the samples are not representative of populations, no population statistics were applied to the results.

5.3.3 Validation of PowerSeq™ MPS mtDNA data

Sequences derived following OREO adjustment were validated for 65 samples for which previous data were available, considering base substitutions lying outside the homopolymeric tracts or the AC-repeat region. Discrepancies were observed in 5/65 samples (Table 5.4), and in all cases these were variants observed in the data generated here that were absent from a reference data source (i.e., potential false positives). For two of these samples, comparative data were available from two sources (1000 Genomes Project Consortium 2015; Batini et al. 2017), one of which agreed with the data generated here. For the remaining three samples (involving six variant sites in total) the discrepancies cannot be resolved. In conclusion, conservatively treating these six sites as errors in the current data, this would yield a false-positive rate of $6/480=1.25\%$; however, it is worth noting that coverage over the discrepant sites in the current data is much higher than in the comparative datasets, even after using OREO to select overarching reads when variants lie in overlapping segments (Table 5.4). No base substitutions observed in the comparative datasets were missed here.

Table 5.4
Discrepancies between the called homoplasmic substitutions and the reference datasets.

Sample	Haplogroup	Variant called and mean coverage	post-OREO variant%, coverage of position	variant in Batini et al. (2017)	variant in 1000 Genomes
CEU-NA12003	T2b	16304C 3315 x	not in overlap	16304C 84 x	NC
CHB-NA18608	M11c	16223T 3815 x	overlap/primer 99.7%, 1872 x	NC 97 x	16223T
gre-12	J1c8	185A 7902 x	not in overlap	185N 13 x	-
		228A 7902 x	overlap/primer 99.7%, 3078 x	NC 13 x	-
		263G 7902 x	overlap/primer 99.9%, 5417 x	NC 13 x	-
		295T 7902 x	not in overlap	NC 13 x	-
		295T 5494 x	not in overlap	NC 67 x	-
ire-0068	J1c3e	16519C 5288 x	not in overlap	NC 22 x	-

NC: not called as different from reference.

5.3.4 Performance of mtDNA amplification across the phylogeny

As previously noted, the diverse sample set studied here (Figure 5.2) is suitable for testing the efficiency of MPS multiplexing in a wide variety of control region sequences. SNPs within primer-binding sites can affect PCR efficiency for individual amplicons, and thus the confidence of variant calling. This could be detected by observing particularly low coverage for an amplicon that is generally well covered in the dataset. As an example, the non-overlapping segment of amplicon #3 (Figure 5.1; also in more detail later in Section 5.3.7; positions 16,248–16,363) in sample tur-16 showed a mean coverage of 159 x, compared with a mean of 2207 x for the same segment in all other samples with the same multiplexing level. The low value of coverage for this segment was also in contrast to the mean coverage of 2046 x for other segments within the same sample. This particular sample belongs to haplogroup HV2a1, and carries two SNPs (16214T and 16217C), which potentially lie under a primer-binding site for amplicon #3. This

example illustrates the fact that sequence variants can affect amplification efficiency for specific amplicons. However, this example involves only one of the 1010 targeted amplicons generated from the sample set, and even in this case the two neighbouring SNP variants decreased, but did not eliminate amplification of the mtDNA template in this region, and therefore the overall performance of the kit in these diverse mtDNAs is robust.

5.3.5 Detection and quantitation of heteroplasmy

Given the issues with reference sequence bias and the difficulty of trimming primer-derived sequences from reads as described above in Sections 5.3.2.1 and 5.3.2.2, it was of interest to ask if the single-reaction multiplex was able to reliably detect and quantify heteroplasmy. For apparent heteroplasmic sites, mixture between different samples could provide a trivial explanation; to address this, the fact that this study used the prototype multiplex kit PowerSeq™ Auto/Mito/Y System (Promega) (Huszar et al. 2018) was exploited, since it could provide evidence of mixtures through examination of autosomal and Y-STR profiles. In the samples with confirmed heteroplasmies, no unexpected mixed STR profiles were observed down to a 1% level of the reads (Chapter 3 and Chapter 4) that could indicate contamination. After variant calling and data processing via OREO, potentially heteroplasmic sites (with both reference and alternative variants) were analysed. Traditionally, based on Sanger sequencing, the levels of heteroplasmy at such sites are described by the minor allele frequency (MAF; $\leq 50\%$) down to a technology-dependent threshold below which heteroplasmy cannot be reliably identified. Here, however, since MPS relies on a reference sequence mapping approach, variants were identified based on alternative allele frequency in which an alternative variant is observed in 10–90% of the reads at a site (corresponding to 10% MAF threshold). When an alternative variant at a given site is present in $<10\%$ or $>90\%$ of the reads, the site is called as homoplasmic (reference or alternative allele, respectively). The 10% MAF threshold represents a similar resolution of heteroplasmy to that of the classical Sanger method (Irwin et al. 2009).

After accounting for the primer-derived sequences at the end of the reads 45 apparent heteroplasmic substitutions were identified, at which the minor allele is either the reference or the alternative variant. Such sites were observed in 39 samples at 33 positions, and each of these samples contained between one and three variants. Two samples were also identified that contained the same single-nucleotide insertion (44.1C) in a heteroplasmic state (showing alternative variant proportions of 69% and 77% in overarching reads). To investigate the status of these apparent heteroplasmic sites, their locations within the control region were considered (Figure 5.5). It is clear from Figure 5.5 that the detected apparently heteroplasmic sites cluster in likely primer-binding sites: thus, despite efforts to remove reference bias by enriching for overarching reads, some bias remained.

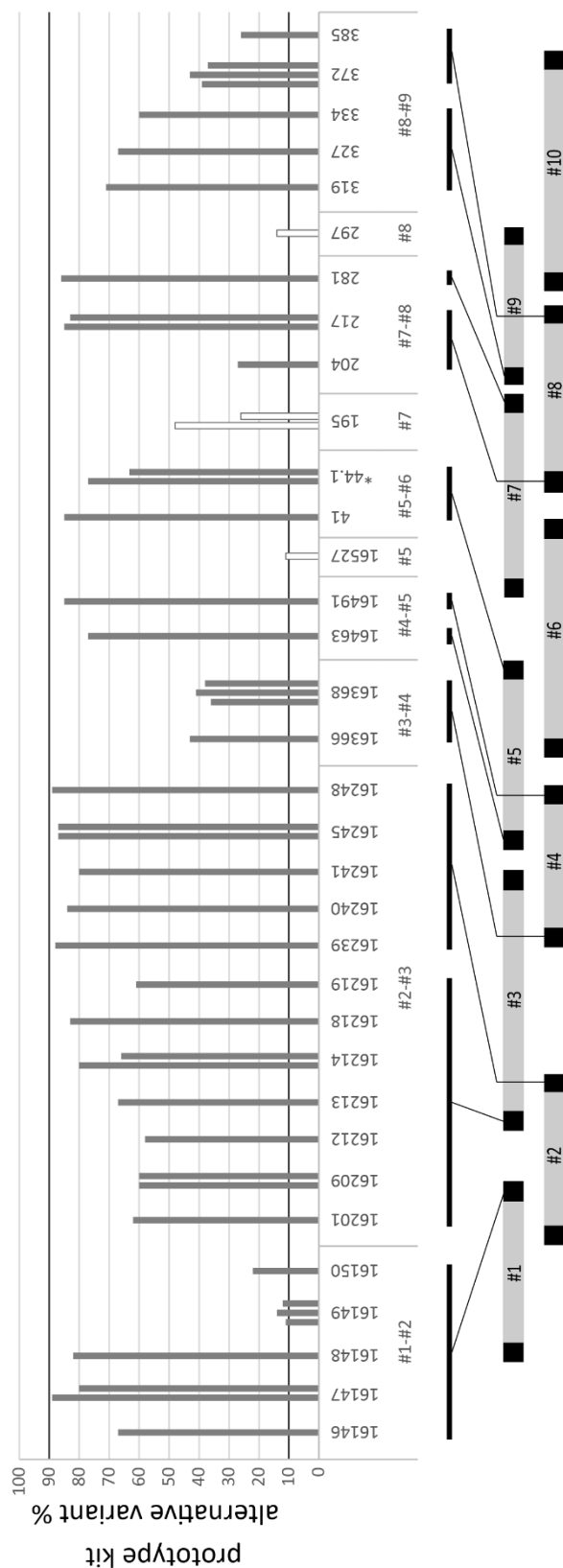


Figure 5.5
Distribution of apparent heteroplasms across the control region.

Percentages of heteroplasmy (as alternative variant percentage, from 10-90%) are shown in the bar charts for each apparently heteroplasmic variant after OREO processing, and arranged by position along the control region. Positions in amplicons are indicated, using the nomenclature of Figure 5.1. The asterisk indicates a site displaying single-base indels - all others are base-substitutions. Black horizontal bars underneath positions indicate probable primer sequences, also represented as black boxes on the schematic amplicons. White bars in the bar charts indicate positions outside probable primer regions. Forty-seven apparent heteroplasms detected using the prototype kit; notably, only four of these survive the conservative criteria of taking a 10% threshold, and not calling sites under potential primer positions.

The OREO approach filters out reads containing likely primer sequences at their ends, so the remaining bias logically must be due to primer sequences that are internal to the reads, most likely via overlap extension (Thornton 2016) - the annealing of overlapping single-stranded PCR products, which can prime synthesis from the 3' end of an already incorporated primer-derived sequence (Figure 5.6).

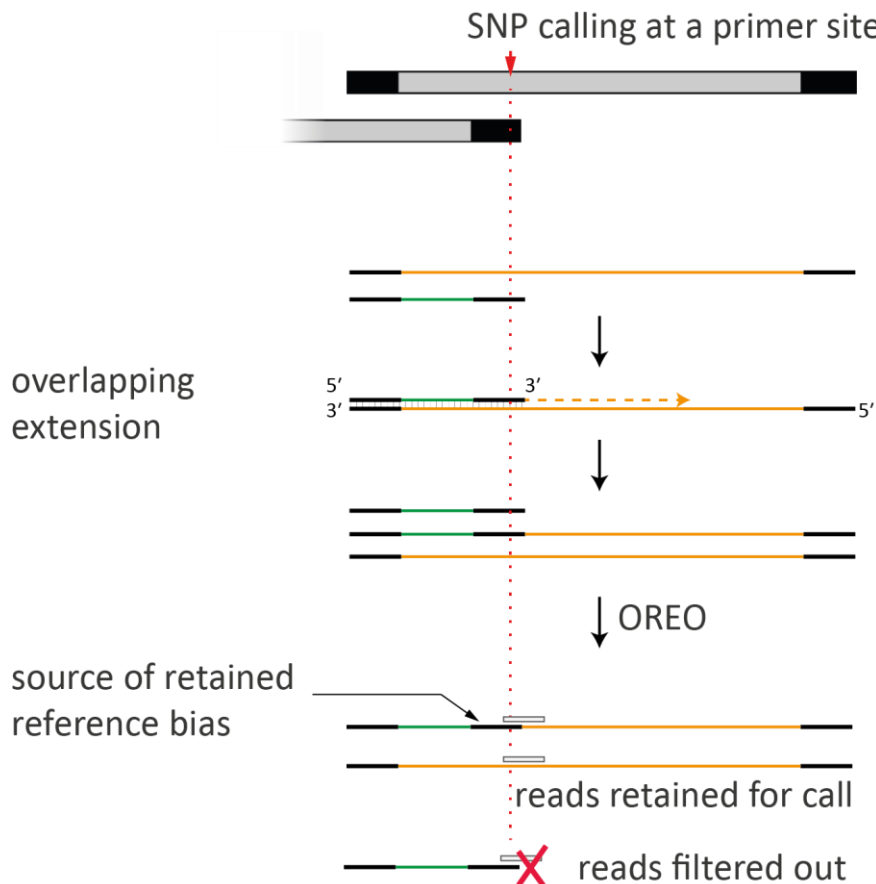


Figure 5.6
Overlap extension of PCR products.

Annealing and extension of overlapping single-stranded PCR products introduces reference sequence bias internal to the reads. Shorter amplicons (green horizontal bars), similarly to primers (black boxes and black thicker bars), can anneal to the designed amplicons (orange horizontal bars) when single-stranded during amplification, and generate products that contain primer-derived sequences internal to the reads. OREO uses probes (white boxes) to filter out reads that do not span over the primer site to reduce reference bias from primers, however the hybrid products of overlapping extension escape this process, and therefore carry over reference bias into the retained reads.

Therefore, heteroplasmy calling is not reliable in regions of likely primer binding sites, even at a conservative level of 10%, and therefore must be omitted. Applying this conservative criterion risks losing some genuine heteroplasmy, and reduces the number of apparent heteroplasmic calls from a total of 47 to four (the white bars in Figure 5.5). Of these, two are at position 195, a well-known heteroplasmy-prone site (Irwin et al. 2009).

Apart from single-nucleotide variants, simple-sequence regions (e.g. homopolymer tracts) were also analysed, including quantitation of length heteroplasmy. Within these regions, 284 instances of variants were found. Each sample contained between one and four such variants (Appendix E shows an example of the principle for estimating length polymorphisms).

To assess the potential ability of other readily available approaches to mitigate the effect of primer-derived sequences in variant and heteroplasmy detection, two tools were also tested that were designed for the analysis of mtDNA MPS data. The online tool mtDNA-Server (Weissensteiner et al. 2016a) acknowledges the problem with overlapping amplicons and heteroplasmy detection, but does not offer an option to correct for the presence of primers in reads. It gave a very high number of heteroplasmic calls (data not shown), comparable to the data obtained here prior to correction with OREO. The trial version of the commercial software GeneMarker[®]HTS v1.2.2.1338 (SoftGenetics[®], LLC) contains the option to input amplicon coordinates, but it is unclear how this information affects data processing; when comparing outputs with and without this option, both gave comparably high numbers of heteroplasmic calls (data not shown), as in the uncorrected data here. Since neither of these tools were aiming to correct for the primer-derived bias that was observed here, OREO remains the most successful tool for decreasing, if not completely removing, bias from the data.

5.3.6 Improved heteroplasmy detection with the CRM Nested System

The improved format of the redesigned PowerSeq™ CRM Nested System multiplexes only mitochondrial amplicons, streamlines the library preparation process and minimises sample loss by using a nested amplification protocol with a single-step PCR including both amplification and incorporation of indexed sequencing adapters. A subset of 57 samples that showed potential heteroplasmy with the prototype kit was analysed using the CRM Nested kit to assess how this improved approach affects the accuracy of detection of heteroplasmic sites. The CRM Nested design not only provides potentially higher coverage per sample, but also has the advantage of preventing the internalisation of the primer sequences (compare Figure 5.6 and Figure 5.7) because overlapping extension from the single-stranded short amplicons is now prevented by the flanking adapter sequences.

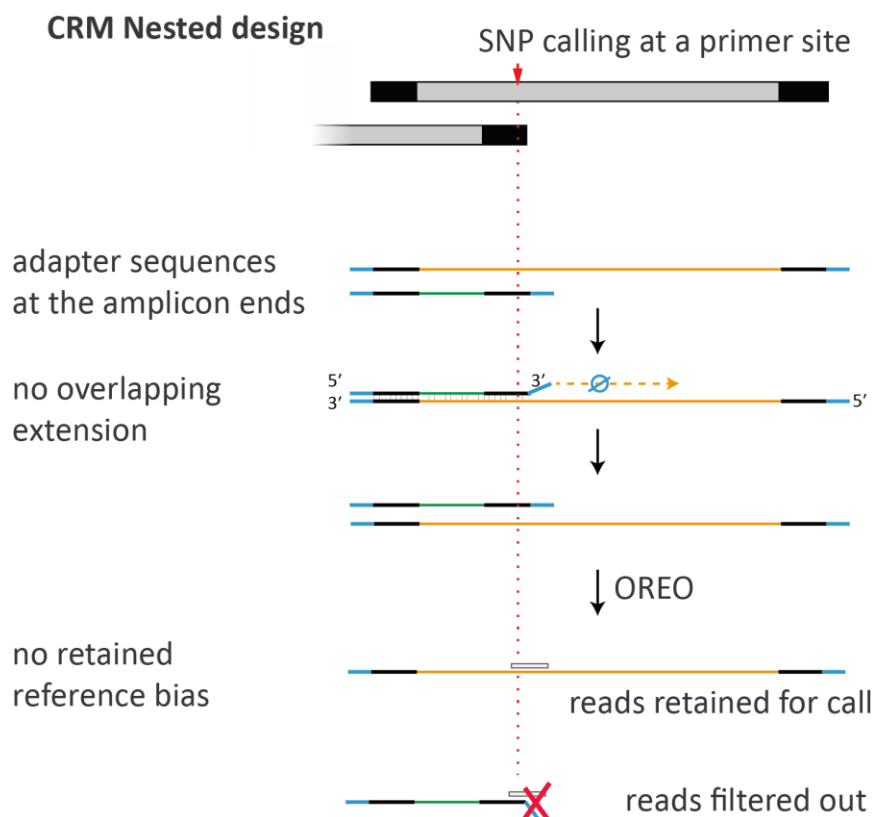


Figure 5.7
The Nested CRM design prevents extension of overlapping single-stranded PCR products and internalisation of primers.

In the CRM Nested design the tailing adapters (blue bars) at the end of the amplicons prevent overlapping extension and internalisation of primer-derived sequences. The remaining primer-derived sequences are successfully filtered out by OREO.

The 47 heteroplasmic calls observed using the prototype kit were re-examined; just using the CRM design resolved nine of these calls even in the raw data as no longer heteroplasmic. These were found to have three reference and six alternative alleles (details shown in Table 5.5), highlighting the effect of previous primer internalisation by overlap extension.

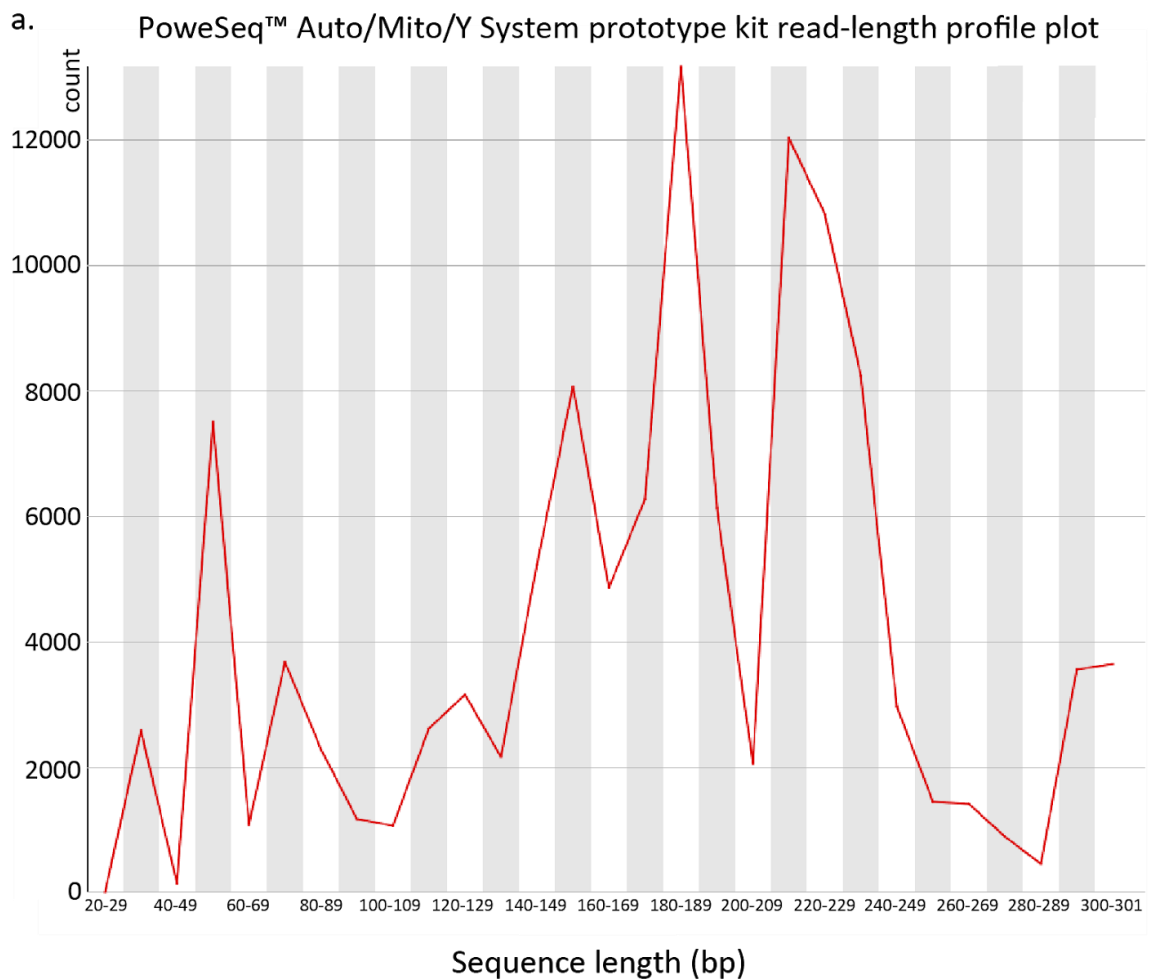
		thresholds: >10% and <90%		thresholds: >5% and <95%		
sample	alternative variant	prototype		CRM Nested		
		Alternative Variant %		Alternative Variant %		
		primer	pType	primer	CRM	CRM + <95 trim + OREO
eng-GB1778	16146G		29.10%		21.44%	15.57%
bhu-1564	16147T		54.49%		83.40%	82.00%
pal-4919	16147T		43.92%		81.50%	77.64%
mbu-33	16148T		3.20%		51.00%	48.46%
YRI-NA18504	16149C		5.09%		0.14%	0.15%
bhu-1611	16149C		6.08%		0.14%	0.16%
CHB-NA18558	16149C		6.80%		0.14%	0.16%
hun-5	16150T		11.10%		11.98%	11.26%
TSI-NA20543	16201T		32.04%		37.46%	36.10%
JPT-NA18940	16209C		24.67%		43.09%	42.16%
CHB-NA18637	16209C		18.34%		39.65%	39.13%
ser-21	16212G		23.63%		46.20%	40.92%
man-231	16213A		23.65%		59.10%	55.17%
bak-41	16214T		26.22%		40.78%	25.11%
bak-55	16214T		22.58%		38.95%	20.46%
bhu-0957	16218T		43.77%		68.76%	65.17%
pal-4919	16219G		23.56%		65.39%	60.40%
bhu-1151	16239T		22.95%		95.57%	95.00%
MXL-NA19664	16240G		11.97%		78.50%	77.97%
bak-35	16241G		22.66%		76.56%	75.24%
ngo-98	16245T		33.84%		91.64%	92.82%
nep-0273	16245T		24.36%		94.07%	93.64%
eng-hgQ-2	16248T		23.01%		77.57%	76.78%
bav-55	16366T		12.38%		77.83%	93.27%
TSI-NA20527	16368C		14.40%		41.34%	71.71%
spa-20	16368C		13.69%		51.31%	74.32%
pal-4929	16368C		13.22%		38.85%	72.01%
bhu-0984	16463G		34.72%		63.30%	64.06%
bhu-1564	16491C		50.33%		96.96%	97.45%
aus-m119	16527T		11.20%		9.14%	9.17%
tur-4	41T		24.26%		99.65%	99.74%
bak-41	44.1C		15.23%		44.71%	71.59%
bak-55	44.1C		20.75%		50.40%	77.36%
bkl-46	195C		48.40%		46.31%	46.31%
eng-hgQ-2	195C		26.01%		27.86%	27.91%
gre-2	204C		6.72%		22.28%	22.30%
ire-94	217C		32.79%		99.78%	99.78%
operator control	217C		21.91%		99.37%	99.50%
YRI-NA19175	281G		51.14%		64.66%	61.22%
bhu-1564	297C		18.73%		14.73%	14.88%
tur-4	319C		7.86%		48.53%	46.10%
ork-007	327T		10.76%		98.03%	98.21%
pal-5232	334C		7.21%		47.13%	44.03%
YRI-NA19098	372C		21.30%		74.00%	76.14%
tur-9	372C		22.73%		79.01%	84.70%
bak-33	372C		21.37%		91.36%	92.28%
pal-4919	385G		13.82%		60.06%	63.60%

code: homoplasmy: potential confirmed
heteroplasmy: potential confirmed

Table 5.5
Effects of data processing steps on 47 heteroplasmic variants.

Alternative variant % is shown for both the prototype (pType) and the new CRM Nested (CRM) design. '<95 trim' denotes the trimming off of short amplicons. Grey bars mark whether the variant positions are likely primer sites, therefore likely to be affected by the bias. (Forward primer site for amplicon #8 was slightly changed from position 203 to 218 from the prototype to the commercial form of the kit.)

While samples were subject to bead-based size selection during library preparation as per the manufacturer's protocol, nevertheless a large proportion of the shorter (52 – 89-bp) amplicons (explained in Chapter 2, Figure 2.5) survived the process and therefore further contributed to the non-uniform coverage. Since the new design generates only mitochondrial amplicons, it gives a better defined read-length profile compared to the prototype kit (compare Figures 5.8a and b), which contained amplicons from variable-length STRs as well.



Cont.

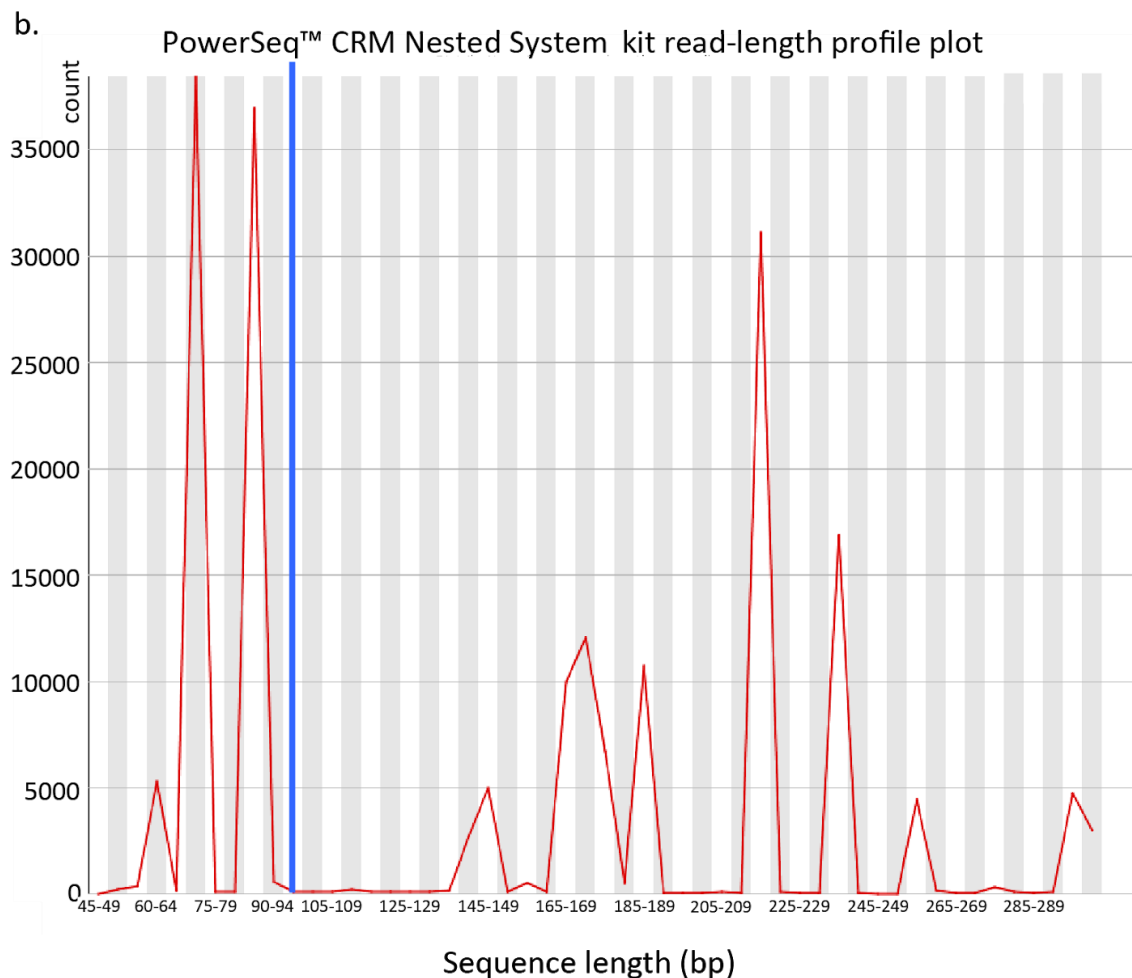
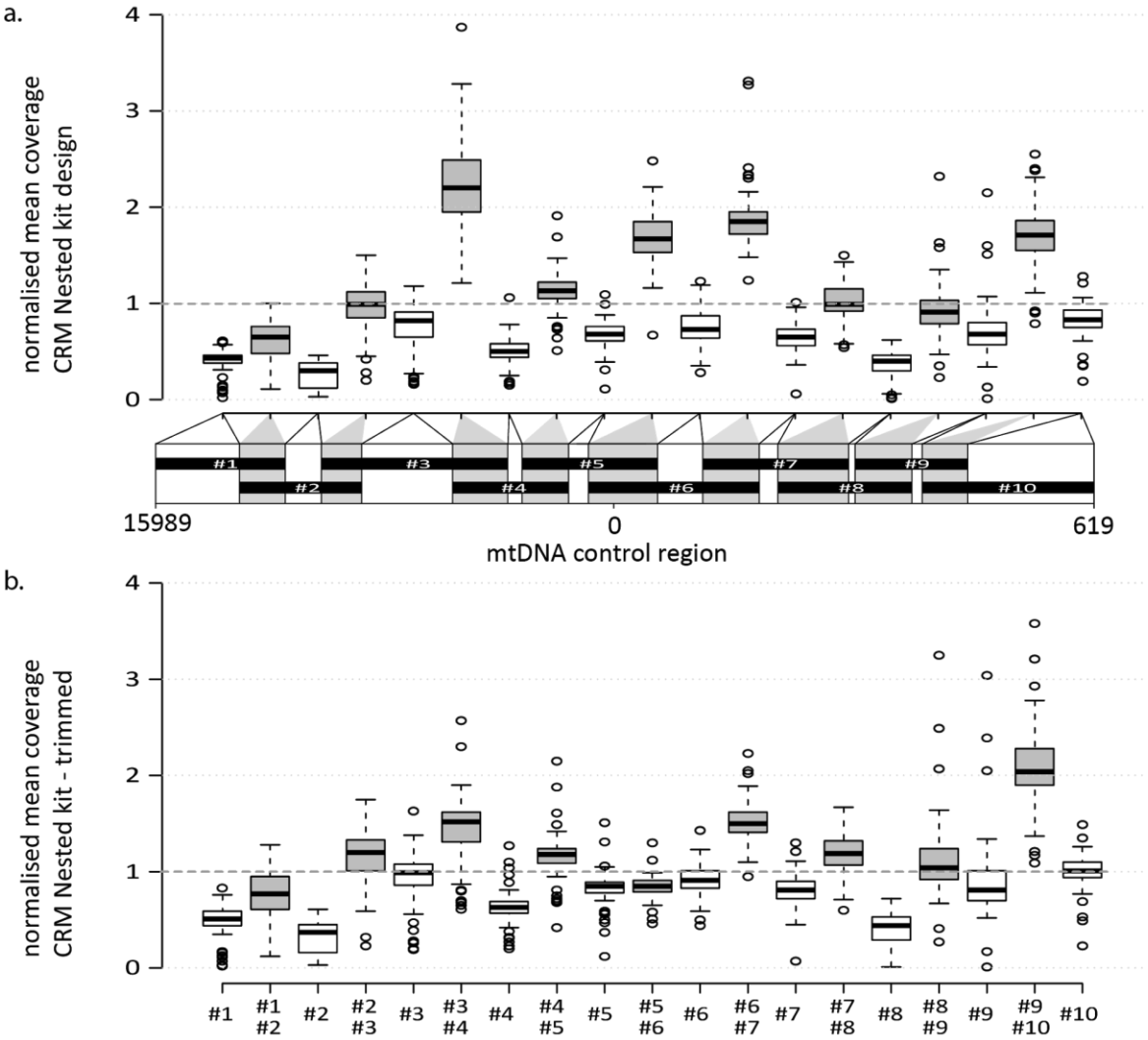


Figure 5.8
Read-length profiles compared for the prototype and the CRM kit.

Compared with the prototype kit (a), the updated CRM design (b) allows the precise removal of reads generated from the short amplicons from the QC-ed FASTQ files using bioinformatic trimming tools. Here, after confirming the ideal threshold via the FastQC software read-length profile function, Trimmomatic was used to remove reads shorter than 95 bases – represented by the blue vertical line in (b) – from sample CHB-NA18608.

This makes it possible to apply *in silico* size selection to the generated data from the CRM kit to minimise the effect of potentially interfering short amplicons. After careful examination of the read-length profile plot for each sample generated by FastQC software (Andrews 2010) a cut-off of 95 nt was defined to remove reads generated from these shorter amplicons (the green reads in Chapter 2, Figure 2.5; Figure 5.6 and Figure 5.7) using Trimmomatic v0.32 software (Bolger et al. 2014).

The generation of these short amplicons is not uniform among overlapping segments, and the removal of these mostly improves the regions of overlap at positions 16,533-52, 16,363-16,509, 109-180 and 16,450-16,509 (Figure 5.9 a,b and c), where they are shown to significantly contribute to the coverage observed in these regions.



Cont.

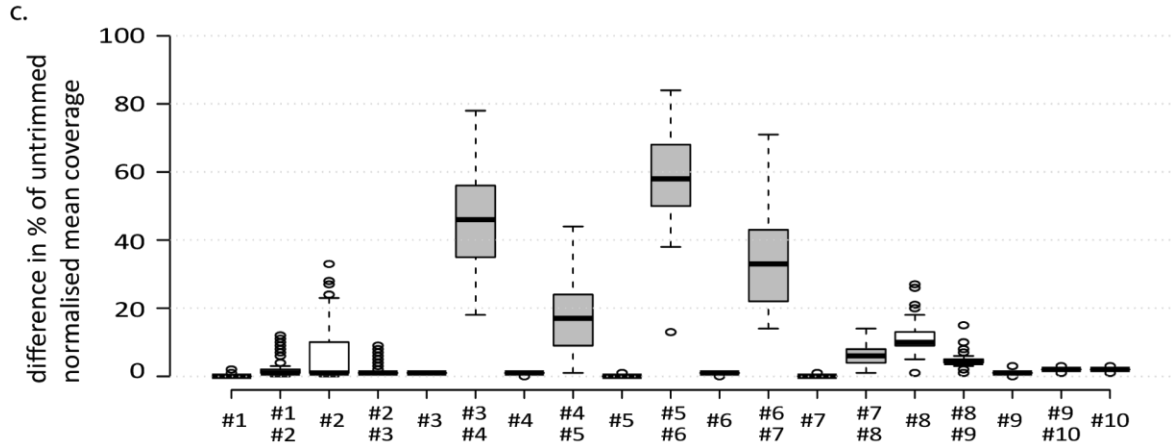


Figure 5.9
Contribution of short amplicons to non-uniform coverage over the control region.

The figure follows Figure 5.3, as described there. Part (a) shows the normalised mean coverage values for the PowerSeq™ CRM Nested kit design before trimming off the short amplicons, (b) shows it after trimming these off. Part (c) plots the percentage of trimmed off reads per region, showing that short amplicons are formed in a non-uniform fashion.

After removing the short amplicons, the remaining 38 calls were still apparently heteroplasmic, although some of the allele proportions changed (in Table 5.5 - column 'CRM+<95trim'). Following the removal of the short reads, OREO was applied as an alternative to primer trimming to bypass primer sites for calling variants. Applying OREO clarified all the 47 heteroplasmic calls (in Table 5.5 - column 'CRM+<95trim+OREO'). Six sites were shown to present clear heteroplasmy, one of which (16150T at 22% in sample hun-5) was successfully identified despite lying under a primer site (Figure 5.10); the remaining 41 occurrences were resolved either as reference (three) or alternative variant (38) alleles. These 38 SNP calls clearly demonstrate the high proportion of reference bias conveyed by the presence of primer sequences in the reads, and the importance of accounting for their presence in the dataset. Comparisons of the effects of the different kits and data processing steps on the 47 initially flagged potential heteroplasmic calls are summarised in Table 5.5.

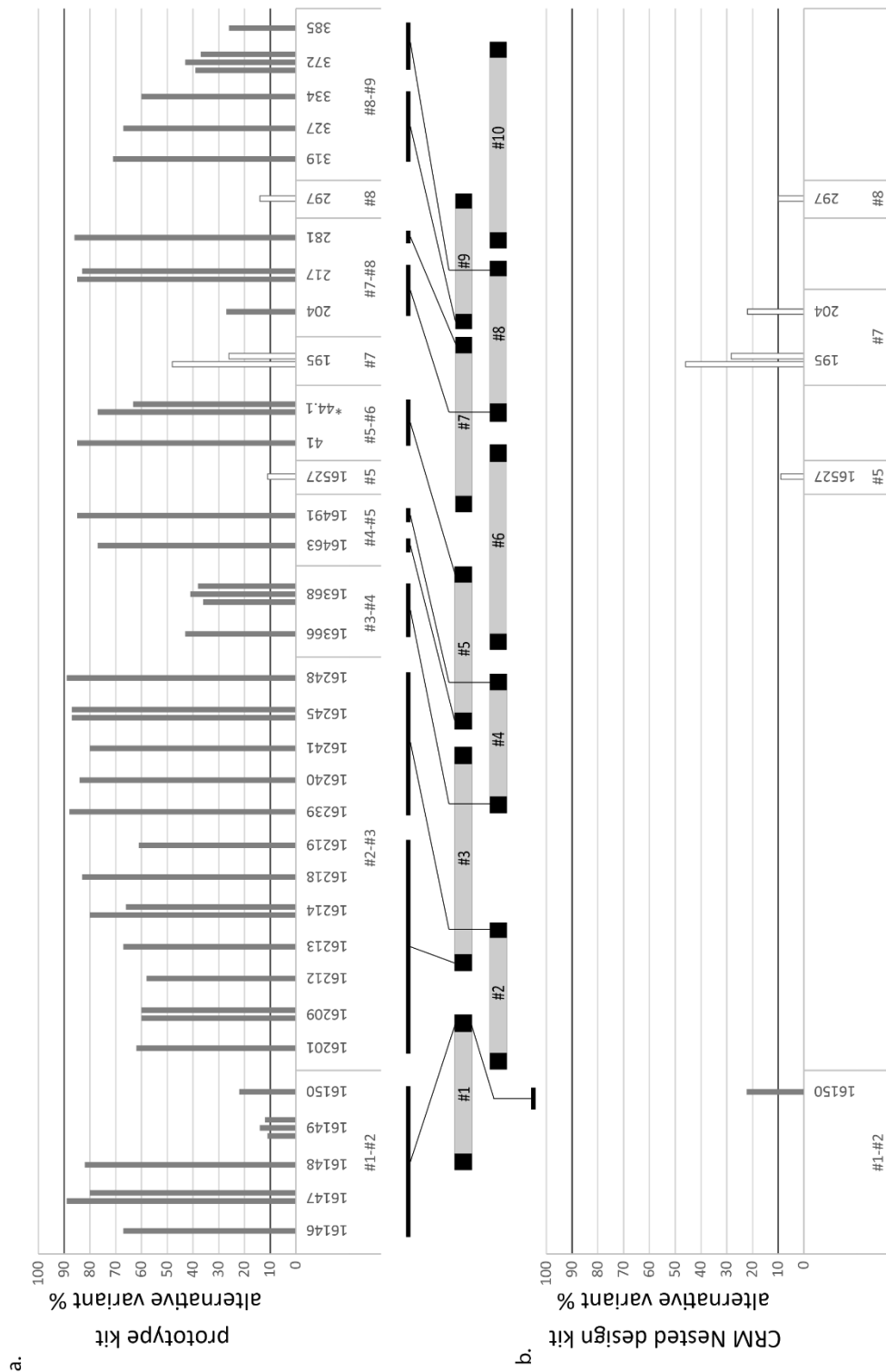


Figure 5.10
Distribution of confirmed heteroplasms across the control region.

The top 'a' follows Figure 5.5, and supplemented with 'b' the confirmed heteroplasms using the improved CRM kit design after data processing (trimming and the application of OREO). Six confirmed heteroplasms were detected; notably, variant at position 16,150 lies under a primer site, and variant at position 204 lies under a primer-binding site in the prototype, but not in the CRM Nested kit design.

The nested approach of the CRM kit prevented the internalisation of primers which had been identified as the likely source of the retained reference bias observed in the prototype. This suggested that the direct trimming of primers from read ends would be more effective with the CRM design; however, if the trimmed length is too short, primer bias will be retained, while if it is too long, coverage gaps will be introduced. In the experiment here, biased SNPs were observed up to 27 bp from the ends of amplicons, suggesting that at least 27 nt should be trimmed from reads to remove primer bias. However, such trimming creates a small gap in the overlap region of amplicon #2 and #3 covering variable positions 16,222 and 16,223, thus introducing false negative calls. Overall, although the conventional trimming approach has the virtue of simplicity, this comes at the cost of introducing coverage gaps, a problem that does not apply to OREO.

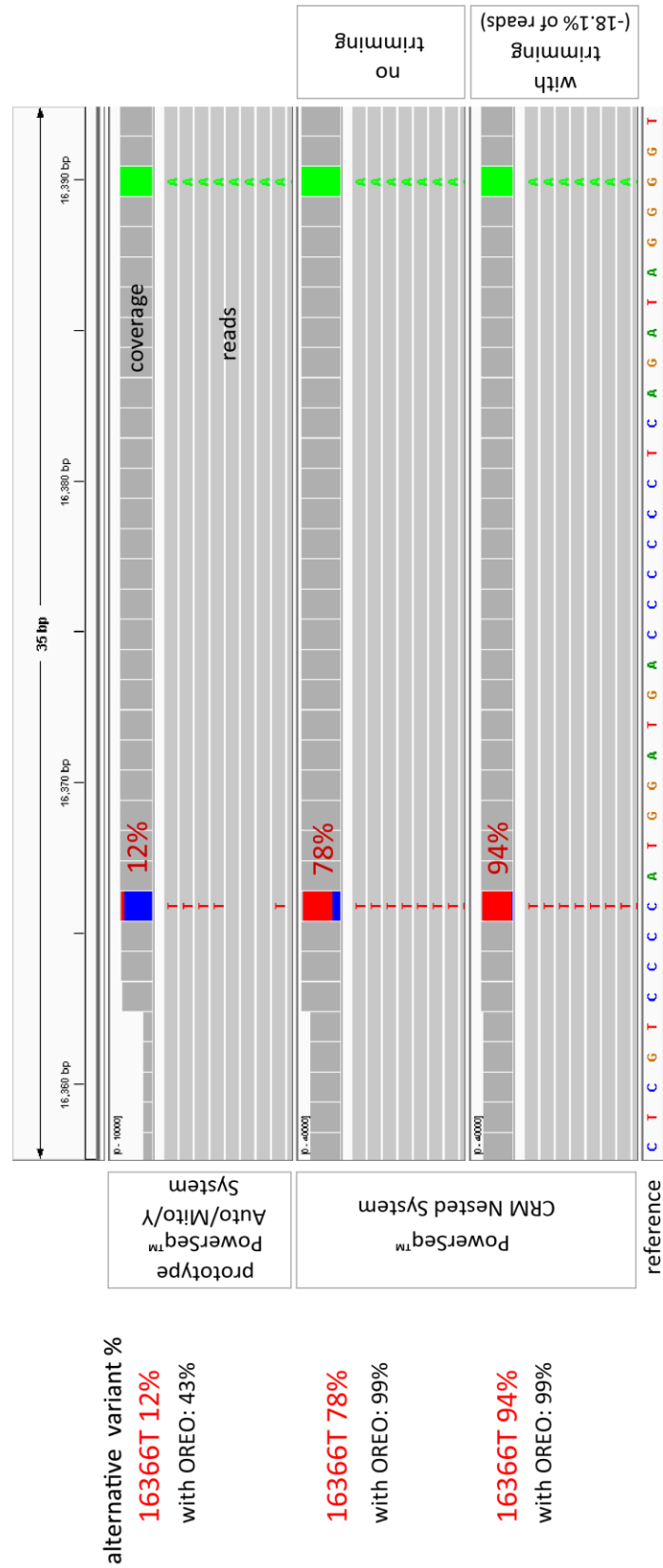


Figure 5.11
Example of the effects of kit design and data processing in mitigating reference sequence bias.

In sample bav-55 at position 16,366 an alternative variant 16366T is observed in different proportions. This position is three bp away from the amplicon/read end, therefore reasonable to consider it being under a primer site. Variant 16390A is clearly not affected, as being further away from the amplicon/read end, thus not showing primer-derived bias. Top track shows data from the prototype kit, lower two tracks show data from the new CRM Nested kit, of which bottom track displays data trimmed of short amplicon derived reads. OREO data processing is shown to improve all three data sets.

Observing the effectiveness of the CRM kit design when followed by appropriate data processing steps (*in silico* size selection and OREO, Table 5.5, Figure 5.11), and permitted by the higher coverage obtained, the heteroplasmy calling threshold was lowered from 10% to 5% (Table 5.3). Applying the 5% threshold flagged 102 sites which were compared between the two kits after data processing, and are shown in Table 5.6; this identified a further six heteroplasmic sites.

Table 5.6
Effects of data processing steps on 102 heteroplasmic variants.

All apparent heteroplasmic variants are listed lowered to the 5% detection level. - Alternative variant %s are shown for both the prototype (pType) and the new CRM Nested (CRM) design. '<95 trim' denotes the trimming off of short amplicons. Grey bars mark whether the variant positions are likely primer sites, and therefore likely to be affected by the bias. (Forward primer site for amplicon #8 was slightly changed from position 203 to 218 from the prototype to the commercial form of the kit.)

		thresholds: >10% and <90%		thresholds: >5% and <95%	
sample	alternative variant	Prototype		CRM Nested	
		Alternative Variant %		Alternative Variant %	
		primer	pType + OREO	primer	CRM + <95 + OREO
pal-5225	16069T		94.17%		87.48%
den-190	16093C		95.99%		97.24%
ser-21	16093C		93.78%		94.80%
tur-9	16093C		94.49%		95.92%
pal-5225	16126C		93.85%		98.53%
den-190	16129A		7.80%		7.71%
bhu-1151	16145A		94.85%		96.83%
pal-5225	16145A		94.33%		99.18%
eng-GB1778	16146G		66.60%		98.84%
bhu-1564	16147T		88.55%		99.66%
pal-4919	16147T		79.80%		99.74%
mbu-33	16148T		82.22%		100.00%
TSI-NA20543	16148T		90.27%		99.92%

Cont.

sample	alternative variant	Prototype		CRM Nested	
		Alternative Variant %		Alternative Variant %	
		primer	pType + OREO	primer	CRM + <95 + OREO
YRI-NA18504	16149C		12.14%		1.98%
bhu-1611	16149C		11.49%		2.96%
CHB-NA18558	16149C		14.03%		3.21%
hun-5	16150T		22.21%		22.26%
bak-25	16153A		5.36%		0.04%
bak-55	16153A		6.94%		0.00%
CEU-NA11992	16153A		7.03%		0.04%
CHB-NA18561	16153A		8.09%		0.10%
ire-0114	16153A		6.53%		0.23%
JPT-NA18974	16153A		5.49%		0.00%
nep-0172	16153A		4.81%		0.00%
nep-0273	16153A		5.10%		0.00%
ork-007	16153A		4.81%		0.05%
pal-4922	16153A		7.40%		0.00%
pal-5232	16153A		6.98%		0.09%
TSI-NA20510	16153A		6.74%		0.04%
bak-33	16175C		5.39%		1.58%
bhu-1611	16175C		6.71%		1.75%
CHB-NA18558	16175C		9.57%		1.39%
CHB-NA18561	16175C		6.85%		0.21%
gre-10	16175C		5.25%		0.30%
hun-29	16175C		5.58%		1.74%
operator control	16175C		6.58%		1.74%
tur-9	16175C		5.24%		1.66%
YRI-NA18504	16175C		7.19%		1.79%
TSI-NA20543	16201T		62.44%		99.86%
JPT-NA18940	16209C		59.91%		99.82%
CHB-NA18637	16209C		59.61%		99.79%
ser-21	16212G		57.86%		99.69%
man-231	16213A		66.71%		99.52%
bak-41	16214T		80.03%		99.64%
bak-55	16214T		65.90%		99.23%
bhu-0957	16218T		83.11%		99.76%

Cont.

sample	alternative variant	Prototype		CRM Nested	
		Alternative Variant %		Alternative Variant %	
		primer	pType + OREO	primer	CRM + <95 + OREO
pal-4919	16219G		61.20%		99.43%
den-190	16221T		92.67%		99.44%
JPT-NA18974	16231C		93.88%		99.80%
bak-55	16234T		95.92%		99.19%
bhu-1892	16234T		94.93%		99.67%
eng-GB1778	16234T		93.21%		99.61%
hun-29	16234T		93.12%		99.53%
nep-0172	16234T		90.77%		99.62%
pal-4922	16234T		90.01%		99.58%
bhu-1151	16239T		88.11%		98.88%
MXL-NA19664	16240G		84.03%		99.57%
bak-35	16241G		79.97%		99.56%
CHB-NA18608	16243C		90.58%		99.88%
ngo-98	16245T		86.98%		99.61%
nep-0273	16245T		86.86%		99.68%
eng-hgQ-2	16248T		89.39%		99.17%
bak-25	16362C		5.50%		0.41%
bak-55	16362C		5.24%		0.53%
CHB-NA18561	16362C		5.74%		0.38%
mbe-237P	16362C		5.22%		0.35%
bav-55	16366T		42.58%		99.56%
TSI-NA20527	16368C		38.07%		99.80%
spa-20	16368C		41.44%		99.86%
pal-4929	16368C		36.03%		99.83%
CHB-NA18561	16449T		5.04%		0.08%
bhu-1564	16452C		90.04%		99.91%
bhu-0984	16463G		77.47%		99.82%
bhu-1564	16491C		85.15%		99.77%
aus-m119	16527T		11.22%		8.10%
tur-4	41T		85.25%		99.60%
bak-41	44.1C		69.00%		99.92%
bak-55	44.1C		77.21%		99.91%
MXL-NA19664	182T		7.27%		9.22%

Cont.

sample	alternative variant	Prototype		CRM Nested	
		Alternative Variant %		Alternative Variant %	
		primer	pType + OREO	primer	CRM + <95 + OREO
bkl-46	195C		48.37%		46.46%
eng-hgQ-2	195C		26.09%		27.95%
bhu-1564	195C		5.41%		2.32%
gre-2	204C		27.02%		22.39%
den-190	207A		95.40%		98.96%
YRI-NA18856	207A		94.22%		94.60%
ire-94	217C		84.65%		99.82%
operator control	217C		82.60%		99.61%
YRI-NA19175	281G		85.88%		99.78%
bhu-1564	297C		13.61%		9.41%
bhu-0984	297C		4.06%		2.07%
nep-0809	303A		6.21%		1.15%
bhu-1564	316C		5.68%		3.92%
ire-94	316C		9.99%		16.14%
tur-4	319C		71.16%		99.88%
CHB-NA18608	326G		92.31%		99.76%
ork-007	327T		67.35%		99.52%
pal-5232	334C		59.60%		99.37%
YRI-NA19098	372C		33.65%		99.73%
tur-9	372C		42.51%		99.73%
bak-33	372C		39.16%		99.75%
pal-4919	385G		25.56%		99.81%
ork-007	456T		95.49%		99.57%

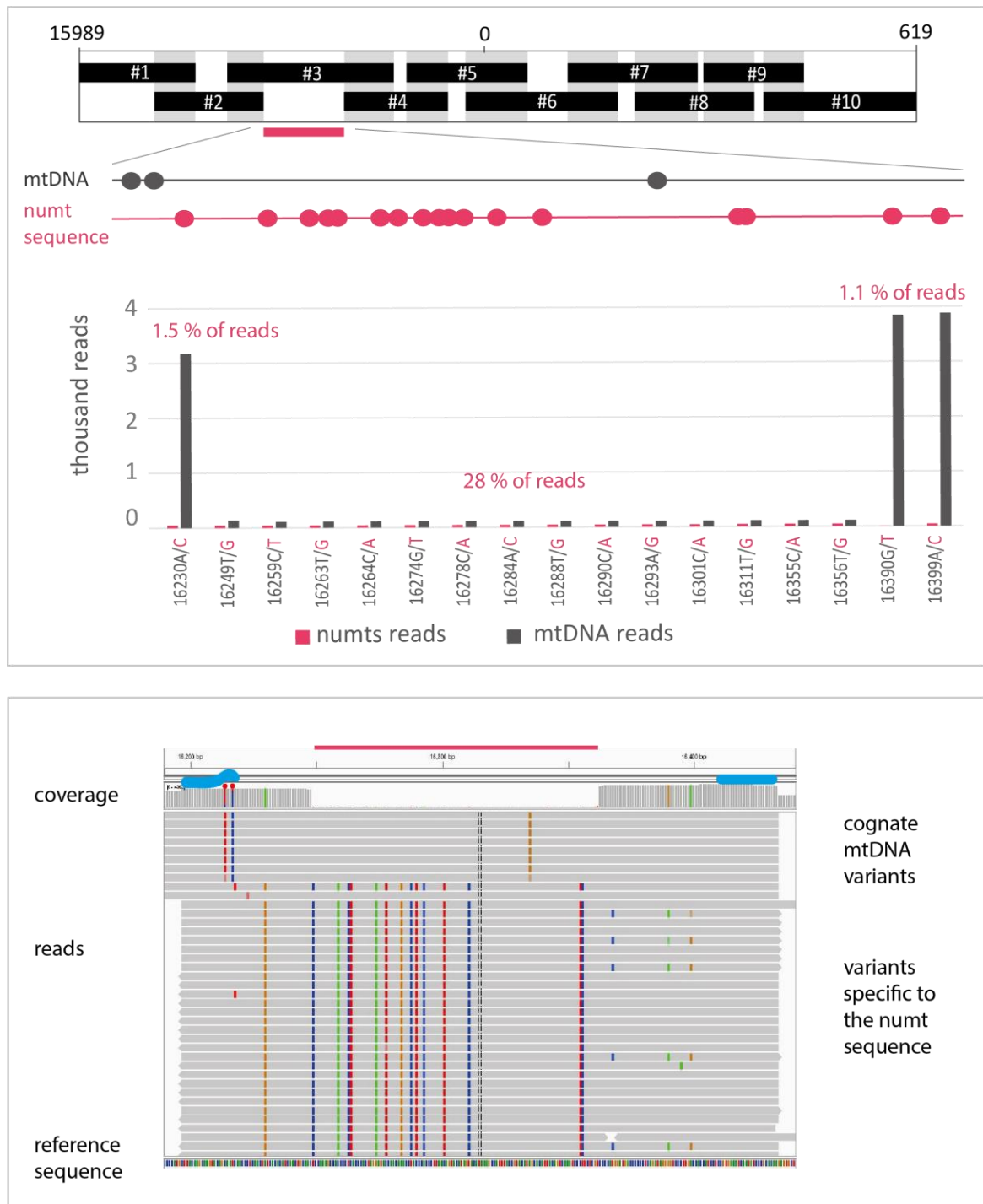
code:	homoplasmy:	potential	confirmed
	heteroplasmy:	potential	confirmed

Heteroplasmy detection thresholds were set to be between 5 and 95% alternative variant proportions. Heteroplasmic sites were detected at eleven different positions within the control region (182, 195, 204, 207, 297, 316, 16,069, 16,093, 16,129, 16,150 and 16,527) in a total of twelve samples, these calls range from 8 to 95% alternative variant proportion (Table 5.6). Several of these positions are often reported elsewhere as heteroplasmic (Holland et al. 2011; Lascaro et al. 2008). For example, position 16,093 in the adjusted CRM data was observed once above (and

also twice below) the 5% reporting threshold. Only two of these are transversions (297C 9% and 316C 16%), both positioned away from primer-binding sites (details in Table 5.6). After applying OREO, the number of homoplasmic variants identified in the control region increased to 197, including five simple indels.

5.3.7 Detection of numt sequences

In a multiplex control-region assay, reduced coverage for a specific amplicon (Eichmann and Parson 2008) could permit the detection of a numt sequence, as a pattern of variants resembling multiple closely linked heteroplasmic sites (Figure 5.12). In the low-coverage interval of the sample (tur-16) described in Section 5.3.4, fourteen sites in a window of 114 bp showed mixed allele calls with the minor component at a level of ~28%. The minor-component sequence was used in a *blastn* query against the nucleotide collections of GenBank, and returned a known polymorphic numt sequence inserted after chr11:49,862,017 (GRCh38) on chromosome 11p11.12, first described by Zischler et al. (1995), and subsequently noted by others (Bintz et al. 2014; Dayama et al. 2014; Lang et al. 2012; Ricchetti et al. 2004; Thomas et al. 1996). Given this knowledge, numt-specific variants were surveyed in the sequence flanking the 114-bp window and presented the same level of numt-specific reads at three additional sites underlying a much higher coverage of cognate mtDNA reads (Figure 5.12).



This numt sequence was identified because of the low-coverage amplicon in tur-16, but having discovered it, it was worth asking whether it also existed as a low-level component of reads in other samples. Similarly low-levels of this sequence were detected (at a mean of 0.67% of the reads) in an additional 43 samples of diverse continental origins, showing that the numt sequence may be polymorphic, and present at high frequency in the diverse sample set surveyed here. This is consistent with previous findings of the geographical distribution of this insertion (Lang et al. 2012; Ricchetti et al. 2004; Thomas et al. 1996), and with the frequency in the 2504 samples of the 1000 Genomes Project (where it is annotated as the structural variant esv3626324 in dbVar, and also as the SNPs rs116522696 and rs115254439). The 29 analysed samples that were also included in the 1000 Genomes Project were inspected: in 1000 Genomes data, 9/29 samples lack the insertion, and the data generated here agree with these. However, the numt sequence is known to be present in all 20 remaining samples, but here it is detected in only four, indicating a high false-negative rate.

This numt sequence was also detected in two-thirds of the samples processed with the PowerSeq™ CRM Nested System, and while again no false positives were detected compared to 1000 Genomes Project data, detection of the presence or absence of the numt sequence between the two kit types was not always consistent.

5.4 Discussion

Here, a set of 101 diverse human DNA samples has been analysed to investigate the calling of variants in the mitochondrial control region using an MPS-based approach. Coverage across the control region is non-uniform in overlapping amplicon approaches (Figure 5.3 and Figure 5.9). Excess coverage may result partly from differences in amplification efficiency among amplicons, but also, because of the single-reaction kit design, from the preferential generation of shorter amplicons which occupy substantial sequencing capacity of an MPS run (Chapter 2, Figure 2.5). Sequence reads from shorter products contain a high proportion of primer-derived sequences which cannot be precisely removed by conventional trimming because the primers are proprietary, and which introduce reference sequence bias in the data. This necessitated the development of an alternative data-processing approach (Overarching Read Enrichment Option, OREO), to bypass the bias introduced by primers at the ends of reads by retaining only overarching reads. Following application of OREO, analysis of the mtDNA control region in a set of 101 samples using the prototype kit design (PowerSeq™ Auto/Mito/Y System) showed generally robust amplification and sequencing, despite the high diversity of analysed mitochondrial genomes, which covers a wide range of the mtDNA phylogeny (Figure 5.2). All 101 samples presented different sequences, defined by variation among 161 SNPs within the control region. Validation against independent data from whole mtDNA sequences for 65/101 samples (1000 Genomes Project Consortium 2015; Batini et al. 2017) showed a high degree of concordance with no false negative variants. A conservative false-positive rate of 1.25% was estimated, but given the high sequence coverage of the data over the relevant sites it is highly likely that the error lies in the comparative data rather than the data generated here, and the false-positive rate is in reality likely to be lower than this.

Heteroplasmy is a ubiquitous phenomenon, and needs to be considered in any forensic analysis of mtDNA (Parson et al. 2014). As with the detection of numt

sequences, the introduction of MPS has improved the sensitivity of detecting lower-level heteroplasmy compared to Sanger-based methods (Gallimore et al. 2018). Application of OREO, an alternative to primer trimming, reduced the number of potential heteroplasmic variants by removing the bias introduced by primer sequences at the ends of reads, but not those that are internalised by overlap extension (Figure 5.6).

The PowerSeq™ CRM Nested design prevents overlap extension by adding adapter sequences onto the ends of the amplicons (Figure 5.7). Data from this kit can then be processed to remove reads from the short amplicons by *in silico* size selection, and further improved by applying OREO when calling variants and quantifying heteroplasms at primer sites (Figure 5.11).

The combination of improved chemistry of the PowerSeq™ CRM Nested System and appropriate data processing with OREO allowed a consideration of the whole control region to accurately call variants and heteroplasms down to the level of 5%, identifying 197 different variants and twelve point heteroplasms in this sample set. Notably, other MPS-based approaches permit lower thresholds, thanks to different kit designs or deep sequencing (Holland et al. 2018; Rathbun et al. 2017); however, for the general purposes of variant and heteroplasmy detection the limit used here of 5% seems sufficient.

Considering kit design more generally, a non-overlapping two primer-mix option (following the original approach; Eichmann and Parson 2008) remains preferable to a single-reaction multiplex. Even though OREO can overcome the artefacts associated with the single-reaction approach, the short amplicons of overlapping regions can take up significant capacity on the surface of the sequencing chip. While this is manageable when processing ideal reference quality samples, as was demonstrated here, the effect will be more detrimental when sample quality is non-ideal: short amplicons will be further elevated, and longer amplicons reduced, enhancing the bias present at the primer sites in variant calls.

The human genome is well known to contain many numt sequences (Hazkani-Covo et al. 2010; Lascaro et al. 2008) and some of these are polymorphic insertions which can be used in human population studies and phylogenetic analyses (Dayama et al. 2014; Lang et al. 2012). These divergent copies of mtDNA fragments are not expected to cause particular concern in short-amplicon sequencing of the control region, due to the multiple copies of cognate mtDNA template, and the extremely high coverage of the resulting cognate reads. Indeed, co-amplified numt sequences are generally below the detection limit of conventional Sanger sequencing-based assays, although they have been detected with denaturing gradient-gel electrophoresis, which was intended to resolve lower level heteroplasmies (Tully et al. 2000). The depth and precision of MPS can allow the detection of low-level numt-derived reads (Bintz et al. 2014), but again these reads do not normally reach the variant calling threshold and are therefore not reported.

In this study, however, reduced amplification efficiency for one amplicon in one individual, due to the presence of SNPs within a primer-binding site, allowed a numt sequence to be detected (Figure 5.12). This prompted a wider screen for the same numt sequence, and it was found to be detectable with both kit types at very low levels in more than half of the sample set (though comparisons to the 1000 Genomes Project data show a high false-negative rate, so its true frequency must be higher). Considering that these nuclear templates are neither specifically targeted nor enriched for detection, but observed only as a byproduct overshadowed by usually overwhelming cognate mitochondrial templates, it is not surprising that they are observed inconsistently. Relevant variables include the individual underlying variants in the sample, the relative amplification success of the nuclear and mitochondrial templates, relative concentration of mtDNA and nuclear DNA in a tissue, the multiplexing level and other run parameters affecting the coverage of the sample. Thus, despite the high population frequency and widespread distribution of this polymorphic numt sequence (Lang et al. 2012; Thomas et al. 1996), it interferes with variant calling in reference samples only rarely. In non-forensic applications numt sequences could interfere with correct reporting of mitochondrial DNA variants (Parr et al. 2006). However, in the analysis

of casework samples where low-level minor components are of particular interest (Bintz et al. 2014), numt sequences could become visible, requiring removal (Ring et al. 2018), or may be falsely interpreted as heteroplasmies or mixtures.

5.4.1 Highlights and Conclusions

The single-reaction multiplex provides a robust tool to analyse mtDNA control region variation across a range of haplogroups. The commercial design of the PowerSeq™ CRM Nested System, provided appropriate data processing is applied to mitigate reference sequence bias, can identify variants throughout the whole control region. With the experimental setup tested here together with the data analysis workflow developed for this study the CRM kit is able to measure heteroplasmies accurately down to a 5% threshold.

To improve the detection and quantification of variants and heteroplasmies a bioinformatic method, OREO, was developed to decrease the effects of non-uniform coverage and reference sequence bias in overlapping amplicon sequencing studies.

CHAPTER 6: Massively parallel sequencing of forensic markers in the People of the British Isles

6.1 Introduction

6.1.1 History of the British Isles

Prehistorically, the British Isles have a long history of occupation by anatomically modern humans dating back to 41-45 KYA (Higham et al. 2011). However, the region was no longer habitable by the start of the Last Glacial Maximum around 26.5 KYA, and it was only by ~11.5 KYA, as the deglaciation commenced and the climate warmed up, that Palaeolithic hunter-gatherers were able to recolonise this area from the continental mainland, via land that connected Britain to the continent of Europe, as shown by radiocarbon dating of human remains (Barton et al. 2003).

From the mesolithic period the oldest complete human skeletal remains in Britain were found in Gough's Cave, Somerset, referred to as the “Cheddar” Man, a Western European mesolithic Hunter-Gatherer (WHG) from about 9.1 KYA (Davies 2000). A recent study sequencing his whole genome to ~2.3 x coverage estimates a phenotype of blue or green eyes, dark hair and dark or very dark skin, consistent with other WHGs of the time period elsewhere, which precedes the introduction of lighter skin pigmentation introduced via the Neolithic Farmer influx. The estimation however compares the genetic variants to the currently known variant pool and their corresponding phenotypes, and thus, it is more uncertain than predicting similar traits from current remains (Brace et al. 2019).

Neolithic Farmers first appear in Britain about 6 KYA via different routes from continental Europe and admix with present WHGs; however, this admixture happens later and to a lesser extent than it does across continental Europe. The introduction of farming caused a transition in population size, material culture and lifestyle (Brace et al. 2019).

It is known from archaeological evidence that a distinctive Bell Beaker pottery spread into Britain about 4.5 KYA, shortly before the start of the Bronze Age. A recent genomic study (Brace et al. 2019; Olalde et al. 2018) using ancient DNA sequences found that Britain underwent significant population replacement at the time marked by the presence of this culture. This transition was accompanied by a massive shift in ancient Y-chromosome sequences, observed as a near replacement of haplogroups I2 and G2 by R1b, which remains the commonest haplogroup in Britain today (Balaesque et al. 2010).

During the Bronze and Iron Ages the Celtic culture from mainland Europe was brought to the British Isles, most likely by a small number of people rather than in the form of an 'invasion' (Searle et al. 2009). What may be referred to as the Celtic people of Britain are the early modern inhabitants, the Britons. They were assumed to be loosely tied by traditions, religions, and the Celtic sub-family of Indo-European languages, which includes Irish Gaelic, Scottish Gaelic, Welsh, Breton, and the now-extinct languages Manx and Cornish (Davies 2000; Forster and Toth 2003).

The Roman occupation (43-410 CE) made Britain part of the Roman Empire under the name of Britannia and is believed to be accompanied by only minimal migration, but marked cultural change that persists in archaeology, roads, town-plans and place names, including that of Leicester. The Roman administration was followed (450-600 CE) by large-scale settlement from the near continental Europe by Angles, Saxons, Frisians and Jutes into Southern and Eastern Britain (Weale et al. 2002). The regions affected by the Anglo-Saxon migration experienced a language shift to the Germanic Old English, and place names were also affected. However, in the Atlantic fringes of Cornwall, Wales, Northwestern England and Scotland the Celtic languages persisted as they remained under the control of Britons.

Orkney and other islands north of Scotland were settled by Vikings from Norway from the late 8th century. Norway annexed Orkney from 875-1472, and thus Orkney was culturally influenced by Scandinavia. Some other Norse Viking settlements were found in the west - the Isle of Man, Ireland and Wales; the raids and large-

scale Viking invasion of Eastern England started from the 9th century, largely from what is now Denmark. From East Anglia to North West England the settlers established a region of administrative control from the 9th - 11th centuries, the Danelaw, witnessed by archaeological finds and preserved in Scandinavian place names, which, for example make up 70% of the minor names of Lincolnshire places (with endings such as “-by”, “-toft”, and “-thorp(e)”) (Smith 1956).

The Normans of northern France invaded Britain in 1066 CE, taking control over England and South Wales and part of Ireland by means of a small elite, and hence accompanied by only minimal migration. The name ‘Norman’ derives from ‘Northmen’, indicating that these people, too, were descended from Scandinavians. The major events described here are also shown on Figure 6.1, reproduced from Leslie et al. (2015).

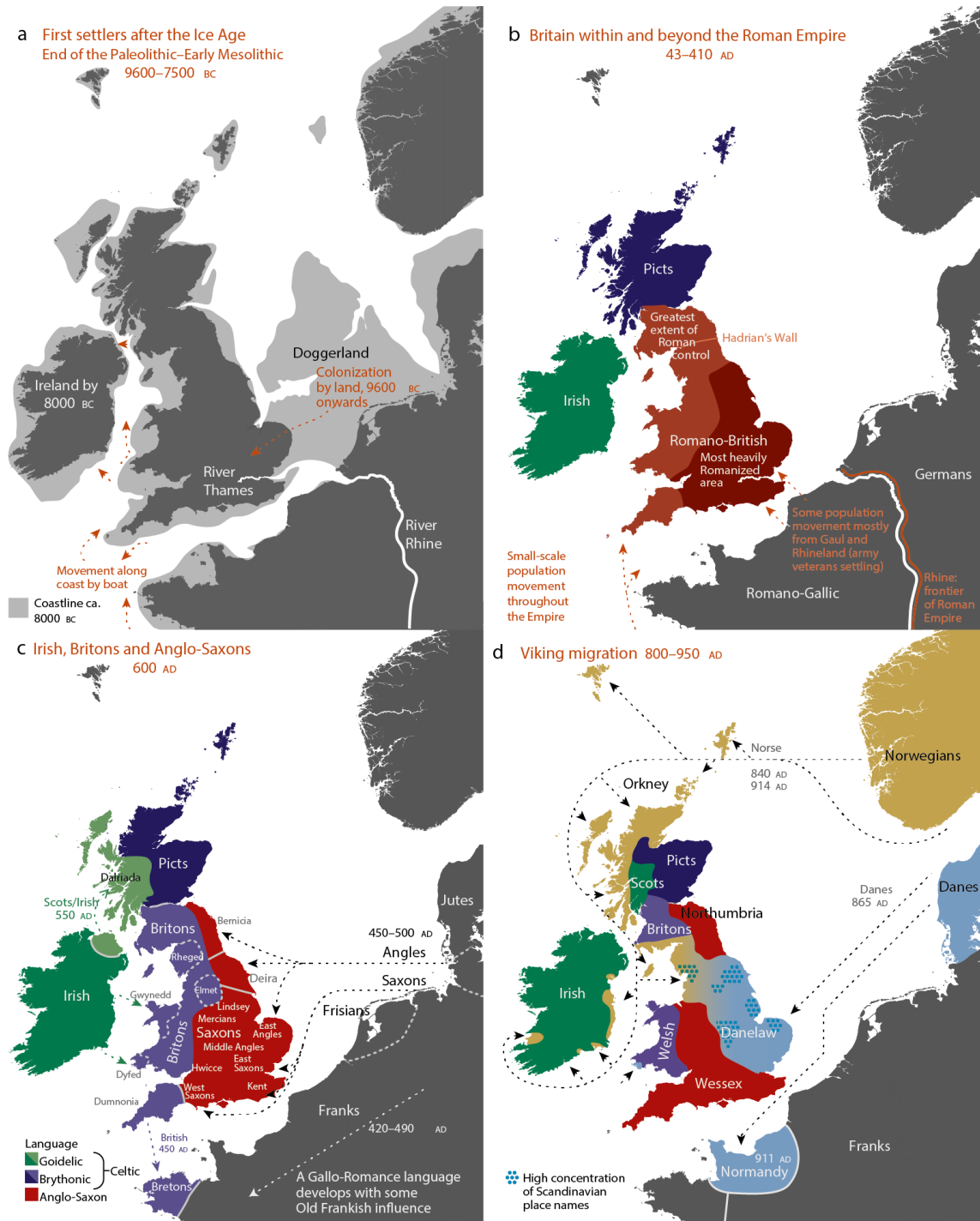


Figure 6.1
Major events in the peopling of the British Isles from Leslie et al. (2015).

6.1.2 Summary of earlier studies of genetic diversity in the British Isles

Uniparental marker studies (Goodacre et al. 2005; Wilson et al. 2001) provided evidence for male Scandinavian contributions from Y-chromosomal DNA in Orkney, but did not find the same from the mtDNA maternal lineage data, suggesting sex-biased processes in settlement and cultural transition.

A Y-focussed study of (Weale et al. 2002) transected the island from England to Wales by sampling appropriately sited small towns, and noted a transition in Y-chromosome types between England and Wales, but no real differentiation between Frisian and the English samples; this was interpreted as the effect of the Anglo-Saxon mass migration shifting the Central English male gene pool. Modelling approaches suggested that the number of migrants needed for this shift was extremely large, giving rise to scepticism from historians. Interestingly; however, in another Y-chromosomal study (Capelli et al. 2003) of Britain and Ireland, they also found high level of Danish/Northern German input in regions with historical Danish influence, but they chose to interpret this as a mark of a Viking presence instead.

In contrast to the several studies of Y-chromosomal diversity in the British Isles, no systematic attention has been paid to the mtDNA lineages, perhaps because widely assumed patrilocal marriage practices are expected to correlate the Y-chromosomes better to geography.

6.1.3 The PoBI project

The differences between genomes are non-randomly distributed with respect to geography, and sampling indigenous populations with shared migration history therefore means the genetic similarity observed within and between them tends to correlate to their geographical proximity due to low ancestral mobility (Ramachandran et al. 2005). The sampling of modern populations in a geographically systematic manner allows inferences about the population structure

within these areas, shaped by migrations, admixture, selection and genetic drift (Winney et al. 2012).

The People of the British Isles project (PoBI) used carefully selected modern samples to infer past population history from autosomal SNP markers across the genome. In order to capture the fine-scale population structure, samples were selected based on the criterion of having all four grandparents born within 50 miles/ 80 km of each other in rural areas, and therefore relatively unaffected by recent major migrations and admixture (Winney et al. 2012).

Conventional methods such as PCA, STRUCTURE (Pritchard et al. 2000) or ADMIXTURE (Alexander et al. 2009) rely on SNP frequencies alone and are ideal to detect large-scale differences between populations, for example at continental levels; however, these are seldom able to detect fine-scale structure within indigenous national populations such as that of Britain beyond the most differentiated regions, for example Orkney and Wales (Leslie et al. 2015). However, more subtle levels of genetic differentiation could be captured by the clustering algorithm, fineSTRUCTURE (Lawson et al. 2012), which also accounts for linkage disequilibrium, the association of SNPs with each other due to coinheritance. Using fineSTRUCTURE the PoBI project analysed over 500,000 common genome-wide autosomal SNPs, and their frequencies and haplotype associations identified 17 main genetic clusters (Figure 6.2; Leslie et al. 2015).

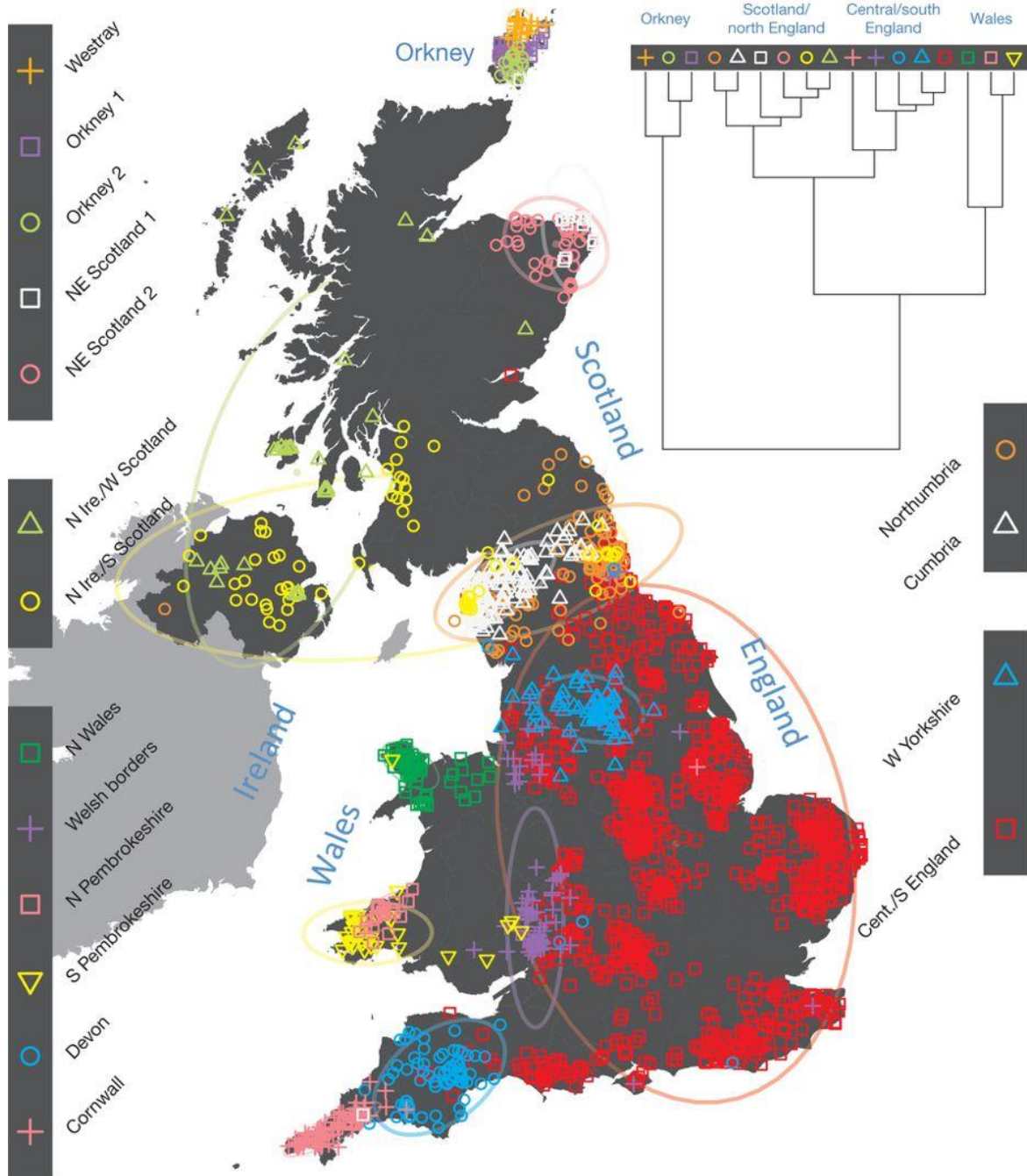


Figure 6.2
Autosomal-based fineSTRUCTURE clusters detected in the PoBI study (from Leslie et al. 2015).

The deepest branching clusters represent the separation of Orkney and Wales, confirming observations from using the less sensitive ADMIXTURE analysis. Although individual genomes were analysed without reference to geography, when

plotting these according to the birthplace of the donor's four grandparents, the clusters showed largely non-overlapping patterns, which demonstrates the power of the method to be able to reveal population structure.

To further understand the composition of these clusters the authors compared them to over six thousand samples from relevant continental sources and defined the contributions of these to the PoBI clusters. Some European groups contribute to most clusters, and thus can be interpreted as more ancient contributions now widespread by standard population movement, but others only provide input to a limited number of PoBI clusters, the relatively recent (medieval) contributors, where the time passed has not yet allowed the spreading of these components across the population. For example, about 25% Norwegian ancestry is estimated for the largest Orkney cluster, aligning well with the known history of these islands being part of the Kingdom of Norway for about six hundred years. The major Central/South England cluster showed 35% ancestry from a North-West German group, while the group did not contribute to Welsh clusters, and therefore this component was considered to be associated with the Saxon migrations.

Interestingly relatively little input from Danish European groups, a proxy for Danish Viking migrants, was found and no Danelaw cluster is differentiated possibly due to free migration in the area of the Central/South England cluster. However, the North-West German group (attributed to the Saxon component) cannot be reliably separated from the Danish European groups, and thus could also reflect the Danish Viking contributions (Kershaw and Røyrvik 2016).

Although these associations are not definitive, the study was groundbreaking in the use of fine-scale genetic differences found in modern samples to understand past migrations and admixture components and how these have affected the genetic landscape of Britain as it is today.

Apart from the reported autosomal variation, the PoBI dataset also contained data on the Y-chromosomal and mtDNA variants from the SNP arrays; although these

data have not been published, the data were available through collaboration (W.Bodmer, J.Wetton personal comm., unpublished data), and could be used here for comparison to MPS data.

6.1.4 Other studies from the British Isles

Early studies of Ireland's population structure showed a general east-to-west cline in Y chromosome haplogroup diversities (Hill et al. 2000) and blood group frequencies (Relethford 1983), but no fine-scale analysis had been done at the time of the PoBI study.

Following the PoBI methodology and focussing on the Republic of Ireland, the study of Gilbert et al. (2017) used the 'Irish DNA Atlas' cohort of 194 samples with four generations of local ancestry, and found ten genetic clusters. Seven of these reflects ancient 'Gaelic' Irish ancestries, while three are shared with the British. The Gaelic Irish clusters sit closer to the Orcadian clusters from PoBI, which may be due to both having high proportions of Norwegian ancestry (up to ~20% in these Irish clusters), which were dated to the time of the Norse Viking activity. A high level of North-West French ancestry from continental Europe was also identified, which was associated in this study with Brittany and with Celtic language ties. A similar study by Byrne et al. (2018) of an Irish population based on 1079 individuals found nine main clusters and a general west-east cline of Celtic-British ancestry, with increased homogeneity in the east which was interpreted as a signal of British admixture due to the geographic proximity.

Gilbert et al. (2019) focussed specifically on Scotland, and performed a similar fineSTRUCTURE analysis on 2554 individuals, including donors from undersampled Scottish areas, the Hebrides, Shetlands, and also the Isle of Man. This study found six main genetic clusters, and a general North East to South West divide.

With sequencing becoming more and more affordable, available genome-wide data is expected to increase. Large datasets, preferably with good geographical

resolution, would be able to further support studies of population history (Leslie et al. 2015). Interestingly, citizen scientists with an interest in genealogy and the will to invest in gathering their own genomic data, and accept the risks and benefits of sharing it with the community, may contribute significantly to this growing source of information, and provide better inferences on population histories beyond the original intention of growing their own family trees (Balanovsky et al. 2017); <https://www2.le.ac.uk/departments/genetics/people/jobling/citizen-science>).

6.1.5 Current ethnic composition of the British Isles

Table 6.1 and Figure 6.3 are compiled from the data of 2011 censuses from England and Wales (<https://www.ons.gov.uk>), from Scotland (<https://www.scotlandscensus.gov.uk>), from Northern Ireland (<https://www.nisra.gov.uk/>), and also supplemented with data for the Isle of Man (<https://www.gov.im>), plus data from the Republic of Ireland (<https://www.cso.ie>). These show the self-reported ethnicities in 2011 in the UK and its countries respectively, the British Crown dependency of the Isle of Man and the Republic of Ireland. For the UK, the majority of the population are self-declared White (87.1%) and the following main ethnicities are above 1%: Asian (all together 7%), Black (3%) and Mixed (2%).

Table 6.1
Self-reported ethnicity components of the British Isles.

Data compiled from the 2011 censuses for the UK, Isle of Man and the Republic of Ireland based on self-reported ethnicities. The forensic STR studies (either using MPS or CE methods) published to date for these ethnic groups are listed on the right.

Ethnicity	UNITED KINGDOM	England	Wales	Scotland	Northern Ireland	ISLE OF MAN	REPUBLIC OF IRELAND	this study	Devesse et al. (2018)	Aliferi et al. (2018)
White	87.1	85.3	95.5	95.9	98.2	96.5	93.6	MPS	MPS	CE
Gypsy/Traveller	0.1	0.1	0.1	0.1	0.1	0	0.7			
Mixed	2	2.3	1	0.4	0.3	0.9	0.9			
Asian (Indian)	2.3	2.6	0.6	0.6	0.3	1.9	0	MPS	CE	CE
Asian (Pakistani)	1.9	2.1	0.4	0.9	0.1	0	0			
Asian (Bangladeshi)	0.7	0.8	0.3	0.1	0	0	0			
Asian (Chinese)	0.7	0.7	0.4	0.6	0.3	0	0.4			
Asian (Other)	1.4	1.5	0.5	0.4	0.3	0	1.5			
Black	3	3.5	0.6	0.7	0.2	0.2	1.4			CE
Other	0.9	1	0.5	0.3	0.1	0.4	1.6			
Population	63,182,178	53,012,456	3,063,456	5,295,403	1,810,863	84,497	4,525,281	n=362	n=400	n=3128

Due to the specific PoBI sampling criteria the most likely sub-population represented by this sequenced dataset is the self-reported 'White' ethnicity. This dataset was generated from geographically distributed rural indigenous people, and somewhat similar (except the stringent sampling criteria) to the 'White British' subset whose forensic markers were sequenced by Devesse et al. (2018) (as shown in Table 6.1 above).

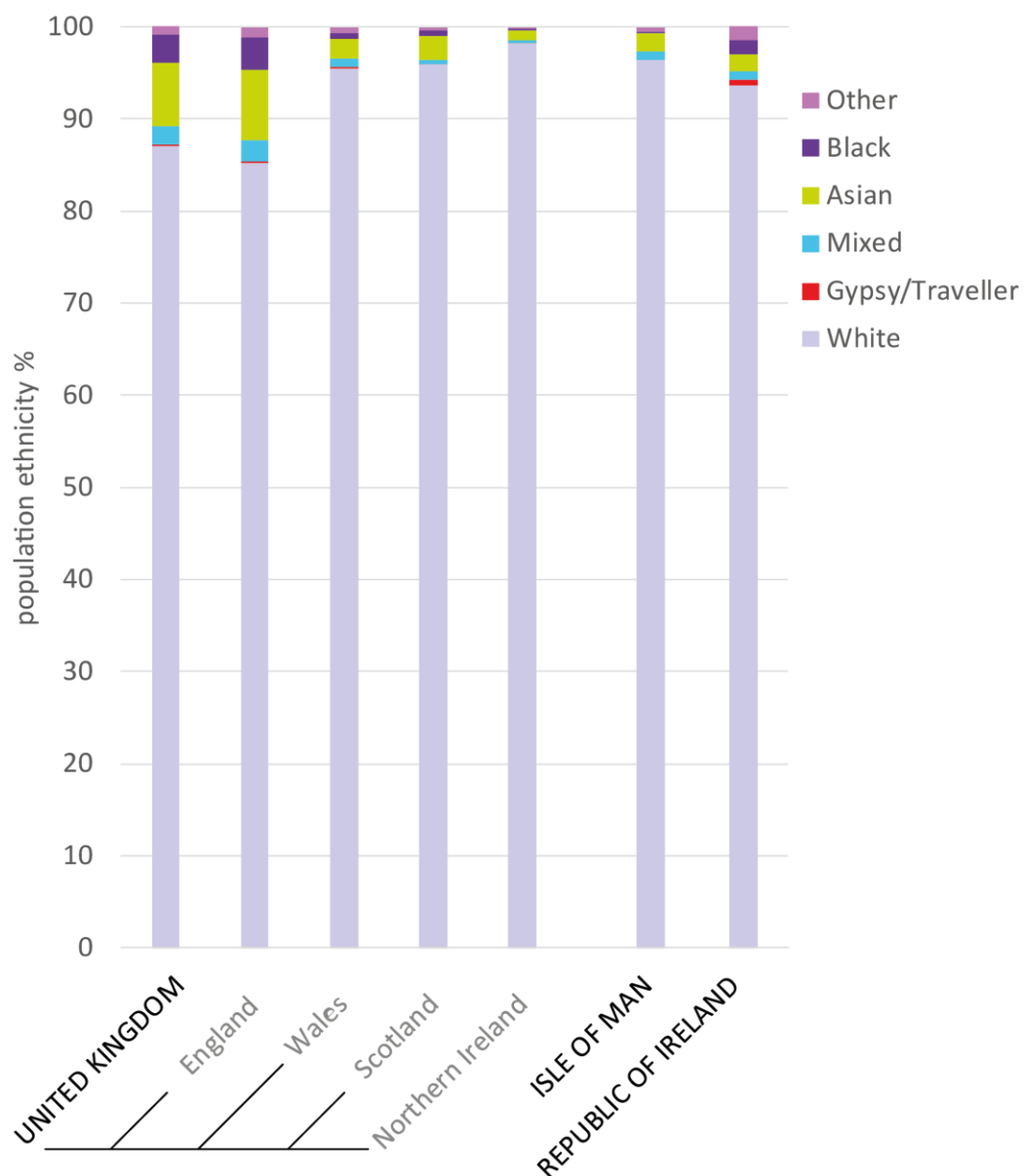


Figure 6.3
The main census-based ethnicity components of the British Isles.

Data compiled from the 2011 censuses.

6.1.6 Terminology of the British Isles

The complex history and administrative evolution of The British Isles has resulted in some variation of the nomenclature to describe its parts. In common usage these terms are often used interchangeably, and also often wrongly, therefore Figure 6.4 from the Encyclopaedia Britannica should help clarify the differences.

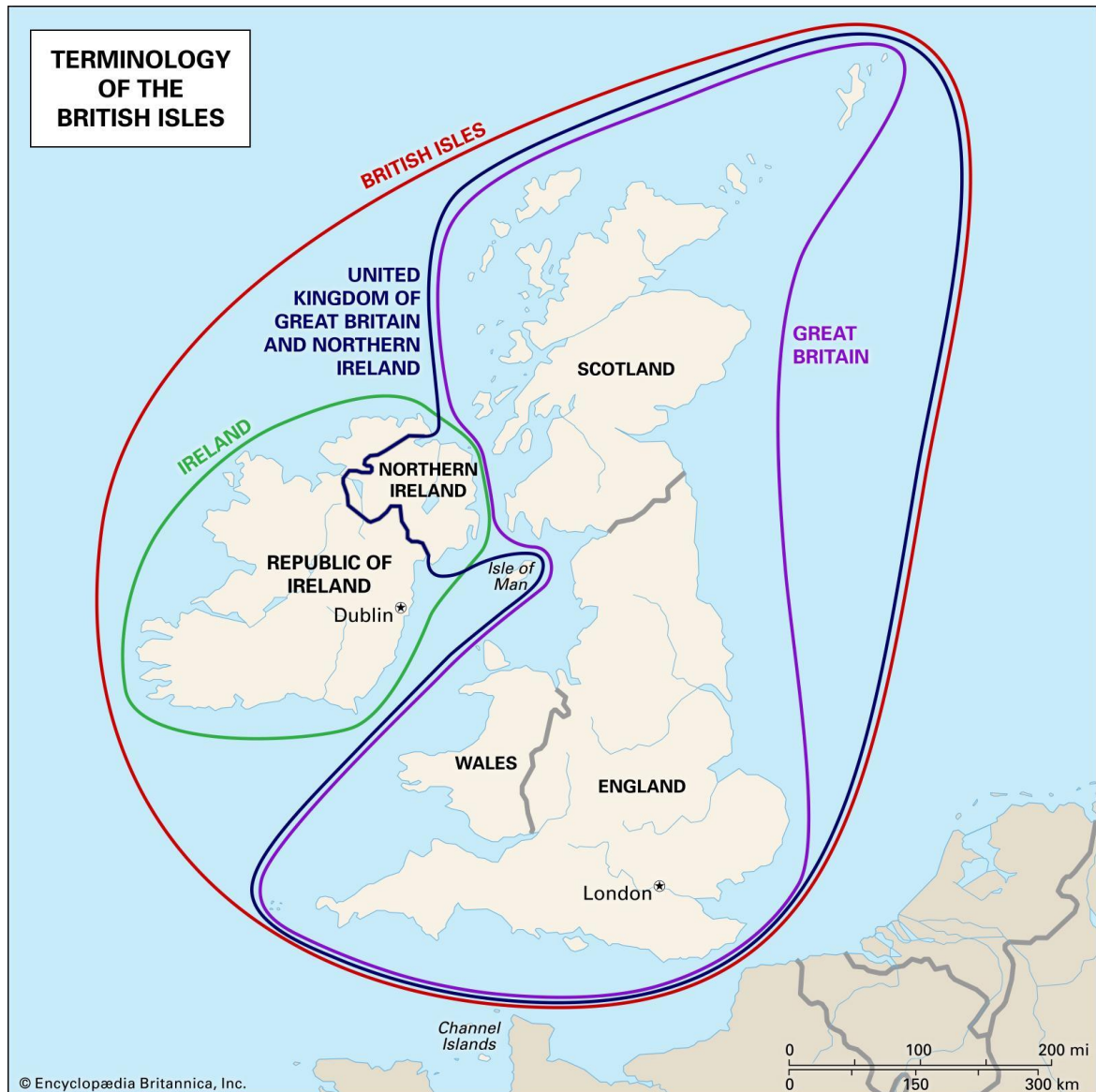


Figure 6.4
Terminology of the British Isles.

From the Encyclopaedia Britannica (<https://www.britannica.com>).

The aim of this chapter is to use MPS of forensic markers (autosomal and Y-STRs and the control region of the mtDNA) as described in earlier chapters to analyse sequences in a panel of indigenous samples from the People of the British Isles cohort and to provide a useful dataset of well-characterised indigenous samples for reference use.

6.2 Materials and Methods

6.2.1 DNA samples

As detailed in Chapter 2, Section 2.1.2 the 362 samples analysed in this study were selected from the males of the People of the British Isles (PoBI) sample set (Leslie et al. 2015; Winney et al. 2012), which aimed to represent local ancestry (i.e. all four grandparents' birthplaces fall within a 50-mile radius in rural areas), relatively unaffected by recent migrations. To emphasise the importance of Y chromosome locality, the selection was prioritised to select samples with known local paternal grandfathers, and where possible available typed SNP array information.

6.2.2 Options for dividing samples

Apart from the originally assigned 44 mostly county-based regions of the samples (Chapter 2, Table 2.2; Figure 6.5), other divisions of the sample set were considered for analysis. These were selected to increase and balance population size in each geographic division or to represent historical, linguistic or previously described genetic differentiation among regions.



Figure 6.5
The locations of the 44 mostly
county-based sampling regions
from The British Isles.

*The map is generated using
Microreact (Argimon et al. 2016).*

6.2.2.1 Thirty-seven regions

To prevent outliers with sampling locations represented with only one or two samples (e.g. Bedfordshire - 1; Buckinghamshire - 2), geographically neighbouring counties were grouped into 37 composite regions with a minimum sample size per region of seven. For example, Bedfordshire, Cambridge and Essex (neighbouring counties in the East of England) were together represented by a total of nine samples (Table 6.2).

Table 6.2**44 sampling regions grouped into 37 balanced composite regions.**











Sampling region	Composite Sample size	Sampling region	Composite Sample size
Argyll and Bute	9	Leicestershire	10
Banff and Buchan	10	Lincolnshire	10
Orkney	10	Derbyshire	13
Scotland	9	Nottinghamshire	10
Isle of Man	8	Northamptonshire	10
Northern Ireland	14	Bedfordshire	9
Republic of Ireland	10	Cambridgeshire	10
Cheshire	10	Essex	10
Cumbria	10	Norfolk	10
Lancashire	10	Suffolk	10
North East	8	Cornwall	10
Yorkshire	10	Devon	9
Mid Wales	10	Dorset	9
North Pembrokeshire	10	Forest of Dean	10
North Wales	7	Gloucestershire	10
South Pembrokeshire	10	Wiltshire	8
Wales	9	Berkshire	10
Herefordshire	9	Buckinghamshire	10
Shropshire	9	Hampshire	10
Staffordshire	11	Kent	10
Warwickshire		Oxfordshire	10
Worcestershire		Sussex	10

6.2.2.2 Ten large geographic, administrative regions

The original 44 sampling regions were then assigned to ten large geographic and administrative regions. This designation is shown in Table 6.3 and Figure 6.6. The ten geographic regions are SCO: Scotland; IRE; Ireland and the Isle of Man; NW: North West; NE: North East; WAL: Wales, WM: West Midlands; EM: East Midlands; EA: East; SW: South West; SE: South East.

Table 6.3

44 sampling regions assigned to ten large geographic and administrative regions.

	SCO	IRE	NW	NE	WAL	WM	EM	EA	SW	SE
										
Sampling region										
Argyll and Bute	✓									
Banff and Buchan	✓									
Orkney	✓									
Scotland	✓									
Isle of Man		✓								
Northern Ireland		✓								
Republic of Ireland		✓								
Cheshire			✓							
Cumbria			✓							
Lancashire			✓							
North East				✓						
Yorkshire				✓						
Mid Wales					✓					
North Pembrokeshire					✓					
North Wales					✓					
South Pembrokeshire					✓					
Wales					✓					
Herefordshire						✓				
Shropshire						✓				
Staffordshire						✓				
Warwickshire						✓				
Worcestershire						✓				
Derbyshire							✓			
Leicestershire							✓			
Lincolnshire							✓			
Nottinghamshire							✓			
Northamptonshire							✓			
Bedfordshire								✓		
Cambridgeshire								✓		
Essex								✓		
Norfolk								✓		
Suffolk								✓		
Cornwall									✓	
Devon									✓	
Dorset									✓	
Forest of Dean									✓	
Gloucestershire									✓	
Wiltshire									✓	
Berkshire										✓
Buckinghamshire										✓
Hampshire										✓
Kent										✓
Oxfordshire										✓
Sussex										✓

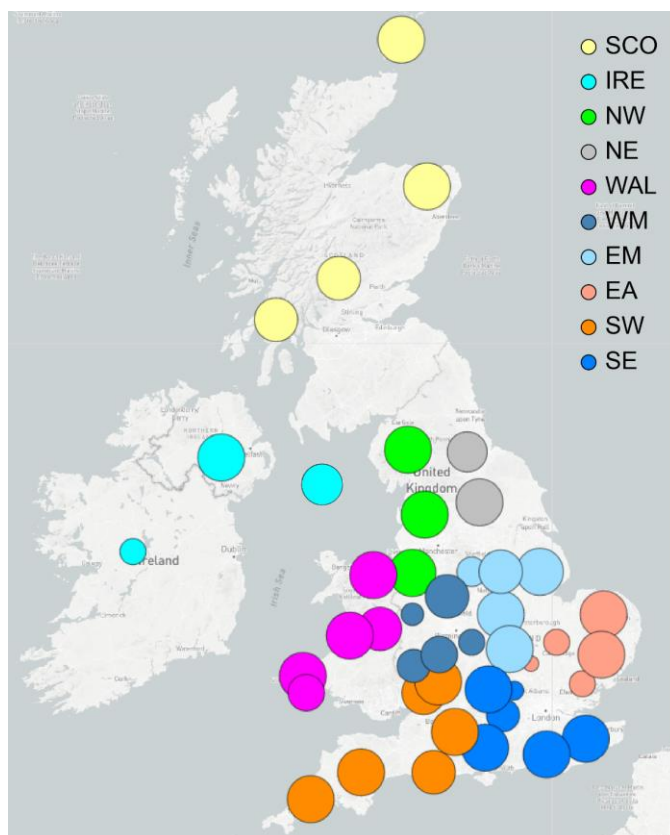


Figure 6.6
The ten large geographic and administrative regions.

The map shows the sampled counties according to their geographic and administrative regions. The sizes of the circles are proportional to sample size. Abbreviations follow that of Table 6.3. The map is generated using Microreact (Argimon et al. 2016).

6.2.2.3 Divisions following the five main PoBI clusters

To follow a genetic subdivision previously supported by fineSTRUCTURE analysis of autosomal SNP array data (Figure 6.2 and Figure 6.7) the following five main clusters which can be represented from this dataset were also tested: Orkney; Wales; Scotland, Ireland, Isle of Man, Cumbria and the North East; Cornwall; and the remaining areas of Central and South East England (Lancashire, Yorkshire, Cheshire, Derbyshire, Nottinghamshire, Lincolnshire, Staffordshire, Shropshire, Leicestershire, Norfolk, Warwickshire, Cambridgeshire, Worcestershire, Northamptonshire, Suffolk, Herefordshire, Bedfordshire, Gloucestershire, Essex, Forest of Dean, Oxfordshire, Buckinghamshire, Berkshire, Wiltshire, Kent, Hampshire, Sussex, Devon and Dorset).

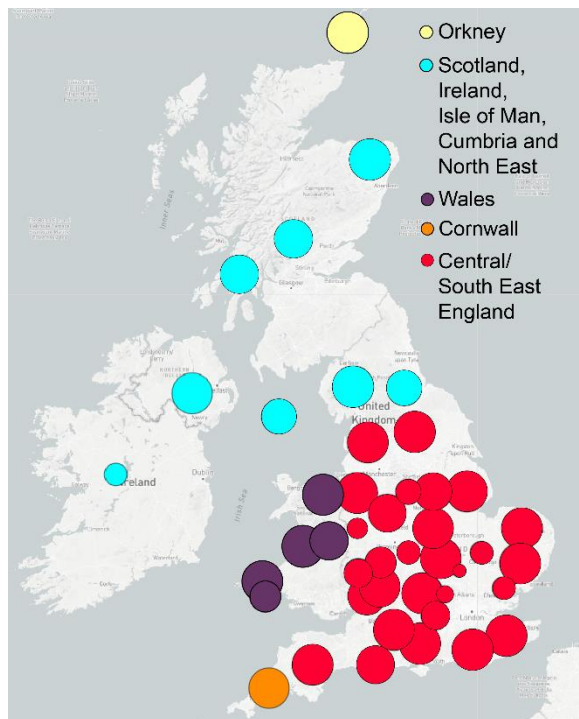


Figure 6.7
Division by the five main autosomal fineSTRUCTURE-based clusters detected in the PoBI study.

The five main differentiated clusters found by Leslie et al. (2015) were also tested in this dataset. The map is generated using Microreact (Argimon et al. 2016).

6.2.2.4 Division following regions with Celtic languages

A simple split of the samples following the language-based division of the Celtic fringe (Figure 6.8) was also tested, represented by samples from Cornwall, Wales, Isle of Man, Ireland and Scotland compared to the rest, which is England except Cornwall.

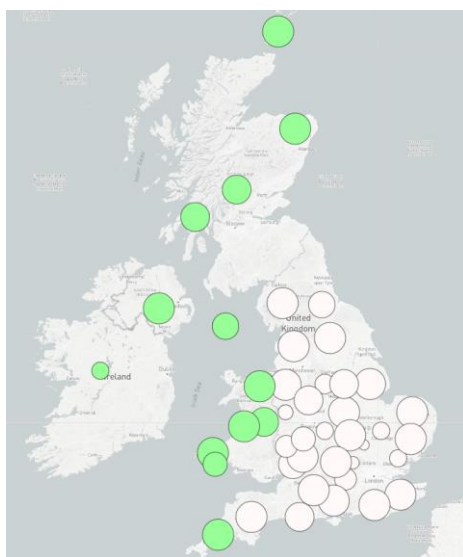


Figure 6.8
Division of the Celtic fringe regions from the rest of the British Isles.

Cornwall, Wales, Isle of Man, Ireland and Scotland highlighted in green represent the areas with current or recent Celtic languages. The map is generated using Microreact (Argimon et al. 2016).

6.2.2.5 Division following the Danelaw region

Another simple split of the populations was along the lines of the historically distinct Danelaw region (Figure 6.9) that was subject to Scandinavian administrative control from the 9th-11th centuries, and which might still be detectable in the current genetic structure. The Danelaw region was represented by samples from Cumbria, Lancashire, Yorkshire, Leicestershire, Lincolnshire, Nottinghamshire, Derbyshire, Northamptonshire, Cambridgeshire, Essex, Bedfordshire, Norfolk and Suffolk compared to the remaining non-Danelaw regions.

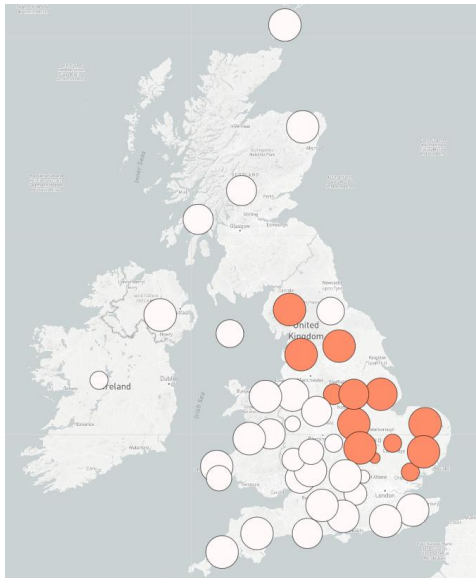


Figure 6.9
Division of the Danelaw region from the rest of the British Isles.

The map is generated using Microreact (Argimon et al. 2016).

6.2.2.6 Division congruent with the PoBI fineSTRUCTURE-based Central/South England cluster

From the five main PoBI clusters tested in 6.2.2.3, the predominant Central/South England cluster (Figure 6.10) is specifically tested in comparison to the rest. This region is represented by samples from Cheshire, Yorkshire, Staffordshire, Worcestershire, Warwickshire, Leicestershire, Lincolnshire, Nottinghamshire, Derbyshire, Northamptonshire, Cambridgeshire, Essex, Bedfordshire, Norfolk, Suffolk, Gloucestershire, Wiltshire, Dorset, Berkshire, Buckinghamshire, Hampshire, Kent, Oxfordshire and Sussex.

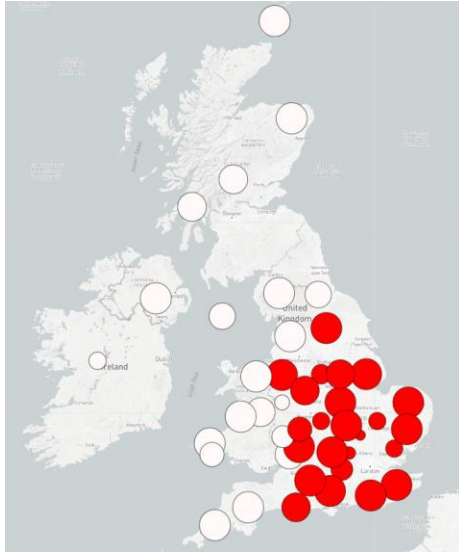


Figure 6.10
The PoBI autosomal Central/South England cluster represented in the samples.

The map is generated using Microreact (Argimon et al. 2016).

6.2.2.7 Division representing the East

A mainly geographic East divide (Figure 6.11) was tested to gauge the difference captured along the West-East axis. This region is represented by samples from Banff and Buchan, North East, Yorkshire, Lincolnshire, Cambridgeshire, Essex, Bedfordshire, Norfolk, Suffolk, Berkshire, Buckinghamshire, Hampshire, Kent, Oxfordshire and Sussex.

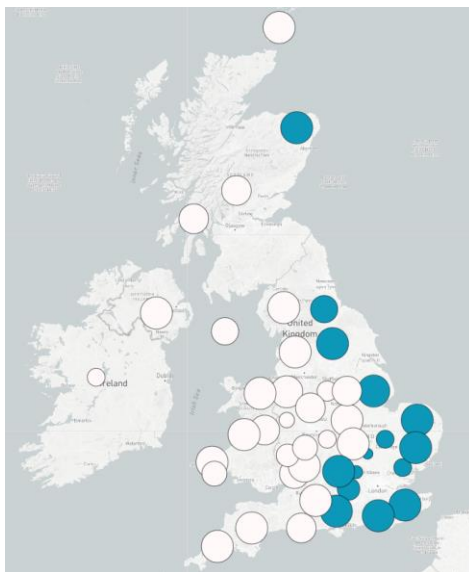


Figure 6.11
Division of the East regions from the rest of the British Isles.

The map is generated using Microreact (Argimon et al. 2016).

6.2.3 DNA quantitation and PCR amplification

DNA samples were quantified as described in Chapter 2, Section 2.2.1, using the Qubit® dsDNA BR assay kit following the manufacturer's recommended protocol. Several samples contained low quantities of DNA and therefore, for better accuracy, required measurements using the Qubit® dsDNA HS assay kit as well.

Targeted STRs (namely the 22 aSTRs and 23 Y-STRs plus the Amelogenin marker listed in Chapter 2, Table 2.3) and the control region of the mtDNA were amplified from these samples in a single multiplex using the Promega PowerSeq™ Auto/Mito/Y System prototype reagents following the manufacturer's recommended protocol, using 0.5 ng genomic DNA as template, as described in Chapter 2, Section 2.2.2.1.

6.2.4 Library preparation and sequencing

The generated amplicons were purified and quantified, then ~500 ng each were taken through TruSeq™ DNA PCR-free HT library preparations as detailed in Chapter 2, Sections 2.2.2.2 through 2.2.2.4. Sequencing of the prepared libraries followed the procedure described in Chapter 2, Section 2.2.4.

6.2.5 Data processing and analyses

The generated SE reads were exported in FASTQ format to be analysed by other software, including quality control (Chapter 2, Section 2.3.2). Mitochondrial reads were analysed for calling variants after applying OREO corrections, detecting numt sequences or length variants, and to generate population genetic data (Chapter 2, Sections 2.3.3.1 through 2.3.3.4 and Section 2.3.3.7). Analysis of autosomal and Y-STRs provided length and sequence information regarding the alleles and population genetic data (Chapter 2, Section 2.3.4).

6.3 Results

6.3.1 Run statistics and coverage values

6.3.1.1 Run statistics

Run statistics as calculated by the MiSeq® for each run are provided in Table 6.4 and the summary graph compares the efficiency of the runs in Figure 6.12.

Table 6.4
Run statistics for the sequencing runs.

	Error Rate (%)	% >= Q30	sample representation (%)			Density (K/mm2)	Clusters PF (%)	Yield Total (Gb)	Reads (M)	Reads PF (M)	% reads Identified (PF)	%reads PhiX (PF)
			Mean	Min	Max							
run1	2.65	68.14	1.512	0.002	14.481	1509 +/- 72	67.80 +/- 4.69	5.53	25.97	17.66	90.8	6.4
run2	2.17	69.89	0.212	0.544	1.775	1331 +/- 26	81.40 +/- 1.15	6.04	23.72	19.31	88.3	9.7
run3	2.52	70.66	0.266	0.498	1.587	1236 +/- 30	82.00 +/- 2.98	5.66	22.04	18.09	88.9	9.2
run4	2.02	71.62	0.225	0.549	1.755	1245 +/- 25	84.22 +/- 1.37	5.93	22.49	18.94	85.9	12.1

PF: passing filter, M: million.

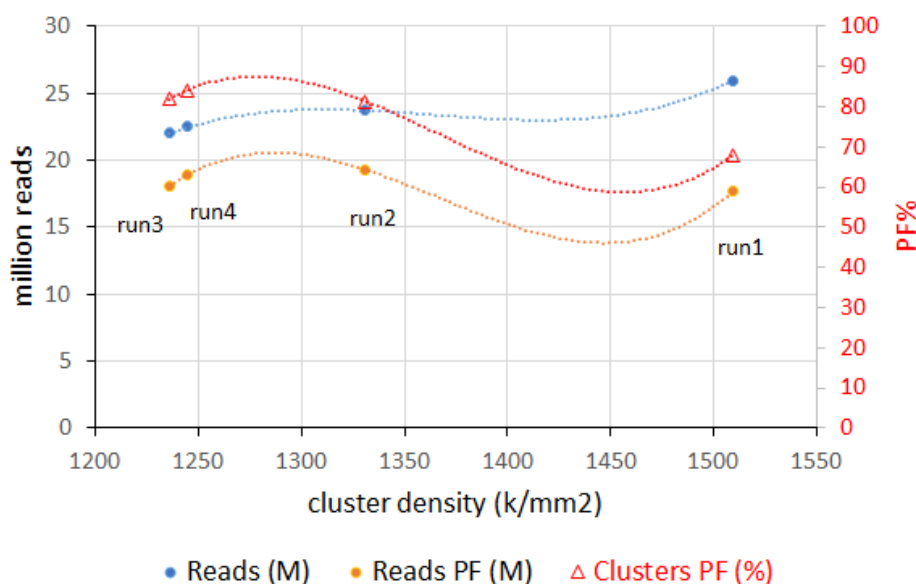


Figure 6.12
Efficiency of the sequencing runs.

Cluster density, Reads and Passing Filter reads with clusters Passing Filter % are plotted. The runs reached a fair consistency in output.

6.3.1.2 Coverage values

Coverage values calculated for each marker type are provided in Table 6.5, and Figure 6.13a specifically shows the normalised coverage values for the mtDNA reads across the control region, demonstrating the characteristic non-uniform coverage of the single reaction multiplex reaction as detailed in Chapter 5 (compare to Chapter 5, Figure 5.3); Figure 6.13b and c shows the same for the STR loci in comparison.

Table 6.5
Coverage values for each marker type.

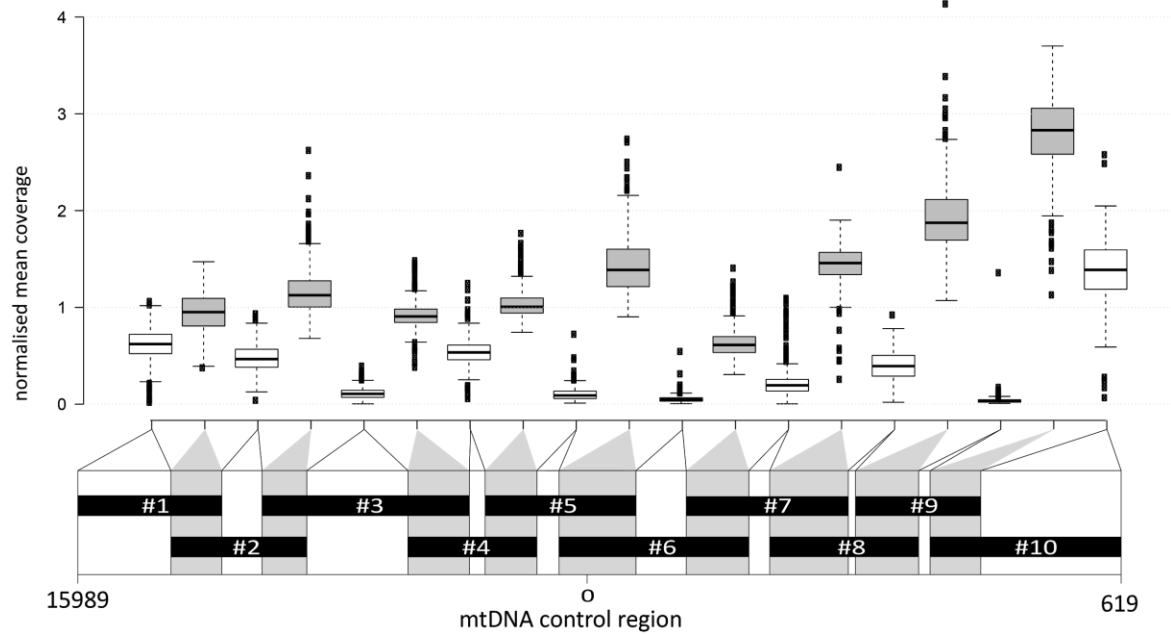
Coverage values for the mtDNA are given for each overlapping or non-overlapping region of the amplicons as shown on the left hand side of the table. Coverage for Amelogenin and the STR loci are given as mean coverage value per locus amplified.

Amplicon(s)	Mean coverage	Standard deviation	Median coverage	Min coverage	Max coverage
#1	3291.2	3262.8	3063.9	48	59855
#1/#2	5135.0	4792.6	4696.1	1041	89400
#2	2526.9	2410.2	2304.3	208	44444
#2/#3	6228.6	4669.0	5578.5	1182	81187
#3	595.5	525.7	522.4	21	7493
#3/#4	4946.6	3765.1	4531.7	835	66956
#4	2886.3	2237.7	2657.6	371	36177
#4/#5	5543.2	3727.5	5011.6	1025	63668
#5	566.3	493.5	452.3	30	4135
#5/#6	7714.4	5315.2	7029.2	1436	88988
#6	274.7	234.1	230.6	24	2615
#6/#7	3417.1	2440.5	3052.2	584	41529
#7	1218.5	1235.8	930.4	22	16494
#7/#8	7853.2	5546.6	7237.6	692	95677
#8	2104.3	1824.6	1861.9	56	29394
#8/#9	10425.2	7339.7	9403.9	1520	124114
#9	234.9	521.2	154.4	22	9547
#9/#10	15228.6	10147.4	14178.5	2276	171424
#10	7422.9	6378.7	6826.3	505	112484

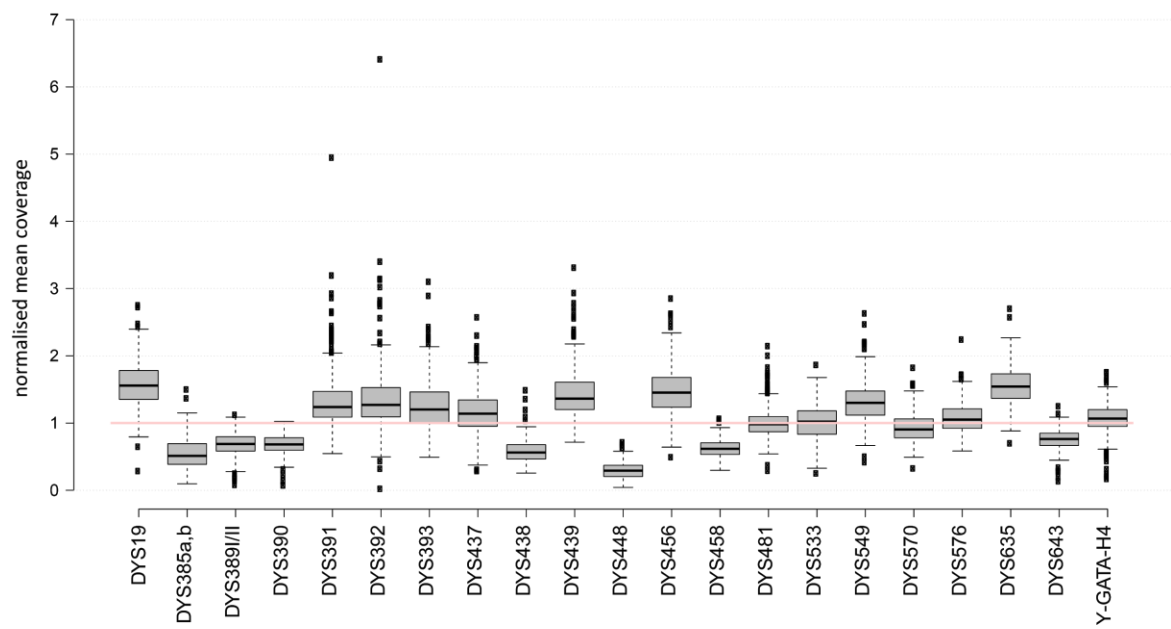
Cont.

Locus	Mean coverage	Standard deviation	Median coverage	Min coverage	Max coverage
Amel	1241.1	1676.3	1114.3	122	31795
CSF1P0	1871.1	1719.4	1760.9	153	31788
D10S1248	1421.4	1157.0	1361.5	198	21081
D12S391	2704.0	2649.8	2523.2	241	49167
D13S317	2602.3	2545.6	2468.0	253	47718
D16S539	2114.5	1970.7	1955.4	181	36388
D18S51	1484.1	1113.6	1386.7	173	19274
D19S433	1384.6	962.9	1308.8	102	16515
D1S1656	4161.9	3774.8	3856.9	319	69854
D21S11	3017.3	2508.9	2911.5	249	45639
D22S1045	2101.7	1655.6	1929.5	92	29502
D2S1338	1243.5	1382.0	1102.4	96	25621
D2S441	2787.0	2809.3	2496.4	480	50448
D3S1358	2200.0	2627.8	2059.7	84	49955
D5S818	3332.4	3345.3	3227.0	241	62891
D7S820	3578.4	3243.6	3369.8	193	60150
D8S1179	3453.5	3199.5	3236.5	519	59377
FGA	2055.7	2014.6	1899.9	312	37305
PentaD	1363.8	1094.3	1211.3	251	18191
PentaE	1951.0	2547.3	1721.7	268	47456
TH01	1563.3	1463.5	1457.0	109	27037
TPOX	2366.0	1921.6	2194.4	187	34682
vWA	2298.2	2816.4	2129.9	284	52848
Locus	Mean coverage	Standard deviation	Median coverage	Min coverage	Max coverage
DYS19	2085.1	2223.9	1947.7	170	41613
DYS385a,b	737.4	733.6	607.6	64	12563
DYS389I	895.0	947.7	831.7	42	17545
DYS389II	919.8	982.6	861.4	47	18246
DYS390	909.0	861.5	880.7	44	15887
DYS391	1710.2	1797.2	1507.5	388	33495
DYS392	1737.1	1708.6	1589.0	24	31736
DYS393	1645.6	1441.6	1495.2	312	25907
DYS437	1583.8	1953.7	1416.1	96	36391
DYS438	778.2	948.5	724.1	57	17847
DYS439	1873.9	1944.5	1718.6	292	36611
DYS448	400.7	502.8	373.4	24	9234
DYS456	1984.3	2278.5	1745.1	300	42723
DYS458	832.3	943.2	777.5	106	17776
DYS481	1328.3	1266.1	1228.2	127	23321
DYS533	1327.3	1066.2	1264.8	105	18874
DYS549	1724.6	1688.3	1580.8	165	31402
DYS570	1216.9	1059.5	1147.1	92	19467
DYS576	1415.3	1386.3	1287.9	251	25945
DYS635	2076.6	2454.8	1917.1	121	46724
DYS643	1018.2	1056.9	938.5	50	19713
Y-GATA-H4	1430.9	1513.9	1324.3	35	28246

a.



b.



Cont.

C.

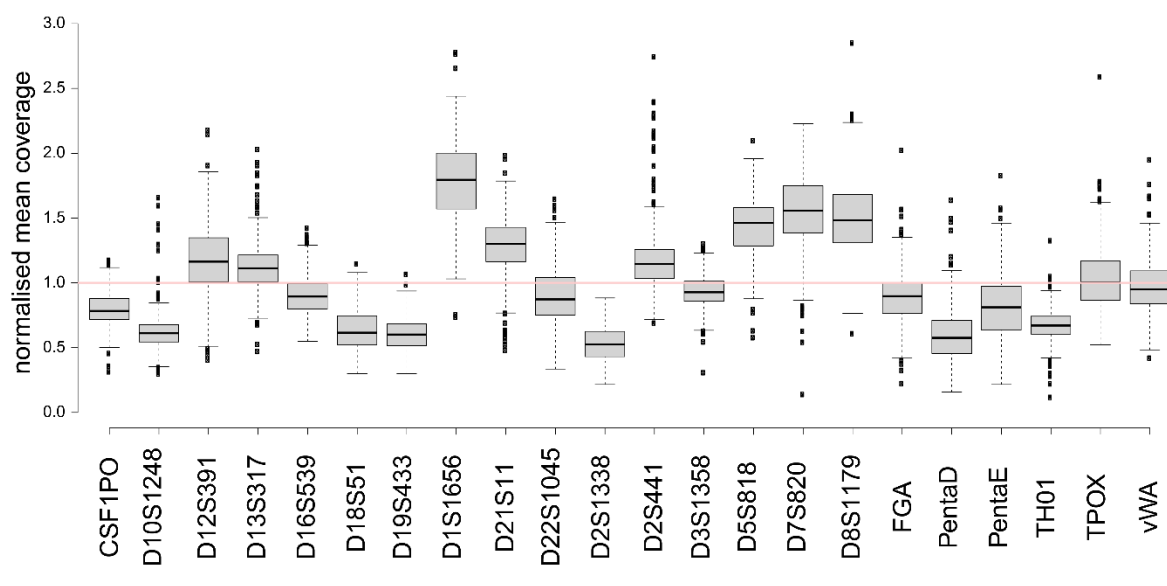


Figure 6.13 Normalised coverage of mtDNA and STR markers sequenced.

The boxplot above shows mean coverage for all 362 samples analysed, for each segment/loci, normalised to sample means. (a) the schematic representation shows the ten designed amplicons across the mtDNA control region, and their overlapping (grey) and non-overlapping (white) segments. Positions and sizes are approximately to scale. (b) normalised mean coverage values of sequenced Y-STR markers, (c) normalised mean coverage values of sequenced aSTR markers. Centre lines indicate the medians; box limits are the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by circles.

6.3.2 MtDNA control region population data in the UK

6.3.2.1 Observed variants and haplotypes

To improve the prediction of haplogroups from the CR sequence data, both EMPOP (Huber et al. 2018; Rock et al. 2013) and HaploGrep2 (Weissensteiner et al. 2016b) were used independently and compared in predictions. These two approaches were found to be concordant: 86% matched exactly (meaning the main and the high definition haplogroup designations matched), and the remaining 14% matched at the broader main haplogroup level, meaning that in these cases one predictor was slightly more specific than the other.

To compare with independently typed data from the same dataset, 268 of these samples were compared to available SNP chip data from PoBI, which included targeted mtDNA variants, and which allowed a broad prediction of mitochondrial haplogroups based on SNPs across the whole mtDNA (J.Wetton personal comm., unpublished data).

Variants located in the control region were compared between the MPS-based and SNP chip-based datasets, and 99% of the samples showed false negative calls and 56% of the samples showed between one and eight false positive calls in the SNP chip-based dataset. This massive difference is likely due to the different techniques, as the probe-design for SNP chips is not aimed to capture all variants, and is likely to be especially affected in the highly variable CR, whereas sequencing is much better suited to capture variation.

However, outside of the control region a SNP chip approach can be more successful in targeting haplogroup defining mutations, and therefore instead of using one-to-one variant comparisons in the CR only, the predicted haplogroups of the mtDNA from sequencing of the CR were compared with the SNP chip-based haplogroup designations from whole mtDNA (Table 6.6). Concordance using this method is less definitive, but overall 96% of the samples were either concordant or broadly correctly predicted (as an example of the latter: H compared to H27, or U

compared to U6). Only 4% of the compared samples showed discordant haplogroups (all related to predictions by SNP chip within the R superhaplogroup, such as different H, HV and B predictions, which were predicted by MPS(CR) to be within V, K, P, U and H, Table 6.6) between SNP-based and MPS-based methods. The discordance is likely due to the limited predictive power of using CR variants only; as characteristic variants in the coding region remain untested. Alternatively discordant results can be a systematic error in prediction tools or errors originating from sample handling.

Table 6.6
Discordant haplogroup predictions.











The table shows the discordant haplogroup predictions from MPS targeting the CR vs SNP chip targeting a defined set of variants along the whole mtDNA.

	MPS (CR)	SNP chip (whole mtDNA)
NIR005	V+@72	H1
ARG002	V19	HV0a
HAM006	V+@72	HV0a
NPE062	V+@16298	HV0a
NPE064	V10a	HV0a
FOD061	U6	HV
SUS443	H3h7	HV
YOR008	H1e1a4	HV
WOR013	H2a2a1	B4b'd'e'j
KEN084	K2a11	H5'36
CHE004	P2	H4a1a1a1
WOR046	P2	H4a1a1

In the 362 analysed samples, 308 distinct mtDNA CR haplotypes were found, belonging to 13 main haplogroups (Figure 6.14). A summary of the main haplogroup frequencies for each of the ten large geographical regions is shown in Table 6.7, and details of haplogroup predictions for each sample are detailed in Appendix F. The range of haplogroups observed in the sample is typical of that observed in previous surveys of western Europe (Richards et al. 2000).

Table 6.7

Frequencies of predicted haplogroups by MPS targeting the CR for the ten large geographical regions.

										
	SCO	IRE	NW	NE	WAL	WM	EM	EA	SW	SE
H(xH1,H2)	0.105	0.091	0.233	0.167	0.152	0.138	0.250	0.207	0.207	0.125
H1	0.289	0.182	0.200	0.167	0.152	0.207	0.205	0.138	0.190	0.208
H2a	0.132	0.182	0.067	0.056	0.087	0.103	0.045	0.172	0.103	0.188
HV	0.026	0.000	0.000	0.000	0.022	0.000	0.045	0.000	0.052	0.000
I	0.026	0.000	0.033	0.000	0.000	0.034	0.000	0.069	0.017	0.042
I1a1	0.000	0.000	0.000	0.000	0.065	0.034	0.000	0.000	0.000	0.021
J1	0.026	0.045	0.000	0.000	0.000	0.034	0.000	0.000	0.000	0.000
J1b	0.026	0.000	0.000	0.000	0.022	0.000	0.000	0.000	0.000	0.000
J1c	0.132	0.091	0.033	0.056	0.109	0.034	0.068	0.034	0.086	0.083
J2	0.000	0.000	0.000	0.000	0.000	0.034	0.023	0.034	0.017	0.042
K	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
K1	0.026	0.136	0.033	0.000	0.043	0.034	0.091	0.034	0.052	0.000
K2	0.000	0.045	0.033	0.056	0.000	0.000	0.023	0.034	0.034	0.042
N1a1	0.000	0.000	0.000	0.000	0.000	0.000	0.023	0.000	0.000	0.021
P2	0.000	0.000	0.033	0.000	0.000	0.034	0.000	0.000	0.000	0.000
R9	0.000	0.000	0.000	0.000	0.000	0.034	0.000	0.000	0.000	0.000
T	0.000	0.000	0.067	0.056	0.087	0.000	0.000	0.034	0.000	0.000
T1	0.000	0.000	0.033	0.167	0.022	0.034	0.000	0.069	0.000	0.021
T2	0.000	0.000	0.033	0.111	0.065	0.034	0.045	0.069	0.086	0.021
U2 U3,U4,U6,U7	0.053	0.045	0.033	0.000	0.022	0.034	0.023	0.034	0.052	0.000
U5a	0.053	0.091	0.067	0.167	0.043	0.034	0.068	0.034	0.052	0.104
U5b	0.053	0.000	0.000	0.000	0.022	0.138	0.091	0.000	0.017	0.042
V	0.026	0.045	0.000	0.000	0.043	0.000	0.000	0.000	0.000	0.021
W	0.000	0.000	0.067	0.000	0.022	0.000	0.000	0.034	0.000	0.021
X2	0.000	0.045	0.033	0.000	0.022	0.000	0.000	0.000	0.034	0.000
n=362	38	22	30	18	46	29	44	29	58	48

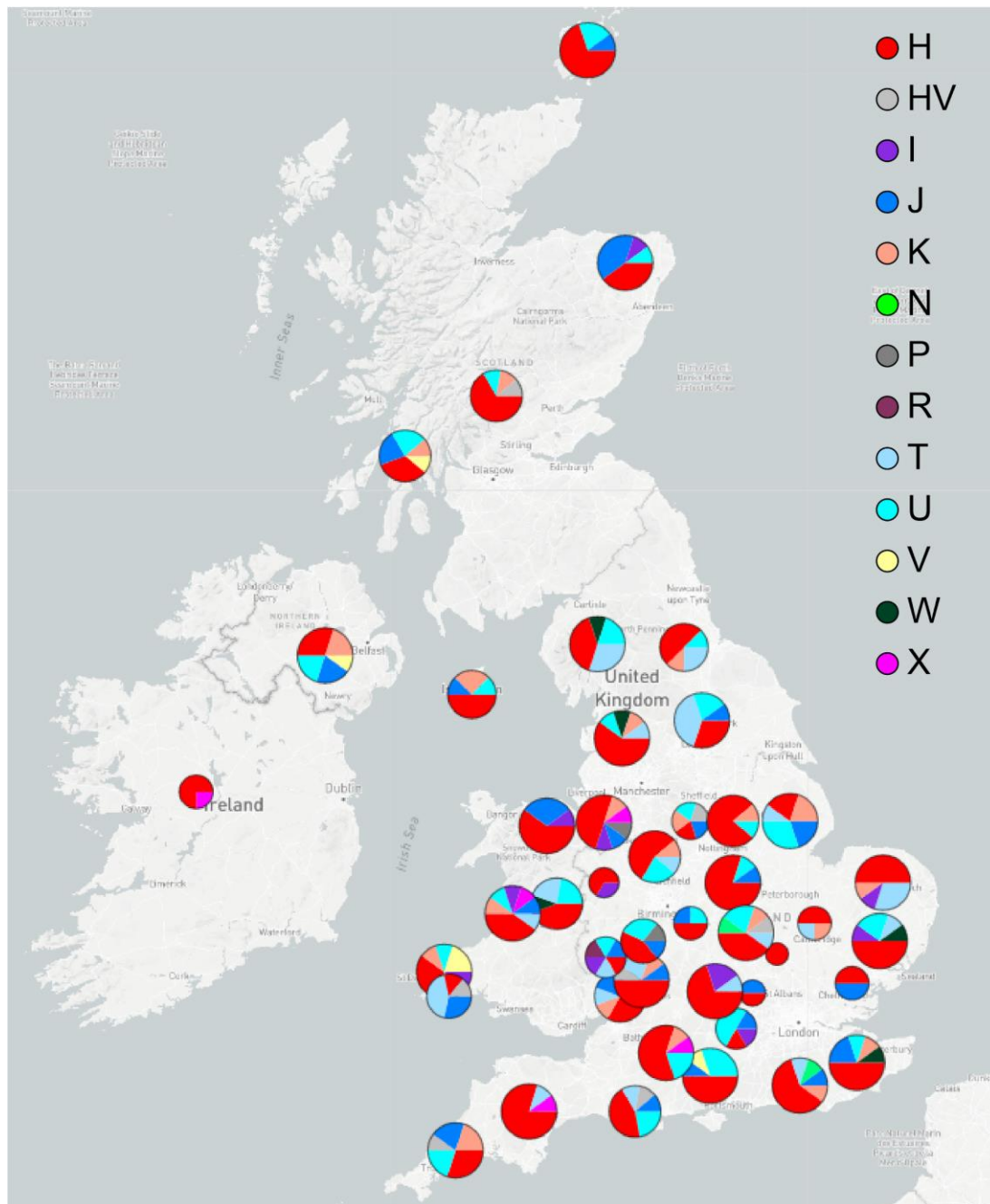


Figure 6.14
Geographical distribution of predicted mtDNA haplogroups
across the British Isles.



The map is generated using Microreact (Argimon et al. 2016) and it can be better explored interactively at <https://microreact.org/project/jlcJm9Kqn> or by scanning this QR code.

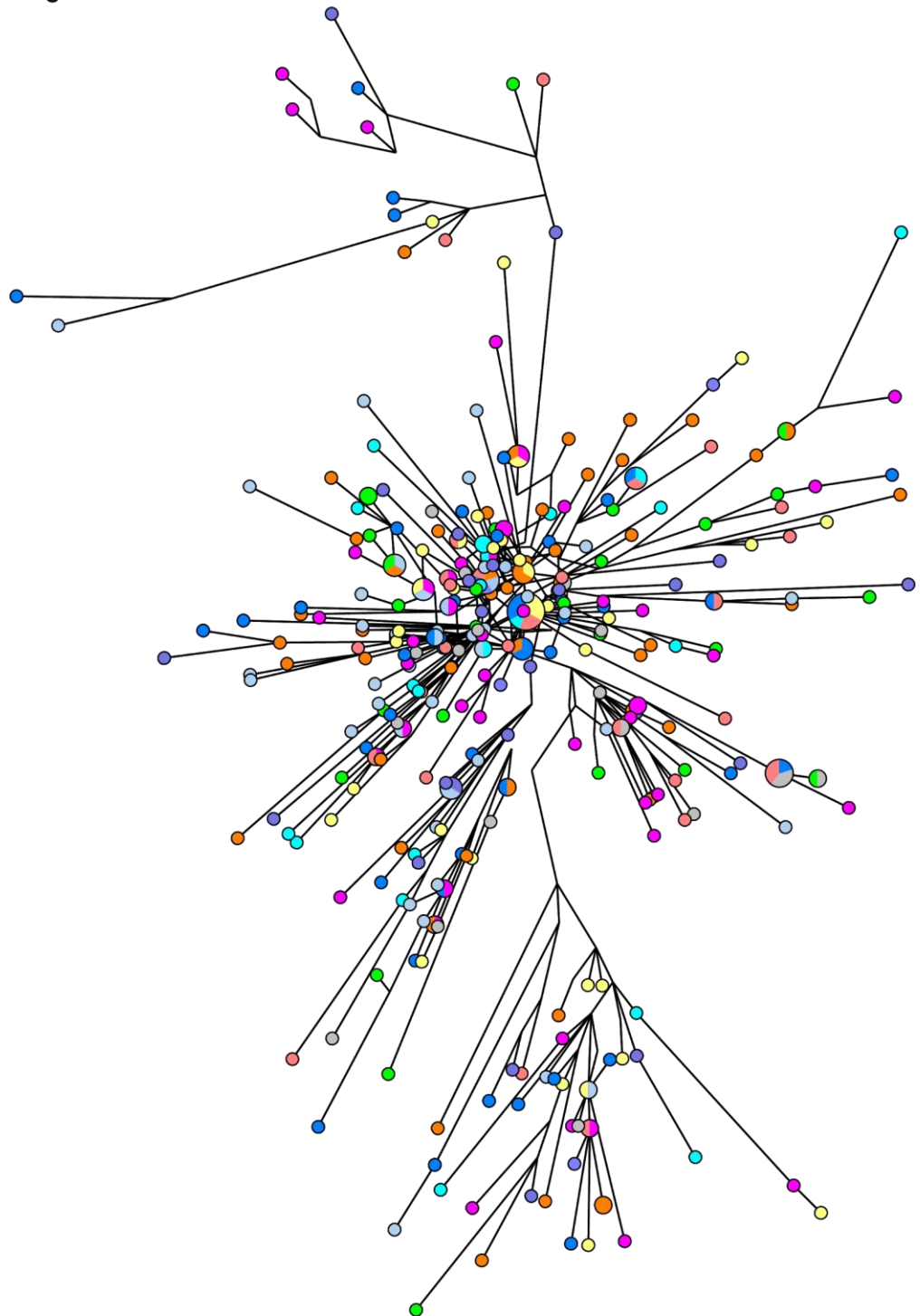
6.3.2.2 Median-joining (MJ) networks of mtDNA CR variants

The observed CR variants were used to build a phylogenetic MJ-network for 362 samples. Variable positions were coded into a 'DNA nucleotide data' input file for the software Network 5.0.1.1. In the built network the circle (node) areas are proportional to the number of samples and the lengths of the branches represent the number of differences between haplotypes. The network was annotated using categories of geographic origin or different levels of haplogroup designations using Network Publisher 2.1.2.5 commercial software. The mtDNA CR variants of 362 samples in a network show no obvious stratification relating to geographic origin (Figure 6.15a), but correlate well with the predicted main haplogroups (Figure 6.15b) and the more resolved haplogroup designations (Figure 6.15c).

a.

Geographic regions

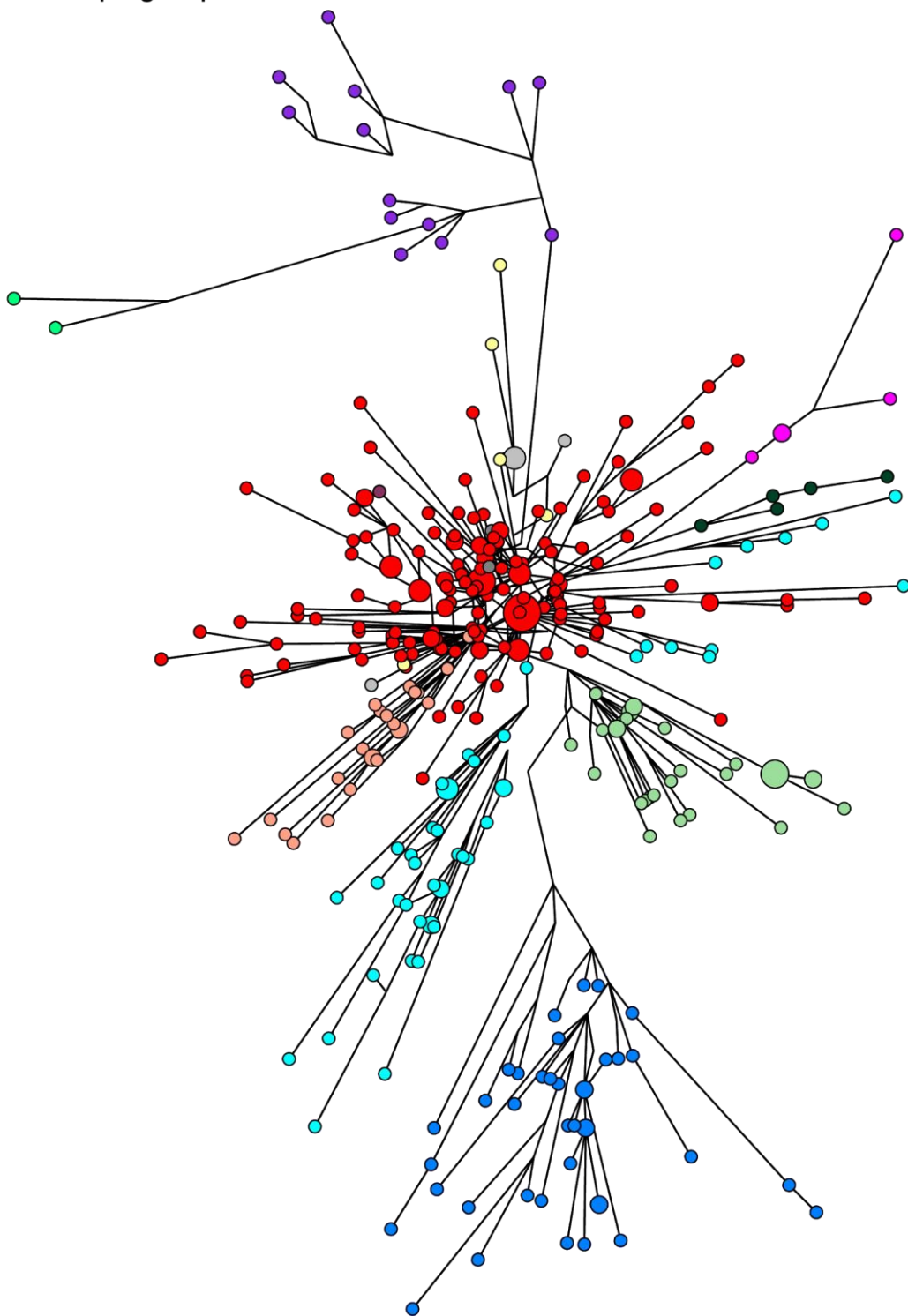
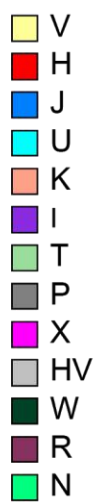
- SCO
- IRE
- NW
- NE
- WAL
- WM
- EM
- EA
- SW
- SE



Cont.

b.

Main mtDNA haplogroups



Cont.

c.

mtDNA haplogroups

- H1
- H2a
- H(xH1,H2)
- HV
- I
- I1a1
- J1
- J1b
- J1c
- J2
- K
- K1
- K2
- N1a1
- P2
- R9
- T
- T1
- T2
- U2,U3,U4,U6,U7
- U5a
- U5b
- V
- W
- X2

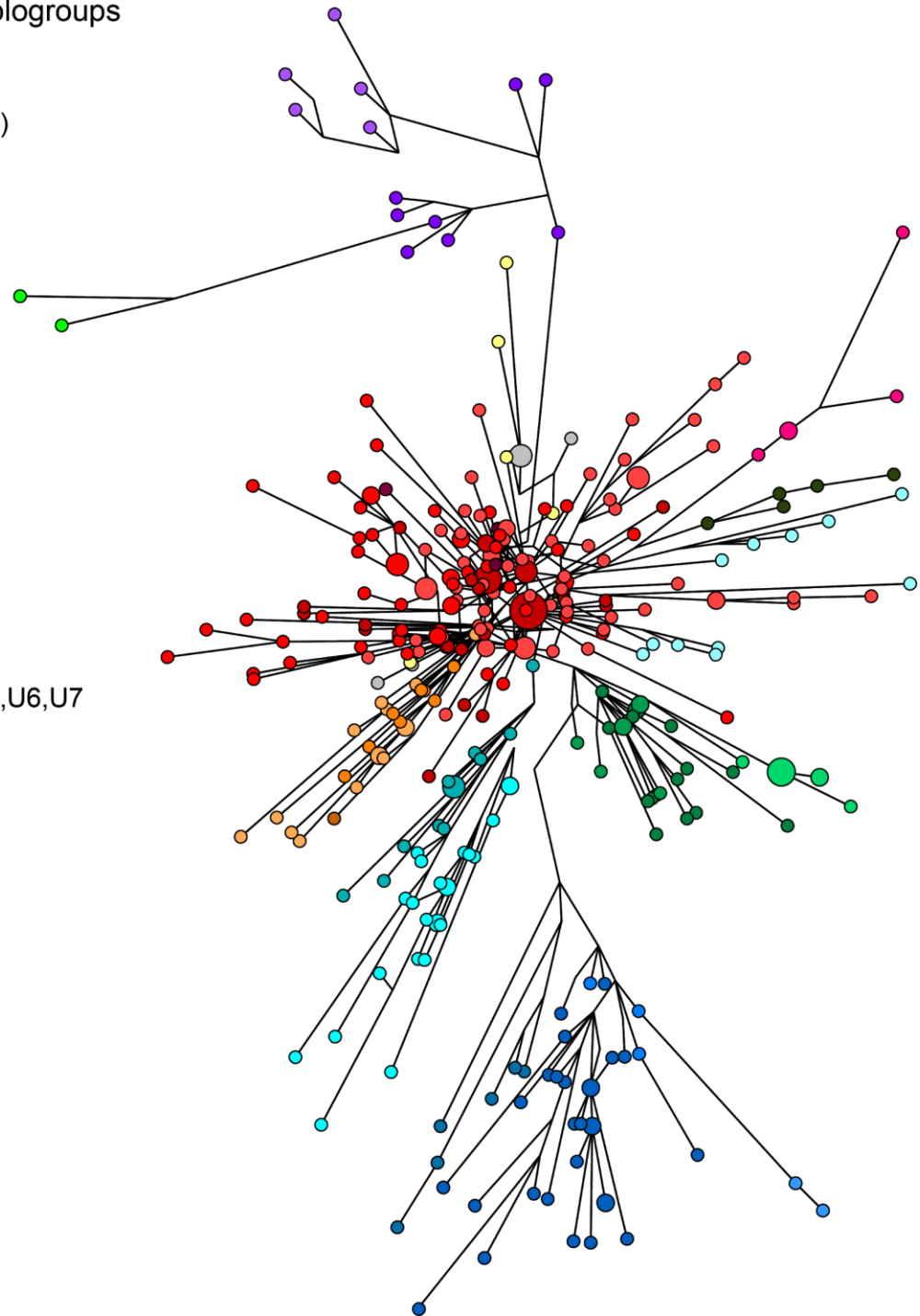


Figure 6.15

Median-joining network of samples with mtDNA CR variants.

(a) is coloured by geographic regions, (b) is coloured by main haplogroups, and (c) is coloured by finer haplogroup levels.

6.3.2.3 Testing population differentiation using mtDNA CR variants

The observed CR variants were used to test population differentiation in the 362 samples. Variable positions were input to the Arlequin v 3.5.2.2 software (Excoffier and Lischer 2010), and summary statistics were generated including molecular diversity indices, F-statistics for population comparisons and differentiation (Table 6.8). The different population divisions tested are detailed in Section 6.2.2.

Table 6.8
MtDNA CR variants molecular statistics, Analysis of MOlecular VAriance (AMOVA) statistics and pairwise F_{ST} values.

Table (a) shows the molecular statistics and (b) shows the AMOVA values using pairwise F_{ST} values for each division. (c) and (d) shows the pairwise F_{ST} and haplotype diversity matrices per multi-component divisions.

a.

Basic statistics	Observed events	Observed sites
Substitutions	200	190
Transitions	178	178
Transversions	22	22
Indels	22	22
Number of polymorphic sites		212
Number of haplotypes	308	
Haplotype diversity(\pm SD)	0.9985 (\pm 0.0005)	
Haplotype match probability	0.0043	

b.

AMOVA	F_{ST}	P	Variation between populations	Variation within populations
37 populations	0.00063	0.44277 (\pm 0.00454)	0.06%	99.94%
10 geographic regions	0	0.53763 (\pm 0.00490)	0%	100%
5 PoBI defined regions	0	0.62495 (\pm 0.00442)	0%	100%
Pairwise	F_{ST}	P		
Celtic fringe vs rest	0.00289	0.07138 (\pm 0.0026)	0.29%	99.71%
Danelaw vs rest	0.00078	0.28868 (\pm 0.0043)	0.08%	99.92%
CSE England vs rest	0	0.54143 (\pm 0.0052)	0%	100%
East vs rest	0.003	0.06039 (\pm 0.0024)	0.30%	99.70%

Cont.

C.

Population pairwise F_{ST} s

Pairwise difference

Significance Level=0.000909 (Bonferroni-corrected)

Number of permutations : 10100

above diagonal:

diagonal:

below diagonal:

P-values (\pm SD)

haplotype diversity (\pm SD)

F_{ST} values

	SCO	IRE	NW	NE	WAL	WM	EM	EA	SW	SE
SCO	HD=0.9957 \pm 0.0079	0.92318 \pm 0.0028	0.13424 \pm 0.0031	0.05712 \pm 0.0023	0.66261 \pm 0.0042	0.55579 \pm 0.0053	0.81725 \pm 0.0034	0.10405 \pm 0.0030	0.64528 \pm 0.0046	0.82922 \pm 0.0034
IRE	0	HD=0.9957 \pm 0.0153	0.53787 \pm 0.0051	0.10524 \pm 0.0027	0.66300 \pm 0.0048	0.64895 \pm 0.0047	0.81259 \pm 0.0038	0.20552 \pm 0.0041	0.96396 \pm 0.0016	0.60123 \pm 0.0047
NW	0.00907	0	HD=0.9977 \pm 0.0094	0.18582 \pm 0.0037	0.23899 \pm 0.0044	0.63330 \pm 0.0048	0.47926 \pm 0.0052	0.51619 \pm 0.0046	0.83576 \pm 0.0040	0.12652 \pm 0.0037
NE	0.02349	0.02038	0.01077	HD=0.9935 \pm 0.0210	0.51242 \pm 0.0048	0.12623 \pm 0.0032	0.10613 \pm 0.0033	0.82507 \pm 0.0041	0.15573 \pm 0.0039	0.21889 \pm 0.0043
WAL	0	0	0.00453	0	HD=0.9981 \pm 0.0050	0.51233 \pm 0.0052	0.34997 \pm 0.0046	0.47035 \pm 0.0047	0.33125 \pm 0.0043	0.80616 \pm 0.0039
WM	0	0	0	0.01448	0	HD=1.0000 \pm 0.0091	0.79220 \pm 0.0041	0.30957 \pm 0.0044	0.66657 \pm 0.0046	0.73735 \pm 0.0047
EM	0	0	0	0.01389	0.00153	0	HD=0.9989 \pm 0.0051	0.06782 \pm 0.0026	0.44946 \pm 0.0047	0.62687 \pm 0.0048
EA	0.01183	0.00834	0	0	0	0.00372	0.01181	HD=0.9951 \pm 0.0106	0.48084 \pm 0.0052	0.51084 \pm 0.0049
SW	0	0	0	0.00984	0.00169	0	0	0	HD=0.9988 \pm 0.0035	0.39778 \pm 0.0047
SE	0	0	0.00832	0.00745	0	0	0	0	0.00056	HD=0.9965 \pm 0.0055

d.

Population pairwise F_{ST} s

Pairwise difference

Significance Level=0.005 (Bonferroni-corrected)

Number of permutations : 10100

above diagonal: P-values (\pm SD)

diagonal: haplotype diversity (\pm SD)

below diagonal: F_{ST} values

	Orkney	Wales	Scotland, Islands and North of England	Cornwall	England (except above)
Orkney	HD=0.9778 \pm 0.0540	0.76547 \pm 0.0043	0.92813 \pm 0.0026	0.56994 \pm 0.0054	0.99238 \pm 0.0008
Wales	0	HD=0.9981 \pm 0.0050	0.80061 \pm 0.0041	0.27720 \pm 0.0040	0.32155 \pm 0.0047
Scotland, Islands and North of England	0	0	HD=0.9991 \pm 0.0027	0.28175 \pm 0.0042	0.59153 \pm 0.0048
Cornwall	0	0.00706	0.00718	HD=0.9778 \pm 0.0540	0.11375 \pm 0.0030
England (except above)	0	0.0011	0	0.01875	HD=0.9984 \pm 0.0007

For the mtDNA CR variants none of the applied divisions showed significant differentiation within the whole PoBI dataset, including the pairwise F_{ST} and haplotype diversity matrix for the 37 population division which is supplied as Appendix G. The mtDNA CR data from 362 samples suggests a high degree of homogeneity, and no stratification across the British Isles.

6.3.3 Y-STR sequence variation population data in the UK

6.3.3.1 Observed variants and haplotypes

Y-STR alleles were called as detailed in Chapter 2, Section 2.3.4. A total of 8334 alleles were analysed in 362 samples. For calculations at the DYS385a,b loci the presence of two alleles was assumed for each homoallelic combination. The length-based frequencies are summarised in Table 6.9. Despite the wide range of alleles sequenced in Chapter 3, there were 49 alleles in this population set which were not observed in the global dataset (Appendix F), sixteen of these are singletons, the result of SNPs in the flanking regions or internal to the arrays, which is a type of variant one might expect to encounter occasionally regardless how thoroughly sampled the populations are. Nine of these alleles displayed allele lengths at the periphery of the previously observed allele size ranges; similarly, such alleles are usually rare and only captured when large numbers of individuals are sampled.

The called variants included eight supernumerary alleles at six loci (Table 6.10). An example of the distribution of observed sequence alleles is shown per geographic region in Table 6.11. Complete tables listing all loci, the visualisation of variation and the sequence strings are provided in Appendix F. The increase in allele diversity achieved by MPS compared to CE is shown in Figure 6.16.

The desktop version of the Y-STR-based NevGen Y-DNA Haplogroup Predictor software (Gentula and Nevski, Serbian DNA Project, 2015, available at: <https://www.nevgen.org/>) was used to predict haplogroups from the 362 samples. To compare with independently typed data from the same dataset, 267 of these samples were compared to available SNP chip data targeting Y-chromosomal variants, which allowed precise haplogroup designations (J.Wetton personal comm., unpublished data). Predictions were completely concordant; however, less than 3% of the samples, all from hg I1, were predicted as I1, but with low probabilities given by the predictor; however, these still matched the SNP-based predictions. This is likely a result of certain hg I1 sub-lineages deviating significantly from the main cluster of I1 haplotypes in the NevGen predictor's comparative

database, making predictions for these uncertain. Altogether, 360 distinct Y-STR haplotypes were found in the 362 analysed samples, belonging to 10 major predicted haplogroups (Table 6.12 and Figure 6.17). Two sample pairs shared variants of both length and sequence.

Table 6.9
Length-based Y-STR allele frequencies from 362 samples.

	DYS19	DYS385a,b	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439
8										0.0028	
9						0.0083				0.0359	
10		0.0055				0.4765				0.2320	0.0635
11	0.0028	0.3191	0.0028			0.4931	0.2541	0.0028		0.0691	0.3301
12		0.0663	0.2017			0.0193	0.0608	0.0414		0.6271	0.4613
12.1		0.0014									
13	0.0166	0.0870	0.5967			0.0028	0.6381	0.8011	0.0028	0.0331	0.1340
14	0.7017	0.3481	0.1906				0.0442	0.1271	0.1796		0.0110
15	0.2127	0.1133	0.0083				0.0028	0.0276	0.6519		
16	0.0635	0.0331							0.1630		
17	0.0028	0.0138							0.0028		
18		0.0097									
19		0.0028									
20											
21					0.0083						
22					0.0994						
23					0.3370						
24					0.4365						
25					0.1133						
26					0.0028						
27				0.0110	0.0028						
28				0.2017							
29				0.4517							
30				0.2390							
31				0.0718							
32				0.0138							
33				0.0083							
34				0.0028							


	DYS448	DYS456	DYS458	DYS481	DYS533	DYS549	DYS570	DYS576	DYS635	DYS643	Y-GATA-H4
8										0.0028	
9					0.0221					0.0525	
10					0.0166	0.0110				0.6630	0.0387
11					0.1989	0.0704				0.0691	0.2928
12					0.5773	0.4365				0.1796	0.5773
12.1											
13		0.0193	0.0028		0.1519	0.3743				0.0331	0.0884
14		0.1575	0.0193		0.0331	0.0967	0.0083	0.0083			0.0028
15		0.3177	0.1575			0.0110	0.0083	0.0359			
16		0.3315	0.2251				0.1257	0.1768			
17	0.0083	0.1492	0.3605				0.4144	0.3011			
18	0.0856	0.0221	0.1768	0.0028			0.2279	0.2901			
19	0.6105	0.0028	0.0497				0.1243	0.1381	0.0110		
20	0.2431		0.0083	0.0166			0.0525	0.0414	0.0276		
21	0.0414			0.0773			0.0331	0.0083	0.1519		
22	0.0110			0.4807			0.0028		0.0663		
23				0.1796			0.0028		0.5497		
24				0.0663					0.1464		
25				0.1133					0.0470		
26				0.0221							
27				0.0249							
28				0.0166							
29											
30											
31											
32											
33											
34											

Table 6.10
Y-STR allele duplications observed in 8 samples.

region	ID	marker	CE	alleles (M1/M2)	% of M1	pos of M2 to M1	N (M1) in set	n of N in set with stutter at pos	mean stutter% at pos in n
WAL	MWA139	DYS389II	29	CE29_TAGA[10]CAGA[3]N[48]TAGA[11]CAGA[5]	100%		140	0	0.0%
			30	CE30_TAGA[10]CAGA[3]N[48]TAGA[12]CAGA[5]	80%	+1			
NW	CHE036	DYS391	10	CE10_TCTA[10]	100%		171	3	0.7%
			11	CE11_TCTA[11]	84%	+1			
EA	NOR163	DYS439	11	CE11_GATA[11]	100%		112	86	1.1%
			12	CE12_GATA[12]	29%	+1			
WAL	NPE029	DYS439	12	CE12_GATA[12]	100%		167	126	1.1%
			13	CE13_GATA[13]	20%	+1			
SCO	BAN008	DYS458	17	CE17_GAAA[17]	100%		131	130	16.5%
			16	CE16_GAAA[16]	49%	-1			
WAL	NPE012	DYS549	11	CE11_GATA[11]	100%		25	0	0.0%
			13	CE13_GATA[13]	3%	+2			
SCO	ARG032	DYS570	17	CE17_TTTC[17]	100%		151	63	1.3%
			18	CE18_TTTC[18]	34%	+1			
EM	LIN700	DYS570	17	CE17_TTTC[17]	100%		151	149	9.7%
			16	CE16_TTTC[16]	40%	-1			

Table 6.11
Example of Y-STR sequence allele variants for the ten large geographical regions.

The complete table listing all loci is in Appendix F



DYS19	CE	SCO	IRE	NW	NE	WAL	WM	EM	EA	SW	SE	ALL
CE11_TCTA[8]CCTA[1]TCTA[3]	11						1					1
CE13_TCTA[10]CCTA[1]TCTA[3]	13				1	1	1	1		1	1	6
CE14_TCTA[11]CCTA[1]TCTA[3]	14	27	15	21	13	35	22	31	17	41	32	254
CE15_TCTA[12]CCTA[1]TCTA[3]	15	6	4	6	4	8	4	9	10	13	12	76
CE15_TCTA[13]CCTA[1]TCTA[2]	15								1			1
CE16_TCTA[13]CCTA[1]TCTA[3]	16	4	3	3		2	1	3	1	3	3	23
CE17_TCTA[14]CCTA[1]TCTA[3]	17	1										1
Sum		38	22	30	18	46	29	44	29	58	48	362

DYS385	CE	SCO	IRE	NW	NE	WAL	WM	EM	EA	SW	SE	ALL
CE10_AAGG[6]GAAA[10]	10	2			2							4
CE11_AAGG[6]GAAA[11]	11	28	12	18	6	33	22	25	17	41	28	230
CE11_AAGG[6]GAAA[11]_+3G>A@18,680,690 _rs369060795_+5A>G@18,680,692 _rs372849212	11					1						1
CE12.1_AAGG[6]GAAA[4]A[1]GAAA[8]	12									1		1
CE12_AAGG[6]GAAA[12]	12	7	6	2	2	5	2	11	3	6	2	46
CE12_AAGG[7]GAAA[11]	12					2						2
CE13_AAGG[6]GAAA[13]	13	3	6	3	5	8	7	8	4	11	8	63
CE14_AAGG[5]GAAA[15]	14									1		1
CE14_AAGG[6]GAAA[14]	14	24	11	24	15	30	18	29	17	44	39	251
CE15_AAGG[5]GAAA[16]	15									1		1
CE15_AAGG[6]GAAA[15]	15	9	7	10	3	7	6	11	10	7	11	81
CE16_AAGG[6]GAAA[16]	16	1	2	1	3	3		1	6	3	3	23
CE16_AAGG[6]GAAA[2]TAAA[1]GAAA[13]	16							1				1
CE17_AAGG[6]GAAA[17]	17	2				2	1			1	4	10
CE18_AAGG[6]GAAA[18]	18			2			2	1	1		1	7
CE19_AAGG[6]GAAA[19]	19					1		1				2
Sum		76	44	60	36	92	58	88	58	116	96	724

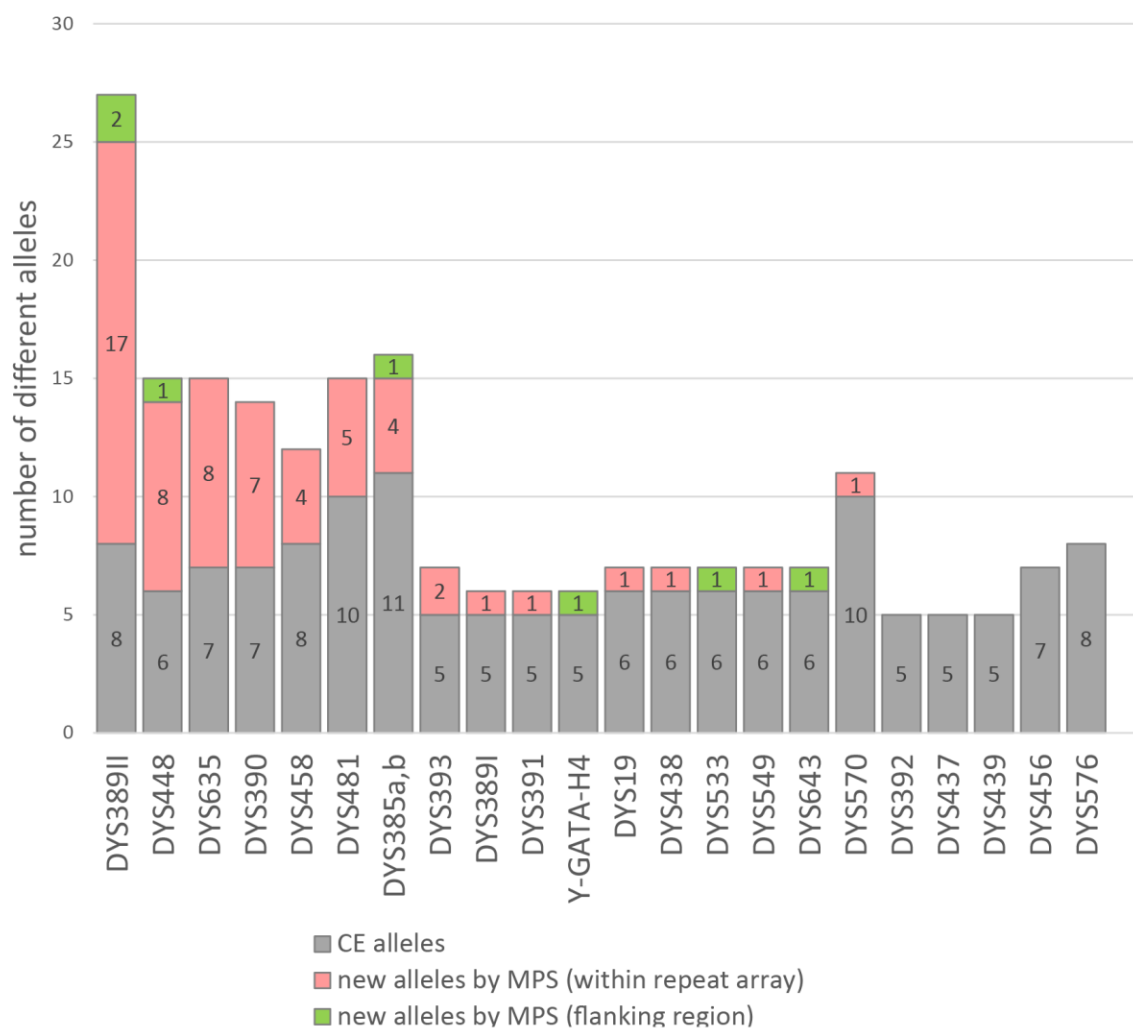



Figure 6.16
Increase in allele diversity of Y-STRs in 362 samples analysed by MPS.

Table 6.12

Frequencies of predicted Y-haplogroups by MPS for the ten large geographical regions.



	SCO	IRE	NW	NE	WAL	WM	EM	EA	SW	SE
E1b	0.000	0.000	0.000	0.056	0.022	0.069	0.023	0.000	0.000	0.021
F	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.021
G2a	0.000	0.091	0.067	0.056	0.022	0.034	0.023	0.034	0.000	0.021
I1	0.053	0.045	0.133	0.222	0.043	0.138	0.136	0.103	0.086	0.208
I2	0.000	0.000	0.033	0.111	0.000	0.000	0.000	0.034	0.000	0.000
I2a	0.026	0.136	0.067	0.111	0.043	0.000	0.023	0.138	0.086	0.104
J2a	0.000	0.000	0.033	0.056	0.022	0.000	0.023	0.034	0.017	0.000
J2b	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.034	0.000	0.042
N1c	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.017	0.000
R1a	0.158	0.045	0.033	0.000	0.022	0.000	0.045	0.034	0.052	0.021
R1b	0.737	0.682	0.633	0.389	0.826	0.759	0.705	0.586	0.741	0.563
T	0.000	0.000	0.000	0.000	0.000	0.000	0.023	0.000	0.000	0.000
n=362	38	22	30	18	46	29	44	29	58	48

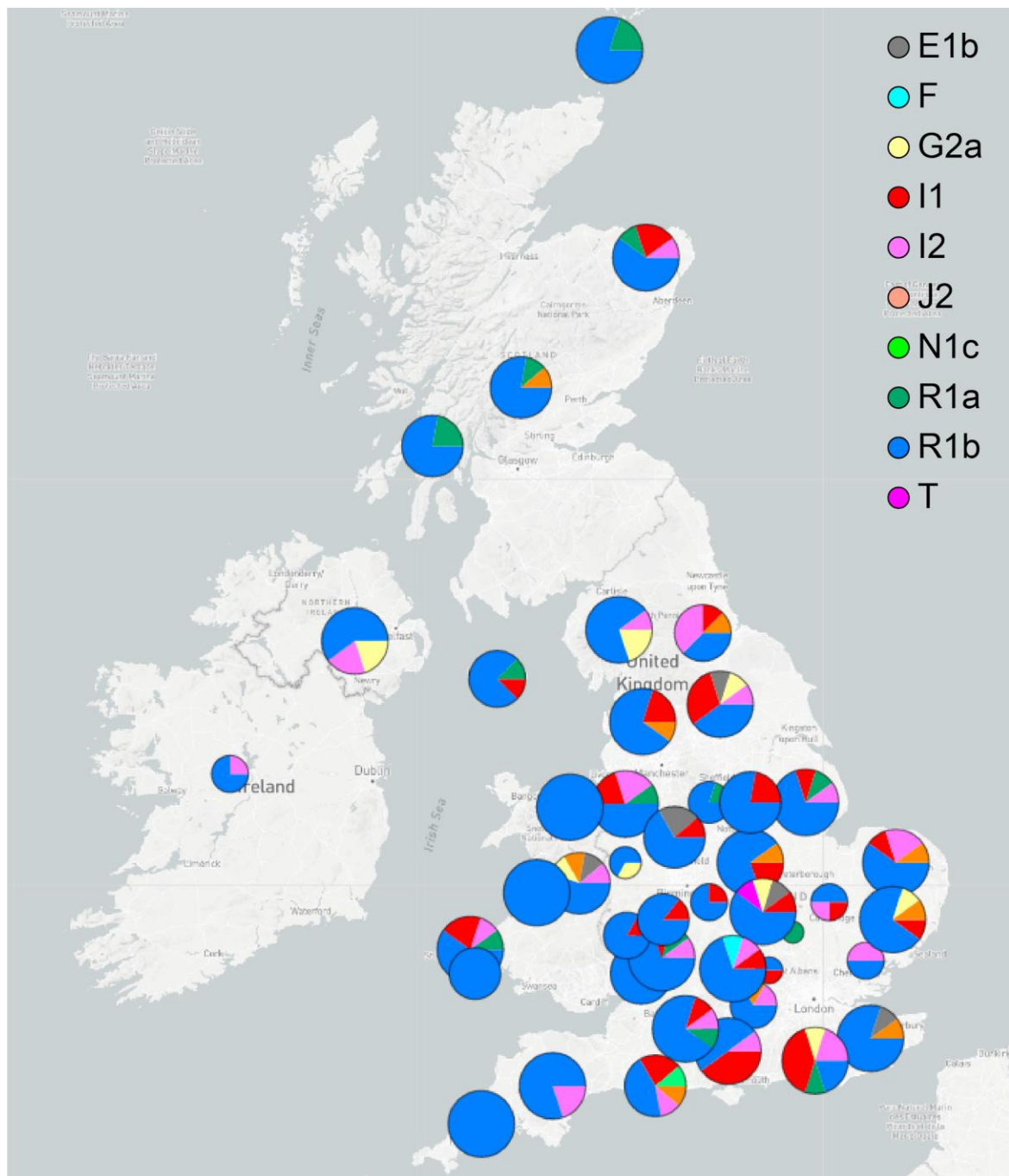


Figure 6.17
Geographical distribution of predicted Y-chromosomal haplogroups across the British Isles.

The map is generated using Microreact (Argimon et al. 2016) and it can be better explored interactively at <https://microreact.org/project/jlcJm9Kqn> or by scanning this QR code.



6.3.3.2 Median-joining (MJ) networks of Y-STR variants

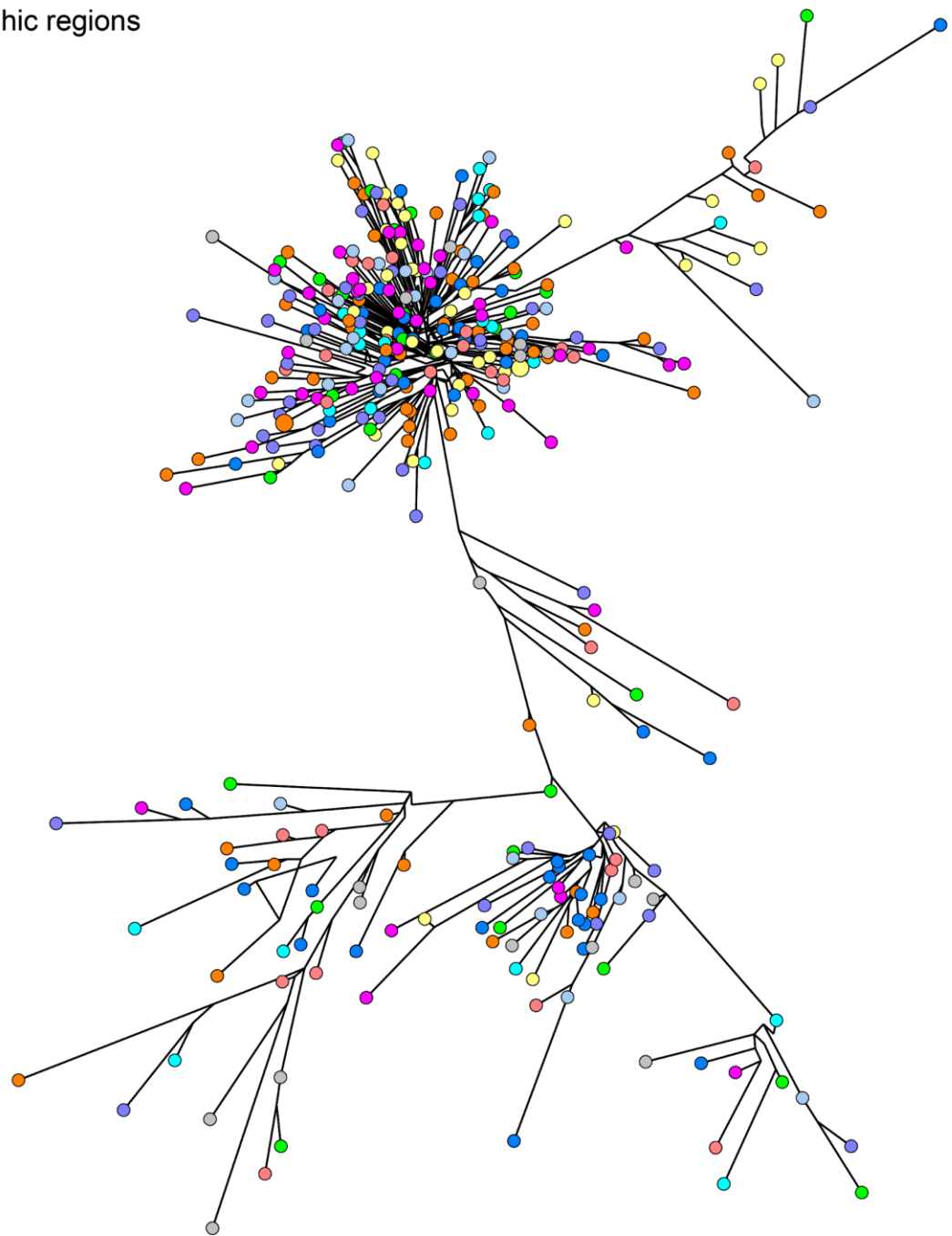
The observed Y-STR allele variants were used to build a MJ-network for 362 samples. Length-based alleles were coded into a 'STR data' input file for the software Network 5.0.1.1. Sequence-based alleles were coded into 'combined data' files of STR and binary data. Networks addressing only the sequence variation of these alleles regardless of the length background were coded into 'binary data' input files. In the built network the circle (node) areas are proportional to the number of samples and the lengths of the branches represent the number of differences between haplotypes. The network was annotated using categories of geographic origin or different levels of predicted haplogroup designations using Network Publisher 2.1.2.5 commercial software.

The Y-STR length-based variants of 362 samples in a network show no obvious stratification relating to the ten large geographic regions (Figure 6.18a), but correlate well with the predicted main haplogroups (Figure 6.18b) and similarly explain well the more resolved haplogroup designations (Figure 6.18c). The resolution of I1 and I2 lineages is not perfect by length and haplogroup E1b is associated with multiple clusters, which is in line with the large diversity observed within this haplogroup, but most major clusters are already well separated on the basis of length.

a.

Geographic regions

- SCO
- IRE
- NW
- NE
- WAL
- WM
- EM
- EA
- SW
- SE

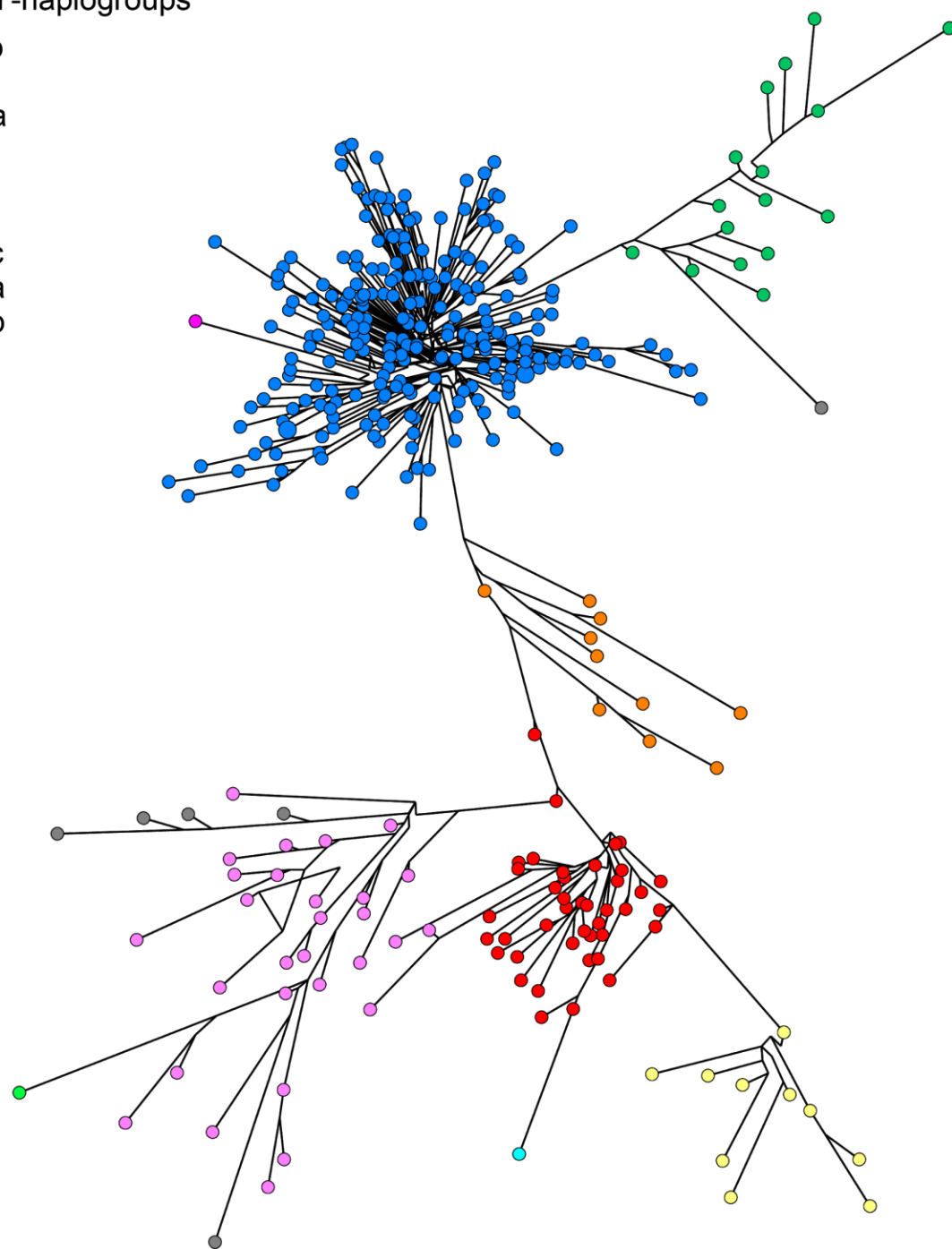


Cont.

b.

Main Y-haplogroups

- E1b
- F
- G2a
- I1
- I2
- J2
- N1c
- R1a
- R1b
- T



Cont.

c.

Y-haplogroups

- E1b1b
- F
- G2a
- G2a2
- G2a2b1
- I1
- I2a1
- I2a1a
- I2a2a
- I2a2b
- I2c1a2
- I2c2
- J2a1
- J2b1
- J2b2
- N1c
- R1a
- R1b
- T

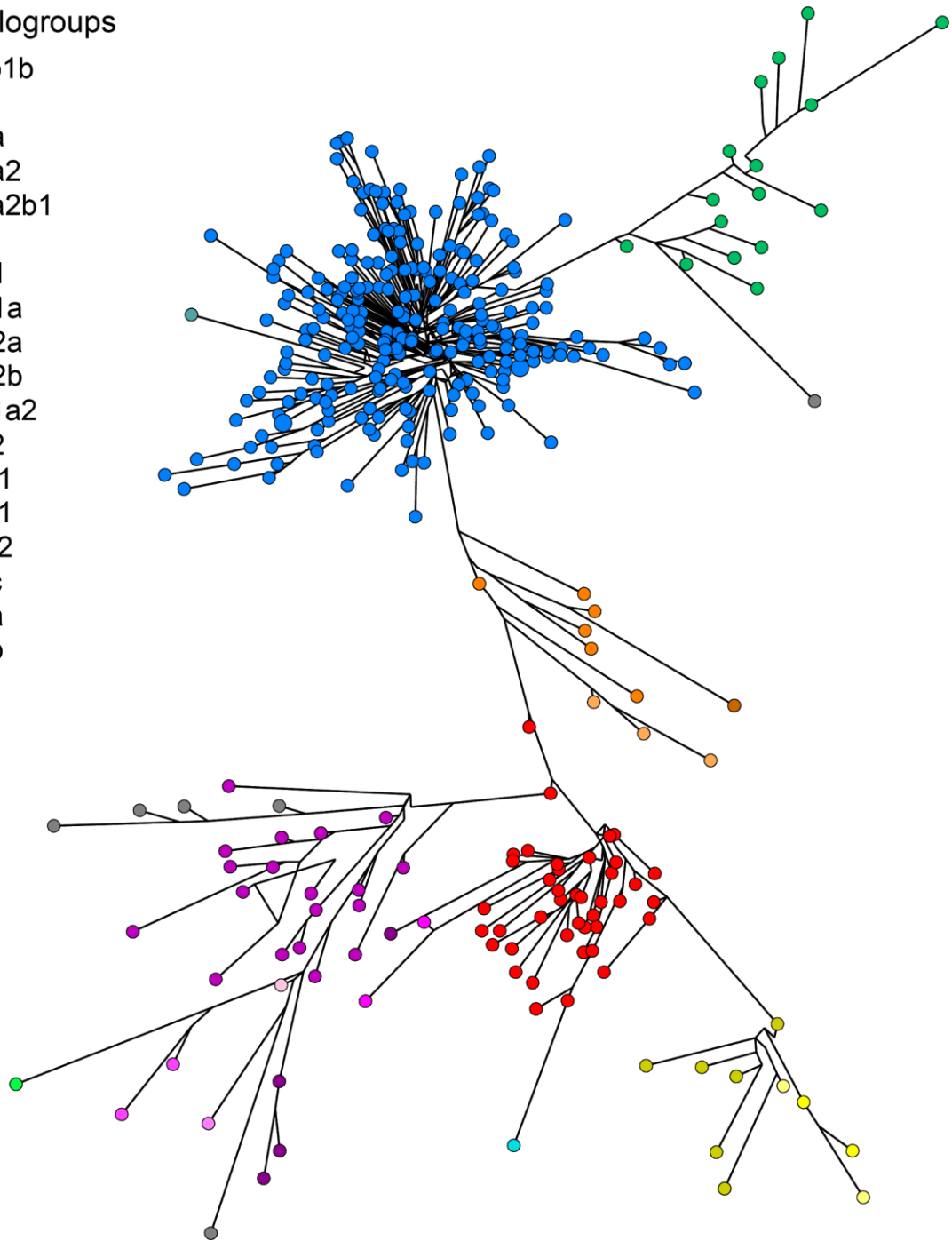


Figure 6.18

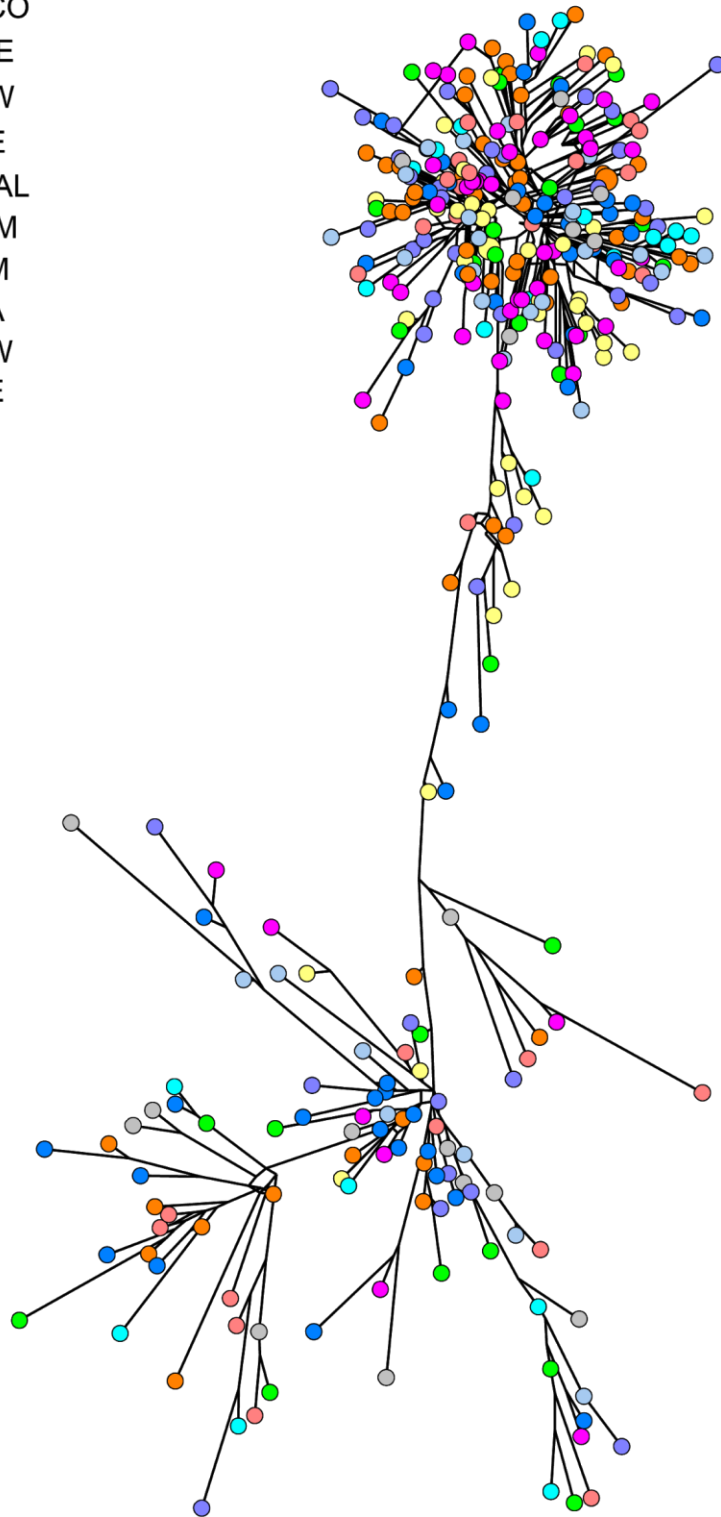
Median-joining network of samples with Y-STR length variants.

(a) is coloured by geographic regions, (b) is coloured by main haplogroups, and (c) is coloured by more highly resolved haplogroup levels.

When sequenced, Y-STR alleles can show a new layer of variation at the array level, repeat pattern variation (RPV), which involves changes in the number of repeat units in the different components of a compound array. These variations are driven by a similar, but slightly slower rate of mutation as the length of individual components is shorter than the full array. The effect of the inclusion of these RPs as variants on a network of these 362 samples is shown in Figure 6.19. Again, there is no obvious stratification relating to the ten large geographic regions (Figure 6.19a), but the clusters correlate well with the predicted main haplogroups (Figure 6.19b) and even better with the more resolved haplogroup designations (Figure 6.19c). With the addition of these RPV variants to the network the haplogroups R1a and R1b became more differentiated from the rest of the network and from each other; R1b particularly shows a slightly more reticulated structure than the length-based cluster. E1b haplotypes show better coherence, clustering nearer to each other when using sequence to analyse the phylogeny. The I1-I2 relationship gets more deeply nested: I1 separates more clearly from the different subclades of I2 by taking a more central position, however I2 subclades that are discernible via prediction do not seem to fully correlate to the subclades in the network.

a.

Geographic regions

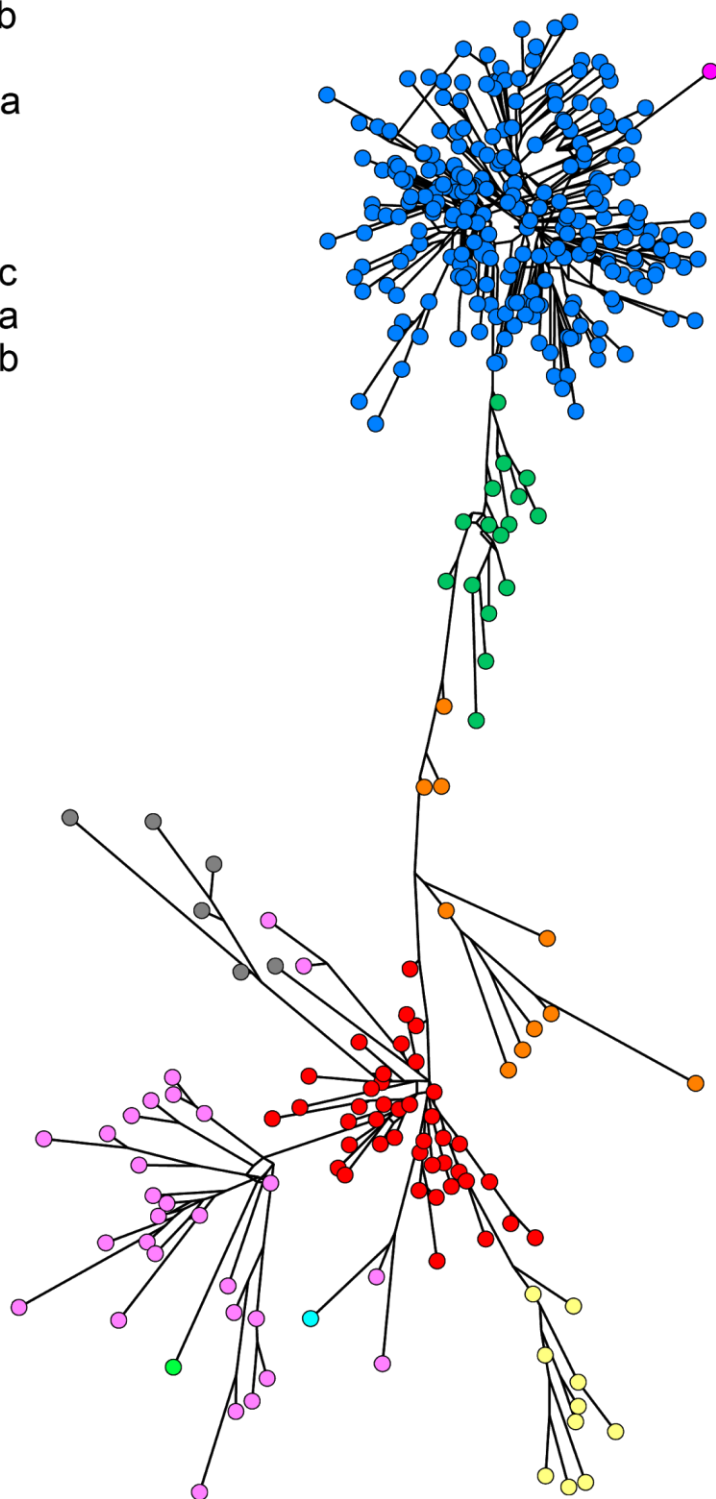


Cont.

b.

Main Y-haplogroups

- E1b
- F
- G2a
- I1
- I2
- J2
- N1c
- R1a
- R1b
- T



Cont.

c.

Y-haplogroups

- E1b1b
- F
- G2a
- G2a2
- G2a2b1
- I1
- I2a1
- I2a1a
- I2a2a
- I2a2b
- I2c1a2
- I2c2
- J2a1
- J2b1
- J2b2
- N1c
- R1a
- R1b
- T

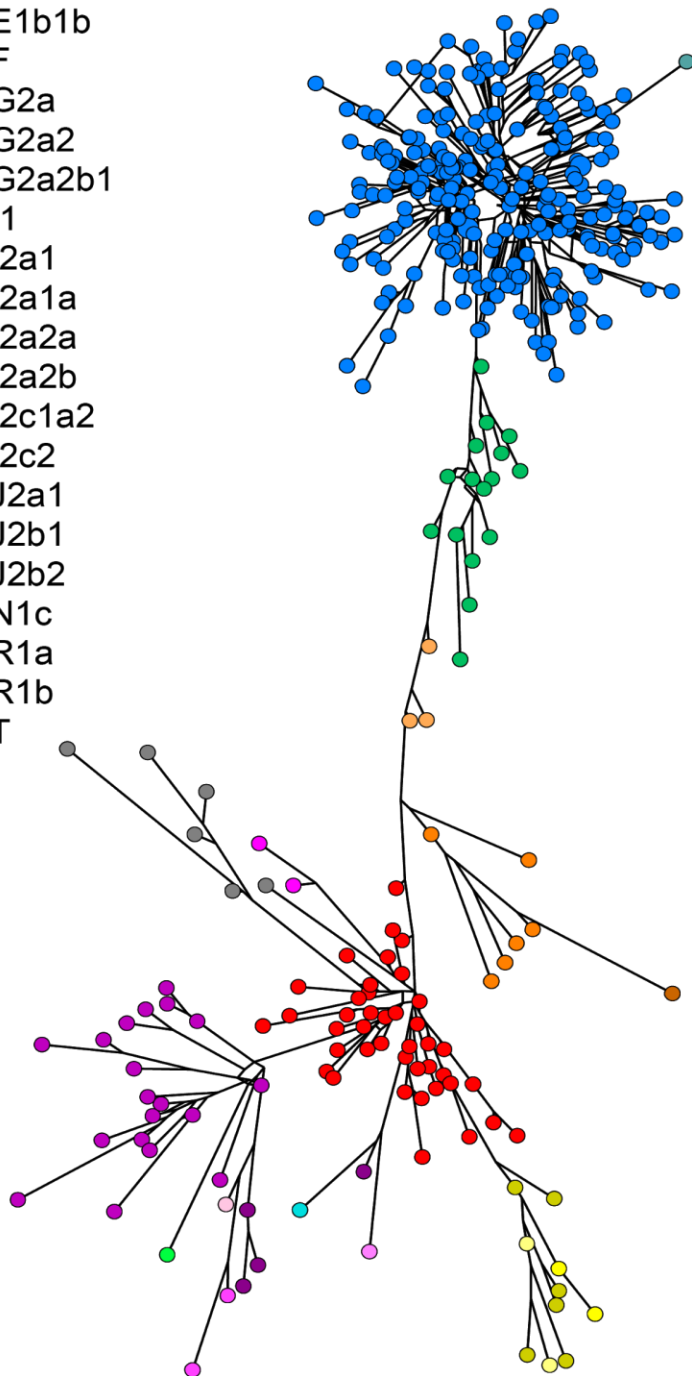


Figure 6.19

Median-joining network of samples with Y-STR sequence array variants.

(a) is coloured by geographic regions, (b) is coloured by main haplogroups, and (c) is coloured by more highly resolved haplogroup levels.

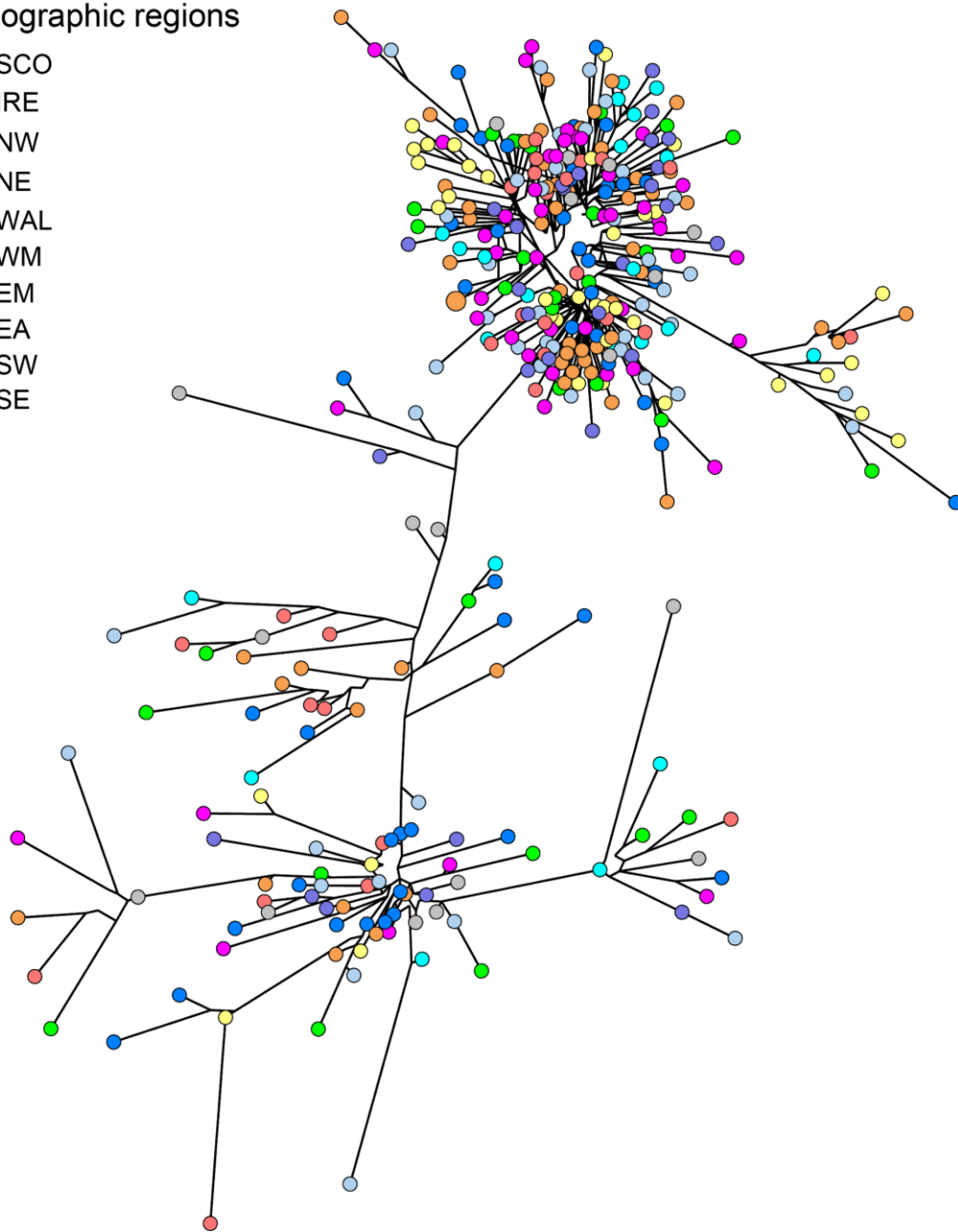
The Y-STR alleles can be further resolved by taking SNPs and indels into consideration; many of these occur in the flanking region outside the array itself. These complete sequence-based alleles were used to build a further network of the 362 samples. Similarly, this resolution level still does not show any obvious stratification relating to the ten large geographic regions (Figure 6.20a), but correlates well with the predicted main haplogroups (Figure 6.20b) and explains even better the more resolved haplogroup designations (Figure 6.20c).

With all the sequence information incorporated into the network, the resolution of the main haplogroups is better than with array sequence, or just length-based information. Clearly the lower mutation rate of SNPs/indels can draw out structures that were not as obvious before, in particular the J2 haplogroup is clearly divided to J2a and J2b, whilst I1 only minimally disrupts the I2 branches, which are the best resolved at this level. E1b remains a largely coherent cluster further away from the rest, and R1a appears as an offshoot cluster from R1b, rather than connecting to the rest of the network. Interestingly within R1b the opening reticulation seen at the previous level is more expanded into an almost ring-like structure.

a.

Geographic regions

- SCO
- IRE
- NW
- NE
- WAL
- WM
- EM
- EA
- SW
- SE

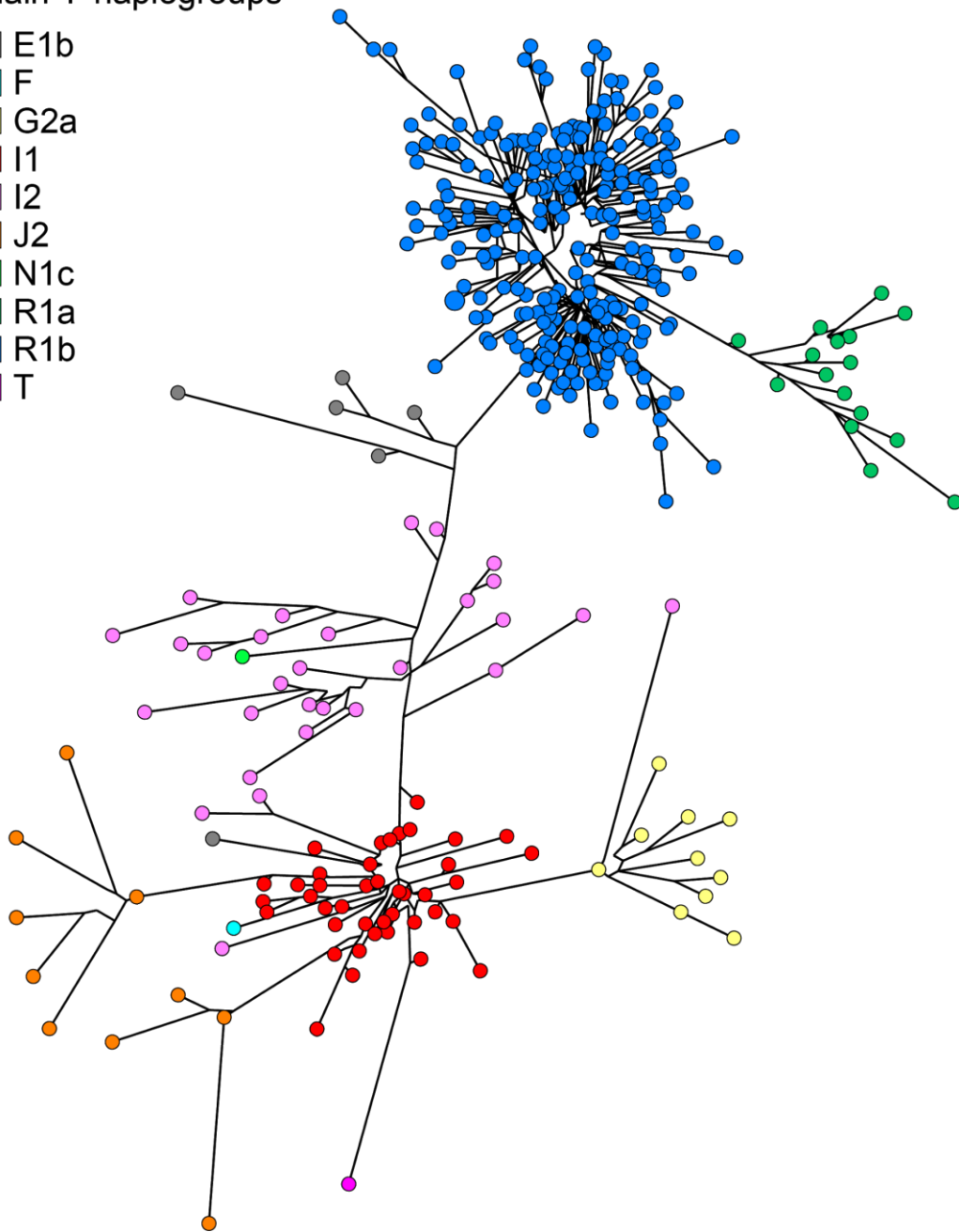


Cont.

b.

Main Y-haplogroups

- E1b
- F
- G2a
- I1
- I2
- J2
- N1c
- R1a
- R1b
- T



Cont.

c.

Y-haplogroups

- E1b1b
- F
- G2a
- G2a2
- G2a2b1
- I1
- I2a1
- I2a1a
- I2a2a
- I2a2b
- I2c1a2
- I2c2
- J2a1
- J2b1
- J2b2
- N1c
- R1a
- R1b
- T

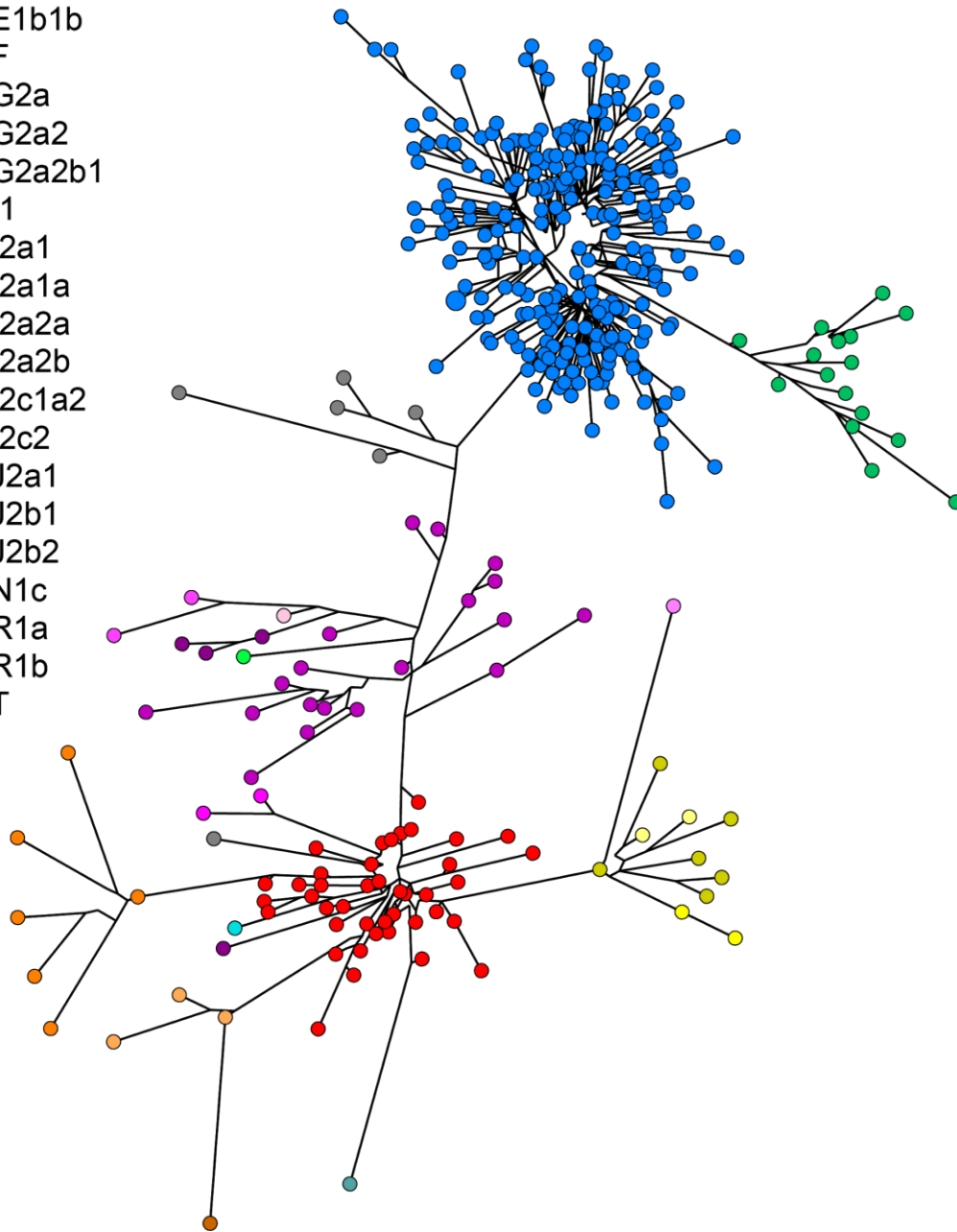


Figure 6.20

Median-joining network of samples with Y-STR sequence array, SNP and indel variants in the flanking region.

(a) is coloured by geographic regions, (b) is coloured by main haplogroups, and (c) is coloured by more highly resolved haplogroup levels.

The NevGen predictor has limited power in predicting haplogroup sub-structure within R1a and R1b with just 23 Y-STRs. To supplement the network, prior sub-haplogroup data from SNP chips (J.Wetton personal comm., unpublished data) was used for R1b, and SNaPshot data from an R1a1-specific study (Lall et al., manuscript in review) was used to sub-classify hg R1a1. Not all samples in this study had these additional data, but where available, these finer classifications were used to reannotate the network for the R1a and R1b clusters (Figure 6.21).

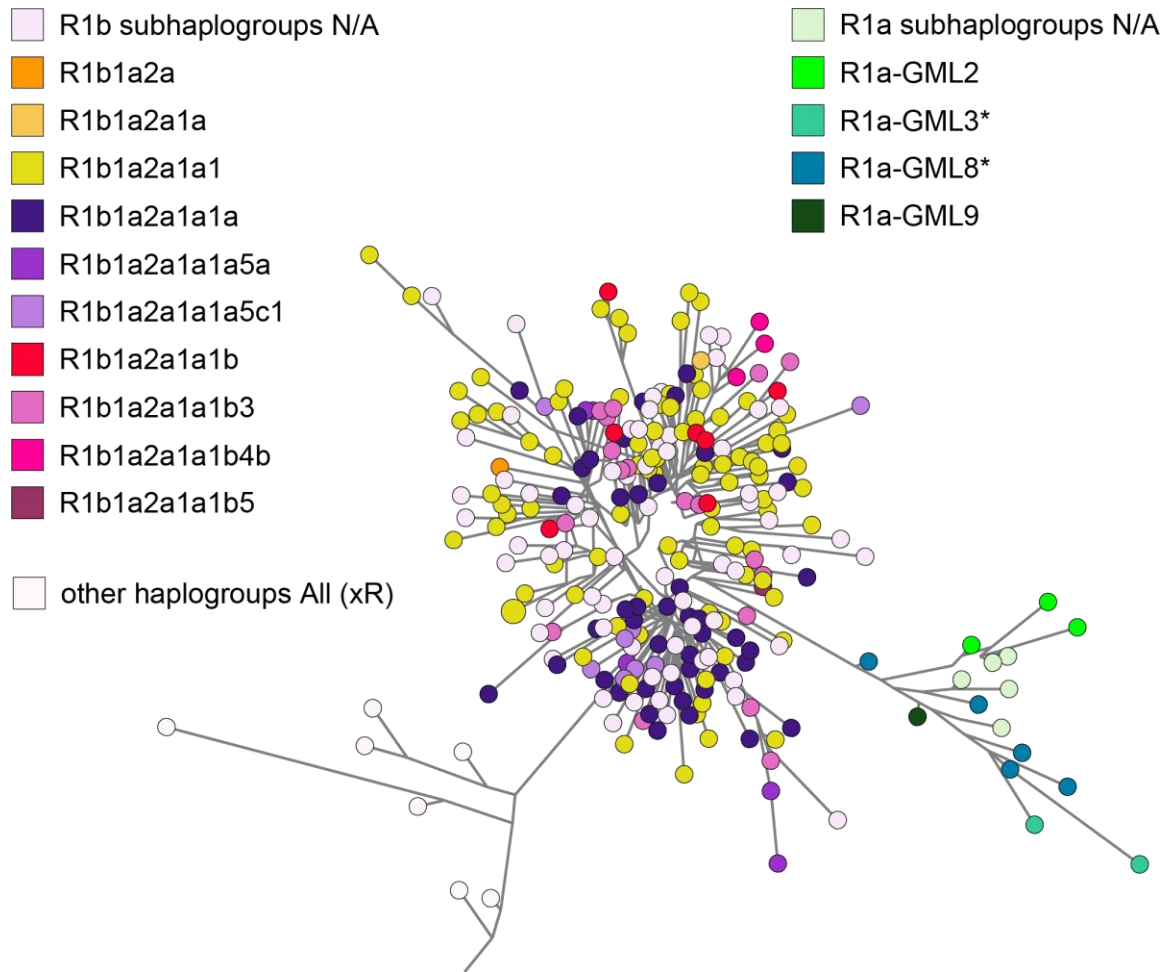


Figure 6.21
Truncated version of the median-joining network focussing on R1a and R1b clusters.

This truncated version of Figure 6.20 displays a finer resolution of haplogroups R1a and R1b.

Annotation of the network by these finer sub-haplogroups showed a possible structure within the R1b cluster at this level. The R1b1a2a1a1a lineage and its derivatives (~purple shades) which fall within the R1b-U106 lineage cluster more at the two opposite sides of the R1b-ring (5-7 and 11-1 o'clock positions) whilst the R1b1a2a1a1b lineage and its derivatives (~red/magenta shades) within the R1b-S116 lineage tend to cluster at the 10-2 o'clock positions of the R1b ring-cluster except for one lineage (R1b1a2a1a1b3) which seems to be uniformly represented along the R1b-ring. These two specific recently-diverged lineages are known to

have opposite frequencies across the UK (Rocca et al. 2012; Valverde et al. 2016); however, there seems to be no clear separation in the network space between these, nor in their allele ranges, or allele sequences.

To better understand the effect of sequence variants, a network using only Y-STR sequence features, flanking SNPs and indels was built and annotated to show the characteristic variants that are tied to certain haplogroups (Figure 6.22). The most interesting is a nested association structure: P superhaplogroup samples carry the insertion of two times two extra repeat units in their DYS635 alleles; within that, R1a samples have a characteristic DYS393 structure where the first nucleotide of the first repeat is changed from A to C; furthermore, as a third layer, a sublineage of R1a (R1a-GML2 or ~R1a1a1a) has a DYS390 structure including a last repeat change from TAGA to GAGA. Similarly, a SNP harboured in the last repeat of DYS458 is characteristic of J2a Y-chromosomes. However, there are sequence features that can arise multiple times, due to the repetitive structure of the STRs; for example, the initial CTG-repeat of DYS481 can expand to (CTG[2]), which is characteristic of all G2a lineages tested here, but also occurred in I2 haplotypes. These Y-haplogroup associated sequence variants were also observed in the same major haplogroups in the global sample set described in Chapter 3. By sequencing further Y-STRs, more specific associations with the phylogeny can be discovered, or singleton observations in small studies such as this, can be shown to be indicative of certain more derived lineages.

6.3.3.3 Testing population differentiation using Y-STR variants

The observed Y-STR length and sequence variants were used to test population differentiation in the 362 samples. Variable length alleles and sequence features were input to the Arlequin v 3.5.2.2 software (Excoffier and Lischer 2010) and summary statistics were generated including molecular diversity indices, F-statistics for population comparisons and differentiation (Table 6.13). The various population divisions tested are detailed in Section 6.2.2. Analysing sequence data strengthens the difference between tested populations, and for the ten geographic regions even reaches significance; otherwise, significantly different comparisons are the divisions based on the area of Celtic-language family, the Central/ South East cluster of the PoBI study and the East region of Britain. The historically distinct Danelaw region did not show significant differentiation either by Y-STR length or by complete sequence data.

Table 6.13
Y-STR length and sequence variants compared using Analysis of MOlecular VAriance (AMOVA) statistics and pairwise R_{ST} values.

Table (a) shows the AMOVA statistics, significant values are in bold, (b) shows the pairwise R_{ST} and haplotype diversity matrices per multi-component division. The pairwise R_{ST} and haplotype diversity matrix for the 37 population divisions is supplied as Appendix G.

a.

AMOVA	By length		By sequence	
	R_{ST}	P	R_{ST}	P
37 populations	0.0274	0.00842+-0.00085	0.04404	0.01119+-0.00115
10 geographic regions	0.02069	0.00293+-0.00164	0.0403	0.00099+-0.00030
5 PoBI defined regions	0.01954	0.00762+-0.00091	0.03183	0.01010+-0.00084
Y-STR	By length		By sequence	
	R_{ST}	P	R_{ST}	P
Celtic fringe vs rest	0.03064	<0.000001 (\pm 0.00001)	0.05131	0.0001 (\pm 0.0001)
Danelaw vs rest	0.00247	0.1877 (\pm 0.0039)	0.00568	0.13632 (\pm 0.0038)
CSE England vs rest	0.01437	0.0001 (\pm 0.0001)	0.03385	0.00109 (\pm 0.0003)
East vs rest	0.03457	<0.000001 (\pm 0.00001)	0.05613	0.0001 (\pm 0.0001)

Cont.

b.

above diagonal P-values (\pm SD)
 diagonal halotype diversity (HD)(\pm SD)
 below diagonal pairwise R_{ST} values
 significance=0.001111 (after Bonferroni correction)
 probabilities based on 10,100 permutations, significant values after Bonferroni correction are in **bold**

By LENGTH	Scotland	Ireland, Isle of Man	North West	North East	Wales	West Midlands	East Midlands	East	South West	South East
Scotland	HD=0.9986 \pm 0.0065	0.23711 \pm 0.0041	0.05000 \pm 0.0023	0.00099 \pm 0.0003	0.14028 \pm 0.0035	0.21186 \pm 0.0043	0.17018 \pm 0.0039	0.02812 \pm 0.0017	0.35412 \pm 0.0052	0.00257 \pm 0.0005
Ireland, Isle of Man	0.00935	HD=1.0000 \pm 0.0137	0.77745 \pm 0.0042	0.09316 \pm 0.0033	0.16276 \pm 0.0036	0.59984 \pm 0.0052	0.44639 \pm 0.0053	0.66241 \pm 0.0050	0.81091 \pm 0.0042	0.25384 \pm 0.0042
North West	0.03084	0	HD=1.0000 \pm 0.0086	0.26245 \pm 0.0041	0.01931 \pm 0.0014	0.44817 \pm 0.0042	0.30017 \pm 0.0044	0.99287 \pm 0.0008	0.19731 \pm 0.0044	0.79101 \pm 0.0043
North East	0.12971	0.03742	0.00915	HD=1.0000 \pm 0.0185	0.00050 \pm 0.0002	0.03911 \pm 0.0019	0.01228 \pm 0.0011	0.44837 \pm 0.0049	0.00703 \pm 0.0009	0.39828 \pm 0.0052
Wales	0.01132	0.0146	0.04517	0.14829	HD=1.0000 \pm 0.0045	0.15909 \pm 0.0037	0.07752 \pm 0.0028	0.01554 \pm 0.0014	0.16761 \pm 0.0038	0.00030 \pm 0.0002
West Midlands	0.00912	0	0	0.05956	0.0119	HD=1.0000 \pm 0.0091	0.92595 \pm 0.0028	0.32264 \pm 0.0049	0.70924 \pm 0.0046	0.12236 \pm 0.0035
East Midlands	0.01028	0	0.00346	0.07329	0.01717	0	HD=1.0000 \pm 0.0048	0.21820 \pm 0.0049	0.52252 \pm 0.0050	0.03287 \pm 0.0018
East	0.04092	0	0	0	0.0476	0.00313	0.00844	HD=1.0000 \pm 0.0091	0.17097 \pm 0.0036	0.68399 \pm 0.0049
South West	0.00106	0	0.00856	0.08541	0.00788	0	0	0.01064	HD=0.9994 \pm 0.0034	0.02000 \pm 0.0013
South East	0.07059	0.00763	0	0	0.08973	0.0178	0.0332	0	0.03392	HD=1.0000 \pm 0.0043

By SEQUENCE	Scotland	Ireland, Isle of Man	North West	North East	Wales	West Midlands	East Midlands	East	South West	South East
Scotland	HD=0.9986 \pm 0.0065	0.10227 \pm 0.0029	0.01158 \pm 0.0011	0.00000 \pm 0.0000	0.11662 \pm 0.0034	0.04089 \pm 0.0021	0.06207 \pm 0.0024	0.00495 \pm 0.0007	0.13811 \pm 0.0036	0.00050 \pm 0.0002
Ireland, Isle of Man	0.03894	HD=1.0000 \pm 0.0137	0.84101 \pm 0.0041	0.02624 \pm 0.0015	0.33710 \pm 0.0044	0.86516 \pm 0.0033	0.97961 \pm 0.0014	0.58143 \pm 0.0051	0.72042 \pm 0.0039	0.28908 \pm 0.0042
North West	0.08271	0	HD=1.0000 \pm 0.0086	0.05900 \pm 0.0022	0.09049 \pm 0.0030	0.50995 \pm 0.0054	0.66954 \pm 0.0043	0.91872 \pm 0.0027	0.25027 \pm 0.0048	0.60717 \pm 0.0047
North East	0.28692	0.12386	0.07365	HD=1.0000 \pm 0.0185	0.00020 \pm 0.0001	0.00802 \pm 0.0008	0.00594 \pm 0.0008	0.15157 \pm 0.0028	0.00089 \pm 0.0003	0.12801 \pm 0.0036
Wales	0.02261	0.00234	0.03912	0.2653	HD=1.0000 \pm 0.0045	0.34749 \pm 0.0046	0.30264 \pm 0.0042	0.03257 \pm 0.0017	0.48728 \pm 0.0046	0.00525 \pm 0.0007
West Midlands	0.05485	0	0	0.16757	0	HD=1.0000 \pm 0.0091	0.87160 \pm 0.0037	0.27631 \pm 0.0048	0.58499 \pm 0.0050	0.13009 \pm 0.0032
East Midlands	0.03573	0	0	0.14475	0.00113	0	HD=1.0000 \pm 0.0048	0.30938 \pm 0.0049	0.66241 \pm 0.0044	0.09464 \pm 0.0030
East	0.10068	0	0	0.03436	0.06534	0.00664	0.00183	HD=1.0000 \pm 0.0091	0.14840 \pm 0.0036	0.82427 \pm 0.0042
South West	0.01764	0	0.00727	0.18835	0	0	0	0.02251	HD=0.9994 \pm 0.0034	0.01475 \pm 0.0011
South East	0.13771	0.00277	0	0.02748	0.09319	0.02739	0.02555	0	0.05364	HD=1.0000 \pm 0.0043

Cont.

C.

above diagonal P-values (\pm SD)
 diagonal halotype diversity (HD)(\pm SD)
 below diagonal pairwise R_{ST} values

significance=0.005 (after Bonferroni correction)

probabilities based on 10,100 permutations, significant values after Bonferroni correction are in **bold**

By LENGTH	Orkney	Wales	Scotland and North	Cornwall	CSE-England
Orkney	HD=1.0000 \pm 0.0447	0.47480 \pm 0.0055	0.31492 \pm 0.0045	0.34442 \pm 0.0042	0.11147 \pm 0.0028
Wales	0	HD=1.0000 \pm 0.0045	0.05821 \pm 0.0022	0.67389 \pm 0.0043	0.00436\pm0.0006
Scotland and North	0.00599	0.01695	HD=0.9996 \pm 0.0026	0.15395 \pm 0.0036	0.35244 \pm 0.0046
Cornwall	0.01376	0	0.02538	HD=1.0000 \pm 0.0447	0.04792 \pm 0.0021
CSE-England	0.03417	0.03737	0.00031	0.05854	HD=1.0000 \pm 0.0004

By SEQUENCE	Orkney	Wales	Scotland and North	Cornwall	CSE-England
Orkney	HD=1.0000 \pm 0.0447	0.11306 \pm 0.0029	0.08049 \pm 0.0026	0.30839 \pm 0.0045	0.01485 \pm 0.0011
Wales	0.06668	HD=1.0000 \pm 0.0045	0.14335 \pm 0.0034	0.34125 \pm 0.0044	0.02762 \pm 0.0016
Scotland and North	0.08169	0.01226	HD=0.9996 \pm 0.0026	0.09494 \pm 0.0029	0.39006 \pm 0.0048
Cornwall	0.11018	0.00765	0.07513	HD=1.0000 \pm 0.0447	0.04277 \pm 0.0018
CSE-England	0.12935	0.03459	0	0.10176	HD=1.0000 \pm 0.0004

In contrast to the mtDNA data which showed no stratification across Britain, the distribution of Y-STR haplotypes revealed a substructuring of male lineages reflecting a general East-West difference.

These underlying genetic structures can be visualised by using the pairwise R_{ST} values of populations and generating Multidimensional Scaling plots, thus representing the genetic distances between these populations. MDS plots were generated for different divisions, Figure 6.23, Figure 6.24 and Figure 6.25 show MDS plots by various divisions with both length-based (a) and sequence-based (b) Y-STR data.

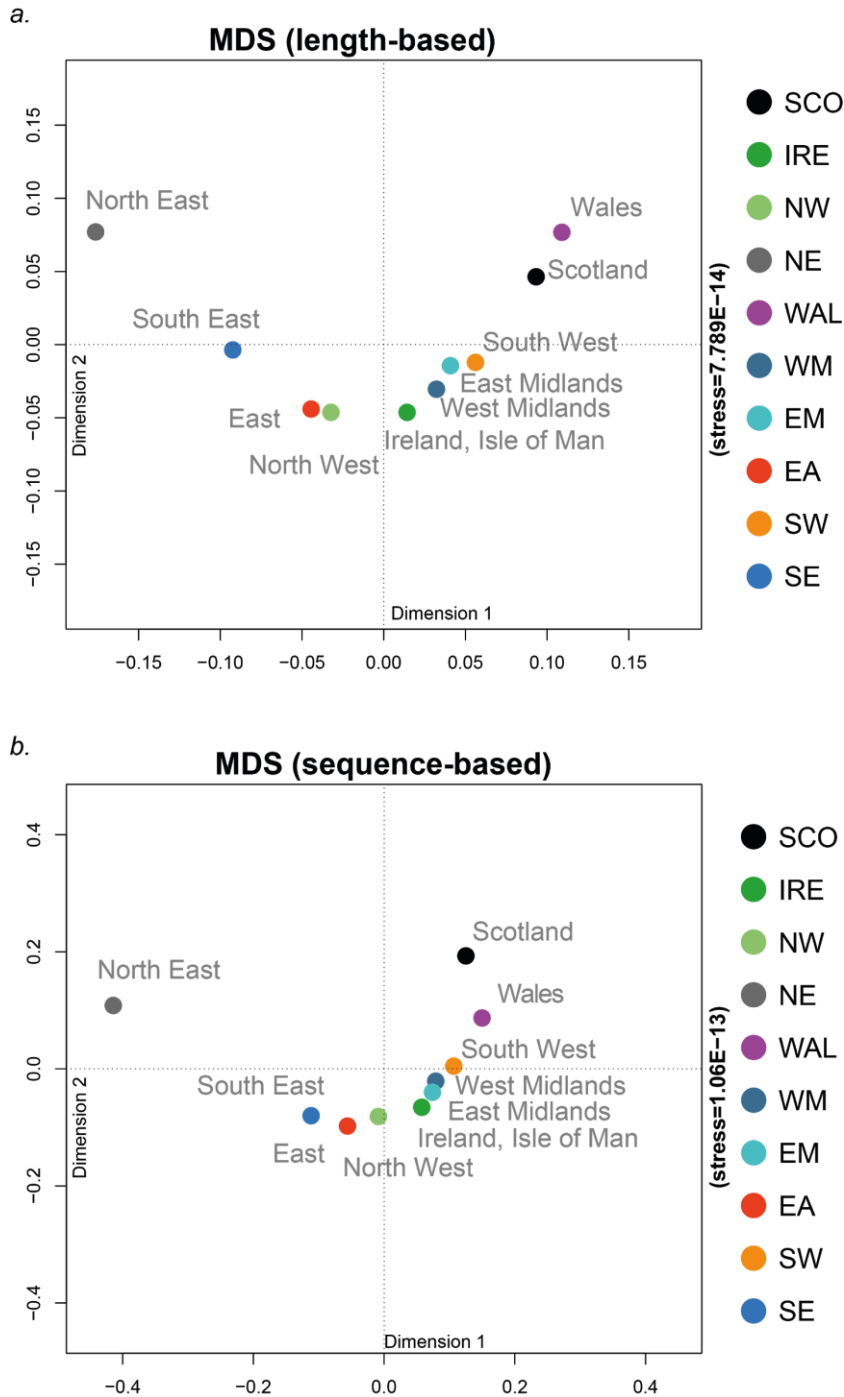
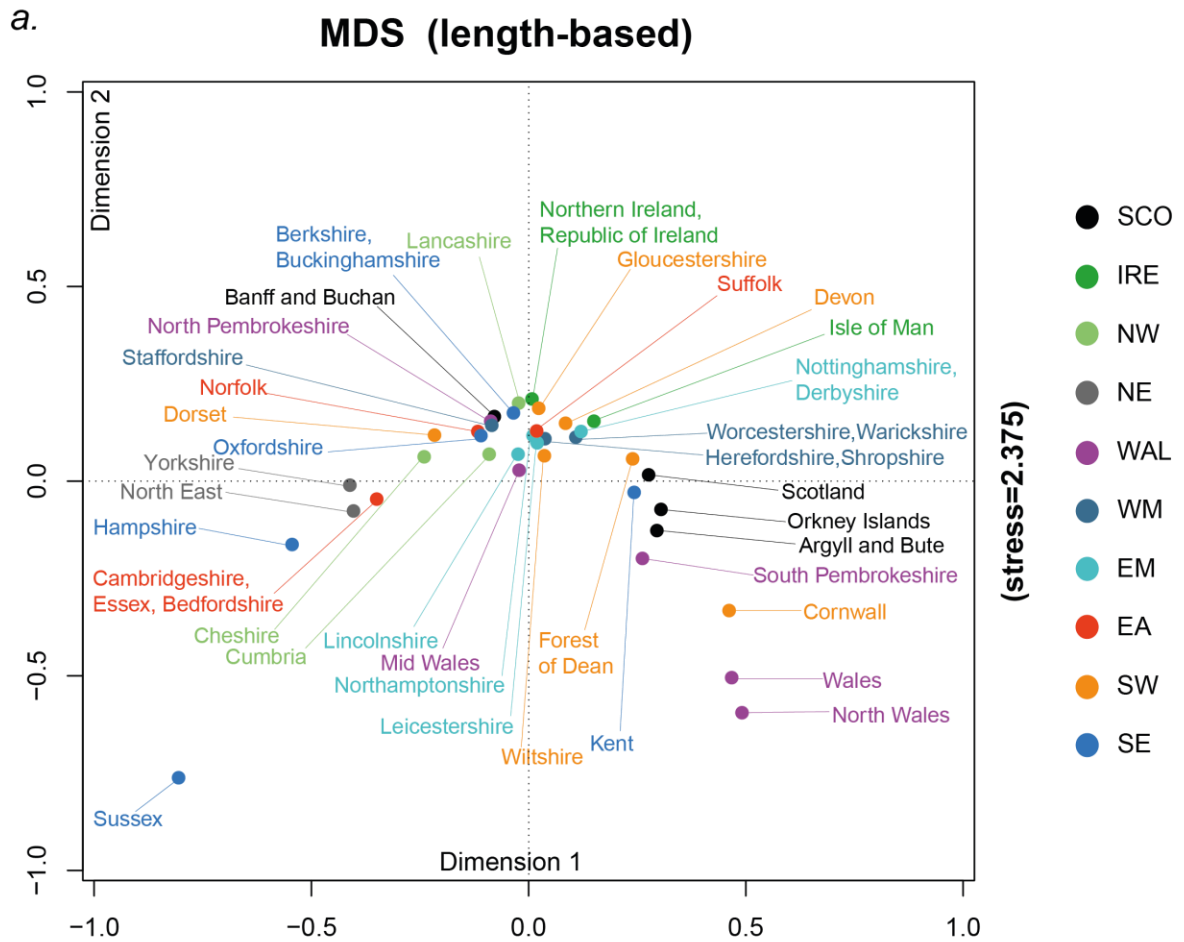


Figure 6.23
Multidimensional Scaling (MDS) plots of pairwise R_{ST} values using Y-STR length/sequence allele variants, by ten regions.

These plots map the ten geographic regions. (a) used length-based and (b) used sequence-based Y-STR data.



Cont.

b.

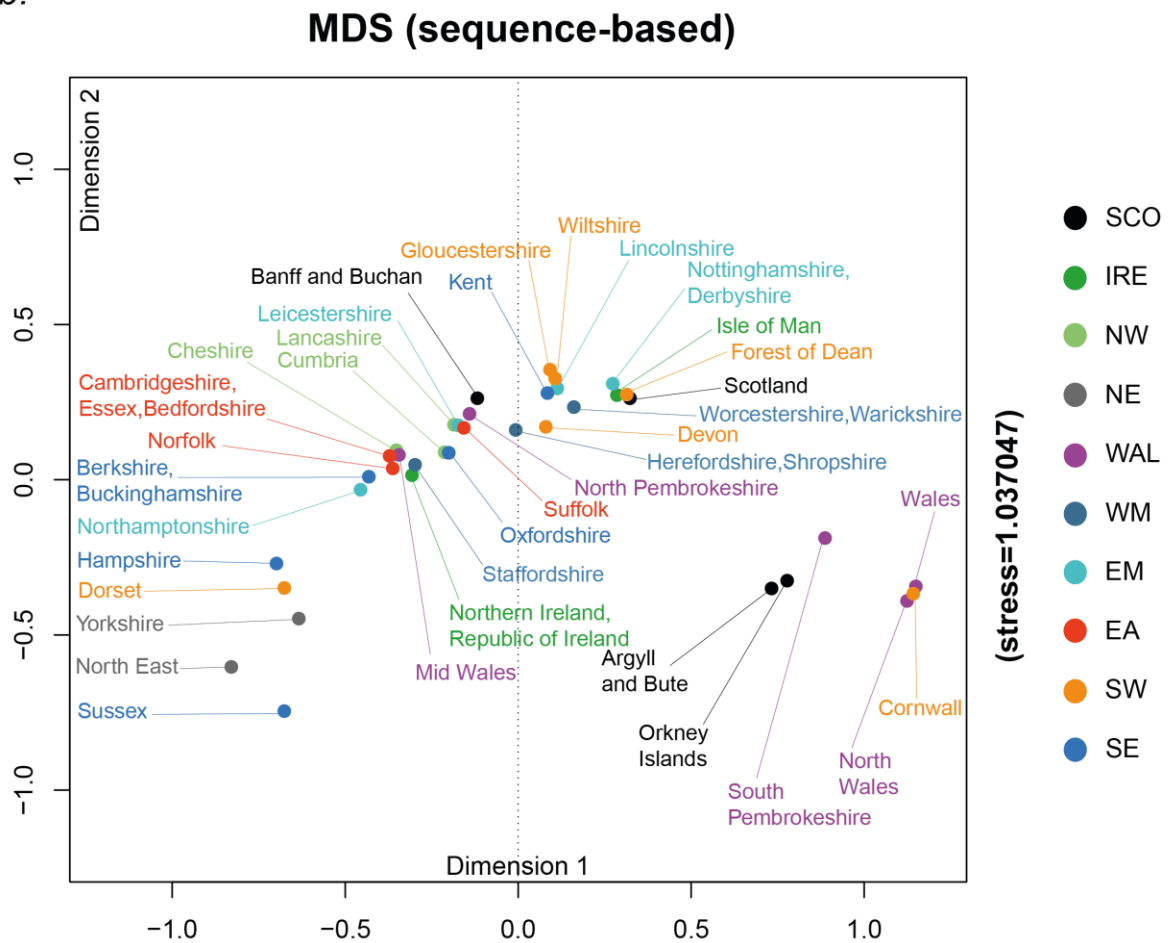
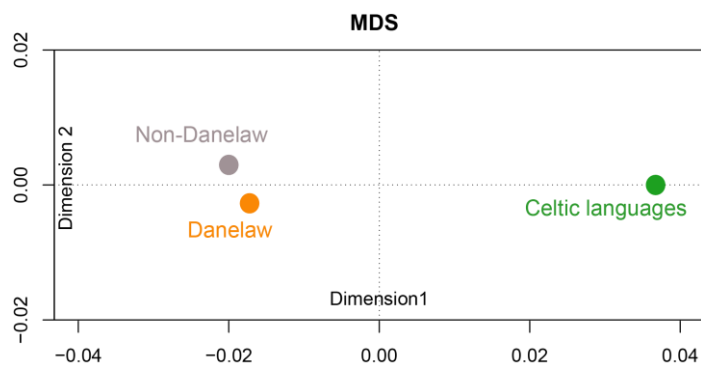
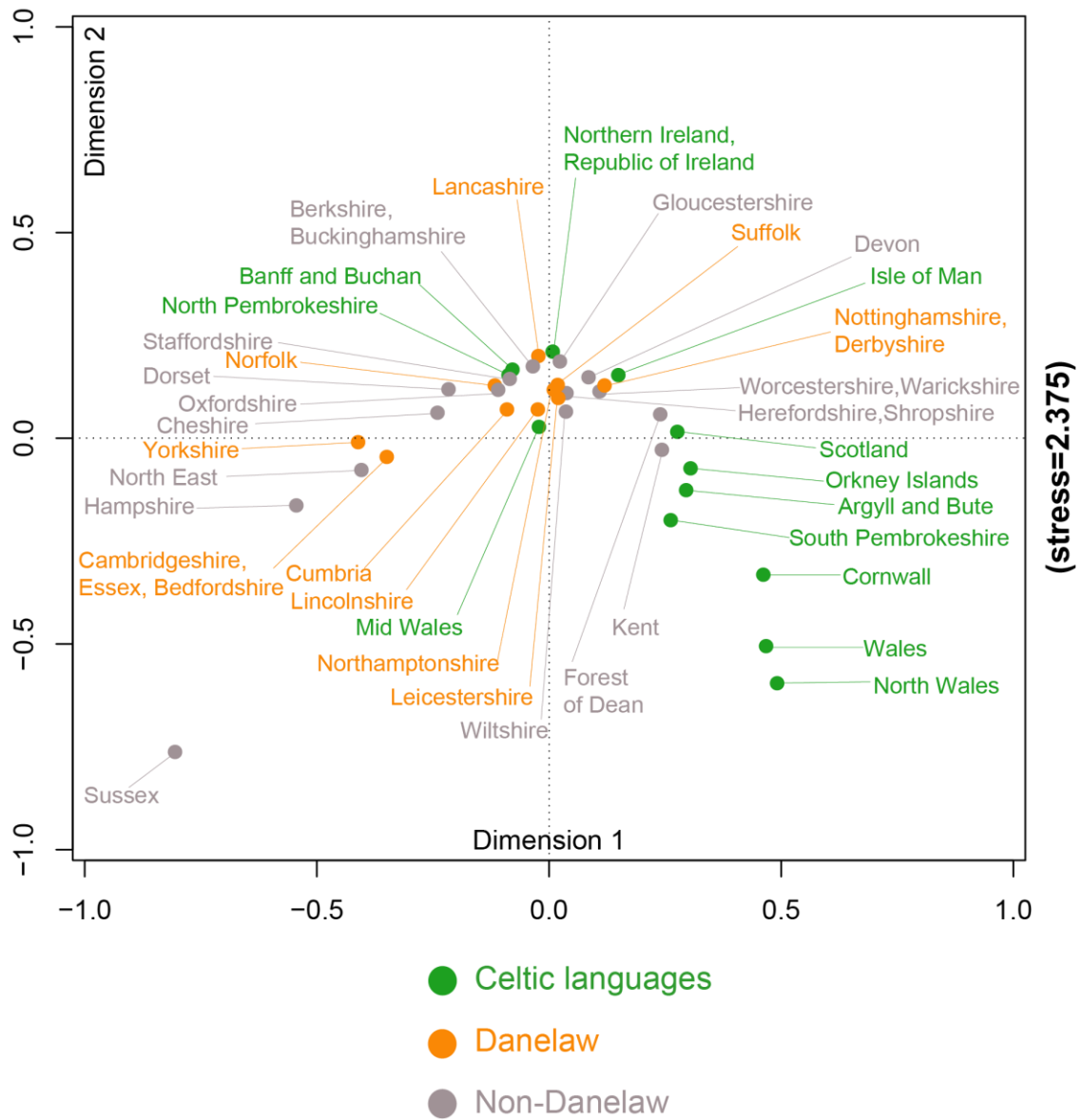


Figure 6.24
Multidimensional Scaling (MDS) plots of pairwise R_{ST} values using Y-STR length/sequence allele variants, by 37 populations.

These plots map the 37 populations coloured by the ten geographic regions. (a) used length-based and (b) used sequence-based Y-STR data.

a.

MDS (length-based)



Cont.

b.

MDS (sequence-based)

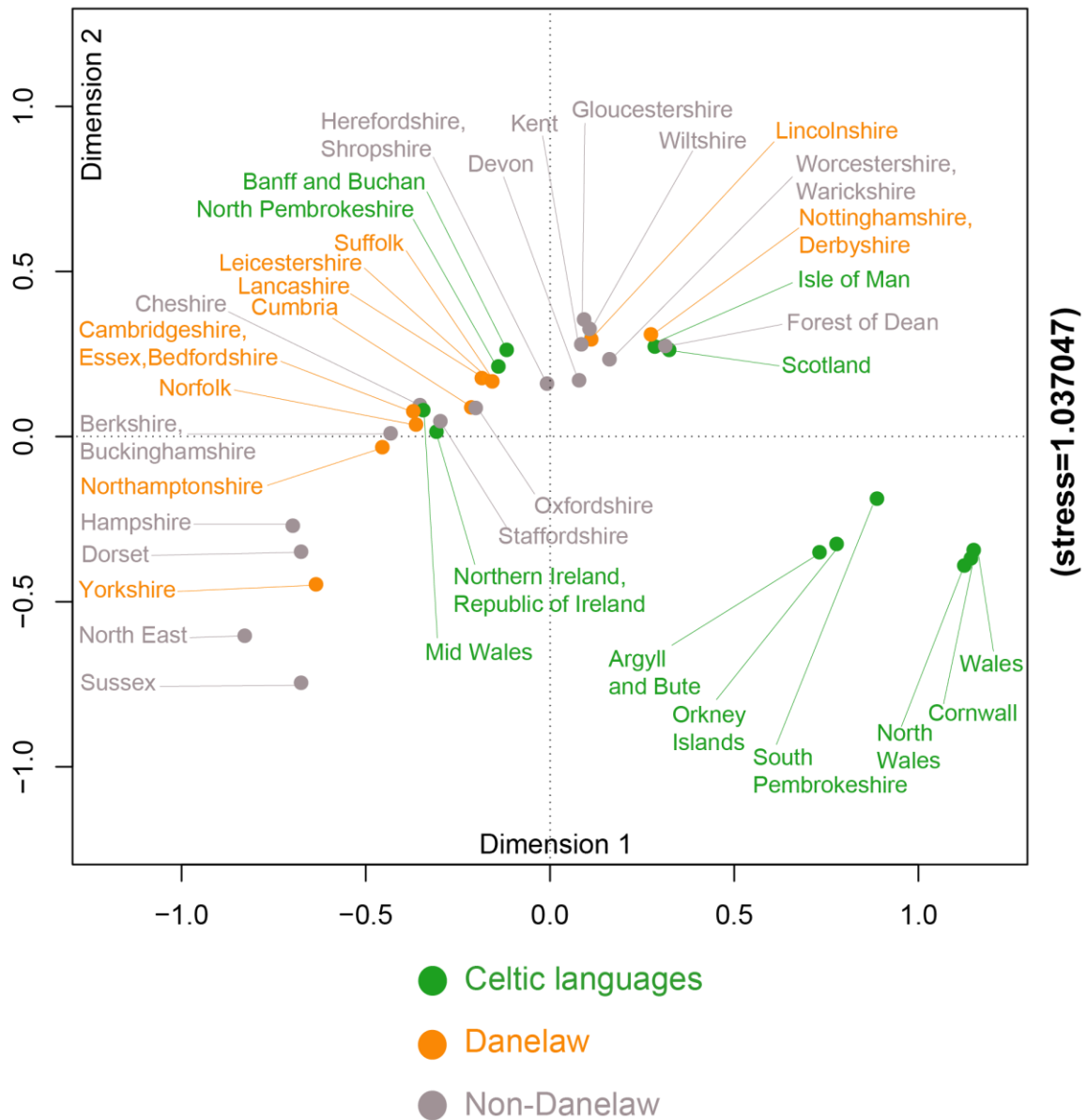


Figure 6.25

Multidimensional Scaling (MDS) plots of pairwise R_{ST} values using Y-STR length/sequence allele variants, by language and historical regions.

These plots map the 37 populations coloured by the Celtic language family and the historically Danelaw regions. (a) used length-based and (b) used sequence-based Y-STR data. (a) also shows a simple three component, zero stress MDS plot of the direct relationship of the three clusters.

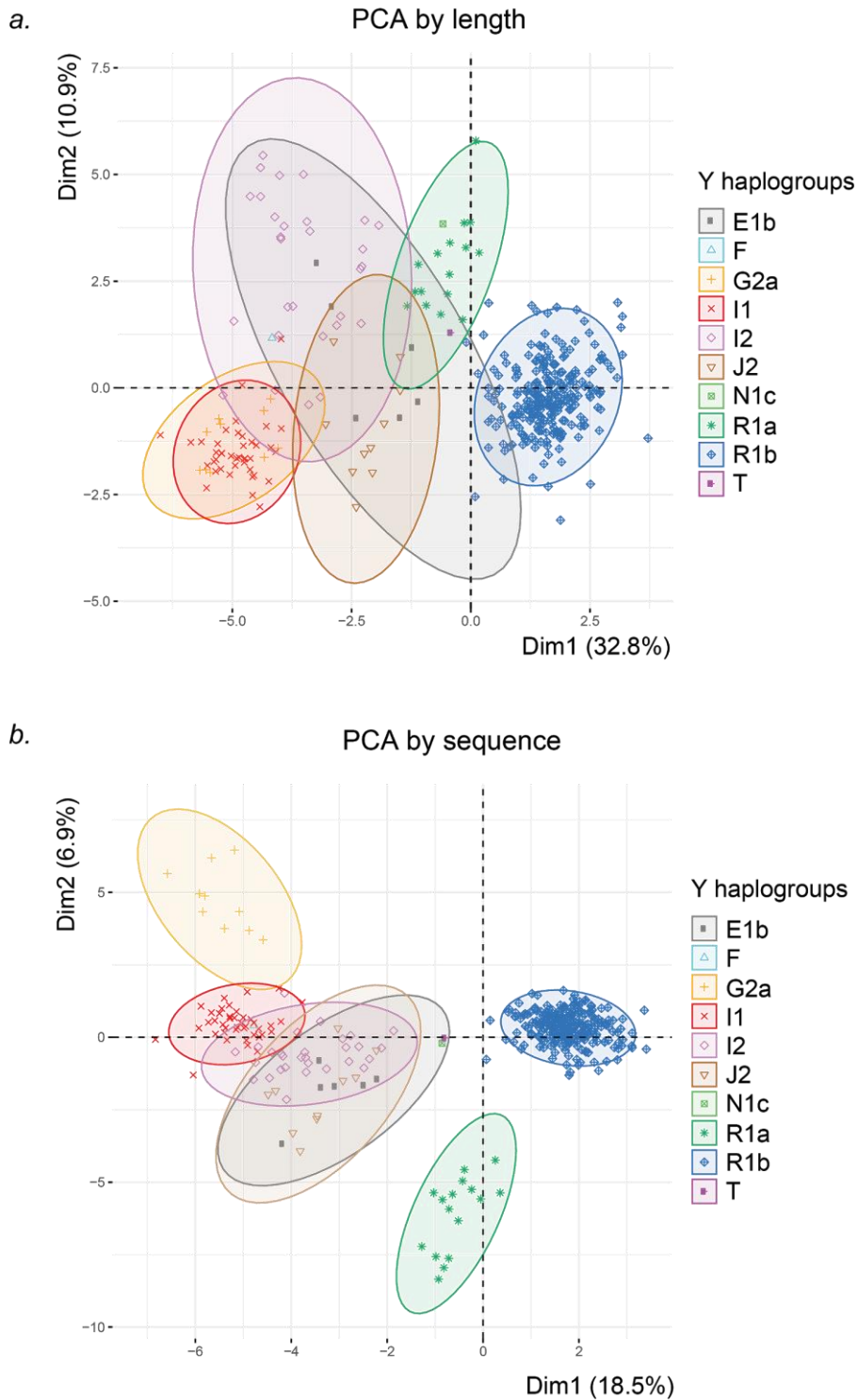
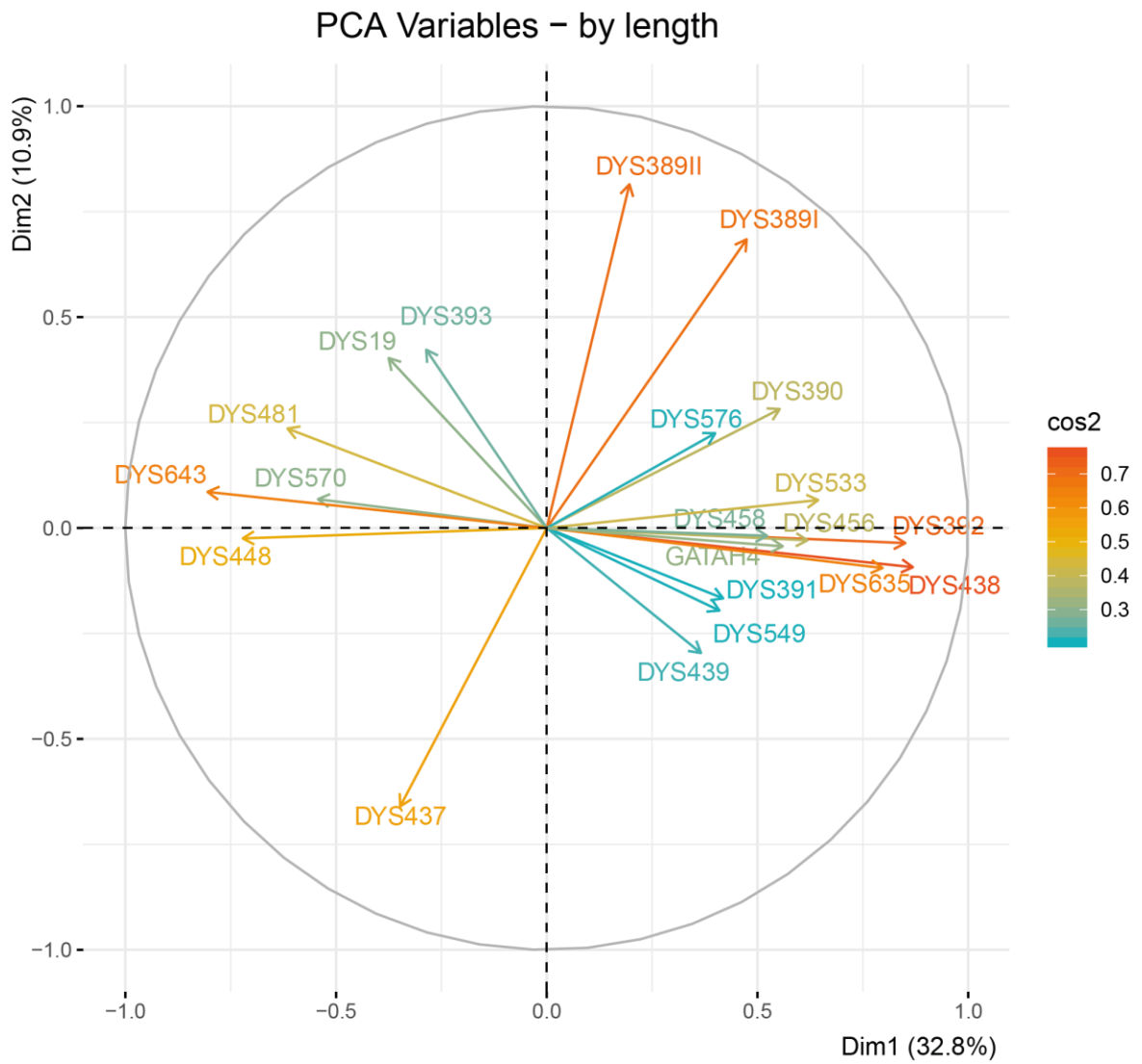


Figure 6.26
Principal Component Analysis (PCA) results plotted using Y-STR
length/sequence allele variants, coloured by Y haplogroups.

PCA plots map all 362 samples using their (a) length-based or (b) sequence-based Y-STR data.

a.



Cont.

b.

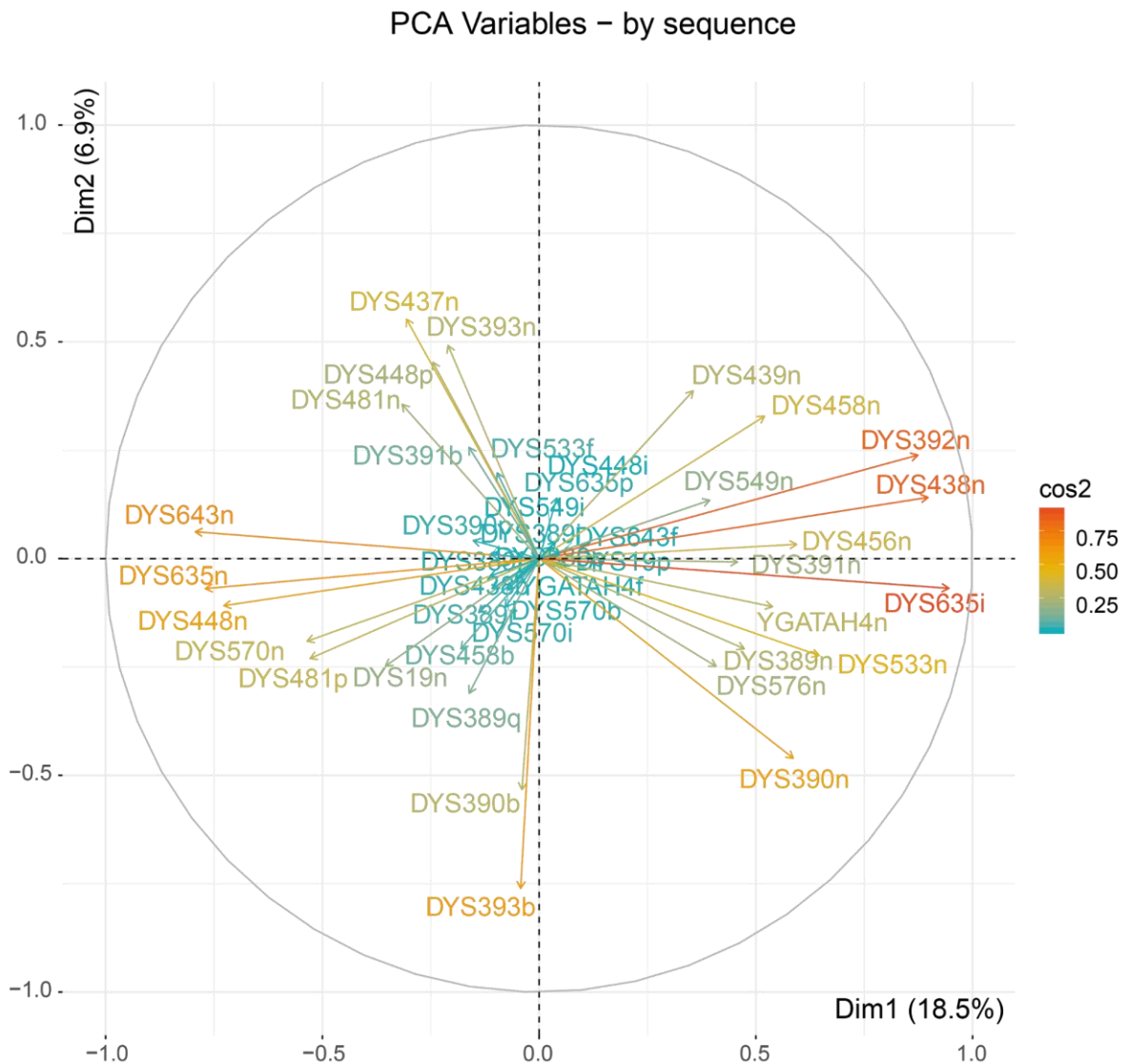


Figure 6.27
Quality of representation of PCA variables in length- and sequence-based analyses.

(a) shows these for the main two components of variation for length-based data and (b) for sequence-based data. The higher \cos^2 values the more represented the variables in the given components.

There are many ways to interrogate this Y-STR-based dataset and to ask how well it represents male population structure. MDS plots showed the expected outliers of Wales and Scotland; however, it also revealed that there is a component which affects the eastern regions, NE, SE, EA, as they tend to cluster away from the central regions. When this was further resolved to 37 populations, others, like Cornwall, were placed distantly and most of the sampling regions of Scotland and Wales were segregated from the central mass. These are all somewhat visible from the plots purely from the length-basis, but all are more marked when looking at sequence-derived data. Interestingly the sampling regions of the East of Britain again showed a differentiation from the main centre, and this substantiated the tests for the division referred to here as the East region. Divisions including the presence of the Celtic languages and the historical Scandinavian administrative region of the Danelaw are displayed together in Figure 6.25. At both length and sequence level the Celtic division is more differentiated from the rest, while the Danelaw cannot be distinguished; this confirms the results of the F-statistics. In Appendix G further Principal Component Analysis (PCA) plots test the dual divides by both length and sequence, and are all in line with the same conclusions. Differentiation is significant in the areas with a Celtic language presence, in the Central/Southeast England cluster of the PoBI study, and in the geographic East division, but not in the area of the historical Danelaw.

PCAs (Figure 6.26) show that the clustering of the 362 Y-STR haplotypes (either length or sequence) correlates well with both the known and predicted haplogroups. Furthermore, it is clear that sequence information of Y-STRs further resolves these haplogroup clusters. Analysing the main components and the haplogroup distributions on the plot, it becomes clear that PC1 of variation correlates with hg R1b. When sequence information is introduced this is more pronounced, and further draws out hg R1a from the main mass of R1b. Oppositely placed are the hgs I1 and I2, J2 and G2a, while hg E1b again shows its variability by encompassing large central areas of the graph, and only showing more of a coherent distribution with the inclusion of the sequence data. These observations confirm the previous MJ network results. The breakdown of the PCA metrics, \cos^2 (Figure 6.27), the quality

of representation of the variables that are transformed into principal components are also helpful to confirm where the main differences lie. It shows, both by length or sequence that the most represented variables in the main component which accounts for 32.8% of the variation, are the loci DYS635, DYS392, DYS438. From just DYS635 alone it is clear how this relates to R1b, as in Section 6.3.3.2., where it was shown that DYS635 contains a structural change within the repeat array that is specific to superhaplogroup P, within which R1 lineages are nested. The second main component from the graph alone suggests it is tied to hg R1a, especially on the sequenced data plot, where the characteristic sequence variant in DYS393 for hg R1a samples correlates with the R1a cluster itself. Hgs G2a and I1 share space on the plot by length, but by sequence they are completely resolved, which is due to the DYS481 sequence variant present in the G2a samples, which was also previously discussed.

In summary, MJ networks, MDS and PCA plots represent distinct but broadly consistent ways to confirm sequence variant associations to haplogroups.

6.3.3.4 Summary of Y-STR forensic statistics

Y-STR related haploid statistics (Table 6.14) were provided by the browser-based application STRAF 1.0.5 (Gouy and Zieger 2017). DYS385a,b was omitted from the dataset.

Table 6.14
Standard Y-STR loci forensic parameters.

locus	N	N _{all}	GD	PIC	PM	PD
DYS19	362	6	0.459375	0.408902	0.541894	0.458106
DYS389I/II	362	8	0.694021	0.644294	0.307897	0.692103
DYS390	362	7	0.674982	0.615719	0.326883	0.673117
DYS391	362	5	0.530846	0.418516	0.470621	0.529379
DYS392	362	5	0.524005	0.464606	0.477443	0.522557
DYS393	362	5	0.340537	0.315594	0.660404	0.339596
DYS437	362	5	0.517592	0.464449	0.483837	0.516163
DYS438	362	6	0.547283	0.496993	0.454229	0.545771
DYS439	362	5	0.657367	0.594599	0.344449	0.655551
DYS448	362	6	0.560506	0.506898	0.441043	0.558957
DYS456	362	7	0.743331	0.697711	0.258722	0.741278
DYS458	362	8	0.76211	0.72366	0.239996	0.760004
DYS481	362	10	0.713824	0.683462	0.288147	0.711853
DYS533	362	6	0.603832	0.5571	0.397836	0.602164
DYS549	362	6	0.657489	0.592858	0.344327	0.655673
DYS570	362	10	0.741831	0.705462	0.260218	0.739782
DYS576	362	8	0.773863	0.73626	0.228274	0.771726
DYS635	362	7	0.647587	0.612679	0.354202	0.645798
DYS643	362	6	0.52102	0.483052	0.480419	0.519581
GATAH4	362	5	0.573193	0.506616	0.42839	0.57161

N	number of samples
N _{all}	number of alleles
GD	Genetic diversity/expected heterozygosity
PIC	Polymorphism Information Content
PM	Match Probability
PD	Power of Discrimination

6.3.4 Population data of aSTR sequence variation in the UK

6.3.4.1 Observed variants

In total, 14,571 distinct aSTR allele calls were observed in this set of 361 samples. Observed allele frequencies by length are given for the 22 STR loci in Table 6.15. Allele frequencies for the sequenced alleles are given in Table 6.16. For each corresponding allele the STRSeq (Gettings et al. 2017) identifier is listed. Allele frequency tables for each of the ten geographical regions are given in Appendix F. Amelogenin was also sequenced and there were no abnormal calls and no indication of sex-chromosomal aneuploidies. However, a SNP was observed in AmelX in one individual; this SNP (rs188865418) has been described previously.

A total of 435 different sequence-based aSTR alleles were identified in this sample set across 22 STRs. Of these 32 have not yet been catalogued in the STRSeq database; fourteen include flanking variants, four carry indels, and 14 are combinations of repeat blocks not yet entered by others into the STR allele catalogue. There are four of these off-catalogue alleles which overlap between the British Isles and the global sample set described in Chapter 4. No duplications or null alleles were detected. No discrepancies were found between the two softwares used for genotyping (FDSTools and STRait Razor 3.0). The sample set contained overall 277 pairs of isometric heterozygotes, discrimination of which improved the overall observed heterozygosity from length-based 80.84% to sequence-based 84.18%: 17.4% of the length homozygotes are heterozygous by sequence. With sequencing allele diversity improves (with an overall 87.5%) compared to length-based designations, with only two loci that did not improve in resolution when sequenced (D22S1045 and PentaE). In this sample set, eight samples are also shared with the 1000 Genomes Project. In the detected alleles 40 calls of 11 different aSTR flanking region SNPs were identified, and all of them were found to be concordant with the 1000 Genomes Project sequence data. A summary table of these SNPs, their alleles and genotypes is provided in Appendix F.

Table 6.15
Length-based aSTR allele frequencies from 361 samples.

	CSF1P0	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	
5								5
6								6
7								7
8				0.0983	0.0097			8
9	0.0208			0.0734	0.1136			9
9.3								9.3
10	0.2396			0.0720	0.0609	0.0083		10
10.3	0.0014							10.3
11	0.3380	0.0028		0.3033	0.3116	0.0125	0.0055	11
11.3				0.0014				11.3
12	0.3130	0.0305		0.2909	0.2853	0.1233	0.0637	12
12.2							0.0014	12.2
13	0.0679	0.3061		0.1025	0.1884	0.1150	0.2341	13
13.2							0.0069	13.2
14	0.0194	0.3463		0.0582	0.0305	0.1939	0.3809	14
14.1								14.1
14.2							0.0152	14.2
14.3								14.3
15		0.1759	0.0360			0.1482	0.1856	15
15.2							0.0402	15.2
15.3								15.3
16		0.1205	0.0291			0.1343	0.0485	16
16.2							0.0083	16.2
16.3								16.3
17		0.0180	0.0928			0.1177	0.0055	17
17.2							0.0028	17.2
17.3			0.0291					17.3
18			0.1828			0.0706		18
18.2							0.0014	18.2
18.3			0.0346					18.3
19			0.1053			0.0499		19
19.2								19.2
19.3			0.0152					19.3
20			0.1302			0.0152		20
20.2								20.2
20.3								20.3
21			0.1080			0.0083		21
21.2								21.2
22			0.1122			0.0014		22
22.2								22.2
23			0.0554			0.0014		23
23.2								23.2
24			0.0346					24
24.2								24.2
25			0.0263					25
25.2								25.2
26			0.0083					26
27								27
28								28
29								29
29.3								29.3
30								30
30.2								30.2
31								31
31.2								31.2
32								32
32.2								32.2
33								33
33.2								33.2
34.2								34.2

Cont.

	D1S1656	D21S11	D22S1045	D2S1338	D2S441	D3S1358	D5S818	
5								5
6								6
7								7
8								8
9					0.0028		0.0346	9
9.3								9.3
10					0.1967		0.0609	10
10.3								10.3
11	0.0803		0.1593		0.3449		0.3546	11
11.3					0.0416			11.3
12	0.1247		0.0042		0.0263		0.3712	12
12.2								12.2
13	0.0568		0.0028		0.0291	0.0028	0.1676	13
13.2								13.2
14	0.0900		0.0332		0.3061	0.1205	0.0111	14
14.1								14.1
14.2								14.2
14.3	0.0028							14.3
15	0.1247		0.3366		0.0457	0.2964		15
15.2								15.2
15.3	0.0873							15.3
16	0.0983		0.3781	0.0568	0.0069	0.2244		16
16.2								16.2
16.3	0.0554							16.3
17	0.0540		0.0776	0.1759		0.2161		17
17.2								17.2
17.3	0.1579							17.3
18	0.0042		0.0083	0.0748		0.1247		18
18.2								18.2
18.3	0.0526							18.3
19				0.1191		0.0125		19
19.2								19.2
19.3	0.0097							19.3
20				0.1510		0.0028		20
20.2								20.2
20.3	0.0014							20.3
21				0.0305				21
21.2								21.2
22				0.0374				22
22.2								22.2
23				0.0970				23
23.2								23.2
24				0.1288				24
24.2		0.0014						24.2
25				0.1066				25
25.2		0.0014						25.2
26		0.0014		0.0222				26
27		0.0332						27
28		0.1510						28
29		0.2230						29
29.3		0.0014						29.3
30		0.2521						30
30.2		0.0291						30.2
31		0.0831						31
31.2		0.0970						31.2
32		0.0139						32
32.2		0.0942						32.2
33		0.0014						33
33.2		0.0139						33.2
34.2		0.0028						34.2

Cont.

	D7S820	D8S1179	FGA	PentaD	PentaE	TH01	TPOX	vWA	
5					0.0859	0.0042			5
6				0.0014		0.2147			6
7	0.0166			0.0111	0.1856	0.2036			7
8	0.1565	0.0180		0.0152	0.0139	0.1080	0.5111		8
9	0.1773	0.0097		0.1981	0.0055	0.1274	0.1150		9
9.3						0.3338			9.3
10	0.2756	0.1122		0.1122	0.0776	0.0083	0.0665		10
10.3									10.3
11	0.1731	0.0942		0.1136	0.0997		0.2659		11
11.3									11.3
12	0.1634	0.1330		0.2285	0.1745		0.0416		12
12.2									12.2
13	0.0332	0.3227		0.2341	0.0859			0.0028	13
13.2									13.2
14	0.0042	0.1842		0.0693	0.0554			0.1162	14
14.1					0.0014				14.1
14.2									14.2
14.3									14.3
15		0.0956		0.0139	0.0568			0.1148	15
15.2									15.2
15.3									15.3
16		0.0277		0.0014	0.0693			0.2075	16
16.2									16.2
16.3									16.3
17		0.0028	0.0014	0.0014	0.0402			0.2573	17
17.2									17.2
17.3									17.3
18			0.0125		0.0222			0.2006	18
18.2									18.2
18.3									18.3
19			0.0693		0.0139			0.0816	19
19.2			0.0014						19.2
19.3									19.3
20			0.1468		0.0069			0.0180	20
20.2			0.0028						20.2
20.3									20.3
21			0.1717		0.0042				21
21.2			0.0014						21.2
22			0.1925		0.0014				22
22.2			0.0097						22.2
23			0.1427						23
23.2			0.0042						23.2
24			0.1150					0.0014	24
24.2									24.2
25			0.0748						25
25.2									25.2
26			0.0457						26
27			0.0083						27
28									28
29									29
29.3									29.3
30									30
30.2									30.2
31									31
31.2									31.2
32									32
32.2									32.2
33									33
33.2									33.2
34.2									34.2

Table 6.16
Sequence-based aSTR allele frequencies from 361 samples.

locus	CE	MPS allele	frequency	STRSeq #
Amel	X	X	0.4986	N/A
		X_11,296,959C>T_rs188865418	0.0014	N/A
	Y	Y	0.5000	N/A
locus	CE	MPS allele	frequency	STRSeq #
CSF1P0	CE9	CE9_ATCT[9]	0.0208	MH085179.1
	CE10	CE10_ATCT[10]	0.2382	MH085181.2
		CE10_ATCT[10]_+86C>A_rs140751340	0.0014	N/A
	CE10.3	CE10.3_ATCT[5]ATC[1]ATCT[5]	0.0014	N/A
	CE11	CE11_ATCT[11]	0.3324	MH085186.2
		CE11_ATCT[11]_+86C>A_rs140751340	0.0055	MH085185.1
	CE12	CE12_ATCT[12]	0.3130	MH085189.2
	CE13	CE13_ATCT[13]	0.0665	MH085190.1
		CE13_ATCT[13]_+86C>A_rs140751340	0.0014	N/A
	CE14	CE14_ATCT[14]	0.0194	MH085191.2
locus	CE	MPS allele	frequency	STRSeq #
D10S1248	CE11	CE11_GGAA[11]	0.0028	MH167058.1
	CE12	CE12_GGAA[12]	0.0305	MH167059.2
	CE13	CE13_GGAA[13]	0.3061	MH167061.1
	CE14	CE14_GGAA[14]	0.3449	MH167062.2
		CE14_GGAA[14]_-1T>A_rs563636310	0.0014	N/A
	CE15	CE15_GGAA[15]	0.1759	MH167063.2
	CE16	CE16_GGAA[16]	0.1205	MH167064.2
	CE17	CE17_GGAA[17]	0.0180	MH167065.1

Cont.

locus	CE	MPS allele	frequency	STRSeq #
D12S391	CE15	CE15_AGAT[8]AGAC[6]AGAT[1]	0.0360	MH167108.1
	CE16	CE16_AGAT[8]AGAC[7]AGAT[1]	0.0014	MH167110.1
		CE16_AGAT[9]AGAC[6]AGAT[1]	0.0277	MH167111.1
	CE17	CE17_AGAT[10]AGAC[6]AGAT[1]	0.0886	MH167114.1
		CE17_AGAT[11]AGAC[5]AGAT[1]	0.0028	MH167115.1
		CE17_AGAT[9]AGAC[7]AGAT[1]	0.0014	MH167113.1
	CE17.3	CE17.3_AGAT[1]GAT[1]AGAT[8]AGAC[7]AGAT[1]	0.0291	MH167119.1
	CE18	CE18_AGAT[10]AGAC[7]AGAT[1]	0.0111	MH167122.1
		CE18_AGAT[11]AGAC[6]AGAT[1]	0.1676	MH167125.1
		CE18_AGAT[12]AGAC[5]AGAT[1]	0.0042	MH167126.1
	CE18.3	CE18.3_AGAT[1]GAT[1]AGAT[9]AGAC[7]AGAT[1]	0.0346	MH167128.1
	CE19	CE19_AGAT[11]AGAC[7]AGAT[1]	0.0235	MH167131.1
		CE19_AGAT[12]AGAC[6]AGAT[1]	0.0803	MH167134.1
		CE19_AGAT[13]AGAC[5]AGAT[1]	0.0014	MH167137.1
	CE19.3	CE19.3_AGAT[1]GAT[1]AGAT[10]AGAC[7]AGAT[1]	0.0097	MH167142.1
		CE19.3_AGAT[5]GAT[1]AGAT[7]AGAC[6]AGAT[1]	0.0055	MH167141.1
	CE20	CE20_AGAT[11]AGAC[8]AGAT[1]	0.0014	MH167145.1
		CE20_AGAT[11]AGAC[9]	0.0277	MH167144.1
		CE20_AGAT[12]AGAC[7]AGAT[1]	0.0457	MH167147.1
		CE20_AGAT[12]AGAC[8]	0.0055	MH167146.1
		CE20_AGAT[13]AGAC[6]AGAT[1]	0.0457	MH167150.1
		CE20_AGAT[13]AGAC[7]	0.0014	MH167149.1
		CE20_AGAT[14]AGAC[5]AGAT[1]	0.0014	MH167152.1
	CE21	CE20_AGGT[1]AGAT[10]AGAC[9]	0.0014	MH167153.1
		CE21_AGAT[10]AGAC[10]AGAT[1]	0.0014	MH167156.1
		CE21_AGAT[11]AGAC[10]	0.0097	MH167157.1
		CE21_AGAT[11]AGAC[9]AGAT[1]	0.0014	MH167158.1
		CE21_AGAT[12]AGAC[8]AGAT[1]	0.0055	MH167160.1
		CE21_AGAT[12]AGAC[9]	0.0623	MH167159.1
		CE21_AGAT[13]AGAC[7]AGAT[1]	0.0083	MH167162.1
	CE22	CE21_AGAT[13]AGAC[8]	0.0152	MH167161.1
		CE21_AGAT[14]AGAC[6]AGAT[1]	0.0042	MH167165.1
		CE22_AGAT[11]AGAC[11]	0.0014	MH167167.1
		CE22_AGAT[12]AGAC[10]	0.0166	MH167169.1
		CE22_AGAT[12]AGAC[9]AGAT[1]	0.0014	MH167170.1
		CE22_AGAT[13]AGAC[8]AGAT[1]	0.0097	MH167172.1
		CE22_AGAT[13]AGAC[9]	0.0596	MH167171.1
	CE23	CE22_AGAT[14]AGAC[7]AGAT[1]	0.0028	MH167174.1
		CE22_AGAT[14]AGAC[8]	0.0194	MH167173.1
		CE22_AGGT[1]AGAT[13]AGAC[7]AGAT[1]	0.0014	MK569935.1
		CE23_AGAT[12]AGAC[10]AGAT[1]	0.0014	N/A
		CE23_AGAT[12]AGAC[11]	0.0014	MH167177.1
		CE23_AGAT[13]AGAC[10]	0.0042	MH167178.1
		CE23_AGAT[13]AGAC[9]AGAT[1]	0.0042	MH167179.1
	CE24	CE23_AGAT[14]AGAC[8]AGAT[1]	0.0111	MH167181.1
		CE23_AGAT[14]AGAC[9]	0.0291	MH167180.1
		CE23_AGAT[15]AGAC[8]	0.0042	MH167182.1
		CE24_AGAT[13]AGAC[10]AGAT[1]	0.0014	N/A
		CE24_AGAT[13]AGAC[11]	0.0028	MH167183.1
		CE24_AGAT[14]AGAC[10]	0.0055	MH167184.1
		CE24_AGAT[14]AGAC[9]AGAT[1]	0.0014	MH167185.1
	CE25	CE24_AGAT[15]AGAC[8]AGAT[1]	0.0111	MH167187.1
		CE24_AGAT[15]AGAC[9]	0.0125	MH167186.1
		CE25_AGAT[14]AGAC[11]	0.0014	MK569939.1
		CE25_AGAT[15]AGAC[10]	0.0069	MH167190.1
		CE25_AGAT[15]AGAC[9]AGAT[1]	0.0042	MH167191.1
		CE25_AGAT[16]AGAC[8]AGAT[1]	0.0097	MH167193.1
		CE25_AGAT[16]AGAC[9]	0.0014	MH167192.1
	CE26	CE25_AGGT[1]AGAT[15]AGAC[8]AGAT[1]	0.0028	MH167194.1
		CE26_AGAT[16]AGAC[10]	0.0014	MH167195.1
		CE26_AGAT[16]AGAC[9]AGAT[1]	0.0014	N/A
		CE26_AGAT[17]AGAC[8]AGAT[1]	0.0042	MH167197.1
		CE26_AGAT[17]AGAC[9]	0.0014	MH167196.1

Cont.

locus	CE	MPS allele	frequency	STRSeq #
D13S317	CE8	CE8_TATC[8]AATC[2]ATCT[3]	0.0983	MH167202.1
	CE9	CE9_TATC[11]AATC[1]ATCT[2]	0.0028	MH167209.1
		CE9_TATC[9]AATC[2]ATCT[3]	0.0706	MH167206.1
	CE10	CE10_TATC[10]AATC[2]ATCT[3]	0.0582	MH167210.1
		CE10_TATC[11]AATC[1]ATCT[3]	0.0111	MH167211.1
		CE10_TATC[12]AATC[1]ATCT[2]	0.0014	N/A
		CE10_TATC[9]AATC[3]ATCT[3]	0.0014	N/A
	CE11	CE11_TATC[11]AATC[2]ATCT[3]	0.1219	MH167218.1
		CE11_TATC[12]AATC[1]ATCT[3]	0.1662	MH167219.1
		CE11_TATC[12]AATC[1]ATCT[3]_-25C>T_rs73250432	0.0152	MH167223.1
	CE11.3	CE11.3_TATC[2]ATC[1]TATC[10]AATC[1]ATCT[3]	0.0014	N/A
	CE12	CE12_TATC[12]AATC[2]ATCT[3]	0.1690	MH167226.1
		CE12_TATC[13]AATC[1]ATCT[3]	0.1150	MH167228.1
		CE12_TATC[13]AATC[1]ATCT[3]_-25C>T_rs73250432	0.0069	MH167230.1
	CE13	CE13_TATC[13]AATC[2]ATCT[3]	0.0582	MH167233.1
		CE13_TATC[14]AATC[1]ATCT[3]	0.0443	MH167234.1
	CE14	CE14_TATC[14]AATC[2]ATCT[3]	0.0471	MH167236.1
		CE14_TATC[15]AATC[1]ATCT[3]	0.0111	MH167237.1
locus	CE	MPS allele	frequency	STRSeq #
D16S539	CE8	CE8_GATA[8]	0.0097	MH167241.1
	CE9	CE9_GATA[9]	0.1122	MH167243.1
		CE9_GATA[9]_-95A>C_rs1728369	0.0014	N/A
	CE10	CE10_GATA[10]	0.0609	MH167249.1
	CE11	CE11_GATA[11]	0.2659	MH167251.1
		CE11_GATA[11]_-95A>C_rs1728369	0.0457	MH167254.1
	CE12	CE12_GATA[12]	0.2258	MH167259.1
		CE12_GATA[12]_-95A>C_rs1728369	0.0596	MH167260.1
	CE13	CE13_GATA[13]	0.1343	MH167261.1
		CE13_GATA[13]_-95A>C_rs1728369	0.0540	MH167262.1
	CE14	CE14_GATA[14]	0.0152	MH167264.1
		CE14_GATA[14]_-95A>C_rs1728369	0.0152	MH167265.1
locus	CE	MPS allele	frequency	STRSeq #
D18S51	CE10	CE10_AGAA[10]	0.0083	MH167284.1
	CE11	CE11_AGAA[11]	0.0125	MH167285.1
	CE12	CE12_AGAA[12]	0.1233	MH167287.1
	CE13	CE13_AGAA[13]	0.1150	MH167288.1
	CE14	CE14_AGAA[1]AGCA[1]AGAA[12]	0.0097	MH167291.1
		CE14_AGAA[14]	0.1842	MH167290.1
	CE15	CE15_AGAA[1]AGCA[1]AGAA[13]	0.0028	MH167294.1
		CE15_AGAA[15]	0.1454	MH167293.1
	CE16	CE16_AGAA[16]	0.1343	MH167296.1
	CE17	CE17_AGAA[17]	0.1177	MH167299.1
	CE18	CE18_AGAA[18]	0.0706	MH167301.1
	CE19	CE19_AGAA[19]	0.0499	MH167302.1
	CE20	CE20_AGAA[20]	0.0152	MH167303.1
	CE21	CE21_AGAA[21]	0.0083	MH167306.1
	CE22	CE22_AGAA[22]	0.0014	MH167309.1
	CE23	CE23_AGAA[23]	0.0014	MH167310.1

Cont.

locus	CE	MPS allele	frequency	STRSeq #
D19S433	CE11	CE11_CCTT[9]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	0.0055	MH174812.1
	CE12	CE12_CCTT[10]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	0.0637	MH174813.1
	CE12.2	CE12.2_CCTT[11]CCTA[1]CCTT[1]TT[1]CCTT[1]	0.0014	MH174815.1
	CE13	CE13_CCTT[11]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	0.2327	MH174816.1
		CE13_CCTT[2]CTTT[1]CCTT[8]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	0.0014	N/A
	CE13.2	CE13.2_CCTT[12]CCTA[1]CCTT[1]TT[1]CCTT[1]	0.0069	MH174819.1
	CE14	CE14_CCTT[12]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	0.3753	MH174820.1
		CE14_CCTT[14]CTTT[1]CCTT[1]	0.0055	MH174821.1
	CE14.2	CE14.2_CCTT[13]CCTA[1]CCTT[1]TT[1]CCTT[1]	0.0152	MH174824.1
	CE15	CE15_CCTT[13]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	0.1856	MH174825.1
	CE15.2	CE15.2_CCTT[14]CCTA[1]CCTT[1]TT[1]CCTT[1]	0.0402	MH174827.1
	CE16	CE16_CCTT[14]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	0.0485	MH174828.1
	CE16.2	CE16.2_CCTT[15]CCTA[1]CCTT[1]TT[1]CCTT[1]	0.0083	MH174829.1
	CE17	CE17_CCTT[15]CCTA[1]CCTT[1]CTTT[1]CCTT[1]	0.0055	MH174830.1
	CE17.2	CE17.2_CCTT[16]CCTA[1]CCTT[1]TT[1]CCTT[1]	0.0028	MH174831.1
	CE18.2	CE18.2_CCTT[17]CCTA[1]CCTT[1]TT[1]CCTT[1]	0.0014	MH174833.1
locus	CE	MPS allele	frequency	STRSeq #
D1S1656	CE11	CE11_CCTA[1]TCTA[10]	0.0014	MH174836.1
		CE11_TCTA[11]	0.0789	MH174837.1
	CE12	CE12_CCTA[1]TCTA[11]	0.0831	MH174838.1
		CE12_TCTA[12]	0.0416	MH174839.1
	CE13	CE13_CCTA[1]TCTA[12]	0.0291	MH174840.1
		CE13_TCTA[13]	0.0277	MH174842.1
	CE14	CE14_CCTA[1]TCTA[13]	0.0886	MH174844.1
		CE14_TCTA[14]	0.0014	MH174845.1
	CE14.3	CE14.3_CCTA[1]TCTA[11]TCA[1]TCTA[2]	0.0028	MH174847.1
	CE15	CE15_CCTA[1]TCTA[14]	0.1191	MH174848.1
		CE15_CTGA[1]TCTA[14]	0.0014	MH174849.1
		CE15_TCTA[15]	0.0042	MH174850.1
		CE15.3_CCTA[1]TCTA[10]TCA[1]TCTA[4]_+6C>T_rs4847015	0.0554	MH174851.1
	CE15.3	CE15.3_CCTA[1]TCTA[11]TCA[1]TCTA[3]_+6C>T_rs4847015	0.0305	MH174852.1
		CE15.3_CCTA[1]TCTA[12]TCA[1]TCTA[2]	0.0014	MK570033.1
		CE16_CCTA[1]TCTA[15]	0.0942	MH174853.1
	CE16	CE16_CTGA[1]TCTA[15]	0.0028	MH174854.1
		CE16_TCTA[16]	0.0014	MH174855.1
		CE16.3_CCTA[1]TCTA[11]TCA[1]TCTA[4]_+6C>T_rs4847015	0.0554	MH174856.1
	CE17	CE17_CCTA[1]TCTA[16]	0.0526	MH174858.1
		CE17_CTGA[1]TCTA[16]	0.0014	MH174859.1
	CE17.3	CE17.3_CCTA[1]TCTA[12]TCA[1]TCTA[4]_+6C>T_rs4847015	0.1579	MH174861.1
	CE18	CE18_CCTA[1]TCTA[17]	0.0042	MH174864.1
	CE18.3	CE18.3_CCTA[1]TCTA[13]TCA[1]TCTA[4]_+6C>T_rs4847015	0.0526	MH174865.1
	CE19.3	CE19.3_CCTA[1]TCTA[14]TCA[1]TCTA[4]_+6C>T_rs4847015	0.0097	MH174866.1
	CE20.3	CE20.3_CCTA[1]TCTA[15]TCA[1]TCTA[4]_+6C>T_rs4847015	0.0014	N/A

Cont.

locus	CE	MPS allele	frequency	STRSeq #
D21S11	CE24.2	CE24.2_TCTA[5]TCTG[6]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[9]	0.0014	MH174712.1
	CE25.2	CE25.2_TCTA[5]TCTG[6]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	0.0014	N/A
	CE26	CE26_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[8]	0.0014	MH174714.1
	CE27	CE27_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[9]	0.0291	MH174720.1
		CE27_TCTA[6]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[8]	0.0042	MH174716.1
	CE28	CE28_TCTA[4]TCTG[6]TCTA[2]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]	0.0028	N/A
		CE28_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	0.1468	MH174728.1
		CE28_TCTA[6]TCTG[4]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	0.0014	N/A
	CE29	CE29_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]	0.1773	MH174737.1
		CE29_TCTA[4]TCTG[7]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	0.0014	MH174738.1
		CE29_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[2]TCA[1]TCTA[2]TCCATA[1]TCTA[11]	0.0014	MH174735.1
		CE29_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	0.0028	MH174736.1
		CE29_TCTA[6]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	0.0402	MH174731.1
	CE29.3	CE29.3_TCTA[6]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[7]TCA[1]TCTA[3]	0.0014	MH174742.1
	CE30	CE30_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]	0.0817	MH174748.1
		CE30_TCTA[4]TCTG[6]TCTA[4]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]	0.0014	N/A
		CE30_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]	0.0360	MH174747.1
		CE30_TCTA[6]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]	0.1316	MH174745.1
		CE30_TCTA[7]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]	0.0014	MH174743.1
	CE30.2	CE30.2_TCTA[5]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]TA[1]TCTA[1]	0.0111	MH174751.1
		CE30.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[10]TA[1]TCTA[1]	0.0166	MH174753.1
		CE30.2_TCTA[6]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[2]TA[1]TCTA[10]	0.0014	N/A
	CE31	CE31_TCTA[4]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[13]	0.0097	MH174761.1
		CE31_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]	0.0374	MH174760.1
		CE31_TCTA[6]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]	0.0277	MH174758.1
		CE31_TCTA[7]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]	0.0083	MH174756.1
	CE31.2	CE31.2_TCTA[5]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]TA[1]TCTA[1]	0.0014	MH174764.1
		CE31.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]TA[1]TCTA[1]	0.0956	MH174766.1
	CE32	CE32_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[13]	0.0069	MH174773.1
		CE32_TCTA[6]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[13]	0.0055	MH174770.1
		CE32_TCTA[8]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[11]	0.0014	MK569900.1
	CE32.2	CE32.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]TA[1]TCTA[1]	0.0942	MH174779.1
	CE33	CE33_TCTA[6]TCTG[5]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[14]	0.0014	MH174781.1
	CE33.2	CE33.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[13]TA[1]TCTA[1]	0.0125	MH174789.1
		CE33.2_TCTA[6]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[12]TA[1]TCTA[1]	0.0014	MH174787.1
	CE34.2	CE34.2_TCTA[5]TCTG[6]TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCATA[1]TCTA[14]TA[1]TCTA[1]	0.0028	MH174797.1

Cont.

locus	CE	MPS allele	frequency	STRSeq #
D22S1045	CE11	CE11_ATT[8]ACT[1]ATT[2]	0.1593	MH167269.1
	CE12	CE12_ATT[9]ACT[1]ATT[2]	0.0042	MH167270.2
	CE13	CE13_ATT[10]ACT[1]ATT[2]	0.0028	MH167273.1
	CE14	CE14_ATT[11]ACT[1]ATT[2]	0.0332	MH167274.2
	CE15	CE15_ATT[12]ACT[1]ATT[2]	0.3366	MH167275.2
	CE16	CE16_ATT[13]ACT[1]ATT[2]	0.3781	MH167279.2
	CE17	CE17_ATT[14]ACT[1]ATT[2]	0.0776	MH167280.2
	CE18	CE18_ATT[15]ACT[1]ATT[2]	0.0083	MH167281.1
locus	CE	MPS allele	frequency	STRSeq #
D2S1338	CE16	CE16_GGAA[10]GGCA[6]_-35C>A_rs6736691	0.0568	MH105114.1
	CE17	CE17_GGAA[10]GGCA[7]_-35C>A_rs6736691	0.0042	MH105119.1
		CE17_GGAA[11]GGCA[6]_-35C>A_rs6736691	0.1717	MH105118.1
	CE18	CE18_GGAA[11]GGCA[7]	0.0249	MH105129.1
		CE18_GGAA[12]GGCA[6]	0.0055	MH105128.1
		CE18_GGAA[12]GGCA[6]_-35C>A_rs6736691	0.0443	MH105124.1
	CE19	CE19_GGAA[11]GGCA[8]	0.0014	MH105136.1
		CE19_GGAA[12]GGCA[7]	0.0956	MH105135.1
		CE19_GGAA[12]GGCA[7]_-35C>A_rs6736691	0.0014	N/A
		CE19_GGAA[13]GGCA[6]	0.0125	MH105134.1
		CE19_GGAA[13]GGCA[6]_-35C>A_rs6736691	0.0083	MH105132.1
	CE20	CE20_GGAA[12]GGGA[1]GGCA[7]	0.0083	MH105148.1
		CE20_GGAA[13]GGCA[7]	0.1094	MH105146.1
		CE20_GGAA[14]GGCA[6]	0.0083	MH105145.1
		CE20_GGAA[2]GGAC[1]GGAA[10]GGCA[7]	0.0249	MH105151.1
	CE21	CE21_GGAA[14]GGCA[7]	0.0180	MH105155.1
		CE21_GGAA[15]GGCA[6]	0.0014	MH105154.1
		CE21_GGAA[2]GGAC[1]GGAA[11]GGCA[7]	0.0097	MH105159.1
		CE21_GGAA[2]GGAC[1]GGAA[12]GGCA[6]	0.0014	MH105158.1
	CE22	CE22_GGAA[15]GGCA[7]	0.0014	MH105161.1
		CE22_GGAA[2]GGAC[1]GGAA[12]GGCA[7]	0.0332	MH105166.1
		CE22_GGAA[2]GGAC[1]GGAA[13]GGCA[6]	0.0028	MH105165.1
	CE23	CE23_GGAA[16]GGCA[7]	0.0014	MH105167.1
		CE23_GGAA[2]GGAC[1]GGAA[13]GGCA[7]	0.0900	MH105171.1
		CE23_GGAA[2]GGAC[1]GGAA[14]GGCA[6]	0.0055	MH105170.1
	CE24	CE24_GGAA[2]GGAC[1]GGAA[13]GGCA[8]	0.0042	MH105176.1
		CE24_GGAA[2]GGAC[1]GGAA[14]GGCA[7]	0.1205	MH105175.1
		CE24_GGAA[2]GGAC[1]GGAA[15]GGCA[6]	0.0042	MH105174.1
	CE25	CE25_GGAA[2]GGAC[1]GGAA[14]GGCA[8]	0.0055	MH105179.1
		CE25_GGAA[2]GGAC[1]GGAA[15]GGCA[7]	0.0983	MH105178.1
		CE25_GGAA[2]GGAC[1]GGAA[16]GGCA[6]	0.0028	MH105177.1
	CE26	CE26_GGAA[2]GGAC[1]GGAA[15]GGCA[8]	0.0014	MH105183.1
		CE26_GGAA[2]GGAC[1]GGAA[16]GGCA[7]	0.0194	MH105182.1
		CE26_GGAA[2]GGAC[1]GGAA[17]GGCA[6]	0.0014	MH105181.1

Cont.

locus	CE	MPS allele	frequency	STRSeq #
D2S441	CE9	CE9_TCTA[9]	0.0028	MH167314.1
	CE10	CE10_TCTA[10]	0.0693	MH167317.2
		CE10_TCTA[8]TCTG[1]TCTA[1]	0.1274	MH167318.2
	CE11	CE11_TCTA[11]	0.3158	MH167320.2
		CE11_TCTA[11]_-25G>A_rs74640515	0.0222	MH167319.1
		CE11_TCTA[8]TTTA[1]TCTA[2]	0.0014	N/A
		CE11_TCTA[9]TCTG[1]TCTA[1]	0.0042	MH167321.2
		CE11_TCTA[3]TCCA[1]TCTA[7]	0.0014	N/A
	CE11.3	CE11.3_TCTA[4]TCA[1]TCTA[7]	0.0416	MH167323.1
		CE12_TCTA[10]TCTG[1]TCTA[1]	0.0014	MH167326.1
	CE12	CE12_TCTA[12]	0.0235	MH167325.1
		CE12_TCTA[9]TTTA[1]TCTA[2]	0.0014	MH167327.1
	CE13	CE13_TCTA[10]TTTA[1]TCTA[2]	0.0277	MH167331.1
		CE13_TCTA[13]	0.0014	MH167329.1
	CE14	CE14_TCTA[11]TTTA[1]TCTA[2]	0.3061	MH167334.2
	CE15	CE15_TCTA[12]TTTA[1]TCTA[2]	0.0457	MH167337.1
	CE16	CE16_TCTA[13]TTTA[1]TCTA[2]	0.0069	MH167338.1

locus	CE	MPS allele	frequency	STRSeq #
D3S1358	CE13	CE13_TCTA[1]TCTG[2]TCTA[10]	0.0028	MH166963.1
	CE14	CE14_TCTA[1]TCTG[1]TCTA[12]	0.0014	MH166964.1
		CE14_TCTA[1]TCTG[2]TCTA[11]	0.1191	MH166965.2
	CE15	CE15_TCTA[1]TCTG[1]TCTA[13]	0.0222	MH166968.2
		CE15_TCTA[1]TCTG[2]TCTA[12]	0.2687	MH166969.2
		CE15_TCTA[1]TCTG[3]TCTA[11]	0.0055	MH166971.1
	CE16	CE16_TCTA[1]TCTG[1]TCTA[14]	0.0125	MH166974.1
		CE16_TCTA[1]TCTG[2]TCTA[13]	0.1399	MH166975.1
		CE16_TCTA[1]TCTG[3]TCTA[12]	0.0693	MH166976.1
		CE16_TCTA[1]TCTG[3]TCTA[1]TCTG[1]TCTA[10]	0.0028	MH166977.1
	CE17	CE17_TCTA[1]TCTG[1]TCTA[15]	0.0069	MH166980.1
		CE17_TCTA[1]TCTG[2]TCTA[14]	0.0970	MH166981.2
		CE17_TCTA[1]TCTG[3]TCTA[13]	0.1066	MH166982.1
		CE17_TCTA[1]TCTG[3]TCTA[13]_-36T>C_rs1045901660	0.0014	N/A
		CE17_TCTA[1]TCTG[4]TCTA[12]	0.0042	MH166983.1
	CE18	CE18_TCTA[1]TCTG[1]TCTA[16]	0.0014	MH166984.1
		CE18_TCTA[1]TCTG[2]TCTA[15]	0.0111	MH166985.1
		CE18_TCTA[1]TCTG[3]TCTA[14]	0.1094	MH166986.2
		CE18_TCTA[1]TCTG[4]TCTA[13]	0.0028	MH166987.1
	CE19	CE19_TCTA[1]TCTG[2]TCTA[16]	0.0014	MK990352.1
		CE19_TCTA[1]TCTG[3]TCTA[15]	0.0111	MH166988.1
	CE20	CE20_TCTA[1]TCTG[3]TCTA[16]	0.0028	MH166990.1

locus	CE	MPS allele	frequency	STRSeq #
D5S818	CE9	CE9_ATCT[10]_+13A>G_rs25768	0.0346	MH166995.1
	CE10	CE10_ATCT[11]_+13A>G_rs25768	0.0222	MH166997.1
		CE10_CTCT[1]ATCT[10]	0.0083	MH166998.1
	CE11	CE10_CTCT[1]ATCT[10]_+13A>G_rs25768	0.0305	MH166999.1
		CE11_ATCT[12]_+13A>G_rs25768	0.0235	MH167001.1
		CE11_CTCT[1]ATCT[11]	0.0873	MH167002.1
		CE11_CTCT[1]ATCT[11]_+13A>G_rs25768	0.2438	MH167004.1
		CE12_ATCT[13]_+13A>G_rs25768	0.0983	MH167005.1
	CE12	CE12_CTCT[1]ATCT[12]	0.0789	MH167007.1
		CE12_CTCT[1]ATCT[12]_+13A>G_rs25768	0.1939	MH167008.1
		CE13_ATCT[14]_+13A>G_rs25768	0.0332	MH167009.1
	CE13	CE13_CTCT[1]ATCT[13]	0.0789	MH167011.1
		CE13_CTCT[1]ATCT[13]_+13A>G_rs25768	0.0554	MH167013.1
		CE14_ATCT[15]_+13A>G_rs25768	0.0014	MH167015.1
	CE14	CE14_CTCT[1]ATCT[14]	0.0069	MH167016.1
		CE14_CTCT[1]ATCT[14]_+13A>G_rs25768	0.0028	MH167017.1

Cont.

locus	CE	MPS allele	frequency	STRSeq #
D7S820	CE7	CE7_TATC[7]_-22T>A_rs7789995	0.0166	MH167024.1
	CE8	CE8_TATC[8]_-22T>A_rs7789995	0.1163	MH167026.1
		CE8_TATC[8]_-22T>A_rs7789995_+9G>A_rs16887642	0.0402	MH167025.1
	CE9	CE9_TATC[9]	0.0028	MH167031.1
		CE9_TATC[9]_-22T>A_rs7789995	0.1690	MH167030.1
		CE9_TATC[9]_-22T>A_rs7789995_+9G>A_rs16887642	0.0042	MH167029.1
	CE10	CE9_TATC[9]_-65A>C_rs7786079_-22T>A_rs7789995	0.0014	MH167032.1
		CE10_TATC[10]	0.0596	MH167035.1
		CE10_TATC[10]_-22T>A_rs7789995	0.2078	MH167034.1
	CE11	CE10_TATC[10]_-65A>C_rs7786079_-22T>A_rs7789995	0.0083	MH167037.1
		CE11_TATC[11]	0.0180	MH167043.1
		CE11_TATC[11]_-22T>A_rs7789995	0.1482	MH167041.1
	CE12	CE11_TATC[11]_-65A>C_rs7786079_-22T>A_rs7789995	0.0069	MH167044.1
		CE12_TATC[12]	0.0402	MH167048.1
		CE12_TATC[12]_-22T>A_rs7789995	0.1233	MH167047.1
	CE13	CE13_TATC[13]	0.0125	MH167051.1
		CE13_TATC[13]_-22T>A_rs7789995	0.0208	MH167050.1
	CE14	CE14_TATC[14]	0.0028	MH167054.1
		CE14_TATC[14]_-22T>A_rs7789995	0.0014	MH167053.1

locus	CE	MPS allele	frequency	STRSeq #
D8S1179	CE8	CE8_TCTA[8]	0.0180	MH105186.1
	CE9	CE9_TCTA[9]	0.0097	MH105187.1
	CE10	CE10_TCTA[10]	0.1108	MH105188.1
		CE10_TCTA[10]_+67C>T_rs138862078	0.0014	MH105189.1
	CE11	CE11_TCTA[11]	0.0942	MH105190.1
	CE12	CE12_TCTA[1]TCTG[1]TCTA[10]	0.0069	MH105196.1
		CE12_TCTA[12]	0.1260	MH105194.1
	CE13	CE13_TCTA[1]TCTG[1]TCTA[11]	0.2673	MH105201.1
		CE13_TCTA[13]	0.0526	MH105197.1
		CE13_TCTA[2]TCTG[1]TCTA[10]	0.0028	MH105198.1
	CE14	CE14_TCTA[1]TCTG[1]TCTA[12]	0.1357	MH105206.1
		CE14_TCTA[1]TCTG[1]TGTA[1]TCTA[11]	0.0028	MH105207.1
		CE14_TCTA[14]	0.0166	MH105204.1
		CE14_TCTA[2]TCTG[1]TCTA[11]	0.0291	MH105205.1
	CE15	CE15_TCTA[1]TCTG[1]TCTA[13]	0.0346	MH105211.1
		CE15_TCTA[15]	0.0042	MH105208.1
		CE15_TCTA[2]TCTG[1]TCTA[12]	0.0568	MH105209.1
	CE16	CE16_TCTA[1]TCTG[1]TCTA[14]	0.0055	MH105216.1
		CE16_TCTA[2]TCTG[1]TCTA[13]	0.0222	MH105214.1
	CE17	CE17_TCTA[1]TCTG[1]TCTA[15]	0.0014	MH105218.1
		CE17_TCTA[2]TCTG[1]TCTA[14]	0.0014	MH105217.1

Cont.

locus	CE	MPS allele	frequency	STRSeq #
FGA	CE17	CE17_GGAA[2]GGAG[1]AAAG[9]AGAA[1]AAAA[1]GAAA[3]	0.0014	MH232603.1
	CE18	CE18_GGAA[2]GGAG[1]AAAG[10]AGAA[1]AAAA[1]GAAA[3]	0.0125	MH232605.1
	CE19	CE19_GGAA[2]GGAG[1]AAAG[11]AGAA[1]AAAA[1]GAAA[3]	0.0693	MH232607.1
	CE19.2	CE19.2_GGAA[2]GGAG[1]AAAG[12]AA[1]AAAA[1]GAAA[3]	0.0014	MH232608.1
	CE20	CE20_GGAA[2]GGAG[1]AAAG[12]AGAA[1]AAAA[1]GAAA[3]	0.1468	MH232609.1
	CE20.2	CE20.2_GGAA[2]GGAG[1]AAAG[13]AA[1]AAAA[1]GAAA[3]	0.0028	MK570043.1
	CE21	CE21_GGAA[2]GGAG[1]AAAG[13]AGAA[1]AAAA[1]GAAA[3]	0.1717	MH232611.1
	CE21.2	CE21.2_GGAA[2]GGAG[1]AAAG[14]AA[1]AAAA[1]GAAA[3]	0.0014	MH232612.1
	CE22	CE22_GGAA[2]GGAG[1]AAAG[14]AGAA[1]AAAA[1]GAAA[3]	0.1925	MH232613.1
	CE22.2	CE22.2_GGAA[2]GGAG[1]AAAG[15]AA[1]AAAA[1]GAAA[3]	0.0097	MH232615.1
	CE23	CE23_GGAA[2]GGAG[1]AAAG[15]AGAA[1]AAAA[1]GAAA[3]	0.1427	MH232617.1
	CE23.2	CE23.2_GGAA[2]GGAG[1]AAAG[16]AA[1]AAAA[1]GAAA[3]	0.0042	MH232620.1
	CE24	CE24_GGAA[2]GGAG[1]AAAG[16]AGAA[1]AAAA[1]GAAA[3]	0.1150	MH232622.1
	CE25	CE25_GGAA[2]GGAG[1]AAAG[1]AAAC[1]AAAG[15]AGAA[1]AAAA[1]GAAA[3]	0.0042	MH232625.1
		CE25_GGAA[2]GGAG[1]AAAG[17]AGAA[1]AAAA[1]GAAA[3]	0.0706	MH232626.1
	CE26	CE26_GGAA[2]GGAG[1]AAAG[1]AAAC[1]AAAG[16]AGAA[1]AAAA[1]GAAA[3]	0.0028	MH232628.1
		CE26_GGAA[2]GGAG[1]AAAG[18]AGAA[1]AAAA[1]GAAA[3]	0.0429	MH232629.1
	CE27	CE27_GGAA[2]GGAG[1]AAAG[1]AAAC[1]AAAG[17]AGAA[1]AAAA[1]GAAA[3]	0.0014	N/A
		CE27_GGAA[2]GGAG[1]AAAG[19]AGAA[1]AAAA[1]GAAA[3]	0.0069	MH232632.1
locus	CE	MPS allele	frequency	STRSeq #
PentaD	CE6	CE6_AAAGA[6]AAAAA[1]	0.0014	MH232673.1
	CE7	CE7_AAAGA[7]AAAAA[1]	0.0111	MH232674.1
	CE8	CE8_AAAGA[8]AAAAA[1]	0.0139	MH232675.2
		CE8_AAAGA[8]AAAAA[1]_+57T>G_rs7279663	0.0014	N/A
	CE9	CE9_AAAGA[9]AAAAA[1]	0.1773	MH232676.2
		CE9_AAAGA[9]AAAAA[1]_+57T>G_rs7279663	0.0208	N/A
	CE10	CE10_AAAGA[10]AAAAA[1]	0.1080	MH232678.1
		CE10_AAAGA[10]AAAAA[1]_+57T>G_rs7279663	0.0042	N/A
	CE11	CE11_AAAGA[11]AAAAA[1]	0.1122	MH232681.2
		CE11_AAAGA[11]AAAAA[1]_+57T>G_rs7279663	0.0014	N/A
	CE12	CE12_AAAGA[12]AAAAA[1]	0.2230	MH232685.1
		CE12_AAAGA[13]	0.0055	MH232686.1
	CE13	CE13_AAAGA[13]AAAAA[1]	0.2299	MH232687.2
		CE13_AAAGA[14]	0.0042	MK990342.1
	CE14	CE14_AAAGA[14]AAAAA[1]	0.0693	MH232690.2
	CE15	CE15_AAAGA[15]AAAAA[1]	0.0139	MH232692.1
	CE16	CE16_AAAGA[16]AAAAA[1]	0.0014	MH232693.1
CE17	CE17_AAAGA[17]AAAAA[1]	0.0014	MH232694.1	

Cont.

locus	CE	MPS allele	frequency	STRSeq #
PentaE	CE5	CE5_TCTTT[5]	0.0859	MH232642.1
	CE7	CE7_TCTTT[7]	0.1856	MH232644.1
	CE8	CE8_TCTTT[8]	0.0139	MH232645.1
	CE9	CE9_TCTTT[9]	0.0055	MH232646.1
	CE10	CE10_TCTTT[10]	0.0776	MH232647.1
	CE11	CE11_TCTTT[11]	0.0997	MH232648.1
	CE12	CE12_TCTTT[12]	0.1745	MH232649.1
	CE13	CE13_TCTTT[13]	0.0859	MH232650.1
	CE14	CE14_TCTTT[14]	0.0554	MH232651.1
	CE14.1	CE14.1_TCTTT[14]_-14.1->A_rs1164568524	0.0014	N/A
	CE15	CE15_TCTTT[15]	0.0568	MH232653.1
	CE16	CE16_TCTTT[16]	0.0693	MH232657.1
	CE17	CE17_TCTTT[17]	0.0402	MH232660.1
	CE18	CE18_TCTTT[18]	0.0222	MH232662.1
	CE19	CE19_TCTTT[19]	0.0139	MH232663.1
	CE20	CE20_TCTTT[20]	0.0069	MH232665.1
	CE21	CE21_TCTTT[21]	0.0042	MH232666.1
	CE22	CE22_TCTTT[22]	0.0014	MH232667.1
locus	CE	MPS allele	frequency	STRSeq #
TH01	CE5	CE5_AATG[5]	0.0042	MH085114.1
	CE6	CE6_AATG[6]	0.2147	MH085115.1
	CE7	CE7_AATG[7]	0.2022	MH085116.1
		CE7_AATG[7]_+136C>T_rs112257019	0.0014	N/A
	CE8	CE8_AATG[8]	0.1080	MH085119.1
	CE9	CE9_AATG[9]	0.1274	MH085121.1
	CE9.3	CE9.3_AATG[6]ATG[1]AATG[3]	0.3338	MH085124.1
	CE10	CE10_AATG[10]	0.0083	MH085125.1
locus	CE	MPS allele	frequency	STRSeq #
TPOX	CE8	CE8_AATG[8]	0.5014	MF044248.2
		CE8_AATG[8]_-52G>T_rs149212737	0.0097	MF044250.1
	CE9	CE9_AATG[9]	0.1150	MF044251.2
	CE10	CE10_AATG[10]	0.0665	MF044253.2
	CE11	CE11_AATG[11]	0.2659	MF044255.2
	CE12	CE12_AATG[12]	0.0416	MF044256.2

Cont.

locus	CE	MPS allele	frequency	STRSeq #
vWA	CE13	CE13_TGGA[2]TAGA[1]TGGA[1]TAGA[8]CAGA[4]TAGA[1]	0.0014	MH167072.1
		CE13_TGGA[2]TAGA[1]TGGA[1]TAGA[9]CAGA[3]TAGA[1]	0.0014	MH167073.1
	CE14	CE14_TGGA[2]TAGA[1]TGGA[1]TAGA[10]CAGA[3]TAGA[1]	0.0249	MH167077.1
		CE14_TGGA[2]TAGA[1]TGGA[1]TAGA[9]CAGA[4]TAGA[1]	0.0028	MH167076.1
		CE14_TGGA[4]TAGA[3]TGGA[1]TAGA[3]CAGA[4]TAGA[1]CAGA[1]TAGA[1]_+72A> T_rs11063969_+77C>T_rs11063970_+90T>C_rs11063971	0.0885	MH167078.1
		CE15_TGGA[2]TAGA[1]TGGA[1]TAGA[10]CAGA[4]TAGA[1]	0.0194	MH167081.1
	CE15	CE15_TGGA[2]TAGA[1]TGGA[1]TAGA[11]CAGA[3]TAGA[1]	0.0913	MH167082.1
		CE15_TGGA[5]TAGA[3]TGGA[1]TAGA[3]CAGA[4]TAGA[1]CAGA[1]TAGA[1]_+72A> T_rs11063969_+77C>T_rs11063970_+90T>C_rs11063971	0.0041	MH167083.1
		CE16_TGGA[2]TAGA[1]TGGA[1]TAGA[11]CAGA[4]TAGA[1]	0.1743	MH167084.1
	CE16	CE16_TGGA[2]TAGA[1]TGGA[1]TAGA[12]CAGA[3]TAGA[1]	0.0332	MH167085.1
		CE17_TGGA[2]TAGA[1]TGGA[1]TAGA[11]CAGA[5]TAGA[1]	0.0028	MH167086.1
	CE17	CE17_TGGA[2]TAGA[1]TGGA[1]TAGA[12]CAGA[4]TAGA[1]	0.2393	MH167088.1
		CE17_TGGA[2]TAGA[1]TGGA[1]TAGA[13]CAGA[3]TAGA[1]	0.0138	MK569948.1
		CE17_TGGA[2]TAGA[1]TGGA[1]TAGA[5]CAGA[1]TAGA[6]CAGA[4]TAGA[1]	0.0014	N/A
	CE18	CE18_TGGA[2]TAGA[1]TGGA[1]TAGA[13]CAGA[4]TAGA[1]	0.1936	MH167093.1
		CE18_TGGA[2]TAGA[1]TGGA[1]TAGA[13]CAGA[4]TAGA[1]_+44C>T_rsN/A_@chr12:5,984,088	0.0028	N/A
		CE18_TGGA[2]TAGA[1]TGGA[1]TAGA[13]CAGA[5]	0.0014	MK569949.1
		CE18_TGGA[2]TAGA[1]TGGA[1]TAGA[14]CAGA[3]TAGA[1]	0.0028	MH167094.1
	CE19	CE19_TGGA[2]TAGA[1]TGGA[1]TAGA[13]CAGA[5]TAGA[1]	0.0028	MH167096.1
		CE19_TGGA[2]TAGA[1]TGGA[1]TAGA[14]CAGA[4]TAGA[1]	0.0775	MH167097.1
		CE19_TGGA[2]TAGA[1]TGGA[1]TAGA[15]CAGA[3]TAGA[1]	0.0014	MH167098.1
	CE20	CE20_TGGA[2]TAGA[1]TGGA[1]TAGA[14]CAGA[5]TAGA[1]	0.0014	MH167100.1
		CE20_TGGA[2]TAGA[1]TGGA[1]TAGA[15]CAGA[4]TAGA[1]	0.0166	MH167101.1
	CE24	CE24_TGGA[2]TAGA[1]TGGA[1]TAGA[3]TGGA[1]TAGA[1]TGGA[1]TAGA[12]CAGA[5]TAGA[1]	0.0014	N/A

6.3.4.2 Testing population differentiation using aSTR variants

The observed aSTR length and sequence variants were used to test population differentiation in the 361 samples. Variable length alleles and sequence features were input to the Arlequin v 3.5.2.2 software (Excoffier and Lischer 2010) and summary statistics were generated including molecular diversity indices, F-statistics for population comparisons and differentiation (Table 6.17). The different population divisions tested are detailed in Section 6.2.2.

Table 6.17
AMOVA statistics and pairwise R_{ST} values for aSTR variants.

Table (a) shows the AMOVA statistics, and some basic statistics, significant values are in bold (b) shows the pairwise R_{ST} and gene diversity matrices per multi-component division. The pairwise R_{ST} and gene diversity matrix for the 37 population divisions is supplied as Appendix G.

a.

AMOVA	R_{ST}	P		Variation between populations	Variation within populations
37 populations	0.00296	0.21822	(± 0.00433)	0.03%	99.97%
10 geographic regions	0.00012	0.36852	(± 0.01226)	0.01%	99.99%
5 PoBI defined regions	0.00354	0.05663	(± 0.00228)	0.35%	99.65%
Pairwise	R_{ST}	P			
Celtic fringe vs rest	0.00019	0.32942	(± 0.01417)	0.02%	99.98%
Danelaw vs rest	0	0.71945	(± 0.00831)	0.00%	100.00%
CSE England vs rest	0.00039	0.13685	(± 0.01002)	0.04%	99.96%
East vs rest	0.00049	0.14174	(± 0.01276)	0.05%	99.95%

Cont.

b.

above diagonal P-values (\pm SD)
 diagonal gene diversity (GD)(\pm SD)
 below diagonal pairwise R_{ST} values
 significance=0.001111 (after Bonferroni correction)
 probabilities based on 10,100 permutations, significant values after Bonferroni correction are in **bold**

By LENGTH	Scotland	Ireland, Isle of Man	North West	North East	Wales	West Midlands	East Midlands	East	South West	South East
Scotland	GD=0.79196 \pm 0.0714	0.37184 +0.0044	0.08227 +0.0032	0.68657 +0.0047	0.24750 +0.0038	0.61509 +0.0045	0.31601 +0.0042	0.07445 +0.0026	0.14048 +0.0033	0.20394 +0.0045
Ireland, Isle of Man	0.00029	GD=0.79642 \pm 0.07811	0.12296 +0.0031	0.81655 +0.0041	0.53618 +0.0047	0.84516 +0.0033	0.50233 +0.0049	0.12920 +0.0034	0.74488 +0.0045	0.06247 +0.0023
North West	0.00261	0.00298	GD=0.78911 \pm 0.07724	0.31314 +0.0041	0.28057 +0.0042	0.97852 +0.0015	0.12543 +0.0042	0.58796 +0.0049	0.92941 +0.0028	0.15800 +0.0039
North East	0	0	0.00115	GD=0.78001 \pm 0.08655	0.69290 +0.0045	0.57677 +0.0049	0.43827 +0.0051	0.35432 +0.0044	0.24413 +0.0040	0.25681 +0.0044
Wales	0.00083	0	0.00091	0	GD=0.79793 \pm 0.06349	0.79151 +0.0036	0.13652 +0.0033	0.07445 +0.0025	0.67152 +0.0049	0.56836 +0.0048
West Midlands	0	0	0	0	0	GD=0.79599 \pm 0.07213	0.61331 +0.0048	0.79081 +0.0041	0.94733 +0.0023	0.55480 +0.0053
East Midlands	0.00073	0.00005	0.00248	0.00053	0.00187	0	GD=0.78748 \pm 0.07122	0.34096 +0.0041	0.88179 +0.0031	0.16830 +0.0039
East	0.00287	0.00306	0	0.00091	0.00286	0	0.00093	GD=0.79253 \pm 0.06942	0.89991 +0.0030	0.65023 +0.0051
South West	0.00145	0	0	0.00145	0	0	0	0	GD=0.79645 \pm 0.06331	0.32264 +0.0049
South East	0.00118	0.00395	0.00196	0.00175	0	0	0.00168	0	0.00053	GD=0.79374 \pm 0.06409

c.

above diagonal P-values (\pm SD)
 diagonal gene diversity (GD)(\pm SD)
 below diagonal pairwise R_{ST} values
 significance=0.005 (after Bonferroni correction)
 probabilities based on 10,100 permutations, significant values after Bonferroni correction are in **bold**

By LENGTH	Orkney	Wales	Scotland and North	Cornwall	CSE-England
Orkney	GD=0.78708 \pm 0.07671	0.81576 \pm 0.0039	0.50975 \pm 0.0048	0.06296 \pm 0.0026	0.87734 \pm 0.0034
Wales	0	GD=0.79793 \pm 0.06349	0.05158 \pm 0.0023	0.00564 \pm 0.0007	0.04307 \pm 0.0020
Scotland and North	0	0.00657	GD=0.79303 \pm 0.07209	0.23849 \pm 0.0048	0.35214 \pm 0.0048
Cornwall	0.03275	0.04823	0.0058	GD=0.79139 \pm 0.07274	0.04445 \pm 0.0020
CSE-England	0	0.0052	0.00041	0.02007	GD=0.7917 \pm 0.06779

In spite of the Y-STR markers showing substructure within the population at both length and sequence levels, the autosomal STR data in the analyzed samples do not show any stratification across the British Isles at length levels, whether it is tested in a history-based, geography- or language-based divide. Similarly to mtDNA, no differentiation is detectable in these markers within the British Isles.

6.3.4.3 Summary of aSTR forensic statistics

aSTR-related statistics (Table 6.18) were provided using the browser-based application STRAF 1.0.5 (Gouy and Zieger 2017).

Table 6.18
Standard aSTR forensic parameters.

locus	N	N _{all}	GD	PIC	PM	PD	H _{obs}	PE	TPI	pHW
CSF1P0	722	7	0.726	0.6753	0.132	0.868	0.7535	0.5157	2.0281	0.7576
D10S1248	722	7	0.7415	0.6972	0.1162	0.8838	0.7645	0.535	2.1235	0.0675
D12S391	722	15	0.8975	0.8874	0.0219	0.9781	0.8726	0.7398	3.9239	0.1126
D13S317	722	8	0.7904	0.7607	0.0768	0.9232	0.8061	0.6104	2.5786	0.3051
D16S539	722	7	0.7694	0.7322	0.0902	0.9098	0.7673	0.5398	2.1488	0.1436
D18S51	722	14	0.8733	0.8586	0.0322	0.9678	0.9003	0.796	5.0139	0.7322
D19S433	722	14	0.7583	0.7236	0.102	0.898	0.8006	0.6001	2.5069	0.2203
D1S1656	722	15	0.9013	0.8915	0.0219	0.9781	0.9141	0.8244	5.8226	0.0701
D21S11	722	16	0.8376	0.8169	0.0472	0.9528	0.8061	0.6104	2.5786	0.2378
D22S1045	722	8	0.7122	0.6617	0.1421	0.8579	0.7313	0.4783	1.8608	0.0468
D2S1338	722	11	0.8843	0.8716	0.027	0.973	0.8643	0.7232	3.6837	0.4452
D2S441	722	9	0.7443	0.7018	0.1082	0.8918	0.723	0.4647	1.805	0.1266
D3S1358	722	8	0.786	0.7515	0.0799	0.9201	0.7701	0.5447	2.1747	0.4329
D5S818	722	6	0.7044	0.651	0.147	0.853	0.7368	0.4875	1.9	0.3789
D7S820	722	8	0.8111	0.7829	0.0633	0.9367	0.7978	0.5949	2.4726	0.6509
D8S1179	722	10	0.8136	0.7899	0.0589	0.9411	0.8116	0.6208	2.6544	0.3685
FGA	722	16	0.8667	0.8508	0.0355	0.9645	0.856	0.7066	3.4712	0.2615
PentaD	722	12	0.8241	0.7992	0.0585	0.9415	0.8449	0.6848	3.2232	0.386
PentaE	722	18	0.8919	0.8812	0.0238	0.9762	0.8504	0.6957	3.3426	0.2727
TH01	722	7	0.7741	0.7382	0.0888	0.9112	0.7645	0.535	2.1235	0.0903
TPOX	722	5	0.6496	0.5987	0.1708	0.8292	0.626	0.3233	1.337	0.2639
vWA	722	9	0.8177	0.7913	0.0624	0.9376	0.8338	0.6632	3.0083	0.2386

N	number of samples
N _{all}	number of alleles
GD	Genetic diversity/expected heterozygosity
PIC	Polymorphism Information Content
PM	match probability
PD	Power of discrimination
H _{obs}	Observed heterozygosity
PE	power of exclusion
TPI	typical paternity index
pHW	Hardy-Weinberg p-value

6.4 Discussion

This chapter extended the work presented in previous chapters on a selected global sample, by focusing on a larger and well-defined sample from one location, the British Isles. In doing this, as expected, increased sequence-based diversity was observed at STRs compared to length-based diversity. Also as expected, the observed range of STR alleles and mtDNA CR sequences was less broad than in the smaller global panel. However, the experimental design in this chapter allowed the tools of population genetics to be applied to address population structure via sequence-based analysis of forensic markers.

The People of the British Isles cohort presents a unique opportunity to study the indigenous population of a country otherwise much affected by past and recent migration from other nations, in Europe and beyond. The sampling strategy of the study allowed the selection of donors with ancestries tied to the same rural areas for the last two generations, by limiting the distance within which each donor's grandparents were born to a 50 mile / 80 km radius (Winney et al. 2012). The majority of the sample donors were over 60 years of age, and thus with two generations back the study randomly sampled the genomes of ancestors who were born on average in the late nineteenth century. From the complete sample collection of 2039 individuals, here a subset of 362 males was selected and their forensic markers sequenced, including 23 Y-STRs, 22 aSTRs, the mtDNA control region and the Amelogenin sex-test marker.

Mitochondrial DNA control region sequencing showed a diverse set of haplotypes, distributed among thirteen major mtDNA haplogroups, but these variants did not show geographical stratification in Britain. The lack of maternal lineage differentiation is consistent with the broader-scale pattern seen in indigenous populations across Western Europe (Richards et al. 2000). It fits a widely accepted picture in which patrilocality (the tendency of women to move to a husband's place of birth on marriage) is predicted to have a homogenising effect on mtDNA landscapes (Wilkins and Marlowe 2006).

Autosomal STR markers were also interrogated to assess population substructure; however, aSTR genotype data also failed to reveal any geographical differentiation in their variants, and showed no differences over historically relevant geographical and linguistic divisions. These markers were originally chosen and developed for individual identification: their high variability arises thanks to high mutation rates, and F_{ST} considering multi-locus genotypes is expected to be low. At inter-continental scales such aSTR genotypes do reveal some differentiation, allowing a noisy ancestry-informative signal to emerge, but within continents population structure is weak or negligible (Phillips et al. 2011). It is therefore unsurprising that no discernible geographical structuring is observed here, at the even smaller geographical scale of the British Isles.

In contrast to the maternal and biparental picture, however, the distribution of Y-STR haplotypes revealed a clear substructuring of male lineages reflecting a general East-West differentiation. The most obvious candidate as a historical cause for this is the mass migration of Anglo-Saxons around 500 CE; this event is generally believed to be responsible for the linguistic divide between English (Germanic) and Celtic languages, and indeed when this linguistically-based division is tested it reveals the same Y-based differentiation. Likewise, the autosomally-defined Central/Southeast England cluster, as identified by the PoBI study, also shows a Y-chromosomal difference. The interpretation by Leslie et al. (2015) of the autosomal major cluster was also based around Anglo-Saxon migration. Previous studies, in particular that of (Weale et al. 2002), have observed East-West differentiation in Y lineages and proposed the Anglo-Saxon influence as an explanation. However, although the current study provides strong evidence of the differentiation and gives some detail on its distribution, it cannot provide any timescale for the events underlying the observed pattern. The best way to date the differentiation would be the haplotyping of samples from ancient DNA in suitable well-dated burial sites, if these can be found. It is also important to consider more general factors, for example the distance to mainland Europe, that could also set a Northwest-Southeast gradient from where a genetically more diverse or different genetic contribution may have been continuously incorporated into the genomes of

Britons. Explicit modelling approaches, such as Approximate Bayesian Computation (Beaumont et al. 2002), could be used to differentiate between such different explanations.

The sequence analysis of Y-STRs not only increased the number of discernible alleles, thereby improving the forensic applicability of these markers, but also unveiled associations of several of these Y-STR sequence features including unique repeat patterns, SNPs and indels, to Y-chromosome haplogroups. This further demonstrates the versatility of these markers, as used by the forensic geneticist and genealogist communities.

6.4.1 Highlights and conclusions

The most commonly used forensic markers, the autosomal and Y-STRs and the mtDNA control region were sequenced and population data was generated from a set of 362 male samples from the People of the British Isles cohort, representative of the indigenous people of Britain.

Sequencing increased the allele diversity of STR alleles and revealed associations between the underlying sequence variants of Y-STRs and the Y-phylogeny.

Population genetic methods found that Y-STRs show a clear stratification of the male lineages, representing an East-West differentiation that could be associated with the Anglo-Saxon mass migrations around 500 CE.

CHAPTER 7: Discussion and future directions

7.1 MPS technology in forensic DNA analysis

With the introduction of massively parallel sequencing, the analysis of forensic STRs has benefited from the potential to describe allelic variants not just by their lengths, but also by their internal array or flanking region variants. The advantages of implementing MPS in the forensic workflow are the increase in allelic diversity, the increase in heterozygosity and the improved resolution of isoalleles (alleles with the same lengths, but different sequences). These advantages can benefit complex mixture deconvolution, and generate more information from a given sample. Implementation of MPS; however, requires investment in new technology platforms, laboratory procedures, analysis and building expertise in interpretation, as well as the generation of new reference databases. These require substantial change compared to the standard CE-based methods, although basic interpretation concepts (e.g. relative fluorescence vs coverage) are still transferable, and MPS-based profiles are back-compatible with CE profiles. With any newly implemented area comes the necessity of a unified nomenclature, which at the moment, is still a collaborative work in progress, but is required to prevent the subjective description of alleles that could lead to potential errors of interpretation.

7.2 Summary of the results

In Chapter 3 of this thesis (Huszar et al. 2018) Y-STRs were analysed using MPS in a diverse set of samples based on a known phylogeny of Y chromosomes in order to capture a wide variety of Y-STR sequences, and place these in an evolutionary context. This resulted in an increase in allele diversity across 19 of the typed 23 Y-STRs, which mostly resulted from intra-array structural variation and to a lesser extent from flanking region variants. The phylogenetically-framed approach resulted in a wider range of allelic variants at the sequence level than is likely to be found in a population-based approach of comparable sample size, and identified 60

alleles (of which at the time of writing 58 are still not catalogued) not described elsewhere. The variants resulting from SNPs, indels and RPVs showed different occurrence patterns within the phylogeny: SNPs and indels, with their lower mutation rates, were mostly monophyletic, while RPVs showed examples of polyphyletic patterns across the Y-phylogeny sampled here. As a result of describing a wider range of repeat-region sequence variants, suggestions to the current nomenclature were made which could be used to unambiguously express sequenced alleles.

In Chapter 4, autosomal STRs were analysed in the same samples using MPS. Compared to length variation, this analysis resulted in an increase in allele diversity across 21 of the 22 typed aSTRs, which mostly resulted from structural variation within repeat arrays; although some loci only presented flanking region variants. Sequence analysis resolved 4.5% of all genotypes as isometric heterozygotes: homozygous by length, but heterozygous by sequence. Most of these pairs were observed in loci with relatively narrow length-based allele ranges combined with several sequence-based variants per length class. The Y-chromosomal diversity-driven sample selection, because of the associated high ethnic diversity, also resulted in a wide range of sequence variants at the autosomal STR loci, and identified 48 alleles not previously described in the STRSeq sequence allele database.

In Chapter 5 (Huszar et al. 2019), mtDNA control region diversity was analysed using MPS in the same samples, showing how mtDNA variant typing behaved within the same workflow. The typing of a panel of highly diverse global samples led to correspondingly high phylogenetic diversity. The work here demonstrates that the MPS approach can provide a robust tool for the analysis of mtDNA control region variation. In this process the effects of single-reaction tiled multiplex typing were considered, and different kit designs were compared for reliability of variant calling and the assessment of heteroplasmy. A key finding was that reference sequence bias, if not properly accounted for, can lead to misleading results on variant calling and heteroplasmy assessment. To mitigate this problem, a

bioinformatic method, OREO, was developed to decrease the effects of non-uniform coverage and reference sequence bias in overlapping amplicon sequencing studies.

In Chapter 6, the described MPS methods were used to analyse sequences for forensic markers in a set of samples from indigenous males from the British Isles. Sequence data were generated for Y-STRs, autosomal STRs and the mtDNA control region in a panel of 362 samples from the People of the British Isles cohort. The sequenced STR alleles were catalogued and compared to known variant alleles. Population genetic methods were used to compare the geographical distributions of Y, autosomal and mitochondrial diversity within the British Isles, and to observe any effects of sex-biased historical processes. The key finding was that population structure for autosomal and mtDNA sequences is negligible, but Y-chromosomal structure is marked, and can be described by a east-west differentiation that could be a reflection of male-mediated Anglo-Saxon mass migration around 500 CE, and correlates with the Celtic vs Germanic linguistic divide between the Atlantic fringe compared to the majority of Central, Eastern Southern Britain. Formal testing of this interpretation will rely on modelling approaches and generation of a catalogue of appropriate ancient DNA evidence from well-dated sources. The generated data from these well-characterised samples will provide a valuable forensic reference dataset for the indigenous people of the British Isles.

7.3 Considerations relating to MPS and future directions

7.3.1 Concordance between CE and MPS

As MPS is set to play a part in future forensic toolkits, it was necessary to compare its compatibility with data generated by standard CE-based STR typing. Discordance between different CE typing kits exists (Drabek et al. 2004; Hill et al.

2007), as a result of different primers targeting the same marker loci, and therefore it was expected there would be similar occasional discordances between CE- and MPS-typed alleles as well. Reported discordances are mainly due to one or other method encompassing flanking region variants, primarily indels that affect the fragment size, and hence the length-based allelic designation, which, depending on primer positions can result in discordant typing between CE to MPS or MPS to MPS kits. To consider the presence of such features, genomic coordinates must be given when reporting MPS alleles, to help identify the potential sources of any such discordances. The presence of variants in the primer-binding regions can affect amplification success, as is the case for CE; however, considering the cost implications in a sequenced analysis, suppliers implement reasonable preventative approaches (e.g. the use of degenerate primers) to account for common variants. Rare variants, however, can decrease amplification success and may even result in a missing 'null' allele. The frequency of such variants can be population specific, and therefore suppliers should consider a wider range of variation beyond the standard forensic cohorts (e.g. in the US, Caucasians, African-Americans, Hispanics). These can include other well studied ethnic groups, but analysis can still be affected by variants in less represented populations. In the UK, several ethnic groups are present in significant proportions, therefore their genetic diversities and structures need to be considered and the variants described in a similar fashion to that of the indigenous population targeted by this study. One study that has addressed this already is that of Devesse et al. (2018), which used the ForenSeq™ DNA Signature Prep Kit to study White British and British Chinese populations of the UK.

7.3.2 Nomenclature adjustments to MPS

For MPS to be implemented fully it needs to use a unified nomenclature for reporting and for international exchange and comparison of sequence-based genetic information. Several different approaches exist and work is in progress to use a system of related nomenclatures that can satisfy the different needs of stakeholders (Gettings et al. 2019).

The basic, unambiguous level of sequence data exchange is the sequence string; however, this is cumbersome, lengthy and non-intuitive for human perception. For compatibility to existing databasing software (i.e. CODIS) a nomenclature is preferred that is short and code-like for efficient storage and query by automated comparison algorithms, e.g. the 'sequence identifier' (SID) approach (Young et al. 2019). For software used in analysis, such as mixture deconvolution or probabilistic genotyping, a similar shorthand version of nomenclature is ideal, being informative about allele structure, but compatible with the existing software, e.g. the 'longest uninterrupted stretch' (LUS) approach (Just and Irwin 2018). There is a need to consider regional legislative restrictions over the reportable regions, and to provide flexibility to include or exclude flanking region variants, while preserving comparability of the STR sequences.

7.3.3 Advantages and disadvantages of MPS over CE

Summarised simply, MPS is generating more information, from less input, across more marker types, to potentially greater detail and resolution. The cost of analysis has become affordable with this technology for complex serious crime investigations, and considering the wealth of information it can produce due to its high throughput, massively parallel approach it is becoming a viable supplement or, some day, maybe even replacement for CE-based forensic DNA typing. With the use of MPS, more information is extracted: isometric heterozygotes can be distinguished, detection thresholds can be lowered (with a sufficiently high coverage) to provide more resolution for minor component detection, mixture deconvolution or lower level heteroplasmy detection. Also, detection, identification and quantitation of numts and other artefacts is feasible.

Profiles can be amplified from degraded samples with higher success rates from shorter STR amplicons, or from iiSNPs, and furthermore phenotypic and ancestry informative markers can also be included, as well as indels and microhaplotypes. Lineage markers can also benefit from the use of MPS - Y-SNPs, Y-STRs or the

control region or entire mtDNA can be more efficiently typed, benefitting from the massively parallel approach.

The current difficulty with MPS is that it is costly to implement as a new technology, and to adopt the different laboratory workflow and analysis procedures requires a different skill set and a good bioinformatic understanding of the processes to analyse the data at the moment, but this should ease with the appearance of more user-friendly commercial software solutions. A significant issue is the requirement for secure storage of very large amounts of sensitive genetic data, and this will need to be part of the restructuring brought about by the introduction of MPS into forensic workflows.

7.3.4 Future directions for MPS and beyond

As of now, the first milestone of using MPS to secure a conviction in a criminal case has been passed in the Netherlands. Also, the FBI approved the use of several MPS kits during 2019 (FBI 2019) the French police have been using automated MPS routinely (Laurent et al. 2017) for typing reference and casework samples for over a year now, and it is expected that MPS will gain more weight worldwide in investigations, from intelligence to casework applications.

Commercial suppliers are offering more kits with combinations of different marker types. This modular approach and scalability is the most suitable for future applications, allowing flexibility. The inclusion of predictive phenotypic and ancestry informative SNP markers, for example in the ForenSeq™ DNA Signature Prep Kit primer set B, opens the way to include probabilistic typing of these markers. Data analysis is still a bottle-neck for many who wish to adopt MPS, and therefore in the future software solutions with clearly auditable analysis steps and modifications will be required. These will help, with lowered cost and more user-friendly analysis, to make MPS more widespread, rather than a specialised service. Although this may be the ultimate goal, it seems unlikely that MPS will supplant CE for a long time, given how well established it is.

MPS could also be used to compare or validate outputs from the emerging field of consumer genomics in the future. The use of direct-to-customer (DTC) companies offering genome-wide SNP testing for genealogical purposes engages citizen scientists (Jobling et al. 2016), and has inadvertently helped to provide investigative leads for criminal investigations of cold cases and missing persons in the US within the last two years (Greytak et al. 2019). The use of forensic genetic genealogy from DTC tests and databases was first publicised in early 2018 by the arrest of Joseph DeAngelo, the alleged Golden State Killer (GSK). With quick succession many similar cases have arisen in the US and companies (Parabon Nanolabs, Bode Technology, Othram) and a non-profit organisation (DNA Doe Project) now offer this type of investigative leads to police forces. There are several questionable aspects of these new developments. As the scientific method has not yet been appropriately tested, validated or published, therefore the reliability of these services offered are uncertain (Phillips 2018), and it may prematurely jeopardise the reputation of a potentially beneficial technique, prior to it being thoroughly developed and validated. Another area of concern are the ethical considerations raised by the use of recreational genetic genealogy data (Syndercombe Court 2018); it is controversial how law enforcement bodies had access to consumer genetic databases without consent from the database donors. As a result several changes have happened since the announcement of the alleged GSK arrest, with databases changing their policies for law enforcement access. The databases currently accessible to law enforcement are FTDNA's own database (provided users opt in to allow this), and GEDmatch, a third party genealogy portal facilitating comparison between raw data from different DTC genomic providers, where again users have to actively opt in to allow law enforcement to use their data for investigations. As of 9th Dec 2019 GEDmatch was purchased by Verogen with the purpose of preserving its genealogical focus, but also allowing better control of law enforcement access and security while committing to develop its available tools. MPS platforms validated with scientific rigour could serve as an affordable intermediate validation tool for data obtained from databases like GEDmatch; also, current or future custom SNP typing panels on MPS platforms could verify the

genotypes obtained from citizen science sources and add more certainty to investigations using these novel approaches. The work presented in this thesis will hopefully play a part in bringing about this next revolution in forensic genetics.

ELECTRONIC APPENDICES:

The published papers from this research project provided as appendices:

- Appendix A. (A_Huszar_et_al_2018.pdf)

Huszar, T.I., Jobling, M.A. and Wetton, J.H. 2018.

A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing.

Forensic Science International Genetics, 35:97-106.

doi: [10.1016/j.fsigen.2018.03.012](https://doi.org/10.1016/j.fsigen.2018.03.012)

- Appendix B (B_Huszar_et_al_2019.pdf)

Huszar, T.I., Wetton, J.H. and Jobling, M.A. 2019.

Mitigating the effects of reference sequence bias in single-multiplex massively parallel sequencing of the mitochondrial DNA control region.

Forensic Science International Genetics, 40:9-17.

doi: [10.1016/j.fsigen.2019.01.008](https://doi.org/10.1016/j.fsigen.2019.01.008)

The following electronic appendices published as:

doi: [10.25392/leicester.data.11663775](https://doi.org/10.25392/leicester.data.11663775)

- Appendix C Global set - Comprehensive Y-STR sequence allele information (C_Global_YSTR.xlsx)
- Appendix D Global set - Comprehensive aSTR sequence allele information (D_Global_aSTR.xlsx)
- Appendix E Global set - Comprehensive database of mtDNA variation, OREO script, probes and examples. (E_Global_mtDNA.xlsx)
- Appendix F PoBI set - sequence allele information for all forensic markers. (F_PoBI_markers.xlsx)
- Appendix G PoBI set - Further population genetics matrices and graphs. (G_PoBI_popgen_graphs.xlsx)

Bibliography:

- 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526: 68-74.
- Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19: 1655-64.
- Alexeyev M, Shokolenko I, Wilson G, LeDoux S (2013) The maintenance of mitochondrial DNA integrity--critical analysis and update. *Cold Spring Harb Perspect Biol* 5: a012641.
- Aliferi A, Thomson J, McDonald A, Paynter VM, Ferguson S, Vanhinsbergh D, Syndercombe Court D, Ballard D (2018) UK and Irish Y-STR population data-A catalogue of variant alleles. *Forensic Sci Int Genet* 34: e1-e6.
- Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290: 457-65.
- Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 23: 147.
- Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Argimon S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG, Yeats CA, Grundmann H, Spratt BG, Aanensen DM (2016) Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microbial Genomics* 2(11) e000093.
- Armour JA, Palla R, Zeeuwen PL, den Heijer M, Schalkwijk J, Hollox EJ (2007) Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Res* 35: e19.
- Athey TW (2005) Haplogroup Prediction from Y-STR Values Using an Allele-Frequency Approach. *J Genet Geneal* 1: 1-7.
- Athey TW (2006) Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach. *J Genet Geneal* 2: 34-39.
- Au CH, Ho DN, Kwong A, Chan TL, Ma ESK (2017) BAMClipper: removing primers from alignments to minimize false-negative mutations in amplicon next-generation sequencing. *Sci Rep* 7: 1567.
- Balanovsky O, Gurianov V, Zaporozhchenko V, Balaganskaya O, Urasin V, Zhabagin M, Grugni V, Canada R, Al-Zahery N, Raveane A, Wen SQ, Yan S, Wang X, Zalloua P, Marafi A, Koshel S, Semino O, Tyler-Smith C, Balanovska E (2017) Phylogeography of human Y-chromosome haplogroup Q3-L275 from an academic/citizen science collaboration. *BMC Evol Biol* 17: 18.
- Balaresque P, Bowden GR, Adams SM, Leung HY, King TE, Rosser ZH, Goodwin J, Moisan JP, Richard C, Millward A, Demaine AG, Barbujani G, Previdere C, Wilson IJ, Tyler-Smith C, Jobling MA (2010) A predominantly neolithic origin for European paternal lineages. *PLoS Biol* 8: e1000285.
- Balaresque P, Bowden GR, Parkin EJ, Omran GA, Heyer E, Quintana-Murci L, Roewer L, Stoneking M, Nasidze I, Carvalho-Silva DR, Tyler-Smith C, de Knijff P, Jobling MA (2008) Dynamic nature of the proximal AZFc region of the human Y chromosome:

- multiple independent deletion and duplication events revealed by microsatellite analysis. *Hum Mutat* 29: 1171-80.
- Balaresque P, Poulet N, Cussat-Blanc S, Gerard P, Quintana-Murci L, Heyer E, Jobling MA (2015) Y-chromosome descent clusters and male differential reproductive success: young lineage expansions dominate Asian pastoral nomadic populations. *Eur J Hum Genet* 23: 1413-22.
- Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, Choi Y, van Duijn K, Vermeulen M, Brauer S, Decorte R, Poetsch M, von Wurmb-Schwark N, de Knijff P, Labuda D, Vezina H, Knoblauch H, Lessig R, Roewer L, Ploski R, Dobosz T, Henke L, Henke J, Furtado MR, Kayser M (2010) Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *Am J Hum Genet* 87: 341-53.
- Ballantyne KN, Keerl V, Wollstein A, Choi Y, Zuniga SB, Ralf A, Vermeulen M, de Knijff P, Kayser M (2012) A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. *Forensic Sci Int Genet* 6: 208-18.
- Ballantyne KN, Ralf A, Aboukhalid R, Achakzai NM, Anjos MJ, Ayub Q, Balazic J, Ballantyne J, Ballard DJ, Berger B, Bobillo C, Bouabdellah M, Burri H, Capal T, Caratti S, Cardenas J, Cartault F, Carvalho EF, Carvalho M, Cheng B, Coble MD, Comas D, Corach D, D'Amato ME, Davison S, de Knijff P, De Ungria MC, Decorte R, Dobosz T, Dupuy BM, Elmrghni S, Gliwinski M, Gomes SC, Grol L, Haas C, Hanson E, Henke J, Henke L, Herrera-Rodriguez F, Hill CR, Holmlund G, Honda K, Immel UD, Inokuchi S, Jobling MA, Kaddura M, Kim JS, Kim SH, Kim W, King TE, Klausriegler E, Kling D, Kovacevic L, Kovatsi L, Krajewski P, Kravchenko S, Larmuseau MH, Lee EY, Lessig R, Livshits LA, Marjanovic D, Minarik M, Mizuno N, Moreira H, Morling N, Mukherjee M, Munier P, Nagaraju J, Neuhuber F, Nie S, Nilasitsataporn P, Nishi T, Oh HH, Olofsson J, Onofri V, Palo JU, Pamjav H, Parson W, Petlach M, Phillips C, Ploski R, Prasad SP, Primorac D, Purnomo GA, Purps J, Rangel-Villalobos H, Rebala K, Rerkamnuaychoke B, Gonzalez DR, Robino C, Roewer L, Rosa A, Sajantila A, Sala A, Salvador JM, Sanz P, Schmitt C, Sharma AK, Silva DA, Shin KJ, et al. (2014) Toward male individualization with rapidly mutating y-chromosomal short tandem repeats. *Hum Mutat* 35: 1021-32.
- Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16: 37-48.
- Barton RNE, Jacobi RM, Stapert D, Street MJ (2003) The Late-glacial reoccupation of the British Isles and the Creswellian. *Journal of Quaternary Science* 18: 631-643.
- Batini C, Hallast P, Vagene AJ, Zadik D, Eriksen HA, Pamjav H, Sajantila A, Wetton JH, Jobling MA (2017) Population resequencing of European mitochondrial genomes highlights sex-bias in Bronze Age demographic expansions. *Sci Rep* 7: 12086.
- Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* 162: 2025-35.
- Bennett P (2000) Demystified ... microsatellites. *Mol Pathol* 53: 177-83.
- Berger C, Parson W (2009) Mini-midi-mito: adapting the amplification and sequencing strategy of mtDNA to the degradation state of crime scene samples. *Forensic Sci Int Genet* 3: 149-53.
- Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, Blanché H, Deleuze J-F, Cann H, Mallick S, Reich D, Sandhu MS, Skoglund P, Scally A, Xue Y, Durbin R, Tyler-Smith C (2019) Insights into human genetic variation and population history from 929 diverse genomes. *bioRxiv*: 674986.

- Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, Jonasdottir A, Magnusson OT, Thorsteinsdottir U, Masson G, Kong A, Gudbjartsson DF, Stefansson K (2016) Multi-nucleotide de novo Mutations in Humans. *PLoS Genet* 12: e1006315.
- Bintz BJ, Dixon GB, Wilson MR (2014) Simultaneous detection of human mitochondrial DNA and nuclear-inserted mitochondrial-origin sequences (NumtS) using forensic mtDNA amplification strategies and pyrosequencing technology. *J Forensic Sci* 59: 1064-73.
- Bodner M, Bastisch I, Butler JM, Fimmers R, Gill P, Gusmao L, Morling N, Phillips C, Prinz M, Schneider PM, Parson W (2016) Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER). *Forensic Sci Int Genet* 24: 97-102.
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114-2120.
- Borsting C, Morling N (2015) Next generation sequencing and its applications in forensic genetics. *Forensic Sci Int Genet* 18: 78-89.
- Brace S, Diekmann Y, Booth TJ, van Dorp L, Faltyskova Z, Rohland N, Mallick S, Olalde I, Ferry M, Michel M, Oppenheimer J, Broomandkhoshbacht N, Stewardson K, Martiniano R, Walsh S, Kayser M, Charlton S, Hellenthal G, Armit I, Schulting R, Craig OE, Sheridan A, Parker Pearson M, Stringer C, Reich D, Thomas MG, Barnes I (2019) Ancient genomes indicate population replacement in Early Neolithic Britain. *Nature Ecol Evol* 3: 765-771.
- Brinkmann B, Klitsch M, Neuhuber F, Huhne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet* 62: 1408-15.
- Budowle B, van Daal A (2008) Forensically relevant SNP classes. *Biotechniques* 44: 603-8, 610.
- Butler JM (2005) *Forensic DNA typing : biology, technology, and genetics of STR markers*, 2nd edn. Elsevier Academic Press, Amsterdam; Boston
- Butler JM (2007) Short tandem repeat typing technologies used in human identity testing. *Biotechniques* 43: ii-v.
- Butler JM (2010) *Fundamentals of forensic DNA typing*. Elsevier Academic Press, Amsterdam; Boston
- Butler JM, Decker AE, Vallone PM, Kline MC (2006) Allele frequencies for 27 Y-STR loci with U.S. Caucasian, African American, and Hispanic samples. *Forensic Sci Int* 156: 250-60.
- Butler JM, Hill CR (2012) Biology and Genetics of New Autosomal STR Loci Useful for Forensic DNA Analysis. *Forensic Sci Rev* 24: 15-26.
- Butler JM, Shen Y, McCord BR (2003) The development of reduced size STR amplicons as tools for analysis of degraded DNA. *J Forensic Sci* 48: 1054-64.
- Byrne RP, Martiniano R, Cassidy LM, Carrigan M, Hellenthal G, Hardiman O, Bradley DG, McLaughlin RL (2018) Insular Celtic population structure and genomic footprints of migration. *PLoS Genet* 14: e1007152.
- Cadenas AM, Regueiro M, Gayden T, Singh N, Zhivotovsky LA, Underhill PA, Herrera RJ (2007) Male amelogenin dropouts: phylogenetic context, origins and implications. *Forensic Sci Int* 166: 155-63.
- Capelli C, Redhead N, Abernethy JK, Gratrix F, Wilson JF, Moen T, Hervig T, Richards M, Stumpf MP, Underhill PA, Bradshaw P, Shaha A, Thomas MG, Bradman N, Goldstein DB (2003) A Y chromosome census of the British Isles. *Curr Biol* 13: 979-84.

- Carracedo A, Bar W, Lincoln P, Mayr W, Morling N, Olaisen B, Schneider P, Budowle B, Brinkmann B, Gill P, Holland M, Tully G, Wilson M (2000) DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing. *Forensic Sci Int* 110: 79-85.
- Chaitanya L, Breslin K, Zuniga S, Wirken L, Pospiech E, Kukla-Bartoszek M, Sijen T, Knijff P, Liu F, Branicki W, Kayser M, Walsh S (2018) The HirisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. *Forensic Sci Int Genet* 35: 123-135.
- Churchill JD, Schmedes SE, King JL, Budowle B (2016) Evaluation of the Illumina® Beta Version ForenSeq DNA Signature Prep Kit for use in genetic profiling. *Forensic Sci Int Genet* 20: 20-29.
- Cloete K, Ehrenreich L, D'Amato ME, Leat N, Davison S, Benjeddou M (2010) Analysis of seventeen Y-chromosome STR loci in the Cape Muslim population of South Africa. *Leg Med (Tokyo)* 12: 42-5.
- D'Amato ME, Ehrenreich L, Cloete K, Benjeddou M, Davison S (2010) Characterization of the highly discriminatory loci DYS449, DYS481, DYS518, DYS612, DYS626, DYS644 and DYS710. *Forensic Sci Int Genet* 4: 104-10.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project Analysis G (2011) The variant call format and VCFtools. *Bioinformatics* 27: 2156-8.
- Davies N (2000) *The Isles: A History*. Macmillan, London
- Dayama G, Emery SB, Kidd JM, Mills RE (2014) The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Res* 42: 12640-9.
- de Knijff P (2019a) The first MPS-based conviction in a criminal case. ISFG 2019 Congress talk, Prague
- de Knijff P (2019b) From next generation sequencing to now generation sequencing in forensics. *Forensic Sci Int Genet* 38: 175-180.
- Devesse L, Ballard D, Davenport L, Riethorst I, Mason-Buck G, Syndercombe Court D (2018) Concordance of the ForenSeq system and characterisation of sequence-specific autosomal STR alleles across two major population groups. *Forensic Sci Int Genet* 34: 57-61.
- Drabek J, Chung DT, Butler JM, McCord BR (2004) Concordance study between Miniplex assays and a commercial STR typing kit. *J Forensic Sci* 49: 859-60.
- Eichmann C, Parson W (2008) 'Mitominis': multiplex PCR analysis of reduced size amplicons for compound sequence analysis of the entire mtDNA control region in highly degraded samples. *Int J Legal Med* 122: 385-8.
- Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10: 564-7.
- FBI (2019) CODIS and NDIS fact sheet.(<https://www.fbi.gov/services/laboratory/biometric-analysis/codis/codis-and-ndis-fact-sheet>).
- Forster P, Kayser M, Meyer E, Roewer L, Pfeiffer H, Benkmann H, Brinkmann B (1998) Phylogenetic resolution of complex mutational features at Y-STR DYS390 in aboriginal Australians and Papuans. *Mol Biol Evol* 15: 1108-14.
- Forster P, Toth A (2003) Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proc Natl Acad Sci U S A* 100: 9079-84.
- Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Ginjupalli S, Gunturi S, Ponnuswamy V, Natarajan S, Nachimuthu PK (2003) A classifier for the SNP-based inference of ancestry. *J Forensic Sci* 48: 771-82.
- Fujii K, Watahiki H, Mita Y, Iwashima Y, Miyaguchi H, Kitayama T, Nakahara H, Mizuno N, Sekiguchi K (2016) Next-generation sequencing analysis of off-ladder alleles

- due to migration shift caused by sequence variation at D12S391 locus. *Leg Med (Tokyo)* 22: 62-7.
- Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezzenov DV (2009) The challenges of sequencing by synthesis. *Nat Biotechnol* 27: 1013-23.
- Gallimore JM, McElhove JA, Holland MM (2018) Assessing heteroplasmic variant drift in the mtDNA control region of human hairs using an MPS approach. *Forensic Sci Int Genet* 32: 7-17.
- Gettings KB, Aponte RA, Vallone PM, Butler JM (2015) STR allele sequence variation: Current knowledge and future issues. *Forensic Sci Int Genet* 18: 118-30.
- Gettings KB, Ballard D, Bodner M, Borsuk LA, King J, Parson W, Phillips C (2019) Report from the STRAND Working Group on the 2019 STR Sequence Nomenclature Meeting. *Forensic Sci Int Genet* 43:102165.
- Gettings KB, Borsuk LA, Ballard D, Bodner M, Budowle B, Devesse L, King J, Parson W, Phillips C, Vallone PM (2017) STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci. *Forensic Sci Int Genet* 31: 111-117.
- Gettings KB, Kiesler KM, Faith SA, Montano E, Baker CH, Young BA, Guerrieri RA, Vallone PM (2016) Sequence variation of 22 autosomal STR loci detected by next generation sequencing. *Forensic Sci Int Genet* 21: 15-21.
- Gilbert E, O'Reilly S, Merrigan M, McGettigan D, Molloy AM, Brody LC, Bodmer W, Hutnik K, Ennis S, Lawson DJ, Wilson JF, Cavalleri GL (2017) The Irish DNA Atlas: Revealing Fine-Scale Population Structure and History within Ireland. *Sci Rep* 7: 17199.
- Gilbert E, O'Reilly S, Merrigan M, McGettigan D, Vitart V, Joshi PK, Clark DW, Campbell H, Hayward C, Ring SM, Golding J, Goodfellow S, Navarro P, Kerr SM, Amador C, Campbell A, Haley CS, Porteous DJ, Cavalleri GL, Wilson JF (2019) The genetic landscape of Scotland and the Isles. *Proc Natl Acad Sci U S A* 116: 19064-19070.
- Gill P (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *Int J Legal Med* 114: 204-10.
- Gill P, Jeffreys AJ, Werrett DJ (1985) Forensic application of DNA 'fingerprints'. *Nature* 318: 577-9.
- Goodacre S, Helgason A, Nicholson J, Southam L, Ferguson L, Hickey E, Vega E, Stefansson K, Ward R, Sykes B (2005) Genetic evidence for a family-based Scandinavian settlement of Shetland and Orkney during the Viking periods. *Heredity (Edinb)* 95: 129-35.
- Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17: 333-51.
- Gouy A, Zieger M (2017) STRAF - A convenient online tool for STR data evaluation in forensic genetics. *Forensic Sci Int Genet* 30: 148-151.
- Grandi FC, An W (2013) Non-LTR retrotransposons and microsatellites: Partners in genomic variation. *Mob Genet Elements* 3: e25674.
- Greytak EM, Moore C, Armentrout SL (2019) Genetic genealogy for cold case and active investigations. *Forensic Sci Int* 299: 103-113.
- Gusmao L, Sanchez-Diz P, Calafell F, Martin P, Alonso CA, Alvarez-Fernandez F, Alves C, Borjas-Fajardo L, Bozzo WR, Bravo ML, Builes JJ, Capilla J, Carvalho M, Castillo C, Catanesi CI, Corach D, Di Lonardo AM, Espinheira R, Fagundes de Carvalho E, Farfan MJ, Figueiredo HP, Gomes I, Lojo MM, Marino M, Pinheiro MF, Pontes ML, Prieto V, Ramos-Luis E, Riancho JA, Souza Goes AC, Santapa OA, Sumita DR, Vallejo G, Vidal Rioja L, Vide MC, Vieira da Silva CI, Whittle MR,

- Zabala W, Zarrabeitia MT, Alonso A, Carracedo A, Amorim A (2005) Mutation rates at Y chromosome specific microsatellites. *Hum Mutat* 26: 520-8.
- Hallast P, Batini C, Zadik D, Maisano Delser P, Wetton JH, Arroyo-Pardo E, Cavalleri GL, de Knijff P, Destro Bisol G, Dupuy BM, Eriksen HA, Jorde LB, King TE, Larmuseau MH, Lopez de Munain A, Lopez-Parra AM, Loutradis A, Milasin J, Novelletto A, Pamjav H, Sajantila A, Schempp W, Sears M, Tolun A, Tyler-Smith C, Van Geystelen A, Watkins S, Winney B, Jobling MA (2015) The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. *Mol Biol Evol* 32: 661-73.
- Hazkani-Covo E, Zeller RM, Martin W (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet* 6: e1000834.
- Hearne CM, Ghosh S, Todd JA (1992) Microsatellites for linkage analysis of genetic traits. *Trends Genet* 8: 288-94.
- Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics* 107: 1-8.
- Helgason A, Einarsson AW, Guethmundsdottir VB, Sigurethsson A, Gunnarsdottir ED, Jagadeesan A, Ebenesersdottir SS, Kong A, Stefansson K (2015) The Y-chromosome point mutation rate in humans. *Nat Genet* 47: 453-7.
- Higham T, Compton T, Stringer C, Jacobi R, Shapiro B, Trinkaus E, Chandler B, Groning F, Collins C, Hillson S, O'Higgins P, FitzGerald C, Fagan M (2011) The earliest evidence for anatomically modern humans in northwestern Europe. *Nature* 479: 521-4.
- Higuchi R, von Beroldingen CH, Sensabaugh GF, Erlich HA (1988) DNA typing from single hairs. *Nature* 332: 543-6.
- Hill CR, Kline MC, Mulero JJ, Lagace RE, Chang CW, Hennessy LK, Butler JM (2007) Concordance study between the AmpFISTR MiniFiler PCR amplification kit and conventional STR typing kits. *J Forensic Sci* 52: 870-3.
- Hill EW, Jobling MA, Bradley DG (2000) Y-chromosome variation and Irish origins. *Nature* 404: 351-2.
- Holland MM, Makova KD, McElhoe JA (2018) Deep-Coverage MPS Analysis of Heteroplasmic Variants within the mtGenome Allows for Frequent Differentiation of Maternal Relatives. *Genes (Basel)* 9.
- Holland MM, McQuillan MR, O'Hanlon KA (2011) Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy. *Croat Med J* 52: 299-313.
- Holland MM, Parsons TJ (1999) Mitochondrial DNA Sequence Analysis - Validation and Use for Forensic Casework. *Forensic Sci Rev* 11: 21-50.
- Hoogenboom J, van der Gaag KJ, de Leeuw RH, Sijen T, de Knijff P, Laros JF (2017) FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise. *Forensic Sci Int Genet* 27: 27-40.
- Huber N, Parson W, Dur A (2018) Next generation database search algorithm for forensic mitogenome analyses. *Forensic Sci Int Genet* 37: 204-214.
- Huszar TI, Jobling MA, Wetton JH (2018) A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing. *Forensic Sci Int Genet* 35: 97-106.
- Huszar TI, Wetton JH, Jobling MA (2019) Mitigating the effects of reference sequence bias in single-multiplex massively parallel sequencing of the mitochondrial DNA control region. *Forensic Sci Int Genet* 40: 9-17.
- International HapMap Consortium (2003) The International HapMap Project. *Nature* 426: 789-96.

- Irwin JA, Parson W, Coble MD, Just RS (2011) mtGenome reference population databases and the future of forensic mtDNA analysis. *Forensic Sci Int Genet* 5: 222-5.
- Irwin JA, Saunier JL, Niederstatter H, Strouss KM, Sturk KA, Diegoli TM, Brandstatter A, Parson W, Parsons TJ (2009) Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. *J Mol Evol* 68: 516-27.
- Jeffreys AJ, Brookfield JF, Semeonoff R (1985a) Positive identification of an immigration test-case using human DNA fingerprints. *Nature* 317: 818-9.
- Jeffreys AJ, Turner M, Debenham P (1991) The efficiency of multilocus DNA fingerprint probes for individualization and establishment of family relationships, determined from extensive casework. *Am J Hum Genet* 48: 824-40.
- Jeffreys AJ, Wilson V, Thein SL (1985b) Hypervariable 'minisatellite' regions in human DNA. *Nature* 314: 67-73.
- Jobling MA (2001) Y-chromosomal SNP haplotype diversity in forensic analysis. *Forensic Sci Int* 118: 158-62.
- Jobling MA, Gill P (2004) Encoded evidence: DNA in forensic analysis. *Nat Rev Genet* 5: 739-51.
- Jobling MA, Lo IC, Turner DJ, Bowden GR, Lee AC, Xue Y, Carvalho-Silva D, Hurles ME, Adams SM, Chang YM, Kraaijenbrink T, Henke J, Guanti G, McKeown B, van Oorschot RA, Mitchell RJ, de Knijff P, Tyler-Smith C, Parkin EJ (2007) Structural variation on the short arm of the human Y chromosome: recurrent multigene deletions encompassing Amelogenin Y. *Hum Mol Genet* 16: 307-16.
- Jobling MA, Pandya A, Tyler-Smith C (1997) The Y chromosome in forensic analysis and paternity testing. *Int J Legal Med* 110: 118-24.
- Jobling MA, Rasteiro R, Wetton JH (2016) In the blood: the myth and reality of genetic markers of identity. *Ethnic and Racial Studies* 39: 142-161.
- Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet* 4: 598-612.
- Jobling MA, Tyler-Smith C (2017) Human Y-chromosome variation in the genome-sequencing era. *Nat Rev Genet* 18: 485-497.
- Just RS, Irwin JA (2018) Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results. *Forensic Sci Int Genet* 34: 197-205.
- Just RS, Irwin JA, Parson W (2015) Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. *Forensic Sci Int Genet* 18: 131-9.
- Just RS, Moreno LI, Smerick JB, Irwin JA (2017) Performance and concordance of the ForenSeq system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens. *Forensic Sci Int Genet* 28: 1-9.
- Kayser M, Brauer S, Weiss G, Schiefenhovel W, Underhill PA, Stoneking M (2001) Independent histories of human Y chromosomes from Melanesia and Australia. *Am J Hum Genet* 68: 173-190.
- Kayser M, Kittler R, Erler A, Hedman M, Lee AC, Mohyuddin A, Mehdi SQ, Rosser Z, Stoneking M, Jobling MA, Sajantila A, Tyler-Smith C (2004) A comprehensive survey of human Y-chromosomal microsatellites. *Am J Hum Genet* 74: 1183-97.
- Kechin A, Boyarskikh U, Kel A, Filipenko M (2017) cutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. *J Comput Biol* 24: 1138-1143.
- Kershaw J, Røyrvik E (2016) The 'People of the British Isles' project and Viking settlement in England. *Antiquity* 90: 1670-1680.

- Khubrani YM, Wetton JH, Jobling MA (2018) Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs. *Forensic Sci Int Genet* 33: 98-105.
- Kim D (2016) Library generation for Next-Generation Sequencing. vol WO/2016/025872
- Kim KS, Sappington TW (2013) Microsatellite data analysis for population genetics. *Methods Mol Biol* 1006: 271-95.
- King TE, Fortes GG, Balaesque P, Thomas MG, Balding D, Maisano Delser P, Neumann R, Parson W, Knapp M, Walsh S, Tonasso L, Holt J, Kayser M, Appleby J, Forster P, Ekserdjian D, Hofreiter M, Schurer K (2014) Identification of the remains of King Richard III. *Nat Commun* 5: 5631.
- Krings M, Geisert H, Schmitz RW, Krainitzki H, Paabo S (1999) DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. *Proc Natl Acad Sci U S A* 96: 5581-5.
- Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Paabo S (1997) Neandertal DNA sequences and the origin of modern humans. *Cell* 90: 19-30.
- Kwon SY, Lee HY, Kim EH, Lee EY, Shin KJ (2016) Investigation into the sequence structure of 23 Y chromosomal STR loci using massively parallel sequencing. *Forensic Sci Int Genet* 25: 132-141.
- Lang M, Sazzini M, Calabrese FM, Simone D, Boattini A, Romeo G, Luiselli D, Attimonelli M, Gasparre G (2012) Polymorphic NumtS trace human population relationships. *Hum Genet* 131: 757-71.
- LaRue BL, Ge J, King JL, Budowle B (2012) A validation study of the Qiagen Investigator DIPplex[®] kit; an INDEL-based assay for human identification. *Int J Legal Med* 126: 533-40.
- Lascaro D, Castellana S, Gasparre G, Romeo G, Saccone C, Attimonelli M (2008) The RHNumtS compilation: features and bioinformatics approaches to locate and quantify Human NumtS. *BMC Genomics* 9: 267.
- Laurent FX, Ausset L, Clot M, Jullien S, Chantrel Y, Hollard C, Pene L (2017) Automation of library preparation using Illumina ForenSeq kit for routine sequencing of casework samples. *Forensic Sci Int Genet Supplement Series* 6: e415-e417.
- Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS Genet* 8: e1002453.
- Leat N, Ehrenreich L, Benjeddou M, Cloete K, Davison S (2007) Properties of novel and widely studied Y-STR loci in three South African populations. *Forensic Sci Int* 168: 154-61.
- Leclair B, Fregeau CJ, Bowen KL, Fourney RM (2004) Systematic analysis of stutter percentages and allele peak height and peak area ratios at heterozygous STR loci for forensic casework and database samples. *J Forensic Sci* 49: 968-80.
- Lee EY, Lee HY, Shin KJ (2016) Off-ladder alleles due to a single nucleotide polymorphism in the flanking region at DYS481 detected by the PowerPlex[®] Y23 System. *Forensic Sci Int Genet* 24: e7-e8.
- Lee EY, Shin KJ, Rakha A, Sim JE, Park MJ, Kim NY, Yang WI, Lee HY (2014) Analysis of 22 Y chromosomal STR haplotypes and Y haplogroup distribution in Pathans of Pakistan. *Forensic Sci Int Genet* 11: 111-6.
- Lee JE, Hong EJ, Kim JH, Shin SY, Kim YY, Han BG (2013) Instability at Short Tandem Repeats in Lymphoblastoid Cell Lines. *Osong Public Health Res Perspect* 4: 194-6.
- Leibelt C, Budowle B, Collins P, Daoudi Y, Moretti T, Nunn G, Reeder D, Roby R (2003) Identification of a D8S1179 primer binding site mutation and the validation of a primer designed to recover null alleles. *Forensic Sci Int* 133: 220-7.

- Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, Hutnik K, Royrvik EC, Cunliffe B, Wellcome Trust Case Control C, International Multiple Sclerosis Genetics C, Lawson DJ, Falush D, Freeman C, Pirinen M, Myers S, Robinson M, Donnelly P, Bodmer W (2015) The fine-scale genetic structure of the British population. *Nature* 519: 309-314.
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-60.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-9.
- Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics* 24: 713-4.
- Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* 17: 10-12.
- Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, Anaya V, Richardson R, Davis J, Genomes Project C, MacArthur DG, Sidow A, Duret L, Gerstein M, Makova KD, Marchini J, McVean G, Lunter G (2013) The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* 23: 749-61.
- Mullis KB (1990) The unusual origin of the polymerase chain reaction. *Sci Am* 262: 56-61, 64-5.
- Myres NM, Ekins JE, Lin AA, Cavalli-Sforza LL, Woodward SR, Underhill PA (2007) Y-chromosome short tandem repeat DYS458.2 non-consensus alleles occur independently in both binary haplogroups J1-M267 and R1b3-M405. *Croat Med J* 48: 450-9.
- Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* 156: 297-304.
- Novroski NMM, King JL, Churchill JD, Seah LH, Budowle B (2016) Characterization of genetic sequence variation of 58 STR loci in four major population groups. *Forensic Sci Int Genet* 25: 214-226.
- Oh YN, Lee HY, Lee EY, Kim EH, Yang WI, Shin KJ (2015) Haplotype and mutation analysis for newly suggested Y-STRs in Korean father-son pairs. *Forensic Sci Int Genet* 15: 64-8.
- Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, Rohland N, Mallick S, Szecsenyi-Nagy A, Mitnik A, Altena E, Lipson M, Lazaridis I, Harper TK, Patterson N, Broomandkhoshbacht N, Diekmann Y, Faltyskova Z, Fernandes D, Ferry M, Harney E, de Knijff P, Michel M, Oppenheimer J, Stewardson K, Barclay A, Alt KW, Liesau C, Rios P, Blasco C, Miguel JV, Garcia RM, Fernandez AA, Banffy E, Bernabo-Brea M, Billoin D, Bonsall C, Bonsall L, Allen T, Buster L, Carver S, Navarro LC, Craig OE, Cook GT, Cunliffe B, Denaire A, Dinwiddy KE, Dodwell N, Ernee M, Evans C, Kucharik M, Farre JF, Fowler C, Gazenbeek M, Pena RG, Haber-Uriarte M, Haduch E, Hey G, Jowett N, Knowles T, Massy K, Pfrengle S, Lefranc P, Lemerrier O, Lefebvre A, Martinez CH, Olmo VG, Ramirez AB, Maurandi JL, Majo T, McKinley JI, McSweeney K, Mende BG, Modi A, Kulcsar G, Kiss V, Czene A, Patay R, Endrodi A, Kohler K, Hajdu T, Szeniczey T, Dani J, Bernert Z, Hoole M, Cheronet O, Keating D, Veleminsky P, Dobes M, Candilio F, Brown F, Fernandez RF, Herrero-Corral AM, Tusa S, Carnieri E, Lentini L, Valenti A, Zanini A, Waddington C, Delibes G, et al. (2018) The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* 555: 190-196.

- Ott J, Wang J, Leal SM (2015) Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* 16: 275-84.
- Parr RL, Maki J, Reguly B, Dakubo GD, Aguirre A, Wittock R, Robinson K, Jakupciak JP, Thayer RE (2006) The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation. *BMC Genomics* 7: 185.
- Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmao L, Hares DR, Irwin JA, King JL, Knijff P, Morling N, Prinz M, Schneider PM, Neste CV, Willuweit S, Phillips C (2016) Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic Sci Int Genet* 22: 54-63.
- Parson W, Bandelt HJ (2007) Extended guidelines for mtDNA typing of population data in forensic science. *Forensic Sci Int Genet* 1: 13-9.
- Parson W, Dur A (2007) EMPOP--a forensic mtDNA database. *Forensic Sci Int Genet* 1: 88-92.
- Parson W, Gusmao L, Hares DR, Irwin JA, Mayr WR, Morling N, Pokorak E, Prinz M, Salas A, Schneider PM, Parsons TJ, Genetics DNACotISfF (2014) DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. *Forensic Sci Int Genet* 13: 134-42.
- Phillips C (2015) Forensic genetic analysis of bio-geographical ancestry. *Forensic Sci Int Genet* 18: 49-65.
- Phillips C (2018) The Golden State Killer investigation and the nascent field of forensic genealogy. *Forensic Science International: Genetics* 36: 186-188.
- Phillips C, Devesse L, Ballard D, van Weert L, de la Puente M, Melis S, Alvarez Iglesias V, Freire-Aradas A, Oldroyd N, Holt C, Syndercombe Court D, Carracedo A, Lareu MV (2018) Global patterns of STR sequence variation: Sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit. *Electrophoresis* 39: 2708-2724.
- Phillips C, Fernandez-Formoso L, Garcia-Magarinos M, Porras L, Tvedebrink T, Amigo J, Fondevila M, Gomez-Tato A, Alvarez-Dios J, Freire-Aradas A, Gomez-Carballa A, Mosquera-Miguel A, Carracedo A, Lareu MV (2011) Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel. *Forensic Sci Int Genet* 5: 155-69.
- Pontes ML, Fondevila M, Lareu MV, Medeiros R (2015) SNP Markers as Additional Information to Resolve Complex Kinship Cases. *Transfus Med Hemother* 42: 385-8.
- Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, Chen Y, Banerjee R, Rodriguez-Flores JL, Cerezo M, Shao H, Gymrek M, Malhotra A, Louzada S, Desalle R, Ritchie GR, Cerveira E, Fitzgerald TW, Garrison E, Marcketta A, Mittelman D, Romanovitch M, Zhang C, Zheng-Bradley X, Abecasis GR, McCarroll SA, Flicek P, Underhill PA, Coin L, Zerbino DR, Yang F, Lee C, Clarke L, Auton A, Erlich Y, Handsaker RE, Genomes Project C, Bustamante CD, Tyler-Smith C (2016) Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat Genet* 48: 593-9.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945-59.
- Promega (2015) User guide, PowerSeq™ Auto/Mito/Y System. pp 1-4
- Promega (2018) User guide: PowerSeq™ CRM Nested System, Custom. pp 1-4

- Purps J, Siegert S, Willuweit S, Nagy M, Alves C, Salazar R, Angustia SM, Santos LH, Anslinger K, Bayer B, Ayub Q, Wei W, Xue Y, Tyler-Smith C, Bafalluy MB, Martinez-Jarreta B, Egyed B, Balitzki B, Tschumi S, Ballard D, Court DS, Barrantes X, Bassler G, Wiest T, Berger B, Niederstatter H, Parson W, Davis C, Budowle B, Burri H, Borer U, Koller C, Carvalho EF, Domingues PM, Chamoun WT, Coble MD, Hill CR, Corach D, Caputo M, D'Amato ME, Davison S, Decorte R, Larmuseau MH, Ottoni C, Rickards O, Lu D, Jiang C, Dobosz T, Jonkisz A, Frank WE, Furac I, Gehrig C, Castella V, Grskovic B, Haas C, Wobst J, Hadzic G, Drobnic K, Honda K, Hou Y, Zhou D, Li Y, Hu S, Chen S, Immel UD, Lessig R, Jakovski Z, Ilievska T, Klann AE, Garcia CC, de Knijff P, Kraaijenbrink T, Kondili A, Miniati P, Vouropoulou M, Kovacevic L, Marjanovic D, Lindner I, Mansour I, Al-Azem M, Andari AE, Marino M, Furfuro S, Locarno L, Martin P, Luque GM, Alonso A, Miranda LS, Moreira H, Mizuno N, Iwashima Y, Neto RS, Nogueira TL, Silva R, Nastainczyk-Wulf M, Edelmann J, Kohl M, Nie S, Wang X, Cheng B, et al. (2014) A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Sci Int Genet* 12: 12-23.
- Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13: 341.
- R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- Rabbani B, Tekin M, Mahdiah N (2014) The promise of whole-exome sequencing in medical genetics. *J Hum Genet* 59: 5-15.
- Ragoussis J (2009) Genotyping technologies for genetic research. *Annu Rev Genomics Hum Genet* 10: 117-33.
- Rajeevan H, Soundararajan U, Kidd JR, Pakstis AJ, Kidd KK (2012) ALFRED: an allele frequency resource for research and teaching. *Nucleic Acids Res* 40: D1010-5.
- Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* 102: 15942-7.
- Rambaut A (2006-2012) Fig.Tree. Tree Figure Drawing Tool, version 1.4.0.
- Rathbun MM, McElhoe JA, Parson W, Holland MM (2017) Considering DNA damage when interpreting mtDNA heteroplasmy in deep sequencing data. *Forensic Sci Int Genet* 26: 1-11.
- Redd AJ, Agellon AB, Kearney VA, Contreras VA, Karafet T, Park H, de Knijff P, Butler JM, Hammer MF (2002) Forensic value of 14 novel STRs on the human Y chromosome. *Forensic Sci Int* 130: 97-111.
- Relethford JH (1983) Genetic structure and population history of Ireland: a comparison of blood group and anthropometric analyses. *Ann Hum Biol* 10: 321-33.
- Reuter JA, Spacek DV, Snyder MP (2015) High-throughput sequencing technologies. *Mol Cell* 58: 586-97.
- Reynolds R, Sensabaugh G, Blake E (1991) Analysis of genetic markers in forensic DNA samples using the polymerase chain reaction. *Anal Chem* 63: 2-15.
- Rhoads A, Au KF (2015) PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* 13: 278-89.
- Ricchetti M, Tekaia F, Dujon B (2004) Continued colonization of the human genome by mitochondrial DNA. *PLoS Biol* 2: E273.
- Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y,

- Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Norby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt HJ (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* 67: 1251-76.
- Ring JD, Sturk-Andreaggi K, Alyse Peck M, Marshall C (2018) Bioinformatic removal of NUMT-associated variants in mitotiling next-generation sequencing data from whole blood samples. *Electrophoresis* 39: 2785-2797.
- Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. *Nat Biotechnol* 29: 24-6.
- Rocca RA, Magoon G, Reynolds DF, Krahn T, Tilroe VO, Op den Velde Boots PM, Grierson AJ (2012) Discovery of Western European R1b1a2 Y chromosome variants in 1000 genomes project data: an online community approach. *PLoS One* 7: e41634.
- Rock AW, Dur A, van Oven M, Parson W (2013) Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA). *Forensic Sci Int Genet* 7: 601-9.
- Roffey PE, Eckhoff CI, Kuhl JL (2000) A rare mutation in the amelogenin gene and its potential investigative ramifications. *J Forensic Sci* 45: 1016-9.
- Rolf B, Wiegand P, Brinkmann B (2002) Somatic mutations at STR loci--a reason for three-allele pattern and mosaicism. *Forensic Sci Int* 126: 200-2.
- Ruitberg CM, Reeder DJ, Butler JM (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Res* 29: 320-2.
- Sanchez JJ, Phillips C, Borsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D, Schneider PM, Carracedo A, Morling N (2006) A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis* 27: 1713-24.
- Santos FR, Pandya A, Tyler-Smith C (1998) Reliability of DNA-based sex tests. *Nat Genet* 18: 103.
- Schlecht J, Kaplan ME, Barnard K, Karafet T, Hammer MF, Merchant NC (2008) Machine-learning approaches for classifying haplogroup from Y chromosome STR data. *PLoS Comput Biol* 4: e1000093.
- Searle JB, Kotlik P, Rambau RV, Markova S, Herman JS, McDevitt AD (2009) The Celtic fringe of Britain: insights from small mammal phylogeography. *Proc Biol Sci* 276: 4287-94.
- Seman A, Bakar ZA, Isa MN (2012) An efficient clustering algorithm for partitioning Y-short tandem repeats data. *BMC Res Notes* 5: 557.
- Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, Lin AA, Mitra M, Sil SK, Ramesh A, Usha Rani MV, Thakur CM, Cavalli-Sforza LL, Majumder PP, Underhill PA (2006) Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am J Hum Genet* 78: 202-21.
- Shadrach B, Commane M, Hren C, Warshawsky I (2004) A rare mutation in the primer binding region of the amelogenin gene can interfere with gender identification. *J Mol Diagn* 6: 401-5.
- Shi W, Ayub Q, Vermeulen M, Shao RG, Zuniga S, van der Gaag K, de Knijff P, Kayser M, Xue Y, Tyler-Smith C (2010) A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Mol Biol Evol* 27: 385-93.

- Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfsing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang SP, Waterston RH, Wilson RK, Rozen S, Page DC (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423: 825-37.
- Smith AH (1956) English place-name elements. University Press, Cambridge Eng.
- Sobrinho B, Brion M, Carracedo A (2005) SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Sci Int* 154: 181-94.
- Sobrinho B, Carracedo A (2005) SNP typing in forensic genetics: a review. *Methods Mol Biol* 297: 107-26.
- Strobl C, Eduardoff M, Bus MM, Allen M, Parson W (2018) Evaluation of the precision ID whole MtDNA genome panel for forensic analyses. *Forensic Sci Int Genet* 35: 21-25.
- Sullivan KM, Mannucci A, Kimpton CP, Gill P (1993) A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin. *Biotechniques* 15: 636-8, 640-1.
- Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, Stefansson K (2012) A direct characterization of human mutation based on microsatellites. *Nat Genet* 44: 1161-5.
- Syndercombe Court D (2018) Forensic genealogy: Some serious concerns. *Forensic Sci Int Genet* 36: 203-204.
- Tamaki K, Jeffreys AJ (2005) Human tandem repeat sequences in forensic DNA typing. *Leg Med (Tokyo)* 7: 244-50.
- Tamura K, Stecher G, Peterson D, Filipowski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 30: 2725-9.
- Thangaraj K, Reddy AG, Singh L (2002) Is the amelogenin gene reliable for gender identification in forensic casework and prenatal diagnosis? *Int J Legal Med* 116: 121-3.
- Thomas R, Zischler H, Paabo S, Stoneking M (1996) Novel mitochondrial DNA insertion polymorphism and its usefulness for human population studies. *Hum Biol* 68: 847-54.
- Thornton JA (2016) Splicing by Overlap Extension PCR to Obtain Hybrid DNA Products. *Methods Mol Biol* 1373: 43-9.
- Torrioni A, Achilli A, Macaulay V, Richards M, Bandelt HJ (2006) Harvesting the fruit of the human mtDNA tree. *Trends Genet* 22: 339-45.
- Tully G, Sullivan KM, Nixon P, Stones RE, Gill P (1996) Rapid detection of mitochondrial sequence polymorphisms using multiplex solid-phase fluorescent minisequencing. *Genomics* 34: 107-13.
- Tully LA, Parsons TJ, Steighner RJ, Holland MM, Marino MA, Prenger VL (2000) A sensitive denaturing gradient-Gel electrophoresis assay reveals a high frequency of heteroplasmy in hypervariable region 1 of the human mtDNA control region. *Am J Hum Genet* 67: 432-43.
- Urquhart A, Kimpton CP, Downes TJ, Gill P (1994) Variation in short tandem repeat sequences--a survey of twelve microsatellite loci for use as forensic identification markers. *Int J Legal Med* 107: 13-20.
- Valverde L, Illescas MJ, Villaescusa P, Gotor AM, García A, Cardoso S, Algorta J, Catarino S, Rouault K, Férec C, Hardiman O, Zarrabeitia M, Jiménez S, Pinheiro MF, Jarreta BM, Olofsson J, Morling N, de Pancorbo MM (2016) New clues to the

- evolutionary history of the main European paternal lineage M269: dissection of the Y-SNP S116 in Atlantic Europe and Iberia. *Eur J Hum Genet* 24: 437-441.
- Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43: 11 10 1-33.
- van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. *Trends Genet* 30: 418-26.
- van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* 30: E386-94.
- van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH (2014) Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. *Hum Mutat* 35: 187-91.
- Vermeulen M, Wollstein A, van der Gaag K, Lao O, Xue Y, Wang Q, Roewer L, Knoblauch H, Tyler-Smith C, de Knijff P, Kayser M (2009) Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms. *Forensic Sci Int Genet* 3: 205-13.
- Wallace H (2006) The UK National DNA Database. Balancing crime detection, human rights and privacy. *EMBO Rep* 7 Spec No: S26-30.
- Wang DY, Green RL, Lagace RE, Oldroyd NJ, Hennessy LK, Mulero JJ (2012) Identification and secondary structure analysis of a region affecting electrophoretic mobility of the STR locus SE33. *Forensic Sci Int Genet* 6: 310-6.
- Warshauer DH, Churchill JD, Novroski N, King JL, Budowle B (2015a) Novel Y-chromosome Short Tandem Repeat Variants Detected Through the Use of Massively Parallel Sequencing. *Genomics Proteomics Bioinformatics* 13: 250-7.
- Warshauer DH, King JL, Budowle B (2015b) STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. *Forensic Sci Int Genet* 14: 182-6.
- Warshauer DH, Lin D, Hari K, Jain R, Davis C, Larue B, King JL, Budowle B (2013) STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic Sci Int Genet* 7: 409-17.
- Weale ME, Weiss DA, Jager RF, Bradman N, Thomas MG (2002) Y chromosome evidence for Anglo-Saxon mass migration. *Mol Biol Evol* 19: 1008-21.
- Weber JL, Wong C (1993) Mutation of human short tandem repeats. *Hum Mol Genet* 2: 1123-8.
- Weiler NE, de Vries G, Sijen T (2016) Development of a control region-based mtDNA SNaPshot selection tool, integrated into a mini amplicon sequencing method. *Sci Justice* 56: 96-103.
- Weissensteiner H, Forer L, Fuchsberger C, Schopf B, Kloss-Brandstatter A, Specht G, Kronenberg F, Schonherr S (2016a) mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. *Nucleic Acids Res* 44: W64-9.
- Weissensteiner H, Pacher D, Kloss-Brandstatter A, Forer L, Specht G, Bandelt HJ, Kronenberg F, Salas A, Schonherr S (2016b) HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res* 44: W58-63.
- Wendt FR, Churchill JD, Novroski NMM, King JL, Ng J, Oldt RF, McCulloh KL, Weise JA, Smith DG, Kanthaswamy S, Budowle B (2016) Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx forensic genomics system. *Forensic Sci Int Genet* 24: 18-23.

- Wendt FR, King JL, Novroski NMM, Churchill JD, Ng J, Oldt RF, McCulloh KL, Weise JA, Smith DG, Kanthaswamy S, Budowle B (2017) Flanking region variation of ForenSeq DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans. *Forensic Sci Int Genet* 28: 146-154.
- Wilkins JF, Marlowe FW (2006) Sex-biased migration in humans: what should we expect from genetic data? *BioEssays* 28: 290-300.
- Willems T, Gymrek M, Highnam G, Genomes Project C, Mittelman D, Erlich Y (2014) The landscape of human STR variation. *Genome Res* 24: 1894-904.
- Wilson JF, Weiss DA, Richards M, Thomas MG, Bradman N, Goldstein DB (2001) Genetic evidence for different male and female roles during cultural transitions in the British Isles. *Proc Natl Acad Sci U S A* 98: 5078-83.
- Winney B, Boumertit A, Day T, Davison D, Echeta C, Evseeva I, Hutnik K, Leslie S, Nicodemus K, Royrvik EC, Tonks S, Yang X, Cheshire J, Longley P, Mateos P, Groom A, Relton C, Bishop DT, Black K, Northwood E, Parkinson L, Frayling TM, Steele A, Sampson JR, King T, Dixon R, Middleton D, Jennings B, Bowden R, Donnelly P, Bodmer W (2012) People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *Eur J Hum Genet* 20: 203-10.
- Woerner AE, King JL, Budowle B (2017) Fast STR allele identification with STRait Razor 3.0. *Forensic Sci Int Genet* 30: 18-23.
- Yang Y, Xie B, Yan J (2014) Application of next-generation sequencing technology in forensic science. *Genomics Proteomics Bioinformatics* 12: 190-7.
- Young B, Faris T, Armogida L (2019) A nomenclature for sequence-based forensic DNA analysis. *Forensic Sci Int Genet* 42: 14-20.
- Yucesan E, Ozten N (2019) Pharmacogenetics: Role of Single Nucleotide Polymorphisms. *Methods Mol Biol* 2054: 137-145.
- Zhao X, Ma K, Li H, Cao Y, Liu W, Zhou H, Ping Y (2015) Multiplex Y-STRs analysis using the ion torrent personal genome machine (PGM). *Forensic Sci Int Genet* 19: 192-196.
- Zidkova A, Horinek A, Kebrdlova V, Korabecna M (2013) Application of the new insertion-deletion polymorphism kit for forensic identification and parentage testing on the Czech population. *Int J Legal Med* 127: 7-10.
- Zischler H, Geisert H, von Haeseler A, Paabo S (1995) A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. *Nature* 378: 489-92.